Chemometrics and Intelligent Laboratory Systems 189 (2019) 96–101

Contents lists available at ScienceDirect

Chemometrics and Intelligent Laboratory Systems

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

iDNA6mA (5-step rule): Identification of DNA N6-methyladenine sites in the rice genome by intelligent computational model via Chou's 5-step rule

Muhammad Tahir [a,b], Hilal Tayara [a,**], Kil To Chong [c,*]

[a] Department of Electronics and Information Engineering, Chonbuk National University, Jeonju, 54896, South Korea
[b] Department of Computer Science, Abdul Wali Khan University, Mardan, 23200, Pakistan
[c] Advanced Electronics and Information Research Center, Chonbuk National University, Jeonju, 54896, South Korea

ARTICLE INFO

Keywords:
Convolution neural network
N^6-methyladenine
Deep learning
DNA

ABSTRACT

DNA methylation is an elementary epigenetic process. The N6-methyladenine is related to a large kind of biological processes i.e., transcription, DNA replication, and repair. In genome, the N6-methyladenine (6 mA) site distribution is non-random; therefore, precise discrimination of 6 mA is necessary to understand its biological functions. Through biochemical experiments, the N6-methyladenine produced a positive outcome, still, these wet lab processes are very time consuming and high pricy. In view of this, it is of high priority to introduce a powerful, accurate, and fast computational model to identify N6-methyladenine sites. In this connection, we propose an intelligent computational model called iDNA6mA (5-step rule) using deep learning approach to identify N6-methyladenine sites from DNA sequences in the rice genome. Existing methods used handcrafted features to identify N6-methyladenine sites; however, the proposed computational model automatically extracts the key features from DNA input sequences via the proposed convolution neural network (CNN) model. The intelligent computational model iDNA6mA (5-step rule) obtained 86.64% of accuracy, 86.70% of sensitivity, 86.59% of specificity, 0.732 of MCC, and 0.931 of auROC. The results demonstrate that the proposed intelligent computational model achieved better performance in terms of all evaluation parameters than existing techniques. It is observed that iDNA6mA (5-step rule) model will become a useful tool in the fields of computational biology, bioinformatics, and for the academic research on N6-methyladenine sites prediction. A user-friendly webserver has been established and freely accessible at https://home.jbnu.ac.kr/NSCL/iDNA6mA.htm.

1. Introduction

DNA methylation is an elementary epigenetic process. In eukaryotes [1], the high recognized DNA modification is 5-methylcytosine (5 mC) sites [2] while in prokaryotes; the most pervasive DNA modification is N^6-methyladenine (6 mA) sites [3]. The abundance and presence of 6 mA in eukaryotes have been reported in early discussion [1,4]. Specific unicellular eukaryotes, green algae, and ciliates consist of both 5 mC and 6 mA in their genomes, however the biological importance of these modifications, for living beings had remained mostly uncharacterized [5]. Most recently, DNA N^6-methyladenine (6 mA) as a non-canonical DNA modification has been reported in three kingdoms of life [6]. The modification of N^6-methyladenine is nearly concerned along with a number of biological processes i.e., transcription [7], DNA replication [8] and repair [9]. To decipher the complete biological functions of 6 mA, it is imperative to identify its location throughout the whole genome.

A number of experimental procedures like single-molecule real-time sequencing (SMRT-seq) [10], methylated DNA Immunoprecipitation sequencing (MeDIP-seq) [11], and capillary electrophoresis and laser-induced fluorescence (CE-LIF) [12] have been reported. A group of Chinese scientists [13] has recently explored the 6 mA profile of rice genome by mass spectrometry analysis and immunoprecipitation followed by sequencing (IP-seq). However, along with productive data information, the experimental procedure also impedes the genome-wide prediction of N^6-methyladenine. Therefore, the development of a computational model using the pre-existed experimental data to identify the N^6-methyladenine site will cope with this issue and provide ease for future studies. Although the biochemical exploratory techniques can give some information related to the 6 mA sites, it is high pricy and labor-intensive. Thus, it is a huge challenge to establish fast and precise

* Corresponding author.
** Corresponding author.
E-mail addresses: hilaltayara@jbnu.ac.kr (H. Tayara), kitchong@jbnu.ac.kr (K.T. Chong).

https://doi.org/10.1016/j.chemolab.2019.04.007
Received 7 February 2019; Received in revised form 2 April 2019;
Available online 20 April 2019

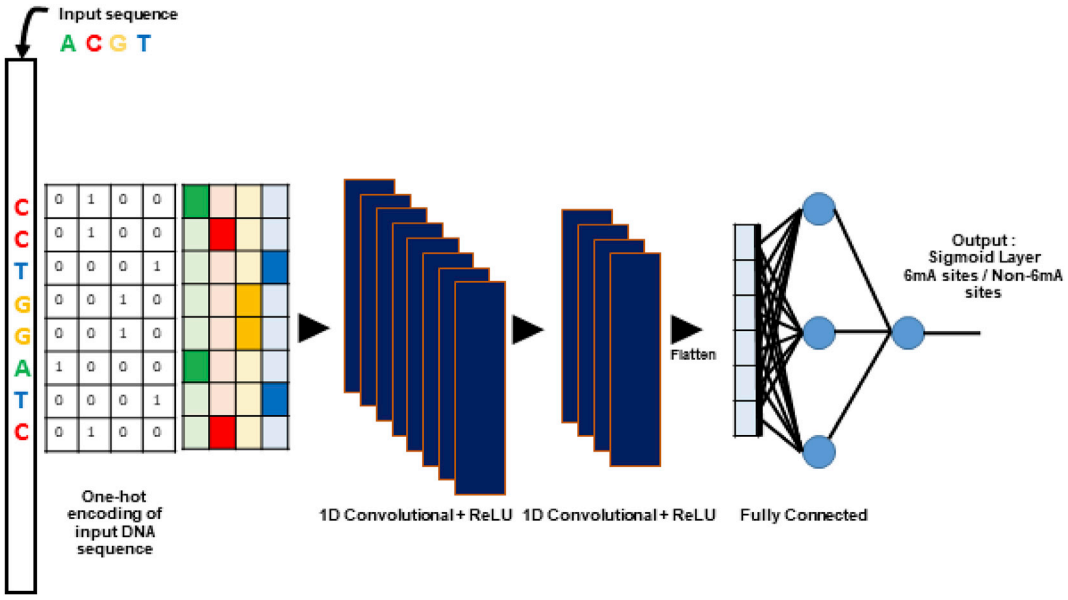

Fig. 1. The architecture of iDNA6mA (5-step rule) model.

computational model to identifying the 6 mA sites.

The ubiquitous and complex post-translational modifications (PTMs) play a key roles in different biological processes, as well as in protein localization and trafficking, protein folding, cell signaling, transcriptional regulation, cell-cell interactions, apoptosis, and regulating cellular dynamic and plasticity [14–40]. The irregularity of the PTMs are nearly related or connected with harmful diseases i.e., Alzheimer's, Parkinson's and cancers. Accordingly, researchers introduced a variety of predictors and techniques for discrimination of PTM sites using protein and RNA/DNA samples [41–45]. Recently, Chen et al. [46] introduced a predictor using machine learning to predict the N^6-methyladenine sites in the rice genome based on support vector machine (SVM) and nucleotide frequency and nucleotide chemical properties as features extraction techniques. The existing computational models need domain knowledge to hand design the input feature spaces. The second step of the guideline of Chou's 5-step rules [17,44,47–59] for developing a useful prediction model is to extract feature space from the RNA/DNA sequences. The computational model may automatically capture the important features of 6 mA sites from input samples. This concept achieved by deep learning to extract the features from multiple levels of abstraction. Deep learning has generated very successful results in natural language processing [60,61], speech recognition [62–64], and image recognition [65–67]. Currently, various predictors have been proposed based on deep learning such as iDeepS [68], branch point selection [69], alternative splicing sites prediction [70], and iRNA-PseKNC(2methyl) prediction model [71].

In this connection, we propose iDNA6mA (5-step rule) model for the identification of DNA N^6-methyladenine sites that is based on the deep learning approach. We observe that our proposed novel deep learning method has superior prediction outcomes compared to the existing machine learning method [46]. In addition, a user-friendly webserver has been established and freely accessible at https://home.jbnu.ac.kr/NSCL/iDNA6mA.htm.

2. Materials and methods

2.1. Benchmark dataset

In accordance to the guideline of chou's 5-step rules for constructing a useful prediction model [34,72–79], the first step is to select/construct a reliable benchmark dataset for training and testing the prediction system effectively. Therefore, we have selected and downloaded a valid benchmark dataset (http://lin-group.cn/server/i6mAPred/data), which was constructed by Chen et al. [46], to show the efficiency of the proposed prediction model. The length of all samples is 41-bp long with the 6 mA site in the center. Thus, the benchmark dataset can be formulated as below:

$$S = S^+ \cup S^- \quad (1)$$

The benchmark dataset S consists of 1760 samples; where S^+ represents the positive subset and contains 880 6 mA sites and S^- is the negative subset and contains 880 non-6mA sites. The $\cup$ is the union in the set theory. The dataset is divided as 70% training, 10% validation and 20% testing.

2.2. The proposed model

Fig. 1 illustrates the framework of iDNA6mA (5-step rule) model that is based on convolutional neural networks (CNN). The CNN is a frequently and widely employed method by various researchers [69,70,80] in the field of bioinformatics. During training the convolutional neural network, automatically learns the primary features from the input samples. The iDNA6mA (5-step rule) model takes a single input of a DNA sequence $s = \{s_1 s_2 s_3 \ldots\ldots s_n\}$, where $s_i \in \{A, C, G, T\}$ and $n = 41$, and produces a real-valued prediction output. We first employ one-hot encoding for the sequences and feed them into CNN to identify 6 mA sites. The length of vector equals to the length of the input sample (e.g. here the length is 41) and 4-channel are A, C, G, and T and represented as (1, 0, 0, 0), (0, 1, 0, 0), (0, 0, 1, 0), and (0, 0, 0, 1), respectively.

Generally, one processing step in convolution neural network named a layer, that is represented by convolution layer, pooling layer, ReLU layer, normalization layer, dropout layer, fully connected layer, etc. Different hyper-parameters have been tuned during learning such as filter size, number of convolution layers and so forth. Table 1 shows the tuned hyper-parameters in the proposed CNN model (see Table 2).

The best hyper-parameters have been chosen on the bases of high

Table 1
The hyper parameters to be tuned in CNN.

Hyper Parameter	Range
No. of convolution layers	[1,2]
Filter size	[2–5,7]
The number of the filters	[2,4,6,8,10,12]
Dropout probability	[0.1, 0.2, 0.3, 0.4]

Table 2
The summary of iDNA6mA (5-step rule) model.

Model Layers	Output Shape
Sample Input	(41,4)
Conv1D(8,5,1)	(41,8)
Conv1D(4,3,1)	(41,4)
Dropout(0.25)	164
Dense(1)	1

success rate of auROC, MCC, accuracy, sensitivity, and specificity. The convolution layer can be numerically expressed as below:

$$Conv(R)_{jk} = \mathrm{Re}LU\left(\sum_{fs=0}^{FS-1}\sum_{f=0}^{F-1} W_{fsf}^{k} R_{j+fs,f}\right) \quad (2)$$

In Equation (2), R represents the input DNA/RNA sample, fs and f represent the filter size and the number of the input channels, respectively, j denotes the index of the output position, and k denotes the index of the filters. ReLU denotes the rectified linear function and numerically represented below:

$$\mathrm{Re}LU(t) = \max(0, t) \quad (3)$$

Where t is the input to the neuron and "max" is an operand that returns the maximum of its inputs. The sigmoid layer is a nonlinear activation function and outputs in the range [0, 1]. It is used for predicting whether the input sequence is an N^6-methyladenine site or not. This function can be mathematically expressed as below:

$$Sigmoid(t) = \frac{1}{1+e^{-t}} \quad (4)$$

Where t is the input to the neuron.

In this paper, Keras framework [81] is used for building the proposed model iDNA6mA (5-step rule). The number of batches is set to 32 and the number of epochs is set to 50. Learning rate is set to 0.001 and optimizer is Adam.

3. Results and discussion

3.1. Performance evaluation

In order to measure the success rate of the prediction system, the following four metrics are used [75,82–89].

$$\begin{cases} MCC = \dfrac{1-\left(\dfrac{E_{-}^{+}+E_{+}^{-}}{E^{-}+E^{+}}\right)}{\sqrt{\left(\dfrac{E_{+}^{-}+E_{-}^{+}}{E^{-}}+1\right)\left(\dfrac{E_{+}^{-}+E_{-}^{+}}{E^{+}}+1\right)}} & -1 \le MCC \le 1 \\ Accuracy = 1-\left(\dfrac{E_{-}^{+}+E_{+}^{-}}{E^{+}+E^{-}}\right) & 0 \le Acc \le 1 \\ Ssensitivity = 1-\left(\dfrac{E_{-}^{+}}{E^{+}}\right) & 0 \le Sn \le 1 \\ Specificity = 1-\left(\dfrac{E_{+}^{-}}{E^{-}}\right) & 0 \le Sp \le 1 \end{cases} \quad (5)$$

The significance and details of these evaluation metrics can be found in Refs. [87–90]. E^{+} denotes the size of the positive dataset samples or N^6-methyladenine sites; while E^{-} denotes the size of the negative dataset samples or non- N^6-methyladenine sites; E_{-}^{+} is the number of non-N^6-methyladenine sites sample that are predicted incorrectly to be N^6-methyladenine sites while E_{+}^{-} is the number of N^6-methyladenine sites samples that are predicted incorrectly to be of non- N^6-methyladenine sites. MCC reflects the performance of proposed prediction model on imbalance dataset, here the ratio of negative and positive sequences is the same. The ROC curve is used to calculate the success rate of proposed intelligent computational model. The auROC(area under the ROC curve) is the most important indicator to measure the prediction quality of a binary classifier.

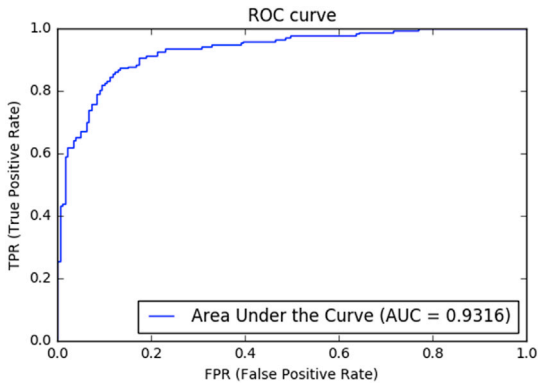

Fig. 2. The auROC curve of the intelligent computational model iDNA6mA (5-step rule).

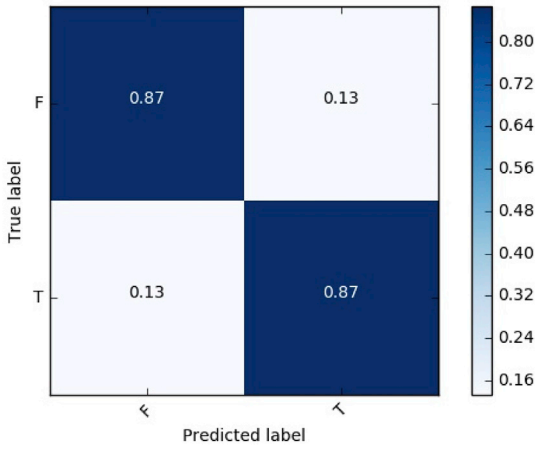

Fig. 3. The visualization of the confusion matrix of the intelligent computational model iDNA6mA (5-step rule).

3.2. Results and Discussion

In this section, we discuss the success rate of the proposed prediction system iDNA6mA (5-step rule) using one benchmark dataset. The proposed predictor obtains 86.64% of accuracy, 86.59% of specificity, 86.70% of sensitivity, 0.732 of MCC, and 0.931 of auROC. The detailed outcomes of the proposed predictor are shown bellows:

$$\begin{cases} Acc = 86.64\% \\ Sen = 86.70\% \\ Sp = 86.59\% \\ MCC = 0.732 \end{cases}$$

Additionally, Fig. 2 shows the performance of the auROC curve of the iDNA6mA (5-step rule) model and the visualization representation of the confusion matrix is shown in Fig. 3.

Table 3
The performance comparison between iDNA6mA (5-step rule) and other with Existing Method.

Methods	Accuracy	Specificity	Sensitivity	MCC	auROC
iDNA6mA (5-step rule)	86.64	86.59	86.70	0.73	0.931
6 mA-Pred [46]	83.13	83.30	82.95	0.66	0.886

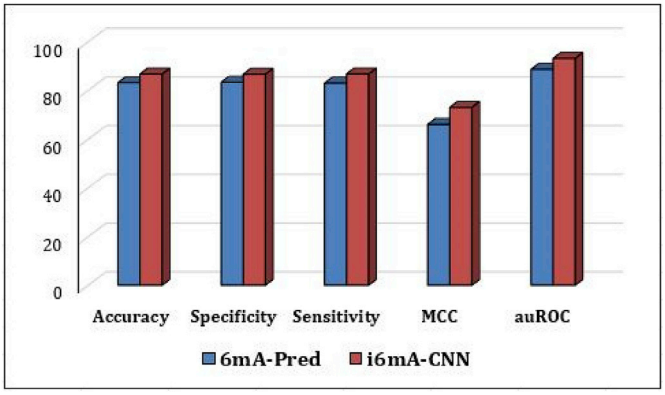

Fig. 4. Performance Comparison of proposed iDNA6mA (5-step rule) with existing method 6 mA-Pred.

The success rate of the iDNA6mA (5-step rule) computational model is compared with existing 6 mA-Pred model [46]. Table 3 shows the performance of the four evaluation metrics of the both models. It is evident that iDNA6mA (5-step rule) model outperforms the 6 mA-Pred model by 3.51% of accuracy, 3.29% of specificity, 3.75% of sensitivity, 7.00% of MCC, and 4.50% of auROC.

The graphical illustration of the experimental outcomes is presented in Fig. 4 in which the iDNA6mA (5-step rule) method obtains remarkable outcomes compared to the existing prediction system.

4. Webserver

In order to make the proposed iDNA6mA (5-step rule) model accessible by other researchers, a user-friendly webserver has been constructed [91–97]. The developed webserver supports finding i6mA sites using either direct input in Fasta format as shown in Fig. 5 or uploading the sequences in on file in Fasta format as shown in Fig. 6. The webserver has been built using Python programming language with Flask library. It is made accessible at https://home.jbnu.ac.kr/NSCL/iDNA6mA.htm.

5. Conclusions

We developed a novel and fully automatic deep learning based computational model namely iDNA6mA (5-step rule) to identify N^6-methyladenine sites from DNA sequences only. We first employed one-hot encoding for the input DNA sequences and fed them into CNN to identify N^6-methyladenine sites. The simulation outcomes showed the effectiveness of our proposed iDNA6mA (5-step rule) prediction model. It

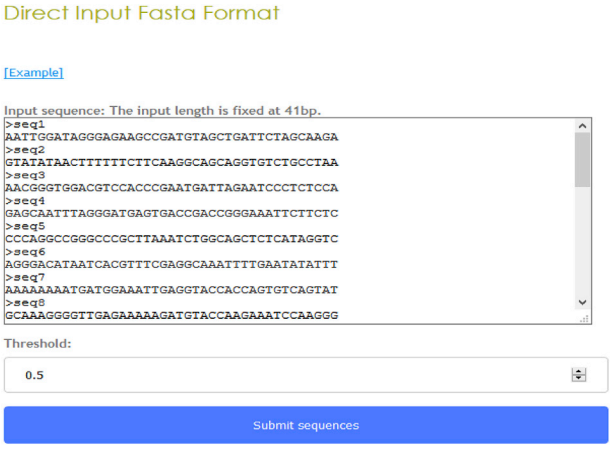

Fig. 5. Direct input sequences for i6mA sites identification using the proposed iDNA6mA (5-step rule) model.

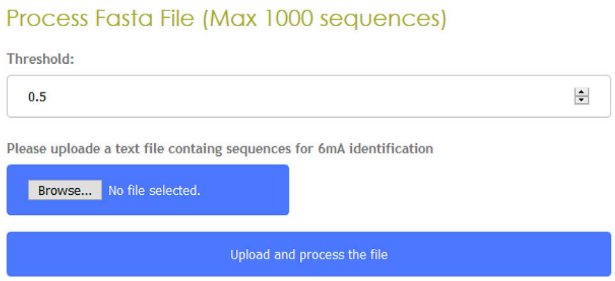

Figure 6. Processing FASTA file containing sequences for i6mA sites identification using the proposed iDNA6mA (5-step rule) model.

provided better prediction success rates in terms of all evaluation metrics (accuracy, specificity, sensitivity, MCC, and ROC) compared with the state-of the-art technique. The proposed methodology can be potentially effective in pharmaceutical industry for drug innovation and design and in the area of the bioinformatics. Finally, a webserver has been built using Python programming language with Flask library and made accessible at https://home.jbnu.ac.kr/NSCL/iDNA6mA.htm. In addition, we have deposited the model with the best learnt weights on the github at https://github.com/hilal-t/iDNA6mA.

Acknowledgements

This research was supported by the Brain Research Program of the National Research Foundation funded by the Korean government (MSIT) (No. NRF-2017M3C7A1044815).

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.chemolab.2019.04.007.

References

[1] B. Vanyushin, S. Tkacheva, A. Belozersky, Rare bases in animal DNA, Nature 225 (1970) 948.
[2] B. Vanyushin, A. Belozersky, N. Kokurina, D. Kadirova, 5-Methylcytosine and 6-methylaminopurine in bacterial DNA, Nature 218 (1968) 1066.
[3] D. Dunn, J. Smith, Occurrence of a new base in the deoxyribonucleic acid of a strain of Bacterium coli, Nature 175 (1955) 336.
[4] G. Unger, H. Venner, Remarks on minor bases in spermatic desoxyribonucleic acid, Hoppe-Seyler's Zeitschrift fur physiologische Chemie 344 (1966) 280–283.
[5] S. Hattman, C. Kenny, L. Berger, K. Pratt, Comparative study of DNA methylation in three unicellular eucaryotes, J. Bacteriol. 135 (1978) 1156–1157.
[6] 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.
[7] J.L. Robbins-Manke, Z.Z. Zdraveski, M. Marinus, J.M. Essigmann, Analysis of global gene expression and double-strand-break formation in DNA adenine methyltransferase-and mismatch repair-deficient Escherichia coli, J. Bacteriol. 187 (2005) 7027–7037.
[8] 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.
[9] 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.
[10] 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.
[11] K.R. Pomraning, K.M. Smith, M. Freitag, Genome-wide high throughput analysis of DNA methylation in eukaryotes, Methods 47 (2009) 142–150.
[12] A.M. Krais, M.G. Cornelius, H.H. Schmeiser, Genomic N6-methyladenine determination by MEKC with LIF, Electrophoresis 31 (2010) 3548–3551.
[13] C. Zhou, C. Wang, H. Liu, Q. Zhou, Q. Liu, Y. Guo, T. Peng, J. Song, J. Zhang, L. Chen, Identification and analysis of adenine N 6-methylation sites in the rice genome, Nature plants 4 (2018) 554.
[14] H.-L. Xie, L. Fu, X.-D. Nie, Using ensemble SVM to identify human GPCRs N-linked glycosylation sites based on the general form of Chou's PseAAC, Protein Engineering, Design & Selection 26 (2013) 735–742.
[15] Y. Xu, J. Ding, L.-Y. Wu, K.-C. Chou, iSNO-PseAAC: predict cysteine S-nitrosylation sites in proteins by incorporating position specific amino acid propensity into pseudo amino acid composition, PLoS One 8 (2013) e55844.

[16] Y. Xu, X.-J. Shao, L.-Y. Wu, N.-Y. Deng, K.-C. Chou, iSNO-AAPair: incorporating amino acid pairwise coupling into PseAAC for predicting cysteine S-nitrosylation sites in proteins, PeerJ 1 (2013) e171.
[17] C. Jia, X. Lin, Z. Wang, Prediction of protein S-nitrosylation sites based on adapted normal distribution bi-profile Bayes and Chou's pseudo amino acid composition, Int. J. Mol. Sci. 15 (2014) 10410–10423.
[18] W.-R. Qiu, X. Xiao, W.-Z. Lin, K.-C. Chou, iMethyl-PseAAC, Identification of protein methylation sites via a pseudo amino acid composition approach, BioMed Res. Int. (2014) 2014.
[19] Y. Xu, X. Wen, X.-J. Shao, N.-Y. Deng, K.-C. Chou, iHyd-PseAAC: predicting hydroxyproline and hydroxylysine in proteins by incorporating dipeptide position-specific propensity into pseudo amino acid composition, Int. J. Mol. Sci. 15 (2014) 7594–7610.
[20] Y. Xu, X. Wen, L.-S. Wen, L.-Y. Wu, N.-Y. Deng, K.-C. Chou, iNitro-Tyr: prediction of nitrotyrosine sites in proteins with general pseudo amino acid composition, PLoS One 9 (2014) e105018.
[21] J. Zhang, X. Zhao, P. Sun, Z. Ma, PSNO: predicting cysteine S-nitrosylation sites by incorporating various sequence-derived features into the general form of Chou's PseAAC, Int. J. Mol. Sci. 15 (2014) 11204–11219.
[22] 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.
[23] K.-C. Chou, Impacts of bioinformatics to medicinal chemistry, Med. Chem. 11 (2015) 218–234.
[24] W.-R. Qiu, X. Xiao, W.-Z. Lin, K.-C. Chou, iUbiq-Lys: prediction of lysine ubiquitination sites in proteins by extracting sequence evolution information via a gray system model, J. Biomol. Struct. Dyn. 33 (2015) 1731–1742.
[25] W. Chen, H. Tang, J. Ye, H. Lin, K.-C. Chou, iRNA-PseU: identifying RNA pseudouridine sites, Mol. Ther. Nucleic Acids 5 (2016).
[26] J. Jia, Z. Liu, X. Xiao, B. Liu, K.-C. Chou, iSuc-PseOpt: identifying lysine succinylation sites in proteins by incorporating sequence-coupling effects into pseudo components and optimizing imbalanced training dataset, Anal. Biochem. 497 (2016) 48–56.
[27] J. Jia, Z. Liu, X. Xiao, B. Liu, K.-C. Chou, pSuc-Lys: predict lysine succinylation sites in proteins with PseAAC and ensemble random forest approach, J. Theor. Biol. 394 (2016) 223–230.
[28] J. Jia, Z. Liu, X. Xiao, B. Liu, K.-C. Chou, iCar-PseCp: identify carbonylation sites in proteins by Monte Carlo sampling and incorporating sequence coupled effects into general PseAAC, Oncotarget 7 (2016) 34558.
[29] J. Jia, L. Zhang, Z. Liu, X. Xiao, K.-C. Chou, pSumo-CD: predicting sumoylation sites in proteins with covariance discriminant algorithm by incorporating sequence-coupled effects into general PseAAC, Bioinformatics 32 (2016) 3133–3141.
[30] Z. Ju, J.-Z. Cao, H. Gu, Predicting lysine phosphoglycerylation with fuzzy SVM by incorporating k-spaced amino acid pairs into Chou' s general PseAAC, J. Theor. Biol. 397 (2016) 145–150.
[31] Z. Liu, X. Xiao, D.-J. Yu, J. Jia, W.-R. Qiu, K.-C. Chou, pRNAm-PC: predicting N6-methyladenosine sites in RNA sequences via physical–chemical properties, Anal. Biochem. 497 (2016) 60–67.
[32] W.-R. Qiu, B.-Q. Sun, X. Xiao, Z.-C. Xu, K.-C. Chou, iHyd-PseCp: identify hydroxyproline and hydroxylysine in proteins by incorporating sequence-coupled effects into general PseAAC, Oncotarget 7 (2016) 44310.
[33] W.-R. Qiu, B.-Q. Sun, X. Xiao, Z.-C. Xu, K.-C. Chou, iPTM-mLys: identifying multiple lysine PTM sites and their different types, Bioinformatics 32 (2016) 3116–3123.
[34] 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.
[35] Z. Ju, J.-J. He, Prediction of lysine crotonylation sites by incorporating the composition of k-spaced amino acid pairs into Chou's general PseAAC, J. Mol. Graph. Model. 77 (2017) 200–204.
[36] L.-M. Liu, Y. Xu, K.-C. Chou, iPGK-PseAAC: identify lysine phosphoglycerylation sites in proteins by incorporating four different tiers of amino acid pairwise coupling information into the general PseAAC, Med. Chem. 13 (2017) 552–559.
[37] W.-R. Qiu, S.-Y. Jiang, B.-Q. Sun, X. Xiao, X. Cheng, K.-C. Chou, iRNA-2methyl: identify RNA 2'-O-methylation sites by incorporating sequence-coupled effects into general PseKNC and ensemble classifier, Med. Chem. 13 (2017) 734–743.
[38] W.-R. Qiu, S.-Y. Jiang, Z.-C. Xu, X. Xiao, K.-C. Chou, iRNAm5C-PseDNC: identifying RNA 5-methylcytosine sites by incorporating physical-chemical properties into pseudo dinucleotide composition, Oncotarget 8 (2017) 41178.
[39] W.R. Qiu, B.Q. Sun, X. Xiao, D. Xu, K.C. Chou, iPhos-PseEvo: identifying human phosphorylated proteins by incorporating evolutionary information into general PseAAC via grey system theory, Molecular Informatics 36 (2017) 1600010.
[40] Y. Xu, Z. Wang, C. Li, K.-C. Chou, iPreny-PseAAC: identify C-terminal cysteine prenylation sites in proteins by incorporating two tiers of sequence couplings into PseAAC, Med. Chem. 13 (2017) 544–551.
[41] S. Akbar, M. Hayat, iMethyl-STTNC: identification of N6-methyladenosine sites by extending the idea of SAAC into Chou's PseAAC to formulate RNA sequences, J. Theor. Biol. 455 (2018) 205–211.
[42] 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.
[43] W. Chen, P. Feng, H. Yang, H. Ding, H. Lin, K.-C. Chou, iRNA-3typeA: identifying three types of modification at RNA's adenosine sites, Mol. Ther. Nucleic Acids 11 (2018) 468–474.
[44] 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.
[45] Z. Ju, S.-Y. Wang, Prediction of citrullination sites by incorporating k-spaced amino acid pairs into Chou's general pseudo amino acid composition, Gene 664 (2018) 78–83.
[46] W. Chen, H. Lv, F. Nie, H. Lin, i6mA-Pred: identifying DNA N6-methyladenine sites in the rice genome, Bioinformatics (2019), btz015.
[47] L. Cai, T. Huang, J. Su, X. Zhang, W. Chen, F. Zhang, L. He, K.-C. Chou, Implications of newly identified brain eQTL genes and their interactors in Schizophrenia, Mol. Ther. Nucleic Acids 12 (2018) 433–442.
[48] 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. (2018), bby028.
[49] Y. Zhang, R. Xie, J. Wang, A. Leier, T.T. Marquez-Lago, T. Akutsu, G.I. Webb, K.-C. Chou, J. Song, Computational analysis and prediction of lysine malonylation sites by exploiting informative features in an integrative machine-learning framework, Briefings Bioinf. (2018) 5.
[50] B. Liu, S. Wang, R. Long, K.-C. Chou, iRSpot-EL: identify recombination spots with an ensemble learning approach, Bioinformatics 33 (2016) 35–41.
[51] 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.
[52] B. Liu, L. Fang, R. Long, X. Lan, K.-C. Chou, iEnhancer-2L: a two-layer predictor for identifying enhancers and their strength by pseudo k-tuple nucleotide composition, Bioinformatics 32 (2015) 362–369.
[53] 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.
[54] J. Song, F. Li, K. Takemoto, G. Haffari, T. Akutsu, K.-C. Chou, G.I. Webb, PREvaIL, an integrative approach for inferring catalytic residues using sequence, structural, and network features in a machine-learning framework, J. Theor. Biol. 443 (2018) 125–137.
[55] X. Cheng, X. Xiao, K.-C. Chou, pLoc-mEuk: predict subcellular localization of multi-label eukaryotic proteins by extracting the key GO information into general PseAAC, Genomics 110 (2018) 50–58.
[56] 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.
[57] J. Wang, J. Li, B. Yang, R. Xie, T.T. Marquez-Lago, A. Leier, M. Hayashida, T. Akutsu, Y. Zhang, K.-C. Chou, Bastion3: a two-layer ensemble predictor of type III secreted effectors, Bioinformatics 10 (2018).
[58] X. Xiao, Z.-C. Xu, W.-R. Qiu, P. Wang, H.-T. Ge, K.-C. Chou, iPSW (2L)-PseKNC: a two-layer predictor for identifying promoters and their strength by hybrid features via pseudo K-tuple nucleotide composition, Genomics (2018), https://doi.org/10.1016/j.ygeno.2018.12.001.
[59] J. Jia, X. Li, W. Qiu, X. Xiao, K.-C. Chou, iPPI-PseAAC (CGR): identify protein-protein interactions by incorporating chaos game representation into PseAAC, J. Theor. Biol. 460 (2019) 195–203.
[60] L. Deng, Y. Liu, Deep Learning in Natural Language Processing, Springer, 2018.
[61] T. Young, D. Hazarika, S. Poria, E. Cambria, Recent trends in deep learning based natural language processing, IEEE Comput. Intell. Mag. 13 (2018) 55–75.
[62] R. Ranjan, V.M. Patel, R. Chellappa, Hyperface: a deep multi-task learning framework for face detection, landmark localization, pose estimation, and gender recognition, IEEE Trans. Pattern Anal. Mach. Intell. 41 (2019) 121–135.
[63] G. Saon, G. Kurata, T. Sercu, K. Audhkhasi, S. Thomas, D. Dimitriadis, X. Cui, B. Ramabhadran, M. Picheny, L.-L. Lim, English Conversational Telephone Speech Recognition by Humans and Machines, 2017. 1703.02136.
[64] M. Gardner, J. Grus, M. Neumann, O. Tafjord, P. Dasigi, N. Liu, M. Peters, M. Schmitz, L. Zettlemoyer, Allennlp: A Deep Semantic Natural Language Processing Platform, 2018. 1803.07640.
[65] X. Yang, W. Liu, D. Tao, J. Cheng, Canonical correlation analysis networks for two-view image recognition, Inf. Sci. 385 (2017) 338–352.
[66] H. Tayara, K. Chong, Object detection in very high-resolution aerial images using one-stage densely connected feature pyramid network, Sensors 18 (2018), 3341-3341.
[67] H. Tayara, K.G. Soo, K.T. Chong, Vehicle detection and counting in high-resolution aerial images using convolutional regression neural network, IEEE Access 6 (2018) 2220–2230.
[68] X. Pan, P. Rijnbeek, J. Yan, H.-B. Shen, Prediction of RNA-protein sequence and structure binding preferences using deep convolutional and recurrent neural networks, BMC Genomics 19 (2018) 511.
[69] I. Nazari, H. Tayara, K.T. Chong, Branch point selection in RNA splicing using deep learning, IEEE Access 7 (2019) 1800–1807.
[70] M. Oubounyt, Z. Louadi, H. Tayara, K.T. Chong, Deep learning models based on distributed feature representations for alternative splicing prediction, IEEE Access 6 (2018) 58826–58834.
[71] M. Tahir, H. Tayara, K.T. Chong, iRNA-PseKNC (2methyl): identify RNA 2'-O-methylation sites by convolution neural network and Chou's pseudo components, J. Theor. Biol. 465 (2019) 1–6.
[72] J. Ge, Y.-T. Yu, RNA pseudouridylation: new insights into an old modification, Trends Biochem. Sci. 38 (2013) 210–218.
[73] A. Basak, C.C. Query, A pseudouridine residue in the spliceosome core is part of the filamentous growth program in yeast, Cell Rep. 8 (2014) 966–973.
[74] K. Blin, C. Dieterich, R. Wurmus, N. Rajewsky, M. Landthaler, A. Akalin, DoRiNA 2.0—upgrading the doRiNA database of RNA interactions in post-transcriptional regulation, Nucleic Acids Res. 43 (2014) D160–D167.

[75] M. Tahir, M. Hayat, S.A. Khan, iNuc-ext-PseTNC: an efficient ensemble model for identification of nucleosome positioning by extending the concept of Chou's PseAAC to pseudo-tri-nucleotide composition, Mol. Genet. Genom. (2018) 1–12.
[76] 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.
[77] 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.
[78] 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.
[79] C.-Q. Feng, Z.-Y. Zhang, X.-J. Zhu, Y. Lin, W. Chen, H. Tang, H. Lin, iTerm-PseKNC: a sequence-based tool for predicting bacterial transcriptional terminators, Bioinformatics (2018), bty827.
[80] M. Tahir, H. Tayara, K.T. Chong, iRNA-PseKNC(2methyl): identify RNA 2'-O-methylation sites by convolution neural network and Chou's pseudo components, J. Theor. Biol. 465 (2019) 1–6.
[81] F. Chollet, others, Keras: deep learning library for theano and tensorflow, 7, https://keras._io/k,7_(2015), 2015.
[82] M. Hayat, A. Khan, Predicting membrane protein types by fusing composite protein sequence features into pseudo amino acid composition, J. Theor. Biol. 271 (2011) 10–17.
[83] 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.
[84] 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.
[85] 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.
[86] 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.
[87] K.C. Chou, Prediction of protein signal sequences and their cleavage sites, Proteins: Structure, Function, and Bioinformatics 42 (2001) 136–139.
[88] K.-C. Chou, Using subsite coupling to predict signal peptides, Protein Eng. 14 (2001) 75–79.
[89] K.-C. Chou, Prediction of signal peptides using scaled window, Peptides 22 (2001) 1973–1979.
[90] 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.
[91] X. Cheng, X. Xiao, K.-C. Chou, pLoc-mVirus: predict subcellular localization of multi-location virus proteins via incorporating the optimal GO information into general PseAAC, Gene 628 (2017) 315–321.
[92] X. Cheng, W.-Z. Lin, X. Xiao, K.-C. Chou, pLocbal-mAnimal: predict subcellular localization of animal proteins by balancing training dataset and PseAAC, Bioinformatics 35 (3) (01 February 2019) 398–406.
[93] X. Cheng, S.G. Zhao, X. Xiao, K.C. Chou, iATC-mISF: a multi-label classifier for predicting the classes of anatomical therapeutic chemicals (vlo 33, pg 341, 2017), Bioinformatics 33 (2017), 2610-2610.
[94] X. Cheng, S.G. Zhao, X. Xiao, K.C. Chou, iATC-mHyb: a hybrid multi-label classifier for predicting the classification of anatomical therapeutic chemicals, Oncotarget 8 (2017) 58494–58503.
[95] K.C. Chou, Some remarks on predicting multi-label attributes in molecular biosystems, Mol. Biosyst. 9 (2013) 1092–1100.
[96] X. Cheng, X. Xiao, K.C. Chou, pLoc-mHum: predict subcellular localization of multi-location human proteins via general PseAAC to winnow out the crucial GO information, Bioinformatics 34 (2018) 1448–1456.
[97] K.C. Chou, An unprecedented revolution in medicinal chemistry driven by the progress of biological science, Curr. Top. Med. Chem. 17 (2017) 2337–2358.