Journal of Theoretical Biology 465 (2019) 1–6

Contents lists available at ScienceDirect

Journal of Theoretical Biology

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

iRNA-PseKNC(2methyl): Identify RNA 2'-O-methylation sites by convolution neural network and Chou's pseudo components

Muhammad Tahir[a,b,1], Hilal Tayara[a,1,*], 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

Article history:
Received 21 November 2018
Revised 15 December 2018
Accepted 23 December 2018
Available online 24 December 2018

Keywords:
Convolution neural network
2'-O-methylation
RNA
Deep learning
CNN
SVM

ABSTRACT

The 2'-O-methylation transferase is involved in the process of 2'-O-methylation. In catalytic processes, the 2-hydroxy group of the ribose moiety of a nucleotide accept a methyl group. This methylation process is a post-transcriptional modification, which occurs in various cellular RNAs and plays a vital role in regulation of gene expressions at the post-transcriptional level. Through biochemical experiments 2'-O-methylation sites produce good results but these biochemical process and exploratory techniques are very expensive. Thus, it is required to develop a computational method to identify 2'-O-methylation sites. In this work, we proposed a simple and precise convolution neural network method namely: iRNA-PseKNC(2methyl) to identify 2'-O-methylation sites. The existing techniques use handcrafted features, while the proposed method automatically extracts the features of 2'-O-methylation using the proposed convolution neural network model. The proposed prediction iRNA-PseKNC(2methyl) method obtained 98.27% of accuracy, 96.29% of sensitivity, 100% of specificity, and 0.965 of MCC on *Home sapiens dataset*. The reported outcomes present that our proposed method obtained better outcomes than existing method in terms of all evaluation parameters. These outcomes show that iRNA-PseKNC(2methyl) method might be beneficial for the academic research and drug design.

1. Introduction

The 2'-O-methylation transferase is involved in the process of 2'-O-methylation. In this whole catalytic process, the 2-hydroxy group of the ribose moiety of a nucleotide accept a methyl group (Kiss, 2002). This methylation process is a post-transcriptional modification which occurs in various cellular RNAs and plays an important role in regulation of gene expressions at the post-transcriptional level (Bachellerie et al., 2002). It has been reported that the accumulation of 2'-O-methylation site around the functional region of ribosomal RNA (rRNA) affects the structure and function of rRNA (Decatur and Fournier, 2002). Moreover, the methylation in the cap structure of mRNA results the RNA sensor Mda5 to distinguish between itself and non-autologous mRNA (Züst et al., 2011). Furthermore, this process of 2'-O-methylation protects the ends of endo-small interfering RNAs, piwi-interacting RNa (piRNA) and microRNA from uridine and exonuclease degradation while regulate specific RNAi pathways (Li et al., 2005). Although the process of 2'-O-methylation in mRNA is still not clearly decipher; therefore, a detailed study is required to explore the mechanism of RNA 2'-O-methylation. The identification of 2'-O-methylation site is the first step in understanding the regulatory mechanism (Ramachandran and Chen, 2008). Recently, many approaches and techniques were developed to analyze the RNA 2'-O-methylation i.e., liquid chromatography coupled with mass spectrometry and two-dimensional thin layer chromatography (Dong et al., 2012). However, a reverse transcription at low dNTP concentration and PCR based method, RTL-P, were proposed to characterize and explore both old and novel 2-O-methylation sites in human rRNA, yeast rRNA, and mouse piRNA.

Through biochemical experiments, the 2'-O-methylation sites achieved good result but the biochemical process and exploratory techniques are expensive and takes a lot of time. Thus, it is needed to propose efficient statistical or computational models to identify 2'-O-methylation sites because of the huge amount of RNA/DNA sequences produced in the post genome area. The main challenge of building computational models for genomic tasks is how to definite the biological sequences with a discrete vector or model as most of machine learning algorithms require vectors representa-

* Corresponding author.
E-mail addresses: hilaltayara@jbnu.ac.kr (H. Tayara), kitchong@jbnu.ac.kr (K.T. Chong).
[1] These authors contributed equally to this work.

https://doi.org/10.1016/j.jtbi.2018.12.034

tions. These algorithms include "Covariance Discriminant" or "CD" algorithm (Chou and Elrod, 2002; Chou and Cai, 2003), "Support Vector Machine" or "SVM" algorithm (Cai et al., 2006), "Nearest Neighbor" or "NN" algorithm (Hu et al., 2011), and "Optimization" algorithm (Zhang and Chou, 1992). The pseudo amino acid composition (PseAAC) (Chou, 2001b) was developed for preventing losing the sequence-pattern information for proteins that may result from defining the vectors in a discrete mode. PseAAC was mostly applied in almost all of the computational proteomics (Chou, 2017; Jia et al., 2014; Ju and He, 2017; Ju and Wang, 2018; Ju et al., 2016; Xie et al., 2013; Zhang et al., 2014) and a result three different open access software have been developed namely 'PseAAC-Builder 'propy' (Cao et al., 2013), 'Builder' (Du et al., 2012), and 'PseAAC-General' (Cao et al., 2013). PseAAC-Builder and propy are used for generating various modes of Chou's special PseAAC (Chou, 2009) while PseAAC-General include not only all the special modes of feature vectors for proteins but, in addition, the higher level feature vectors such as "Gene Ontology" mode, "Sequential Evolution" or "PSSM" mode, and "Functional Domain" mode (Chou, 2011). Pseudo K-tuple Nucleotide Composition (PseKNC) (Chen et al., 2014) was introduced for producing different feature vectors for DNA/RNA sequences (Chen et al., 2015a; Liu et al., 2017b; Liu et al., 2016a; Liu et al., 2017c). A very powerful webservers, for generating any required feature vectors for RNA/DNA sequences and peptide/protein according to the users' requirement or their own definition, have been developed namely: 'Pse-in-One' One' (Liu et al., 2015) and its updated version 'Pse-in-One2.0′ (Liu et al., 2017a).

Post-translational modifications (PTM) are the different type of alterations, which support to diversify the some number of genome of the living organisms. In this regard, many researchers developed various computational method to identifying PTM sites functions using protein and RNA/RNA sequences (Chen et al., 2015b; Chen et al., 2016a; Chen et al., 2018b; Chou, 2015; Feng et al., 2017; Feng et al., 2018; Jia et al., 2016a; Jia et al., 2016b; Jia et al., 2016c; Jia et al., 2016d; Liu et al., 2016b; Qiu et al., 2014; Qiu et al., 2015; Qiu et al., 2016a; Qiu et al., 2016b; Qiu et al., 2016c; Qiu et al., 2017a; Qiu et al., 2017b; Qiu et al., 2018; Qiu et al., 2017c; Xu and Chou, 2016; Xu et al., 2013a; Xu et al., 2017; Xu et al., 2013b; Xu et al., 2014a; Xu et al., 2014b). In the previous works, there are several predictors were developed using machine learning techniques to detecting 2'-O-methylation sites (Chen et al., 2016b; Sun et al., 2015). The existing methods need specific and basic information to design the input features, according to 5-step rules of Chou's, the second step of Chou's rules is to extract numerical values from the input samples (Chou, 2011). As 2'-O-methylation sites is affected by RNA sequences, the system may automatically learn the feature of 2'-O-methylation sites from RNA sequences. The idea is obtained from deep learning, to fully extract the features from multiple levels of abstraction. In natural language processing (Collobert et al., 2011), information retrieval (Qu et al., 2017), image processing (Tayara and Chong, 2018; Tayara et al., 2018) and so on the deep learning approaches produced better outcomes. Currently, various genomics computational methods have been introduced used deep learning approaches (Aoki and Sakakibara, 2018; Nazari et al., 2018; Oubounyt et al., 2018; Pan et al., 2018b; Yang et al., 2017).

In this study, we used both machine learning and deep learning methods to develop computational method for 2'-O-methylation sites. The deep learning method produces better results as compared to machine learning method. The proposed iRNA-PseKNC(2methyl) method use the five-fold cross validation test and convolution neural networks (CNN). The proposed method is a simple and efficient architecture for 2'-O-methylation sites. The proposed prediction model is evaluated on one dataset and produces more efficient outcome than existing method published recently in the literature (Chen et al., 2016b). According to series of publications (Cai et al., 2018; Song et al., 2018; Zhang et al., 2018), the investigator have widely emphasized the guidelines of Chou's 5-step rules are: (1) dataset selection or construction; (2) convert the raw samples into feature vector; (3) classification algorithm; (4) cross-validation test; (5) develop a web-server. Below, these rules are address one by one.

2. Materials and methods

In this study, we use both machine learning and deep learning approach to distinguish 2'-O-methylation sites and not 2'-O-methylation sites. In machine learning approach two various feature extraction methods namely: multivariate mutual information (MMI) and *n*-Gram (Nanni, 2005; Pan et al., 2018a) are used for sample formulation and SVM is used for classification. While the deep learning approach is based on CNN model to identify 2'-O-methylation sites from raw genomic sequences, CNN model learns automatically the most important features from the input sequences during training.

2.1. Dataset

In this work, the dataset was downloaded from RMBase which contained 2'-O-methylaiton sites in *H. sapiens* (Chen et al., 2016b). Chen et al., predictor based on RMBase, to reduce homolog deviation and avoid redundancy, using the CD-HIT procedure to remove 80% similarity of RNA sequences (Chen et al., 2016b). Finally, 147 sequences of 2'-O-methylaiton sites yielded and treated as positive sequences. The length of each sequence is 41-nt with the 2'-O-methylation site in the center. The negative sequences were obtained by selecting the length of 41-nt sequences, in which the center nucleotides were not 2'-O-methylated and a huge number of imbalanced dataset for negative sequences are produced. Therefore, randomly 147 negative sequences were selected to balance negative and positive sequence. The benchmark dataset can be mathematically represented as:

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

where the subset D^{+}contains 147 true and, D^{-} contains 147 false 2'-O-methylation site samples, and the symbol ∪for union in set theory.

2.2. Machine learning approach

2.2.1. n-*Grams*

The *n*-Grams feature extraction method is a pair of values (v_i, c_i), the feature represented by v_i and the total number of the feature represented by c_i in a DNA/RNA or protein samples (Nanni, 2005). For example, to express RNA samples with 3-gram, *v* regards to the set of 3-nucleotide pairs and *c* represents the total numbers of pairs occurrence within entire sample. The mathematical representation of *n*-Gram method can be express as:

$$\begin{aligned} G &= G_1 \cup G_2 \cup G_3 \\ &= \{R_i\} \cup \left\{R_i R_j\right\} \cup \left\{R_i R_j R_l\right\} \\ &= \{A, C, U, G, GA, GC, AG, \ldots..GG, ACA, \ldots\ldots GGG\} \end{aligned} \tag{2}$$

where G denotes the list of nucleotide combination, G_1, G_2 and G_3, and represents 1-Gram, 2-Gram and 3-Gram with 4, 16 and 64 features, respectively, R_i, R_j, $R_l \in \{A, C, U, G\}$, and resulting in 84-D.

2.2.2. Multivariate mutual information

In the previous work (Pan et al., 2018a), multivariate mutual information (MMI) has been widely utilized to extract feature spaces from protein sample. Accordingly, the nucleotides sample can be expressed by applying multivariate mutual information method. In

this method, the DNA/RNA sample is defined by 2-tuple and 3-tuple nucleotide composition set K2 and K3 as follows.

$$K2 = \{AA, AC, AU, AG, CC, CU, CG, \text{UU}, \text{UG}, \text{GG}\}$$
$$K3 = \{AAA, AAC, AAU, AAG, ACC, ACU, ACG, AUU, AUG, AGG, CCC, CCU, CCG, CUU, CUG, CGG, \text{UUU}, \text{UUG}, \text{UGG}, \text{GGG}\} \quad (3)$$

The order of nucleotides is not important because MMIs in a tuple have no relationship, therefore if 2-tuple have unique constant but the order are not same, they may have the similar information and be assigned as single tuple. The K2 and K3 have 10 elements and 20 elements, respectively, but there does not exist the same composition with various order tuples

2.2.3. Support vector machine

SVM is a supervised learning technique and widely employed in the area of bioinformatics and system biology, and obtained better outcomes than other machine learning classification algorithms (Tahir and Hayat, 2016; Tahir et al., 2017; Tahir et al., 2018a; Tahir et al., 2018b). The main concept of support vector machine is to convert the data into a feature space with high-dimensional and then define the best separating hyperplane. SVM used a kernel function namely Gaussian radial basis function (RBF) to its better performance in non-linear classification. In this study, we use RBF kernel function and LIBSVM package to implement SVM model. The RBF kernel function has two parameters: the kernel width parameter g and the regularization parameter C. Through the optimization technique, their real score will be determined using grid search approach. In this work, we used the value of c is 0.0025 and the value of g is 7.5, respectively.

2.3. Deep learning approach

In previous works and Section 2.2 of this study, used machine learning approaches for the identification of 2'-O-methylation sites. In this work, we use deep learning approach i.e., CNN for the identification of 2'-O-methylation sites from raw genomic sequences. CNN learns automatically the most important features from the input sequences during training. In this regard, a computational automated method iRNA-PseKNC(2methyl) is proposed to identify the 2'-O-methylation sites in genomes accurately.

The iRNA-PseKNC(2methyl) method takes a single RNA sequences $R = \{R_1R_2R_3.....R_n\}$as an input, where $n = 41$ and$R_i \in \{A, C, U, G\}$, and produces a real valued. The input of the proposed method is one-hot encoded and represented as one-dimensional vector with four channels. The length of the vector is 41 and the four channel are A, G, C, and U. For more explanation, A, C, U, and G are denoted as (1, 0, 0, 0), (0, 1, 0, 0), (0, 0, 0, 1), and (0, 0, 1, 0) respectively. The Fig. 1 presents the one-hot encode of an input sequence and full architecture of proposed iRNA-PseKNC(2methyl) method.

Usually one processing step in deep learning network is known as a layer, the layer can be a convolution, pooling, ReLU, normalization, dropout, loss, fully connected layer, and so on. During learning various hyper parameters have been tuned, the tuned parameters are: number of filters, size of the filters, number of convolution layers, the number of the neurons of the dense layer and dropout probability after dense and convolution layers. In Table 1, present the list of these hyper parameters which are used in CNN model. The detailed architecture of the selected 2'-O-methylation sites method is described in Table 2. The optimal performing parameters have been chosen on the base of maximum accuracy. The convolution layer is mathematically represented and computed as

$$Conv(R)_{jf} = \text{Re}LU\left(\sum_{fs=0}^{FS-1}\sum_{f=0}^{F-1} W_{mn}^{f} R_{j+m,f}\right) \quad (4)$$

Table 1
The hyper parameters in CNNs to be tuned.

Hyper Parameter	Range
Number of convolution layers	[1,2,3]
The number of the filters	[4,8,12]
Filter size	[2,4]
Dropout	[0.2, 0.25, 0.3,0.35,0.5]

Table 2
The Detailed architecture of iRNA-PseKNC(2methyl) model.

Layer	Output Shape
Input	(41,4)
Conv1D(8,4,2)	(19,8)
Conv1D(4,2,2)	(9,4)
Dropout(0.35)	36
Dense(1)	1

Table 3
Success rate of machine learning method using SVM.

Feature Extraction	Classifier	Accuracy	Sensitivity	Specificity	MCC
n-Gram	SVM	81.23	81.63	80.95	0.625
MMI	SVM	77.89	74.14	81.62	0.559

Where R denotes the input of RNA sequence, f represents the index of the filter and j represents the index of the output position. ReLU represents the rectified linear function and mathematically can be defined as

$$\text{Re}LU(y) = \begin{cases} y \text{ if } y \geq 0 \\ 0 \text{ if } y < 0 \end{cases} \quad (5)$$

The sigmoid layer that makes predictions weather a given sequence is a 2'-O-methylation sites or not 2'-O-methylation sites.

$$Sigmoid(y) = \frac{1}{1 + e^{-y}} \quad (6)$$

The sigmoid () function outputs normalized class probabilities for a given input. The output of this layer is scaled to the [0, 1] by sigmoid function.

In this study, the proposed model is implemented by Keras framework (Chollet, 2015). Adam optimizer is used for optimization and the learning rate is set to 0.0001. Batch size is set to 10. Number of Epochs is set to 50.

2.4. Cross-validation test

In deep learning and machine learning approaches, error rate is used as an attribute to determine the performance of classification algorithms. Thus to find error rate, the relevant benchmark dataset is partitioned into various folds. For this purpose, the cross validation techniques have been used, where the whole dataset is partitioned into mutually exclusive folds. This method is also called subsampling or k-fold cross validation test. One fold is reserved for testing purpose, while remaining k-1 folds are used for training purpose, and the whole process repeated k-times. Accordingly (Hayat and Khan, 2011; Hayat and Tahir, 2015), odd number or 10-folds are widely used. Therefore, we applied 5-fold to randomly distribute the benchmark dataset into five equal size subsets and evaluate the performance of the prediction method.

2.5. Evaluation matrices

In previous studies, the following four statistical measurements matrices were mostly applied to evaluate the success rate of the computational systems. They are specificity (Sp), accuracy (Acc),

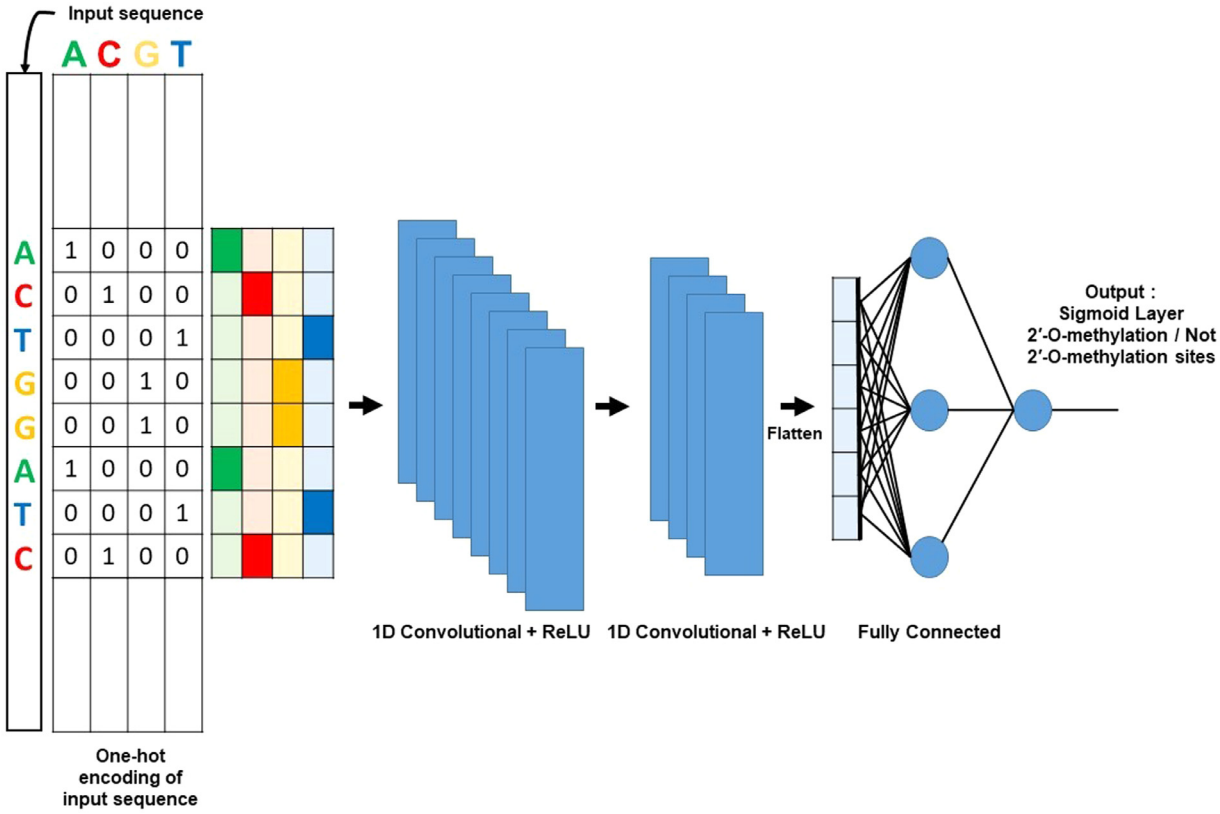

Fig. 1. Illustrate of iRNA-PseKNC(2methyl) model.

sensitivity (Sn), and Matthews's correction coefficient (MCC). However, their formulas were exactly taken from mathematical books are not intuitive and most biologists are not easy to understand. Fortunately, the following four evaluation matrices were derived (Chen et al., 2013), from the studied of signal peptides based on Chou's symbols (Chou, 2001a), the four evaluation matrices can be mathematically expressed as:

$$\begin{cases} Sp = 1 - \dfrac{S_+^-}{S^-} & 0 \le Sp \le 1 \\ Sn = 1 - \dfrac{S_-^+}{S^+} & 0 \le Sn \le 1 \\ Acc = 1 - \dfrac{S_+^- + S_-^+}{S^+ + S^-} & 0 \le Acc \le 1 \\ MCC = \dfrac{1 - \dfrac{S_+^- + S_-^+}{S^+ + S^-}}{\sqrt{\left(\dfrac{S_+^- + S_-^+}{S^-} + 1\right)\left(\dfrac{S_+^- + S_-^+}{S^+} + 1\right)}} & -1 \le MCC \le 1 \end{cases} \quad (7)$$

Where $S^+$and S^-denote the total number of 2'-O-methylation sites samples and not 2'-O-methylation sites samples, respectively; S_-^+represents the number of 2'-O-methylation sites samples incorrectly predicted as not 2'-O-methylation samples while S_+^- represents the number of the not 2'-O-methylation sites samples incorrectly predicted as 2'-O-methylation sites samples. The set of evaluation metrics (Eq. (7)) have been broadly employed in computational biology and bioinformatics (see, e.g., (Chen et al., 2018a; Jia et al., 2019; Li et al., 2018a; Li et al., 2018b; Liu et al., 2018b; Song et al., 2018; Yang et al., 2018)). More explanation, for single-label systems that conventional mathematical formulates (Tahir and Hayat, 2016; Tahir and Hayat, 2017) are valid; for multi-label system, whose presence has become more frequent in biomedicine, system medicine, and system biology (Cheng et al., 2018b; Cheng et al., 2017b; Chou et al., 2018; Xiao et al., 2018), a completely different set of metrics as defined in (Cheng et al., 2017b) is absolutely needed.

3. Results and discussion

3.1. Machine learning method

The experimental results of machine learning methods using two various feature extraction techniques namely: *n*-Gram and MMI, and support vector machine for classification are shown in Table 3. In the case of n-Gram method, the SVM obtained 81.23% of accuracy, 81.63% of sensitivity, 80.95% of specificity, and 0.625 of MCC. Similarly, MMI feature space with SVM achieved 74.14% of sensitivity, 81.62% of specificity, 77.89% of accuracy, and 0.559 of MCC.

3.2. Deep learning method

The measure of efficiency of deep learning method is better than of machine learning ones. In the case of deep learning method, we obtained 98.27% of accuracy, 100.00% of specificity, 96.29% of sensitivity, and 0.965 of MCC. The results of deep learning method as bellows:

$$\begin{cases} Acc = 98.27\% \\ Sen = 96.29\% \\ Sp = 100.00\% \\ MCC = 0.965 \end{cases}$$

3.3. Performance comparison of iRNA-PseKNC(2methyl) with exiting method

The success rate of the proposed iRNA-PseKNC(2methyl) computational method compared with the state-of-the-arts (Chen et al., 2016b) is more robust. The proposed prediction iRNA-PseKNC(2methyl) method produces more efficient improvement in sensitivity, specificity, accuracy, and MCC as shown in Table 4. The

Table 4
Comparison with existing method.

Methods	Accuracy	Sensitivity	Specificity	MCC
Proposed Method	**98.27**	**96.29**	**100.00**	**0.965**
Chen et al. (2016b)	95.58	92.52	98.64	0.91

experimental outcomes presented that the proposed prediction iRNA-PseKNC(2methyl) method obtained significant results compared to the existing method. This remarkable success is ascribed to the convolution neural network.

According to (Chou and Shen, 2009), publicly available web-servers will be the future direction for reporting many predictors for system biology (Chen et al., 2018a; Cheng et al., 2017a; Cheng et al., 2018a; Cheng et al., 2018b; Cheng et al., 2017b; Chou et al., 2018; Jia et al., 2019; Liu et al., 2017c; Liu et al., 2018a; Xiao et al., 2018). Indeed, they have highly improved the powers of system biology on driving medical science into an unprecedented revolution (Chou, 2017) and medical science (Chou, 2015), in future work, we shall effort to develop a web-server for the proposed method introduced in this paper.

4. Conclusion

In this work, we proposed an intelligent and robust method namely: iRNA-PseKNC(2methyl) for 2'-O-methylation sites prediction. We employed both machine leaning and deep learning approaches but the deep learning approach obtained better performance than machine learning. In machine learning, we used two feature extraction methods namely: multivariate mutual information and *n*-Gram to extract feature from RNA sequences, and support vector machine for classification. The deep learning method is based on CNN. Unlike the previous works that use handcrafted features for classification, iRNA-PseKNC(2methyl) automatically extracts the features from RNA sequences. The success rate shows that the iRNA-PseKNC(2methyl) prediction method is more efficient than the existing methods in terms of all evaluation metrics. The proposed method might be helpful in academia research and drug design.

Conflict of interest

No interest.

Acknowledgements

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

References

Aoki, G., Sakakibara, Y., 2018. Convolutional neural networks for classification of alignments of non-coding RNA sequences. Bioinformatics 34, i237–i244.
Bachellerie, J.-P., Cavaille, J., Hüttenhofer, A., 2002. The expanding snoRNA world. Biochimie 84, 775–790.
Cai, L., Huang, T., Su, J., Zhang, X., Chen, W., Zhang, F., He, L., Chou, K.-C., 2018. Implications of newly identified brain eQTL genes and their interactors in Schizophrenia. Mol. Ther.-Nucleic Acids 12, 433–442.
Cai, Y.-D., Feng, K.-Y., Lu, W.-C., Chou, K.-C., 2006. Using LogitBoost classifier to predict protein structural classes. J. Theor. Biol. 238, 172–176.
Cao, D.-S., Xu, Q.-S., Liang, Y.-Z., 2013. propy: a tool to generate various modes of Chou's PseAAC. Bioinformatics 29, 960–962.
Chen, W., Lin, H., Chou, K.-C., 2015a. Pseudo nucleotide composition or PseKNC: an effective formulation for analyzing genomic sequences. Mol. Biosyst. 11, 2620–2634.
Chen, W., Feng, P.-M., Lin, H., Chou, K.-C., 2013. iRSpot-PseDNC: identify recombination spots with pseudo dinucleotide composition. Nucleic Acids Res. 41, e68.
Chen, W., Lei, T.-Y., Jin, D.-C., Lin, H., Chou, K.-C., 2014. PseKNC: a flexible web server for generating pseudo K-tuple nucleotide composition. Anal. Biochem. 456, 53–60.
Chen, W., Feng, P., Ding, H., Lin, H., Chou, K.-C., 2015b. iRNA-methyl: identifying N6-methyladenosine sites using pseudo nucleotide composition. Anal. Biochem. 490, 26–33.
Chen, W., Tang, H., Ye, J., Lin, H., Chou, K.-C., 2016a. iRNA-PseU: identifying RNA pseudouridine sites. Mol. Ther.-Nucleic Acids 5.
Chen, W., Feng, P., Tang, H., Ding, H., Lin, H., 2016b. Identifying 2'-O-methylationation sites by integrating nucleotide chemical properties and nucleotide compositions. Genomics 107, 255–258.
Chen, W., Ding, H., Zhou, X., Lin, H., Chou, K.-C., 2018a. iRNA (m6A)-PseDNC: identifying N6-methyladenosine sites using pseudo dinucleotide composition. Anal. Biochem. 561, 59–65.
Chen, W., Feng, P., Yang, H., Ding, H., Lin, H., Chou, K.-C., 2018b. iRNA-3typeA: identifying three types of modification at RNA's adenosine sites. Mol. Ther.-Nucleic Acids 11, 468–474.
Cheng, X., Xiao, X., Chou, K.-C., 2017a. pLoc-mHum: predict subcellular localization of multi-location human proteins via general PseAAC to winnow out the crucial GO information. Bioinformatics 34, 1448–1456.
Cheng, X., Xiao, X., Chou, K.-C., 2018a. pLoc-mEuk: predict subcellular localization of multi-label eukaryotic proteins by extracting the key GO information into general PseAAC. Genomics 110, 50–58.
Cheng, X., Xiao, X., Chou, K.-C., 2018b. pLoc-mGneg: predict subcellular localization of Gram-negative bacterial proteins by deep gene ontology learning via general PseAAC. Genomics 110, 231–239.
Cheng, X., Zhao, S.-G., Lin, W.-Z., Xiao, X., Chou, K.-C., 2017b. pLoc-mAnimal: predict subcellular localization of animal proteins with both single and multiple sites. Bioinformatics 33, 3524–3531.
Chollet, F. Keras: Deep learning library for theano and tensorflow URL: 7.
Chou, K.-C., 2001a. Prediction of signal peptides using scaled window. Peptides 22, 1973–1979.
Chou, K.-C., 2009. Pseudo amino acid composition and its applications in bioinformatics, proteomics and system biology. Curr. Proteomics 6, 262–274.
Chou, K.-C., 2011. Some remarks on protein attribute prediction and pseudo amino acid composition. J. Theor. Biol. 273, 236–247.
Chou, K.-C., 2015. Impacts of bioinformatics to medicinal chemistry. Med. Chem. 11, 218–234.
Chou, K.-C., 2017. An unprecedented revolution in medicinal chemistry driven by the progress of biological science. Curr. Top. Med. Chem. 17, 2337–2358.
Chou, K.-C., Elrod, D.W., 2002. Bioinformatical analysis of G-protein-coupled receptors. J. Proteome Res. 1, 429–433.
Chou, K.-C., Shen, H.-B., 2009. Recent advances in developing web-servers for predicting protein attributes. Nat. Sci. 1, 63.
Chou, K.-C., Cheng, X., Xiao, X., 2018. pLoc_bal-mHum: predict subcellular localization of human proteins by PseAAC and quasi-balancing training dataset. Genomics.
Chou, K.C., 2001b. Prediction of protein cellular attributes using pseudo-amino acid composition. Proteins 43, 246–255.
Chou, K.C., Cai, Y.D., 2003. Prediction and classification of protein subcellular location—sequence-order effect and pseudo amino acid composition. J. Cell. Biochem. 90, 1250–1260.
Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., Kuksa, P., 2011. Natural language processing (almost) from scratch. J. Mach. Learn. Res. 12, 2493–2537.
Decatur, W.A., Fournier, M.J., 2002. rRNA modifications and ribosome function. Trends Biochem. Sci. 27, 344–351.
Dong, Z.-W., Shao, P., Diao, L.-T., Zhou, H., Yu, C.-H., Qu, L.-H., 2012. RTL-P: a sensitive approach for detecting sites of 2'-O-methylation in RNA molecules. Nucleic Acids Res. 40, e157.
Du, P., Wang, X., Xu, C., Gao, Y., 2012. PseAAC-Builder: a cross-platform stand-alone program for generating various special Chou's pseudo-amino acid compositions. Anal. Biochem. 425, 117–119.
Feng, P., Ding, H., Yang, H., Chen, W., Lin, H., Chou, K.-C., 2017. iRNA-PseColl: identifying the occurrence sites of different RNA modifications by incorporating collective effects of nucleotides into PseKNC. Mol. Ther.-Nucleic Acids 7, 155–163.
Feng, P., Yang, H., Ding, H., Lin, H., Chen, W., Chou, K.-C., 2018. iDNA6mA-PseKNC: Identifying DNA N6-methyladenosine sites by incorporating nucleotide physicochemical properties into PseKNC. Genomics.
Hayat, M., Khan, A., 2011. Predicting membrane protein types by fusing composite protein sequence features into pseudo amino acid composition. J. Theor. Biol. 271, 10–17.
Hayat, M., Tahir, M., 2015. PSOFuzzySVM-TMH: identification of transmembrane helix segments using ensemble feature space by incorporated fuzzy support vector machine. Mol. Biosyst. 11, 2255–2262.
Hu, L., Huang, T., Shi, X., Lu, W.-C., Cai, Y.-D., Chou, K.-C., 2011. Predicting functions of proteins in mouse based on weighted protein-protein interaction network and protein hybrid properties. PLoS One 6, e14556.
Jia, C., Lin, X., Wang, Z., 2014. 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, 10410–10423.
Jia, J., Liu, Z., Xiao, X., Liu, B., Chou, K.-C., 2016a. iSuc-PseOpt: identifying lysine succinylation sites in proteins by incorporating sequence-coupling effects into pseudo components and optimizing imbalanced training dataset. Anal. Biochem. 497, 48–56.
Jia, J., Liu, Z., Xiao, X., Liu, B., Chou, K.-C., 2016b. pSuc-Lys: predict lysine succinylation sites in proteins with PseAAC and ensemble random forest approach. J. Theor. Biol. 394, 223–230.
Jia, J., Liu, Z., Xiao, X., Liu, B., Chou, K.-C., 2016c. iCar-PseCp: identify carbonylation

sites in proteins by Monte Carlo sampling and incorporating sequence coupled effects into general PseAAC. Oncotarget 7, 34558.
Jia, J., Zhang, L., Liu, Z., Xiao, X., Chou, K.-C., 2016d. pSumo-CD: predicting sumoylation sites in proteins with covariance discriminant algorithm by incorporating sequence-coupled effects into general PseAAC. Bioinformatics 32, 3133–3141.
Jia, J., Li, X., Qiu, W., Xiao, X., Chou, K.-C., 2019. iPPI-PseAAC (CGR): identify protein-protein interactions by incorporating chaos game representation into PseAAC. J. Theor. Biol. 460, 195–203.
Ju, Z., He, J.-J., 2017. 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, 200–204.
Ju, Z., Wang, S.-Y., 2018. Prediction of citrullination sites by incorporating k-spaced amino acid pairs into Chou's general pseudo amino acid composition. Gene 664, 78–83.
Ju, Z., Cao, J.-Z., Gu, H., 2016. Predicting lysine phosphoglycerylation with fuzzy SVM by incorporating k-spaced amino acid pairs into Chou's general PseAAC. J. Theor. Biol. 397, 145–150.
Kiss, T., 2002. Small nucleolar RNAs: an abundant group of noncoding RNAs with diverse cellular functions. Cell 109, 145–148.
Li, F., Li, C., Marquez-Lago, T.T., Leier, A., Akutsu, T., Purcell, A.W., Ian Smith, A., Lithgow, T., Daly, R.J., Song, J., 2018a. Quokka: a comprehensive tool for rapid and accurate prediction of kinase family-specific phosphorylation sites in the human proteome. Bioinformatics.
Li, F., Wang, Y., Li, C., Marquez-Lago, T.T., Leier, A., Rawlings, N.D., Haffari, G., Revote, J., Akutsu, T., Chou, K.-C., 2018b. Twenty years of bioinformatics research for protease-specific substrate and cleavage site prediction: a comprehensive revisit and benchmarking of existing methods. Brief. Bioinform.
Li, J., Yang, Z., Yu, B., Liu, J., Chen, X., 2005. Methylation protects miRNAs and siRNAs from a 3′-end uridylation activity in Arabidopsis. Curr. Biol. 15, 1501–1507.
Liu, B., Wu, H., Chou, K.-C., 2017a. 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, 67.
Liu, B., Yang, F., Chou, K.-C., 2017b. 2L-piRNA: a two-layer ensemble classifier for identifying piwi-interacting RNAs and their function. Mol. Ther.-Nucleic Acids 7, 267–277.
Liu, B., Wang, S., Long, R., Chou, K.-C., 2016a. iRSpot-EL: identify recombination spots with an ensemble learning approach. Bioinformatics 33, 35–41.
Liu, B., Yang, F., Huang, D.-S., Chou, K.-C., 2017c. iPromoter-2L: a two-layer predictor for identifying promoters and their types by multi-window-based PseKNC. Bioinformatics 34, 33–40.
Liu, B., Li, K., Huang, D.-S., Chou, K.-C., 2018a. iEnhancer-EL: identifying enhancers and their strength with ensemble learning approach. Bioinformatics.
Liu, B., Weng, F., Huang, D.-S., Chou, K.-C., 2018b. iRO-3wPseKNC: identify DNA replication origins by three-window-based PseKNC. Bioinformatics 1, 8.
Liu, B., Liu, F., Wang, X., Chen, J., Fang, L., Chou, K.-C., 2015. Pse-in-One: a web server for generating various modes of pseudo components of DNA, RNA, and protein sequences. Nucleic Acids Res. 43, W65–W71.
Liu, Z., Xiao, X., Yu, D.-J., Jia, J., Qiu, W.-R., Chou, K.-C., 2016b. pRNAm-PC: predicting N6-methyladenosine sites in RNA sequences via physical–chemical properties. Anal. Biochem. 497, 60–67.
Nanni, L., 2005. Hyperplanes for predicting protein–protein interactions. Neurocomputing 69, 257–263.
Nazari, I., Tayara, H., Chong, K.T., 2018. Branch point selection in RNA splicing using deep learning. IEEE Access 1. doi:10.1109/ACCESS.2018.2886569.
Oubounyt, M., Louadi, Z., Tayara, H., Chong, K.T., 2018. Deep learning models based on distributed feature representations for alternative splicing prediction. IEEE Access 6, 58826–58834.
Pan, G., Jiang, L., Tang, J., Guo, F., 2018a. A novel computational method for detecting DNA methylation sites with DNA sequence information and physicochemical properties. Int. J. Mol. Sci. 19, 511.
Pan, X., Rijnbeek, P., Yan, J., Shen, H.-B., 2018b. Prediction of RNA-protein sequence and structure binding preferences using deep convolutional and recurrent neural networks. BMC Genomics 19, 511.
Qiu, W.-R., Xiao, X., Lin, W.-Z., Chou, K.-C., 2014. iMethyl-PseAAC: identification of protein methylation sites via a pseudo amino acid composition approach. BioMed Res. Int. 2014.
Qiu, W.-R., Xiao, X., Lin, W.-Z., Chou, K.-C., 2015. iUbiq-Lys: prediction of lysine ubiquitination sites in proteins by extracting sequence evolution information via a gray system model. J. Biomol. Struct. Dyn. 33, 1731–1742.
Qiu, W.-R., Xiao, X., Xu, Z.-C., Chou, K.-C., 2016a. iPhos-PseEn: identifying phosphorylation sites in proteins by fusing different pseudo components into an ensemble classifier. Oncotarget 7, 51270.
Qiu, W.-R., Sun, B.-Q., Xiao, X., Xu, Z.-C., Chou, K.-C., 2016b. iPTM-mLys: identifying multiple lysine PTM sites and their different types. Bioinformatics 32, 3116–3123.
Qiu, W.-R., Sun, B.-Q., Xiao, X., Xu, Z.-C., Chou, K.-C., 2016c. iHyd-PseCp: identify hydroxyproline and hydroxylysine in proteins by incorporating sequence-coupled effects into general PseAAC. Oncotarget 7, 44310.
Qiu, W.-R., Jiang, S.-Y., Xu, Z.-C., Xiao, X., Chou, K.-C., 2017a. iRNAm5C-PseDNC: identifying RNA 5-methylcytosine sites by incorporating physical-chemical properties into pseudo dinucleotide composition. Oncotarget 8, 41178.
Qiu, W.-R., Jiang, S.-Y., Sun, B.-Q., Xiao, X., Cheng, X., Chou, K.-C., 2017b. iRNA-2methyl: identify RNA 2'-O-methylation sites by incorporating sequence–coupled effects into general PseKNC and ensemble classifier. Med. Chem. 13, 734–743.
Qiu, W.-R., Sun, B.-Q., Xiao, X., Xu, Z.-C., Jia, J.-H., Chou, K.-C., 2018. iKcr-PseEns: identify lysine crotonylation sites in histone proteins with pseudo components and ensemble classifier. Genomics 110, 239–246.
Qiu, W.R., Sun, B.Q., Xiao, X., Xu, D., Chou, K.C., 2017c. iPhos-PseEvo: identifying human phosphorylated proteins by incorporating evolutionary information into general PseAAC via grey system theory. Mol. Inf. 36, 1600010.
Qu, W., Wang, D., Feng, S., Zhang, Y., Yu, G., 2017. A novel cross-modal hashing algorithm based on multimodal deep learning. Sci. China Inf. Sci. 60, 092104.
Ramachandran, V., Chen, X., 2008. Degradation of microRNAs by a family of exoribonucleases in Arabidopsis. Science 321, 1490–1492.
Song, J., Wang, Y., Li, F., Akutsu, T., Rawlings, N.D., Webb, G.I., Chou, K.-C., 2018. iProt-Sub: a comprehensive package for accurately mapping and predicting protease-specific substrates and cleavage sites. Brief. Bioinform.
Sun, W.-J., Li, J.-H., Liu, S., Wu, J., Zhou, H., Qu, L.-H., Yang, J.-H., 2015. RMBase: a resource for decoding the landscape of RNA modifications from high-throughput sequencing data. Nucleic Acids Res. 44, D259–D265.
Tahir, M., Hayat, M., 2016. 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, 2587–2593.
Tahir, M., Hayat, M., 2017. Machine learning based identification of protein–protein interactions using derived features of physiochemical properties and evolutionary profiles. Artif. Intell. Med. 78, 61–71.
Tahir, M., Hayat, M., Kabir, M., 2017. Sequence based predictor for discrimination of enhancer and their types by applying general form of Chou's trinucleotide composition. Comput. Methods Programs Biomed. 146, 69–75.
Tahir, M., Hayat, M., Khan, S.A., 2018a. 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, 6719–6727.
Tahir, M., Hayat, M., Khan, S.A., 2018b. 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. Genomics 1–12.
Tayara, H., Chong, K., 2018. Object detection in very high-resolution aerial images using one-stage densely connected feature pyramid network. Sensors 18, 3341.
Tayara, H., Soo, K.G., Chong, K.T., 2018. Vehicle detection and counting in high-resolution aerial images using convolutional regression neural network. IEEE Access 6, 2220–2230.
Xiao, X., Cheng, X., Chen, G., Mao, Q., Chou, K.-C., 2018. pLoc_bal-mGpos: predict subcellular localization of Gram-positive bacterial proteins by quasi-balancing training dataset and PseAAC. Genomics.
Xie, H.-L., Fu, L., Nie, X.-D., 2013. 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, 735–742.
Xu, Y., Chou, K.-C., 2016. Recent progress in predicting posttranslational modification sites in proteins. Curr. Top. Med. Chem. 16, 591–603.
Xu, Y., Ding, J., Wu, L.-Y., Chou, K.-C., 2013a. iSNO-PseAAC: predict cysteine S-nitrosylation sites in proteins by incorporating position specific amino acid propensity into pseudo amino acid composition. PLoS One 8, e55844.
Xu, Y., Wang, Z., Li, C., Chou, K.-C., 2017. iPreny-PseAAC: identify C-terminal cysteine prenylation sites in proteins by incorporating two tiers of sequence couplings into PseAAC. Med. Chem. 13, 544–551.
Xu, Y., Shao, X.-J., Wu, L.-Y., Deng, N.-Y., Chou, K.-C., 2013b. iSNO-AAPair: incorporating amino acid pairwise coupling into PseAAC for predicting cysteine S-nitrosylation sites in proteins. PeerJ 1, e171.
Xu, Y., Wen, X., Shao, X.-J., Deng, N.-Y., Chou, K.-C., 2014a. iHyd-PseAAC: Predicting hydroxyproline and hydroxylysine in proteins by incorporating dipeptide position-specific propensity into pseudo amino acid composition. Int. J. Mol. Sci. 15, 7594–7610.
Xu, Y., Wen, X., Wen, L.-S., Wu, L.-Y., Deng, N.-Y., Chou, K.-C., 2014b. iNitro-Tyr: Prediction of nitrotyrosine sites in proteins with general pseudo amino acid composition. PLoS One 9, e105018.
Yang, B., Liu, F., Ren, C., Ouyang, Z., Xie, Z., Bo, X., Shu, W., 2017. BiRen: predicting enhancers with a deep-learning-based model using the DNA sequence alone. Bioinformatics 33, 1930–1936.
Yang, H., Qiu, W.-R., Liu, G., Guo, F.-B., Chen, W., Chou, K.-C., Lin, H., 2018. iRSpot-Pse6NC: Identifying recombination spots in Saccharomyces cerevisiae by incorporating hexamer composition into general PseKNC. Int. J. Biol. Sci. 14, 883.
Zhang, C.T., Chou, K.C., 1992. An optimization approach to predicting protein structural class from amino acid composition. Protein Sci. 1, 401–408.
Zhang, J., Zhao, X., Sun, P., Ma, Z., 2014. 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, 11204–11219.
Zhang, Y., Xie, R., Wang, J., Leier, A., Marquez-Lago, T.T., Akutsu, T., Webb, G.I., Chou, K.-C., Song, J., 2018. Computational analysis and prediction of lysine malonylation sites by exploiting informative features in an integrative machine-learning framework. Brief. Bioinform. 5.
Züst, R., Cervantes-Barragan, L., Habjan, M., Maier, R., Neuman, B.W., Ziebuhr, J., Szretter, K.J., Baker, S.C., Barchet, W., Diamond, M.S., 2011. Ribose 2'-O-methylation provides a molecular signature for the distinction of self and non-self mRNA dependent on the RNA sensor Mda5. Nat. Immunol. 12, 137.