Chemometrics and Intelligent Laboratory Systems 193 (2019) 103811

Contents lists available at ScienceDirect

Chemometrics and Intelligent Laboratory Systems

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

Tutorial article

iN6-Methyl (5-step): Identifying RNA N6-methyladenosine sites using deep learning mode via Chou's 5-step rules and Chou's general PseKNC

Iman Nazari [a], Muhammad Tahir [a,b], Hilal Tayara [a,**], Kil To Chong [a,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:
N6-methyladenosine site
Deep learning
Convolution neural network
RNA methylation
word2vec

ABSTRACT

N6-methyladenosine (m^6A) is an RNA methylation modification and it is involved in various biological progresses such as translation, alternative splicing, degradation, stability, etc. Therefore, it is highly recommended to develop computational models for detecting N6-methyldenosine sites in RNA as experimental technologies, such as m^6A-seq and MeRIP-Seq, are both expensive and time consuming. Previous works start with features design step, which requires domain knowledge, followed by a classifier or cascade of classifiers for m^6A sites identification. In this paper, on the other hand, we utilize an automatic feature learning approach based on the widely used natural language technique "word2vec". The learnt features are extracted automatically from the human genome without any explicit definition. Then, these learnt features are fed to a simple convolution neural network model for classification. The proposed model is denoted as "iN6-Methyl (5-step)". It has been evaluated on three publicly available benchmark datasets and outperformed the current state-of-the-art methods. It is anticipated that the proposed model could be helpful for both academia and drug discovery. Finally, a user-friendly web-server has been established and made freely available at: https://home.jbnu.ac.kr/NSCL/iN6-Methyl.htm.

1. Introduction

N6-methyladenosine (m^6A) is the most frequent RNA modification that exists in various species [1,2]. It plays important roles in various biological processes such as alternative splicing [3], regulation of circadian clock [4], cell differentiation and reprogramming [5], primary microRNA processing [6], and RNA structural dynamics [7]. The m^6A is found at mRNA [8], tRNA, rRNA, small nuclear RNA, and long non-coding RNA [2,9,10]. It also exists in archaea, viruses, bacteria, and most eukaryotes such as yeast, plants, and mammals [11–15] Therefore, identifying m^6A is important to understand their functional mechanisms. Recently, high-throughput experiments such as m^6A-seq [16] and MeRIP-Seq [17] provided a genome-wide m^6A profiles for various species such as *Homo sapiens*, *Mus musculus* [18], and *Saccharomyces cerevisiae* [19]. Based on these experimental findings, it was revealed that m^6A sites are more likely to occur within long internal exons, in 3' UTR, and near the stop codon, [17,18]. In addition, the nonrandom existence of m^6A sites across the genome is conserved from yeast to human. Therefore, it is an essential and important for species [18,19]. On the other hand m^6A-seq and MeRIP-Seq experiments are expensive and not accurate enough. Therefore, it is important to develop reliable computational tools for identifying m^6A sites. In recent years, several types of post transcription modification (PTM) have been studied such as ([20–46]).

Recently, machine learning based approaches have been used for developing computational tools for m^6A site identification. "iRNA-Methyl" was developed by Ref. [47] for m^6A site identification. In this method, sequence-order information using PseDNC (pseudo dinucleotide composition) [48] and physicochemical properties are used for feature extraction then followed by support vector machine [49,50]. More physicochemical properties have been added with a scalable transformation algorithm for a better feature extraction by Ref. [51]. It was suggested by Refs. [52,53] that using different types of feature descriptors could improve the performance of m^6A site identification models. Jia et al. [52] improved the performance by incorporating three types of feature descriptors such as dinucleotide composition, bi-profile Bayes, and KNN scores. On the other hand, Xiang et al. [53] merged k-mer frequency and binary encoding scheme to improve the

* Corresponding author. Department of Electronics and Information Engineering, Chonbuk National University, Jeonju, 54896, South Korea.
** 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.2019.103811
Received 7 February 2019; Received in revised form 2 May 2019; Accepted 23 July 2019
Available online 26 July 2019

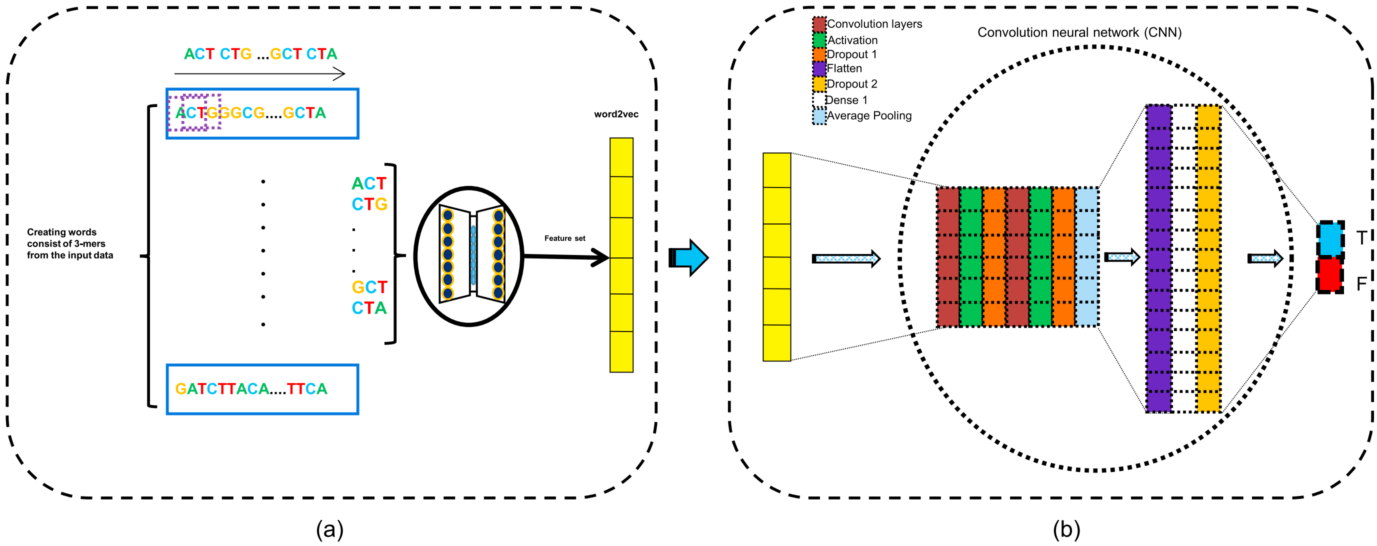


Fig. 1. Illustration of iN6-Methyl (5-step) model. (a) Feature representation by Word2vec model (b) Convolution neural network for classification task.

performance. Recently, a powerful tool, "SRAMP", was proposed by Ref. [54]. In this work various feature extraction techniques have been utilized such as k-nearest neighbor encoding, secondary structure pattern, positional binary encoding of nucleotide sequence, and binary representation of nucleotide sequence. Then random forest model was trained based on the extracted features and the performance outperformed the other methods. Xiang et al. [53] proposed "RNAMethyPre" predictor that was based on position-specific and compositional information for m^{6}A sites on both mouse and human. Most of the previously mentioned predictors are species-specific. However, Qiang et al. [55] proposed a multiple species predictor for m^{6}A sites. They used Local position-specific dinucleotide frequency and dinucleotide binary encoding as features extraction and enhanced them using sequential forward search and F-score algorithm. Then, XGBoost algorithm was used to construct the predictive model. Generally, several m^{6}A predictors have been proposed such as m6Apred M6APred-EL [56], RFAthM6A [57], iRNA(m6A)-PseDNC [58], iRNA-PseColl [37],iRNA-3typeA [59], DeepM6ASeq [60], SRAMP [54], RNAMethPre [53], and BERMP [61].

In general, all of the proposed predictors require domain knowledge to manually design the features. These features should be designed in a way that the sequence-pattern information is preserved. For instance, pseudo amino acid composition [62] or PseAAC [63] is a good example for feature extraction technique. The popularity of this concept has led to developing open source soft-wares such as 'PseAAC-Builder' [64], 'propy' [65], and 'PseAAC-General' [66]. Later, PseAAC was extended to PseKNC (Pseudo K-tuple Nucleotide Composition) [67] to obtain numerical features from DNA/RNA sequences [68,69]. The PseKNC has been constructed in web-servers such as Pse-in-One [70] and 'Pse-in-One2.0' [71].

On the other hand, deep learning based predictors enable designing powerful tools from raw RNA/DNA sequences without handcrafting the features such as DeepCpG [72], iDeepS [73], branch point selection [74], alternative splicing sites prediction [75], 2′-Omethylation sites prediction [76], and other biological processes [77–80]. Deep learning based predictors for m^{6}A such as DeepM6ASeq [60] and BERMP [61] have extracted the features from the raw m6A sites using CNN and RNN. However, we learn the new representation for the m6A sites using word2vec algorithm and then utilize the new representation for m6A identification. The learnt features from word2vec are more comprehensive as they are based on the whole mRNA rather than small set of RNA/DNA samples. In this paper, we propose a novel multiple-species sequence-based predictor, namely "iN6-Methyl (5-step)", for identifying m^{6}A sites in RNA sequences. It consists of two steps. The first step is the feature representation stage in which each sequence is divided into words (3-mer) then a natural language processing models called word2vec is applied in order to map each word to its corresponding feature representation. The second step is a deep learning computational model that predicts the m^{6}A sites based on the generated features of the first step word2vec. The achieved results outperform the state-of-the-art methods in all evaluation metrics. In addition, a user-friendly webserver for m^{6}A prediction is established and made available at: https://home.jbnu.ac.kr/NSCL/iN6-Methyl.htm.

In this work, we follow the Chou's 5-step rules [81] similar to the previous studies [82–98]. The 5-step rules are benchmark dataset construction [82,83,92], mathematical formulation of the samples of the dataset, prediction engine design, performing cross-validation tests for evaluating the performance of the predictor engine, and finally, web-server construction.

2. Materials and methods

2.1. Benchmark datasets

In order to predict m^{6}A sites in multiple species, we use three benchmark datasets for three different species namely *Saccharomyces cerevisiae* (S51) [47], Homo sapiens (H41) [99], and Mus musculus (M41) [18]. The datasets S51, H41, and M41 contain 2614, 2260, and 1450 samples, respectively, and the length of each sample in S51 dataset is 51 nt and it is 41 nt for H41 and M41 datasets. Each sample of these datasets is centered on the m^{6}A site for the positive sequences, whilst the negative sequences prepared by adenines at the center without having biologically m^{6}A peak. As a quality control, we utilize 10-fold cross-validation in the training process. In this case, we randomly split the dataset into 10 folds. Nine folds are used for training and early stopping and the remaining fold is used for testing.

2.2. Methodology

We present a novel method in order to finding and predicting m^{6}A sites in different species called iN6-Methyl (5-step) model. Our proposed method consists of two major steps. The first step is the feature representation stage in which each sequence is divided into words (3-mer) then a natural language processing models called word2vec is applied in order to map each word to its corresponding feature representation. The

Table 1
Word2vec training parameters.

Parameters	Word2vec model
Training Method	CBOW
Vector Size	100
Corpus	Human Genome
Context Words	3-mers
Window Size	5
Minimum Count	5
Negative sampling	5
Epochs	20

second step is a deep learning computational model that predicts the m^{6}A sites based on the generated features of the first step. This process is illustrated in Fig. 1 and is described in details in the following sections.

2.2.1. Distributed feature representation

The existing approaches for m^{6}A sites identification require domain-knowledge to hand-craft the input features of the classification models. In this work, we aim to build a computational model that can learn features representation automatically based on the genomic data. This technique helps in obtaining more optimal features by reducing the noise in the data and, consequently, improving the performance of the final computational model.

Genetic data is considered as a language, that is represented in DNA and RNA sequences, by which the information passes within and between the cells [100–102]. It is based on a continues chain of nucleotides (A, C, G, and T). In addition, NLP techniques have been used successfully in various biological problems such as alternative splicing site prediction [75].

Thus, we utilize NLP model "word2vec" to get interpretable representations for m^{6}A sites Fig. 1 (a). The first step in word2vec is corpus construction. In this step we split the continuous genomic sequences into words represented by overlapped k-mer to break its continuity. In our model we empirically set $k = 3$. This selection performs better than using other values of k such as 4-mer,5-mer, 6-mer, etc. This selection confirms the previous findings of [75,103] in which setting $k = 3$ was the best choice. In addition, 3-mer has been widely used in DNA/RNA sequence formulation [104,105]. Thus, The constructed corpus has four different nucleotides (A, C, G, and T) and consequently forms 64 unique words ($4^3 = 64$). For instance, the biological sequence {ACAGAATG} results in the following words {ACA, CAG, AGA, GAA, AAT, and ATG}. The generated corpus for each sequence is used for training the word2vec model.

Generally, we use human mRNA from GenBank which is available at: hgdownload.soe.ucsc.edu. The genome assembly is divided into 21 chromosomes (Chr1, Chr2, …, X, and Y) and each chromosome is then divided into sentences with length of 100 nt. Finally, each sentence is cut into overlapping 3-mer to create the words. Continuous bag-of-words (CBOW) method is used for training word2vec model. CBOW method predicts the current word w(t) based on the surrounding context words in a predefined window. The detailed parameters that are used for training word2vec model are given in Table 1. These parameters are widely used in genomic data [103]. As a result of using word2vec, each word (3-mer) is represented by a 100-dimensional vector and each sequence with length L is represented by an array of shape $(L-2) \times 100$. Dominissini et al. [18] showed that mammalian m^{6}A have a DRACH consensus motif (D = U, G or A; R = G or A; H = U, C or A) which can denoted in an overlap 3-mer as {AAA, AGA, GGA, UGA,UAA, AAC, GAC, ACA, ACC, ACU, GAA}. These 3-mers are shown in the 2d-space in Fig. 2 from the learnt representation using word2vec. Fig. 2 is obtained by using t-distributed stochastic neighbor embedding (t-SNE) [106].

2.2.2. Deep learning model

The extracted feature representation for each sequence from the first step is used for training the proposed deep learning model which is a a simple and efficient convolution neural network as shown in Fig. 1(b). The grid search algorithm is utilized for searching the best hyper-parameters. The input shape of the proposed model is $(L-2) \times 100$ where L is the length of the input sequence. It consists of two dilated convolution layers [107] where the number of the filters is 32 and the size of the filter is 5 for both of them. The dilation rate is set to $d = 1$ and $d = 2$ for the first and the second convolution layers, respectively. Dilated convolution produces exponentially larger receptive field with less number of layers with comparison to conventional convolution layers. Each layer is followed by rectified linear unit (ReLU) activation function [108] where $ReLU(x) = max(x, 0)$. Alpha dropout is used in order to retain the variance and the mean of the inputs to their original values after applying dropout [109]. The dropout probability is set to 0.2. The generated features of the dilated convolution layers are averaged using average pooling operator with window size equals to 4 and then passed to two fully connected layers. The first layer has 128 nodes and followed by ReLU activation function and alpha dropout with probability of 0.2. On the other hand, the second fully connected layer has only one node with sigmoid activation function for prediction.

3. Results and discussion

In this section we introduce evaluation metrics, the obtained results, and the comparison with the state-of-the-art methods.

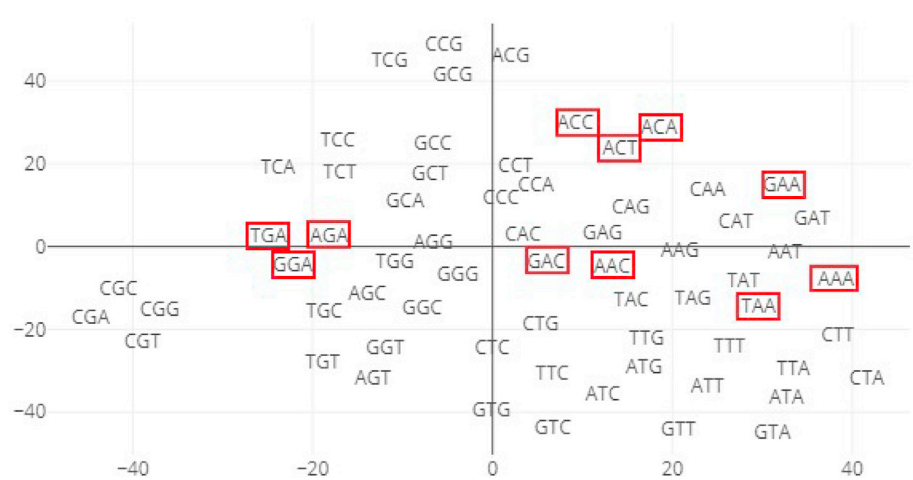


Fig. 2. Visualization of word2vec features using tSNE. The highlighted words show the important 3-mer in m^{6}A sites prediction.

Table 2
The performance of the proposed model using different values of Kmer

Dataset	K-mers	ACC	Sn	Sp	MCC
S51	3-mers	75.38%	76.15%	74.62%	0.5078
M41		89.51%	78.87%	100.0%	0.8079
H41		91.11%	82.14%	100.0%	0.8354
S51	4-mers	70.0%	70.77%	69.23%	0.40
M41		88.81%	79.17%	98.59%	0.7918
H41		90.62%	82.14%	99.11%	0.8244
S51	5-mers	66.92%	69.23%	64.62%	0.3388
M41		88.19%	79.17%	97.22%	0.7767
H41		90.18%	82.14%	98.21%	0.8142
S51	6-mers	68.73%	72.09%	65.38%	0.3756
M41		88.28%	79.10%	97.15%	0.7720
H41		89.33%	81.42%	97.32%	0.7971

3.1. Evaluation metrics

In this work, we use accuracy (ACC), sensitivity (Sn), specificity (Sp), and Matthew correlation coefficient (MCC) based on Chou's symbols that were introduced in Refs. [62,110] and derived in Refs. [48,111]. These metrics were widely used in the recent publications [24,28,37,48,69,82,90–92,112–119].

$$Sn = 1 - \frac{P_-^+}{P^+} \tag{1}$$

$$Sp = 1 - \frac{P_+^-}{P^-} \tag{2}$$

$$ACC = 1 - \frac{P_-^+ + P_+^-}{P^+ + P^-} \tag{3}$$

$$MCC = \frac{1 - \frac{P_-^+ + P_+^-}{P^+ + P^-}}{\sqrt{\left(1 + \frac{P_+^- - P_-^+}{P^+}\right)\left(1 + \frac{P_-^+ - P_+^-}{P^-}\right)}} \tag{4}$$

where P^+ is the total portion of m^{6}A investigated while P_-^+ is the portion of m^{6}A incorrectly predicted as non m^6 sequences. P^- is the total portion of non m^{6}A investigated while P_+^- is the portion of non m^{6}A sequences incorrectly predicted as m^6 ones.

In addition, The area under receiver operating characteristic (ROC-AUC) curves, a graphical form for visualizing the performance of the proposed models, is used. The larger the AUC the better model's performance.

3.2. Results and comparison

As described in Section 2.1, the proposed model is evaluated on three datasets S51, H41, and M41. In order to study the effect of using different values of k-mer we test 3-mer, 4-mer, 5-mer, and 6-mer as shown in Table 2. The results show that using 3-mer produces the best performance on the three datasets compared with the other values of k-mer. These results confirm the finding of the previous studies in which 3-mer was the best performing selection [75,103]. Fig. 3 shows the confusion matrix results of S51, H41, and M41. It can be seen that iN6-Methyl (5-step) model performs better in the case of H41 and M41 datasets than S51 dataset. Fig. 4 shows the achieved AUC for S51, H41, and M41. It can be observed that H41 and M41 have AUC of 90.30% and 91.33%, receptively. while the AUC of S51 is 80.31%.

In addition we compare the results of the proposed model with the state-of-the-art-models pRNAm-PC [51] and M6AMRFS [55] using the same 10-fold cross-validation tests. Fig. 5 show the performance of the proposed model with comparison with other classifiers in terms of ACC, SP, SN, and MCC. It can be seen that iN6-Methyl (5-step) outperforms the other methods as shown in Fig. 5 and Table 3.

More specifically, the accuracy of iN6-Methyl (5-step) is improved by 1.13%, 0.09%, and 1.12% for S51, H41, and M41 datasets, respectively. The sensitivity is improved by 0.94% and 0.10% for S51 and H41 datasets, respectively. The specificity is improved by 1.32% for S51. MCC is also improved by 2.26%, 0.15%,and 21.99% for S51, H41, and M41 datasets, respectively. Thus, we achieve a big improvement in the case of M41 dataset.

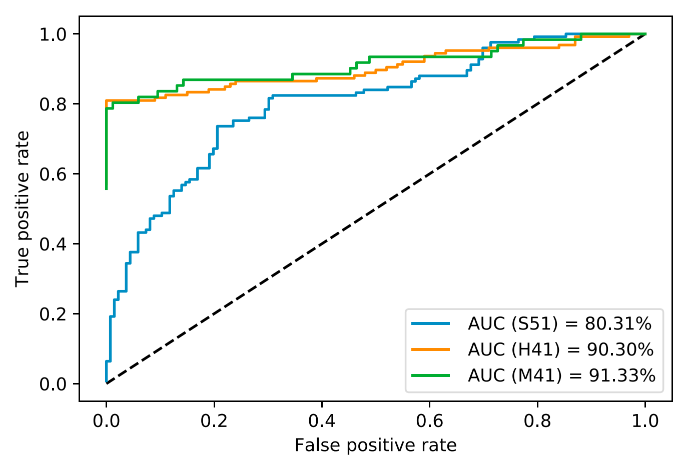
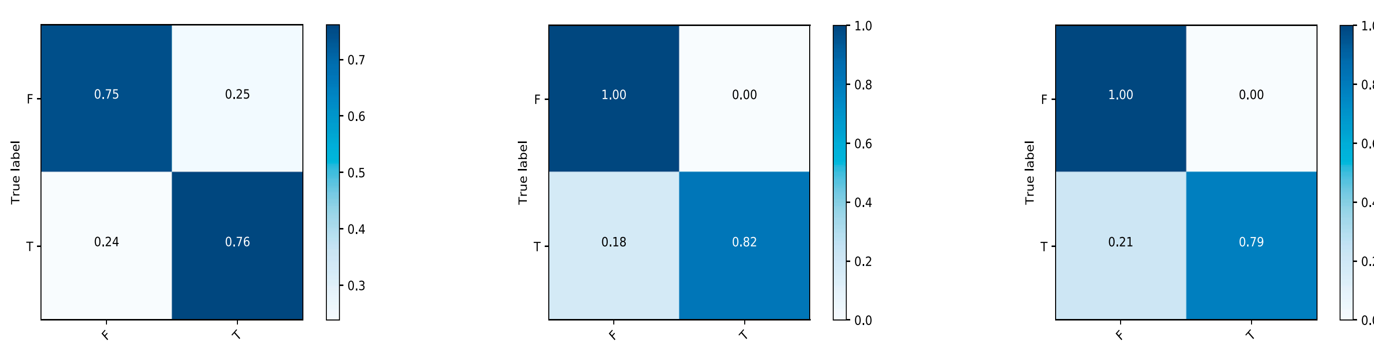


Fig. 4. The AUC curves the proposed model iN6-Methyl (5-step) on three benchmarks S51, H41, and M41.



(a) (b) (c)

Fig. 3. Confusion matrix of the proposed model iN6-Methyl (5-step) on three benchmarks (a) S51, (b) H41, and (c) M41.

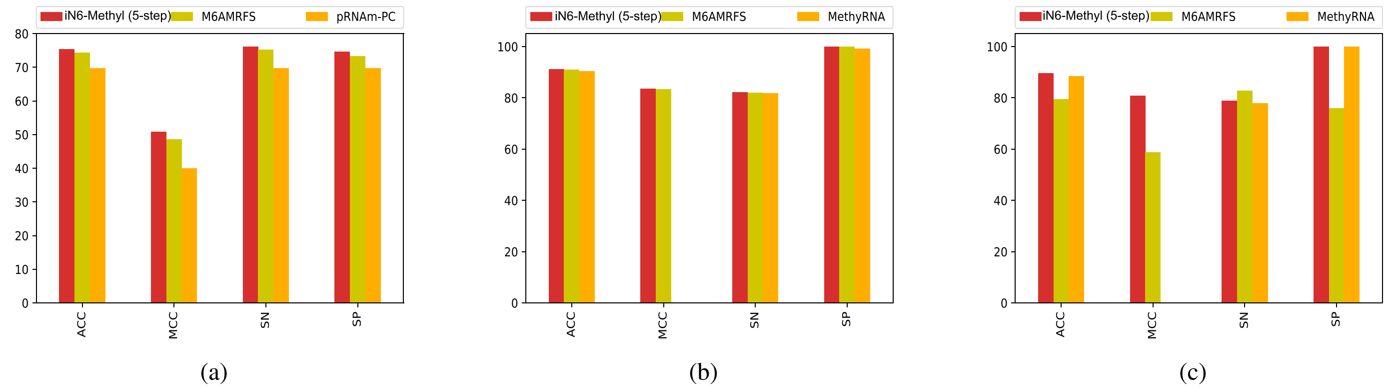


Fig. 5. Performance of iN6-Methyl (5-step) model and other classifiers on three benchmarks (a) S51, (b) H41, and (c) M41.

Table 3
Performances of iN6-Methyl (5-step) and other algorithms.

Dataset	Model	ACC	SN	SP	MCC
S51	pRNAm-PC	69.74%	69.72%	69.75%	0.40
	M6AMRFS	74.25%	75.21%	73.30%	0.4852
	iN6-Methyl (5-step)	**75.38%**	**76.15%**	**74.62%**	**0.5078**
H41	MethyRNA	90.38%	81.68%	99.11%	N.A%
	M6AMRFS	91.02%	82.04%	100.0%	0.8339
	iN6-Methyl (5-step)	**91.11%**	**82.14%**	**100.0%**	**0.8354**
M41	MethyRNA	88.39%	77.79%	100.0%	N.A%
	M6AMRFS	79.33%	82.81%	75.84%	0.5880
	iN6-Methyl (5-step)	**89.51%**	**78.87%**	**100.0%**	**0.8079**

These results indicate that using word2vec to extract the feature from raw genomic sequences enhances the performance of m^6A prediction model. The learnt features using word2vec are extracted from the whole mRNA which are more comprehensive compared with the hand-crafted features used by the previous state-of-the-art models such as pRNAm-PC [51] and M6AMRFS [55].

3.3. Web server

It is highly recommended to construct a web-server that makes the developed tool accessible by the research community [37,41,42,48,82,113,116–118,120–122,122–129]. Therefore, we have developed a user-friendly and easy-to-use web-server and made it available at https://home.jbnu.ac.kr/NSCL/iN6-Methyl.htm. The web-server has been built by Python and Flask library.

4. Conclusion

In this study, we have proposed a novel deep learning based model, called iN6-Methyl (5-step), for the identification of m^6A sites in multiple species. It consists of two steps namely features extraction and classification. We have adopted wrod2vec in order to automatically extract the features form raw genmoics sequences then a simple and efficient deep learning model based on dilated convolution neural network has been used for classifying the m^6A sites. The obtained results outperformed the state-of-the-art models in all evaluation metrics i.e. accuracy, sensitivity, specificity, and Matthew correlation coefficient. Finally, a user friendly webservr is made available for m^6A sites identification in multiple species at https://home.jbnu.ac.kr/NSCL/iN6-Methyl.htm.

Conflicts of interest

The authors declare no conflict of interest.

Funding

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

Data availability statement

The datasets generated for this study is freely available at: https://home.jbnu.ac.kr/NSCL/iN6-Methyl.htm.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.chemolab.2019.103811.

References

[1] T. Chen, Y.-J. Hao, Y. Zhang, M.-M. Li, M. Wang, W. Han, Y. Wu, Y. Lv, J. Hao, L. Wang, et al., m6 a rna methylation is regulated by micrornas and promotes reprogramming to pluripotency, Cell Stem Cell 16 (3) (2015) 289–301.
[2] B. Maden, The numerous modified nucleotides in eukaryotic ribosomal rna, in: Progress in Nucleic Acid Research and Molecular Biology, vol 39, Elsevier, 1990, pp. 241–303.
[3] N. Liu, Q. Dai, G. Zheng, C. He, M. Parisien, T. Pan, N 6-methyladenosine-dependent rna structural switches regulate rna–protein interactions, Nature 518 (7540) (2015) 560.
[4] J.-M. Fustin, M. Doi, Y. Yamaguchi, H. Hida, S. Nishimura, M. Yoshida, T. Isagawa, M.S. Morioka, H. Kakeya, I. Manabe, et al., Rna-methylation-dependent rna processing controls the speed of the circadian clock, Cell 155 (4) (2013) 793–806.
[5] S. Geula, S. Moshitch-Moshkovitz, D. Dominissini, A.A. Mansour, N. Kol, M. Salmon-Divon, V. Hershkovitz, E. Peer, N. Mor, Y.S. Manor, et al., m6a mrna methylation facilitates resolution of naïve pluripotency toward differentiation, Science 347 (6225) (2015) 1002–1006.
[6] C.R. Alarcón, H. Lee, H. Goodarzi, N. Halberg, S.F. Tavazoie, N 6-methyladenosine marks primary micrornas for processing, Nature 519 (7544) (2015) 482.
[7] C. Roost, S.R. Lynch, P.J. Batista, K. Qu, H.Y. Chang, E.T. Kool, Structure and thermodynamics of n6-methyladenosine in rna: a spring-loaded base modification, J. Am. Chem. Soc. 137 (5) (2015) 2107–2115.
[8] Y. Yue, J. Liu, C. He, Rna n6-methyladenosine methylation in post-transcriptional gene expression regulation, Genes Dev. 29 (13) (2015) 1343–1355.
[9] J. Gu, J.R. Patton, S. Shimba, R. Reddy, Localization of modified nucleotides in schizosaccharomyces pombe spliceosomal small nuclear rnas: modified nucleotides are clustered in functionally important regions, RNA 2 (9) (1996) 909.
[10] P.F. Agris, F.A. Vendeix, W.D. Graham, trna's wobble decoding of the genome: 40 years of modification, J. Mol. Biol. 366 (1) (2007) 1–13.
[11] K. Beemon, J. Keith, Localization of n6-methyladenosine in the rous sarcoma virus genome, J. Mol. Biol. 113 (1) (1977) 165–179.
[12] M.J. Clancy, M.E. Shambaugh, C.S. Timpte, J.A. Bokar, Induction of sporulation in saccharomyces cerevisiae leads to the formation of n 6-methyladenosine in mrna: a potential mechanism for the activity of the ime4 gene, Nucleic Acids Res. 30 (20) (2002) 4509–4518.
[13] R. Desrosiers, K. Friderici, F. Rottman, Identification of methylated nucleosides in messenger rna from novikoff hepatoma cells, Proc. Natl. Acad. Sci. 71 (10) (1974) 3971–3975.

[14] R. Levis, S. Penman, 5'-terminal structures of poly (a)+ cytoplasmic messenger rna and of poly (a)+ and poly (a)- heterogeneous nuclear rna of cells of the dipteran drosophila melanogaster, J. Mol. Biol. 120 (4) (1978) 487–515.
[15] J. Nichols, 'cap'structures in maize poly (a)-containing rna, Biochim. Biophys. Acta Nucleic Acids Protein Synth. 563 (2) (1979) 490–495.
[16] D. Dominissini, S. Moshitch-Moshkovitz, M. Salmon-Divon, N. Amariglio, G. Rechavi, Transcriptome-wide mapping of n 6-methyladenosine by m 6 a-seq based on immunocapturing and massively parallel sequencing, Nat. Protoc. 8 (1) (2013) 176.
[17] K.D. Meyer, Y. Saletore, P. Zumbo, O. Elemento, C.E. Mason, S.R. Jaffrey, Comprehensive analysis of mrna methylation reveals enrichment in 3' utrs and near stop codons, Cell 149 (7) (2012) 1635–1646.
[18] D. Dominissini, S. Moshitch-Moshkovitz, S. Schwartz, M. Salmon-Divon, L. Ungar, S. Osenberg, K. Cesarkas, J. Jacob-Hirsch, N. Amariglio, M. Kupiec, et al., Topology of the human and mouse m 6 a rna methylomes revealed by m 6 a-seq, Nature 485 (7397) (2012) 201.
[19] S. Schwartz, S.D. Agarwala, M.R. Mumbach, M. Jovanovic, P. Mertins, A. Shishkin, Y. Tabach, T.S. Mikkelsen, R. Satija, G. Ruvkun, et al., High-resolution mapping reveals a conserved, widespread, dynamic mrna methylation program in yeast meiosis, Cell 155 (6) (2013) 1409–1421.
[20] 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 Eng. Des. Sel. 26 (11) (2013) 735–742.
[21] 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 (6) (2014) 10410–10423.
[22] 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), 947416.
[23] 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 (5) (2014) 7594–7610.
[24] 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 (8) (2014), e105018.
[25] 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 (7) (2014) 11204–11219.
[26] 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 (8) (2015) 1731–1742.
[27] W. Chen, H. Tang, J. Ye, H. Lin, K.-C. Chou, irna-pseu: identifying rna pseudouridine sites, Mol. Ther. Nucleic Acids 5 (2016) e332.
[28] 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.
[29] 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.
[30] 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 (23) (2016) 34558.
[31] 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 (20) (2016) 3133–3141.
[32] 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.
[33] 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 (28) (2016) 44310.
[34] 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 (20) (2016) 3116–3123.
[35] W.-R. Qiu, X. Xiao, Z.-C. Xu, K.-C. Chou, iphos-pseen: identifying phosphorylation sites in proteins by fusing different pseudo components into an ensemble classifier, Oncotarget 7 (32) (2016) 51270.
[36] Y. Xu, K.-C. Chou, Recent progress in predicting posttranslational modification sites in proteins, Curr. Top. Med. Chem. 16 (6) (2016) 591–603.
[37] 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.
[38] 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.
[39] 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 (6) (2017) 552–559.
[40] 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 (8) (2017) 734–743.
[41] 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 (25) (2017) 41178.
[42] 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, Mol. Inf. 36 (5–6) (2017) 1600010.
[43] 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 (6) (2017) 544–551.
[44] 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.
[45] A. Chandra, A. Sharma, A. Dehzangi, S. Ranganathan, A. Jokhan, K.-C. Chou, T. Tsunoda, Phoglystruct: prediction of phosphoglycerylated lysine residues using structural properties of amino acids, Sci. Rep. 8 (1) (2018) 17923.
[46] L. Wang, R. Zhang, Y. Mu, Fu-sulfpred: identification of protein s-sulfenylation sites by fusing forests via chou's general pseaac, J. Theor. Biol. 461 (2019) 51–58.
[47] 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.
[48] W. Chen, P.-M. Feng, H. Lin, K.-C. Chou, irspot-psednc: identify recombination spots with pseudo dinucleotide composition, Nucleic Acids Res. 41 (6) (2013) e68–e68.
[49] S. Ding, X. Zhang, Y. An, Y. Xue, Weighted linear loss multiple birth support vector machine based on information granulation for multi-class classification, Pattern Recognit. 67 (2017) 32–46.
[50] X. Zhang, S. Ding, Y. Xue, An improved multiple birth support vector machine for pattern classification, Neurocomputing 225 (2017) 119–128.
[51] 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.
[52] C.-Z. Jia, J.-J. Zhang, W.-Z. Gu, Rna-methylpred: a high-accuracy predictor to identify n6-methyladenosine in rna, Anal. Biochem. 510 (2016) 72–75.
[53] S. Xiang, K. Liu, Z. Yan, Y. Zhang, Z. Sun, Rnamethpre: a web server for the prediction and query of mrna m6a sites, PLoS One 11 (10) (2016) 1–13, https://doi.org/10.1371/journal.pone.0162707. URL, https://doi.org/10.1371/journal.pone.0162707.
[54] 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 (10) (2016) e91, https://doi.org/10.1093/nar/gkw104, arXiv:/oup/backfile/content_public/journal/nar/44/10/10.1093_nar_gkw104/4/gkw104.pdf, https://doi.org/10.1093/nar/gkw104. URL.
[55] X. Qiang, H. Chen, X. Ye, R. Su, L. Wei, M6amrfs: robust prediction of n6-methyladenosine sites with sequence-based features in multiple species, Front. Genet. 9 (2018) 495, https://doi.org/10.3389/fgene.2018.00495. URL, https://www.frontiersin.org/article/10.3389/fgene.2018.00495.
[56] L. Wei, H. Chen, R. Su, M6apred-el: a sequence-based predictor for identifying n6-methyladenosine sites using ensemble learning, Mol. Ther. Nucleic Acids 12 (2018) 635–644.
[57] X. Wang, R. Yan, Rfathm6a: a new tool for predicting m 6 a sites in arabidopsis thaliana, Plant Mol. Biol. 96 (3) (2018) 327–337.
[58] 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.
[59] 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.
[60] Y. Zhang, M. Hamada, Deepm6aseq: prediction and characterization of m6a-containing sequences using deep learning, BMC Bioinf. 19 (19) (2018) 524.
[61] Y. Huang, N. He, Y. Chen, Z. Chen, L. Li, Bermp: a cross-species classifier for predicting m6a sites by integrating a deep learning algorithm and a random forest approach, Int. J. Biol. Sci. 14 (12) (2018) 1669.
[62] K.-C. Chou, Prediction of protein cellular attributes using pseudo-amino acid composition, Proteins: Struct. Funct. Bioinform. 43 (3) (2001) 246–255.
[63] K.-C. Chou, Using amphiphilic pseudo amino acid composition to predict enzyme subfamily classes, Bioinformatics 21 (1) (2004) 10–19.
[64] P. Du, X. Wang, C. Xu, Y. Gao, Pseaac-builder: a cross-platform stand-alone program for generating various special chou's pseudo-amino acid compositions, Anal. Biochem. 425 (2) (2012) 117–119.
[65] D.-S. Cao, Q.-S. Xu, Y.-Z. Liang, propy: a tool to generate various modes of chou's pseaac, Bioinformatics 29 (7) (2013) 960–962.
[66] P. Du, S. Gu, Y. Jiao, Pseaac-general: fast building various modes of general form of chou's pseudo-amino acid composition for large-scale protein datasets, Int. J. Mol. Sci. 15 (3) (2014) 3495–3506.
[67] 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.
[68] W. Chen, H. Lin, K.-C. Chou, Pseudo nucleotide composition or pseknc: an effective formulation for analyzing genomic sequences, Mol. Biosyst. 11 (10) (2015) 2620–2634.
[69] B. Liu, F. Yang, D.-S. Huang, K.-C. Chou, ipromoter-2l: a two-layer predictor for identifying promoters and their types by multi-window-based pseknc, Bioinformatics 34 (1) (2017) 33–40.
[70] 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 (W1) (2015) W65–W71.
[71] B. Liu, H. Wu, K.-C. Chou, Pse-in-one 2.0: an improved package of web servers for generating various modes of pseudo components of dna, rna, and protein sequences, Nat. Sci. 9 (04) (2017) 67.

[72] C. Angermueller, H.J. Lee, W. Reik, O. Stegle, Deepcpg: accurate prediction of single-cell dna methylation states using deep learning, Genome Biol. 18 (1) (2017) 67.

[73] 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 (1) (2018) 511.

[74] I. Nazari, H. Tayara, K.T. Chong, Branch point selection in rna splicing using deep learning, IEEE Access 7 (2019) 1800–1807, https://doi.org/10.1109/ACCESS.2018.2886569.

[75] 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, https://doi.org/10.1109/ACCESS.2018.2874208.

[76] 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. URL, https://doi.org/10.1016/j.jtbi.2018.12.034, http://www.sciencedirect.com/science/article/pii/S0022519318306349.

[77] 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. (2019).

[78] M. Tahir, H. Tayara, K. T. Chong, Ipseu-Cnn: Identifying Rna Pseudouridine Sites Using Convolutional Neural Networks, Molecular Therapy-Nucleic Acids.

[79] H. Tayara, M. Tahir, K. T. Chong, Iss-Cnn: Identifying Splicing Sites Using Convolution Neural Network, Chemometr. Intell. Lab. Syst..

[80] M. Oubounyt, Z. Louadi, H. Tayara, K.T. Chong, Deepromoter: robust promoter predictor using deep learning, Front. Genet. 10 (2019).

[81] K.-C. Chou, Some remarks on protein attribute prediction and pseudo amino acid composition, J. Theor. Biol. 273 (1) (2011) 236–247.

[82] 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 (1) (2019) 96–102.

[83] 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.

[84] 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 (1) (2018) 50–58.

[85] 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, et al., Quokka: a comprehensive tool for rapid and accurate prediction of kinase family-specific phosphorylation sites in the human proteome, Bioinformatics 34 (24) (2018) 4223–4231.

[86] 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 (2) (March 2019) 638–658.

[87] J. Wang, J. Li, B. Yang, R. Xie, T.T. Marquez-Lago, A. Leier, M. Hayashida, T. Akutsu, Y. Zhang, K.-C. Chou, et al., Bastion3: a two-layer ensemble predictor of type iii secreted effectors, Bioinformatics 10 (2019).

[88] 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, https://doi.org/10.1016/j.ygeno.2018.12.001.

[89] 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. 5 (2018).

[90] Y.D. Khan, M. Jamil, W. Hussain, N. Rasool, S.A. Khan, K.-C. Chou, pssbond-pseaac: prediction of disulfide bonding sites by integration of pseaac and statistical moments, J. Theor. Biol. 463 (2019) 47–55.

[91] 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.

[92] M. Zhang, F. Li, T.T. Marquez-Lago, A. Leier, C. Fan, C.K. Kwoh, K.-C. Chou, J. Song, C. Jia, Multiply: a novel multi-layer predictor for discovering general and specific types of promoters, Bioinformatics (2019), btz016, https://doi.org/10.1093/bioinformatics/btz016.

[93] 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 (8) (2016) 2587–2593.

[94] 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.

[95] 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. 294 (1) (2019) 199–210.

[96] 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 (12) (2018) 6719–6727.

[97] 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.

[98] M. Hayat, M. Tahir, Psofuzzysvm-tmh: identification of transmembrane helix segments using ensemble feature space by incorporated fuzzy support vector machine, Mol. Biosyst. 11 (8) (2015) 2255–2262.

[99] W. Chen, H. Tang, H. Lin, Methyrna: a web server for identification of n6-methyladenosine sites, J. Biomol. Struct. Dyn. 35 (3) (2017) 683–687.

[100] C. Emmeche, J. Hoffmeyer, From language to nature: the semiotic metaphor in biology, Semiotica 84 (1–2) (1991) 1–42.

[101] D.B. Searls, String variable grammar: a logic grammar formalism for the biological language of dna, J. Log. Program. 24 (1–2) (1995) 73–102.

[102] M.D. Yandell, W.H. Majoros, Genomics and natural language processing, Nat. Rev. Genet. 3 (2002) 601. EP –, review Article. URL, https://doi.org/10.1038/nrg861.

[103] E. Asgari, M.R. Mofrad, Continuous distributed representation of biological sequences for deep proteomics and genomics, PLoS One 10 (11) (2015), e0141287.

[104] 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 35 (9) (1 May 2019) 1469–1477.

[105] 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 (12) (June 2019) 2075–2083.

[106] L. v. d. Maaten, G. Hinton, Visualizing data using t-sne, J. Mach. Learn. Res. 9 (Nov) (2008) 2579–2605.

[107] F. Yu, V. Koltun, Multi-scale Context Aggregation by Dilated Convolutions, arXiv preprint arXiv:1511.07122.

[108] V. Nair, G.E. Hinton, Rectified linear units improve restricted Boltzmann machines, in: Proceedings of the 27th International Conference on Machine Learning, ICML-10, 2010, pp. 807–814.

[109] G. Klambauer, T. Unterthiner, A. Mayr, S. Hochreiter, Self-normalizing neural networks, CoRR abs/1706.02515, arXiv:1706.02515. URL, http://arxiv.org/abs/1706.02515.

[110] K.-C. Chou, Using subsite coupling to predict signal peptides, Protein Eng. 14 (2) (2001) 75–79.

[111] Y. Xu, X.-J. Shao, L.-Y. Wu, N.-Y. Deng, K.-C. Chou, iso-aapair: incorporating amino acid pairwise coupling into pseaac for predicting cysteine s-nitrosylation sites in proteins, PeerJ 1 (2013) e171.

[112] W. Hussain, Y.D. Khan, N. Rasool, S.A. Khan, K.-C. Chou, Sprenylc-pseaac: a sequence-based model developed via chou's 5-steps rule and general pseaac for identifying s-prenylation sites in proteins, J. Theor. Biol. 468 (2019) 1–11.

[113] 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 (21) (2014) 12961–12972.

[114] C.-J. Zhang, H. Tang, W.-C. Li, H. Lin, W. Chen, K.-C. Chou, iori-human: identify human origin of replication by incorporating dinucleotide physicochemical properties into pseudo nucleotide composition, Oncotarget 7 (43) (2016) 69783.

[115] W. Chen, H. Ding, P. Feng, H. Lin, K.-C. Chou, iacp: a sequence-based tool for identifying anticancer peptides, Oncotarget 7 (13) (2016) 16895.

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

[117] B. Liu, S. Wang, R. Long, K.-C. Chou, irspot-el: identify recombination spots with an ensemble learning approach, Bioinformatics 33 (1) (2016) 35–41.

[118] 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 (3) (2017) 4208.

[119] A. Ehsan, K. Mahmood, Y.D. Khan, S.A. Khan, K.-C. Chou, A novel modeling in mathematical biology for classification of signal peptides, Sci. Rep. 8 (1) (2018) 1039.

[120] K.-C. Chou, H.-B. Shen, Recent advances in developing web-servers for predicting protein attributes, Nat. Sci. 1 (02) (2009) 63.

[121] W.-R. Qiu, B.-Q. Sun, X. Xiao, Z.-C. Xu, J.-H. Jia, K.-C. Chou, ikcr-pseens: identify lysine crotonylation sites in histone proteins with pseudo components and ensemble classifier, Genomics 110 (5) (2018) 239–246.

[122] X. Cheng, X. Xiao, K.-C. Chou, ploc-mgneg: predict subcellular localization of gram-negative bacterial proteins by deep gene ontology learning via general pseaac, Genomics 110 (4) (2018) 231–239.

[123] X. Xiao, X. Cheng, G. Chen, Q. Mao, K.-C. Chou, ploc_bal-mgpos: predict subcellular localization of gram-positive bacterial proteins by quasi-balancing training dataset and pseaac, Genomics 111 (4) (July 2019) 886–892.

[124] X. Xiao, X. Cheng, S. Su, Q. Mao, K.-C. Chou, ploc-mgpos: incorporate key gene ontology information into general pseaac for predicting subcellular localization of gram-positive bacterial proteins, Nat. Sci. 9 (09) (2017) 330.

[125] X. Cheng, S.-G. Zhao, X. Xiao, K.-C. Chou, iatc-misf: a multi-label classifier for predicting the classes of anatomical therapeutic chemicals, Bioinformatics 33 (3) (2016) 341–346.

[126] W.-R. Qiu, X. Xiao, K.-C. Chou, irspot-tncpseaac: identify recombination spots with trinucleotide composition and pseudo amino acid components, Int. J. Mol. Sci. 15 (2) (2014) 1746–1766.

[127] X. Cheng, W.-Z. Lin, X. Xiao, K.-C. Chou, ploc_bal-manimal: predict subcellular localization of animal proteins by balancing training dataset and pseaac, Bioinformatics 35 (3) (2018) 398–406.

[128] X. Cheng, X. Xiao, K.-C. Chou, ploc_bal-mgneg: predict subcellular localization of gram-negative bacterial proteins by quasi-balancing training dataset and general pseaac, J. Theor. Biol. 458 (2018) 92–102.

[129] X. Cheng, X. Xiao, K.-C. Chou, ploc_bal-mplant: predict subcellular localization of plant proteins by general pseaac and balancing training dataset, Curr. Pharmaceut. Des. 24 (34) (2018) 4013–4022.