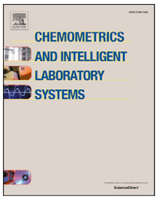
Chemometrics and Intelligent Laboratory Systems 204 (2020) 104102

Contents lists available at ScienceDirect

Chemometrics and Intelligent Laboratory Systems

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



i6mA-DNC: Prediction of DNA N6-Methyladenosine sites in rice genome based on dinucleotide representation using deep learning

Sehi Park[a,1], Abdul Wahab[a,1], Iman Nazari[a], Ji Hyoung Ryu[b,*], Kil To Chong[a,c,**]

[a] Department of Electronics and Information Engineering, Jeonbuk National University, Jeonju, 54896, South Korea
[b] Electronics and Telecommunication Research Institute, Gwangju, 17611, South Korea
[c] Advanced Electronics and Information Research Center, Jeonbuk National University, Jeonju, 54896, South Korea

ARTICLE INFO

Keywords:
DNA N6-methyladenine
Deep learning
Convolution neural network
Dinucleotide

ABSTRACT

DNA methylation is a crucial epigenetic process. DNA N6-methyladenine is closely related to a variety of biological processes such as DNA replication, transcription, repair and cellular defense. In genome, N6-methyladenine (6 mA) sites are not uniformly distributed; therefore, it is required to determine the genomic locations of 6 mA for better comprehension of its biological functions. Although various experimental procedures have been used to identify 6 mA sites and yielded positive results, these biochemical techniques are expensive and time-consuming. In order to solve this problem and provide ease for future researches, it is indispensable to develop a robust and accurate computational model to find N6-methyladenine sites. With this regard, we introduce a deep learning-based computational model called i6mA-DNC to detect the N6-methyladenine sites in the rice genome. We split the DNA sequences into dinucleotide components and feed them to the model. This model automatically extracts optimal features from the pre-processed data using convolution neural network (CNN). Our proposed model i6mA-DNC obtained 89.20% of specificity, 88.01% of sensitivity, 88.60% of accuracy, and 0.772 of MCC. These results prove that our intelligent model achieved better success rates in all evaluation metrics than existing methods. Our model i6mA-DNC is expected to become a useful tool for academic research on N6-methyladenine sites identification. A user-friendly webserver has been established and made freely accessible at https://home.jbnu.ac.kr/NSCL/i6mA-DNC.htm.

1. Introduction

DNA methylation refers to an important epigenetic process by which a methyl groups are added to the DNA bases. It often modifies the activity of the genes without changing the sequences. Among DNA's four bases, the cytosine and adenine bases can be methylated. In eukaryotes, 5-methylcytosine (5 mC) is the most common type of DNA modification [1,2]. In prokaryotes, however, 6-methyadenine (6 mA) is the most dominant DNA modification [3]. As a non-canonical DNA modification, DNA N6-methyladenine (6 mA) has been detected in three kingdoms of life: bacteria, archaea, and eukaryotes [4]. The 6 mA modification is highly correlated with a wide-ranging biological processes, such as DNA replication [5], transcription [6], repair [7], and cellular defense [8–10]. In genome, the 6 mA sites distribution is not uniform. Therefore, in order to fully understand the biological functions of 6 mA, it is essential to precisely identify its genomic location.

To this end, a variety of experimental methods have been introduced to identify 6 mA sites, such as ultra-high performance liquid chromatography coupled with mass spectrometry (UHPLC-Ms/Ms) [11], single-molecule real-time sequencing (SMRT-seq) [12], methylated DNA Immunoprecipitation sequencing (MeDIP-seq) [13], and capillary electrophoresis and laser-induced fluorescence (CE-LIF) [14]. Although the biochemical experimental methods provide positive data information related to the 6 mA sites, it is costly and time-consuming to use these wet lab techniques. To deal with this issue, it is required to design a powerful, precise, and fast computational model using the pre-existing experimental data to detect the N6-methyladenine sites.

Machine learning methods were designed to identify N6-methyladenine sites in the rice genome. A group of biologists [15] obtained the 6 mA profile of rice genome by using mass spectrometry

[*] Corresponding author.,
[**] Corresponding author. Department of Electronics and Information Engineering, Jeonbuk National University, Jeonju, 54896, South Korea.
E-mail addresses: jihyooung@etri.re.kr (J.H. Ryu), kitchong@jbnu.ac.kr (K.T. Chong).
[1] Sehi Park and Abdul Wahab have contributed equally.

https://doi.org/10.1016/j.chemolab.2020.104102
Received 26 September 2019; Received in revised form 13 May 2020; Accepted 1 July 2020
Available online 20 July 2020

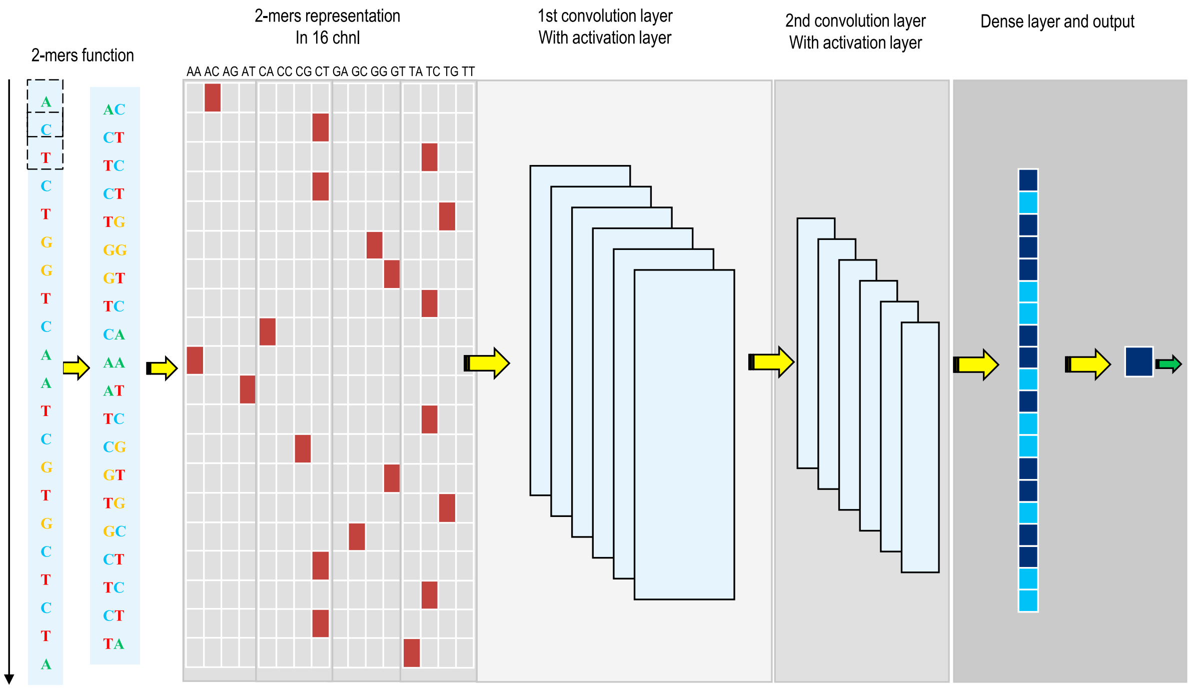


Fig. 1. Illustration of 2-mer representation and model architecture of i6mA-DNC.

Table 1
Hypter-parameter selection.

Parameters	Range
Then number of convolution layers	[1–3]
The number of filters in each convolution layer	[5,7,9,11,16,32,64]
The size of the filters of the convolution layer	[3,5,7,9]
Dropout value after convolution layers	[0,2,0.3,0.4,0.6]

Table 2
Detailed configuration of the proposed CNN model.

Layer	Output shape
Input	(n,40,16)
Conv1D(32,5)	(40,32)
ReLU	(40,32)
Dropout(0.4)	(40,32)
Conv1D(16,5)	(36,16)
ReLU	(36,16)
MaxPool1D(4,2)	(17,16)
Flatten	272
Dropout(0.4)	272
Dense(1)	1
Sigmoid	1

Table 3
Performance comparision between i6mA-DNC with other existing method.

Model	ACC	Sn	Sp	MCC	auRoc
6 mA-Pred	83.13	82.95	83.30	0.66	0.886
iDNA6mA(5step rule)	86.64	86.70	86.59	0.73	0.931
iN6-methylat(5-step)	87.78	86.48	89.09	0.75	0.80
i6mA-DNC	**88.60**	**88.01**	**89.20**	**0.77**	**0.959**

analysis and Immunoprecipitation followed by sequencing (IP-seq). The information acquired in that research made it possible to develop computational models based on machine learning approaches. Chen et al. [16] proposed the first computational model based on support vector machine (SVM) to identify the N6-methyladenine sites in the rice genome using nucleotide frequency and nucleotide chemical properties as features extraction techniques. Their model achieved an accuracy of 83.13% and 0.66 of Matthews's correlation coefficient (MCC). A few months later, Le, N. Q. K [17]. designed another SVM-based model using a continuous bag of nucleobases via Chou's 5- step rule. With cross-validation, this model obtained accuracy of 87.78% and 0.756 of MCC.

Another group of researchers proposed a computational model based on a convolution neural network (CNN) to identify the N6-methyladenine sites in the rice genome. They used the one-hot encoding algorithm for the sequences and fed them into the CNN model. Unlike Chen's SVM model, their model extracts features from DNA sequences automatically. The model, iDNA6mA [18], gained accuracy of 86.64% and 0.732 of MCC. iDNA6mA used one-hot encoding for representing the input sequences. However, one-hot encoding is a low dimensional feature representation that cannot capture the hidden information in the input sequence. Therefore, in order to find the significant information between successive nucleotides we used a higher-order coding in order to extract sequence features and significantly improving the final prediction performance.

In order to determine DNA N6-methyladenine sites, we proposed the i6mA-DNC model for identification of 6 mA using dinucleotide composition that is based on the convolution neural network (CNN). We divided the raw DNA sequences into dinucleotide components and fed them to the CNN model. The novel deep learning model that we proposed has superior prediction outcomes compared to the existing models. In addition, a user-friendly webserver has been established and made freely accessible at https://home.jbnu.ac.kr/NSCL/i6mA-DNC.htm.

2. Materials and methods

2.1. Benchmark dataset

First of all, we selected a valid benchmark dataset to predict and test our proposed model effectively. The dataset for rice genome was constructed by Chen et al. [16]. All samples have 41 nucleotides with the 6 mA site in the center. Therefore, the dataset is represented as:

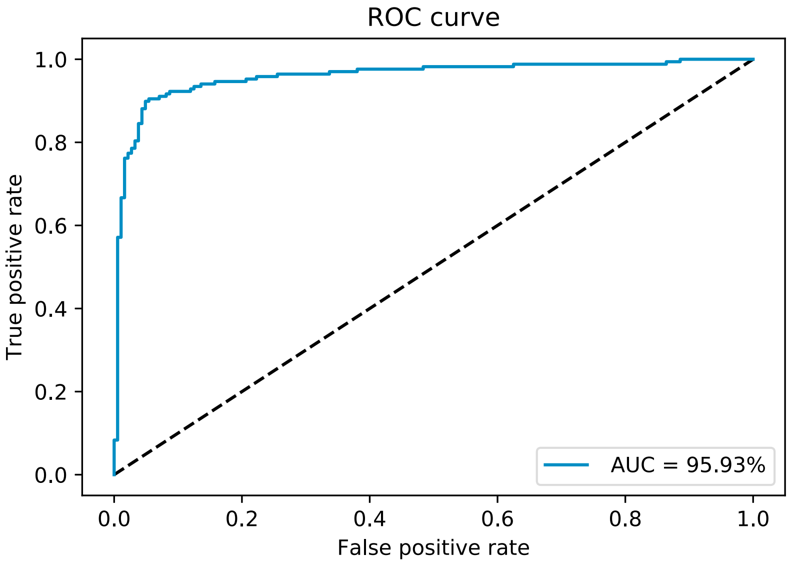


Fig. 2. The auROC curve performance of the computational model i6mA-DNC.

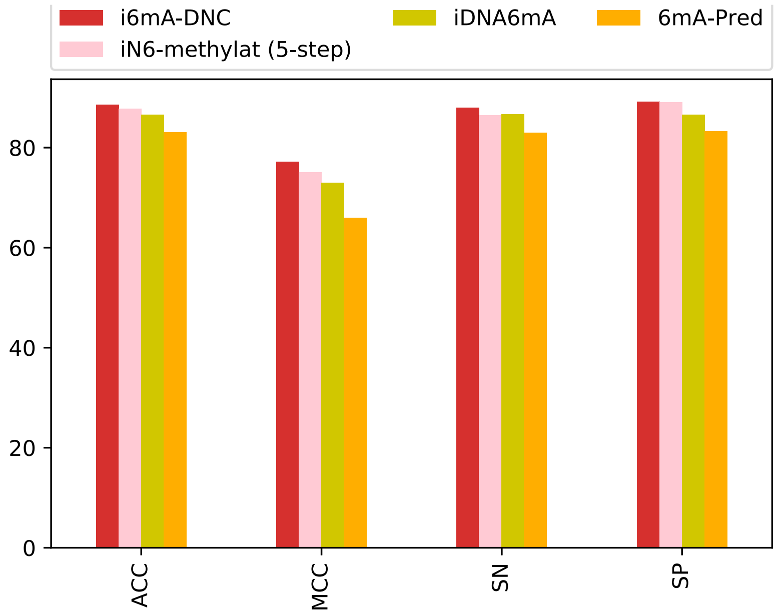


Fig. 3. The results achived by computational model i6mA-DNC compared to other state-of-the-art models.

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

Where R represents the whole dataset comprising of 1760 samples. The subset R^{+} have 880 sequences with 6 mA sites while the subset R^{-} consists of 880 sequences with non-6mA sites. The $\cup$ denotes the union of the two subsets.

2.2. Proposed model

Fig. 1 shows the framework of i6mA-DNC model. It is a model to predict the 6 mA sites using convolutional neural network (CNN).

First of all, the proposed model takes a single input of a DNA sample. A given sequence S is represented as $S = \{N_1, N_2, \ldots N_i\}$ where N_1 denotes the first nucleotide of the sequence, N_2 denotes the second nucleotide and so forth. Every nucleotide belongs to one of these four A, C, G, T. $i =$ 41 and represents the length of the sample.

It is important to convert the raw genomic sequence into a vector form that deep learning can recognize. It is also needed to consider the loss of pattern sequence information while converting sequence data into vector forms, as it mostly happens in the discrete models. To avoid the pattern loss, many techniques have been introduced including pseudo amino acid composition (PseAAC) [19] which was widely used in proteomics. Some vigorous software regarding PseAAC has been developed as an open source, such as PseAAC-Builder [20], protein in python (Propy) [21]. Another approach, known as Pseudo K-tuple nucleotide composition (PseKNC), was introduced to extract different feature vectors from RNA and DNA sequences. This method has been widely used in many research works [22–28].

In order to transform the DNA sequence into vector forms, the proposed model divides all the sequences into overlapping 2-mer compo-

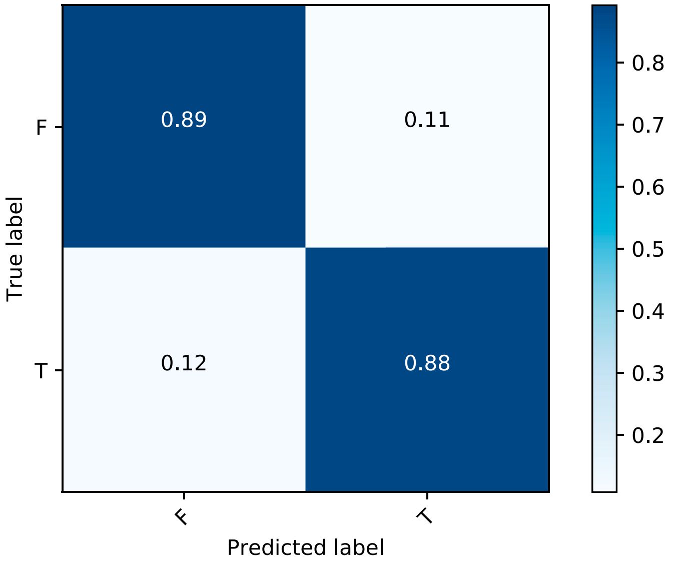


Fig. 4. The visualization of the confusion matrix of computational model i6mA-DNC.

nents by sliding a window of size 2 across the sequences. For example, with 2-mer representation, a biological sequence AGTTCA will result in the following subsequences of 2mers AG, GT, TT, TC, CA. Since a sequence of length L has L-k+1 k-mers, all the samples with 41 nt produce 40 components of 2mers. These 2mer components are then embedded into vector space of 16 dimensions, as all the four single nucleotides (A, C, G and T) can combine each other to produce $4^2 = 16$ possible combination of di-nucleotides. As shown in Fig. 1, AA, AC, TT, for example, are represented as [1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0], [0, 1, 0, 0, 0, 0, 0, 0, 0 0, 0, 0, 0, 0, 0, 0], and [0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1], respectively. Thus, each preprocessed sample has the length of 40 and 16 channels.

After the proposed model converts the DNA samples into vector forms with 2-mer representation, it feeds them to CNN to detect 6 mA sites. The CNN is a method widely used in diverse research areas of bioinformatics [29–39]. While being training, CNN automatically extracts the primary features from the input data.

Convolution neural network is composed of various layers such as convolution layer, pooling layer, ReLU layer, dropout layer, and fully connected layer. Each layer contains different hyper-parameters to be tuned while learning. The best combination of the hyper-parameters of each layer was chosen based on a grid search approach. The tuned hyperparameters are the number of convolution layer, the number of filters in the convolution layers, the size of filters in the convolution layer, and the dropout rate after convolution layers. Table 1 represents the hyperparameter selection of the model.

Table 2 shows the architecture of the proposed model. The Conv1D (f, s) is a one-dimensional convolution layer where f is the number of filters and s is the size of the filter. Every convolution layer is followed by a nonlinear activation function known as rectified linear unit (ReLU).

The Dropout (p) is used as an operator with a probability of p to prevent a model from overfitting. Maxpool1D (m, d) reduces the dimensionality by selecting the maximum value within the window m where m is the pool-size and d is stride. Dense (n) is a fully connected layer with n number of nodes. The last layer is Sigmoid $()$ function which is a nonlinear activation function that squeezes the output values between 0 and 1. We used this layer to predict if the input sequence has an m6A site or non-m6A site.

Our proposed model used Keras framework [40]. The number of epoch, the learning rate, and the batch size were set to 50, 0.005, and 32, respectively. The Adam optimizer was used [41].

2.3. Performance evaluation

In order to evaluate the success rate of the trained model, we selected and performed proper cross-validation. Four metrics are usually used to calculate the performance of prediction system. They are accuracy (ACC), Mathew's correlation coefficient (MCC), specificity (Sp), and sensitivity (Sn). Followings are the mathematical formulation of the four metrics which based on Chou's symbols and already discussed in those papers [42–44].

$$ACC = 1 - \left(\frac{M_{-}^{+} + M_{+}^{-}}{M^{+} + M^{-}}\right) 0 \leq ACC \leq 1$$

$$Sn = 1 - \left(\frac{M_{-}^{+}}{M^{+}}\right) 0 \leq Sn \leq 1$$

$$Sp = 1 - \left(\frac{M_{+}^{-}}{M^{-}}\right) 0 \leq Sp \leq 1$$

$$MCC = \frac{1 - \left(\frac{M_{-}^{+} + M_{+}^{-}}{M^{+} + M^{-}}\right)}{\sqrt{(}1 + \frac{M_{+}^{-} - M_{-}^{+}}{M^{+}}\Big) \sqrt{(}1 + \frac{M_{-}^{+} - M_{+}^{-}}{M^{-}}\Big)} - 1 \leq MCC \leq 1$$

Where M^{+} represents the total count of positive samples that contains methyladenosine sites, while M_{-}^{+} is the count of positive samples that are incorrectly classified as negative ones by predictors. M^{-} represents the total count of negative samples that contain non-methyladenosine sites, while M_{+}^{-} is the count of negative samples that are incorrectly classified as positive ones by the predictor. These metrics are broadly used in bioinformatics researches e.g. Ref. [45–51]. However it supports only single-label systems.

2.3.1. Cross-validation

Choosing a precise cross-validation method is also a foremost part for evaluating the performance of the proposed model. There are three basic cross validation methods: independent dataset test, k-fold cross-validation, and jackknife test [52]. Among these, the k-fold cross-validation is the most robust method used for performance evaluation of predictors. In this work, we used 10-fold cross-validation by dividing the whole dataset randomly into 10 partitions of approximately equal sizes of 10 subsets. A single subset was used as a test dataset to evaluate the proposed model while the remaining 9 subsets were retained as training datasets. The cross-validation process was repeated 10 times, with every subset used once as the test data. Then we calculated the average from the 10 results of all folds.

3. Results and discussion

In this part, we discuss the achievement of the proposed model using benchmark dataset via the k-fold cross-validation test. Table 3 shows the comparison of success rate between the proposed model i6mA-DNC and the existing state-of-the-art models 6 mA-Pred [16], iDNA6mA(5step rule) [18], and iN6-methylat(5-step) [17] for identifying N6-methyladenosine sites in the rice genome.

The proposed model obtains 0.77 of the MCC which is 2% increase from the iN6methylat(5-step) model. The Accuracy is 86.60 which is 1.18% more than the last state-of-the-art method. Whereas Sp, Sn has 89.20 and 88.01 successive values with the increment of 0.11% and 1.5% 3 from the iN6methylat(5-step), respectively.

The area under the receiver operating characteristics curve (auROC) is used to evaluate the performance quality of a classifier. Fig. 2 represents the auROC of i6mA-DNC, which is 95.93. Fig. 3 is a graphical illustration that compares the performance between the four prediction systems: i6mA-DNC, iN6-methylat (5-step) [17], iDNA6mA [18], and

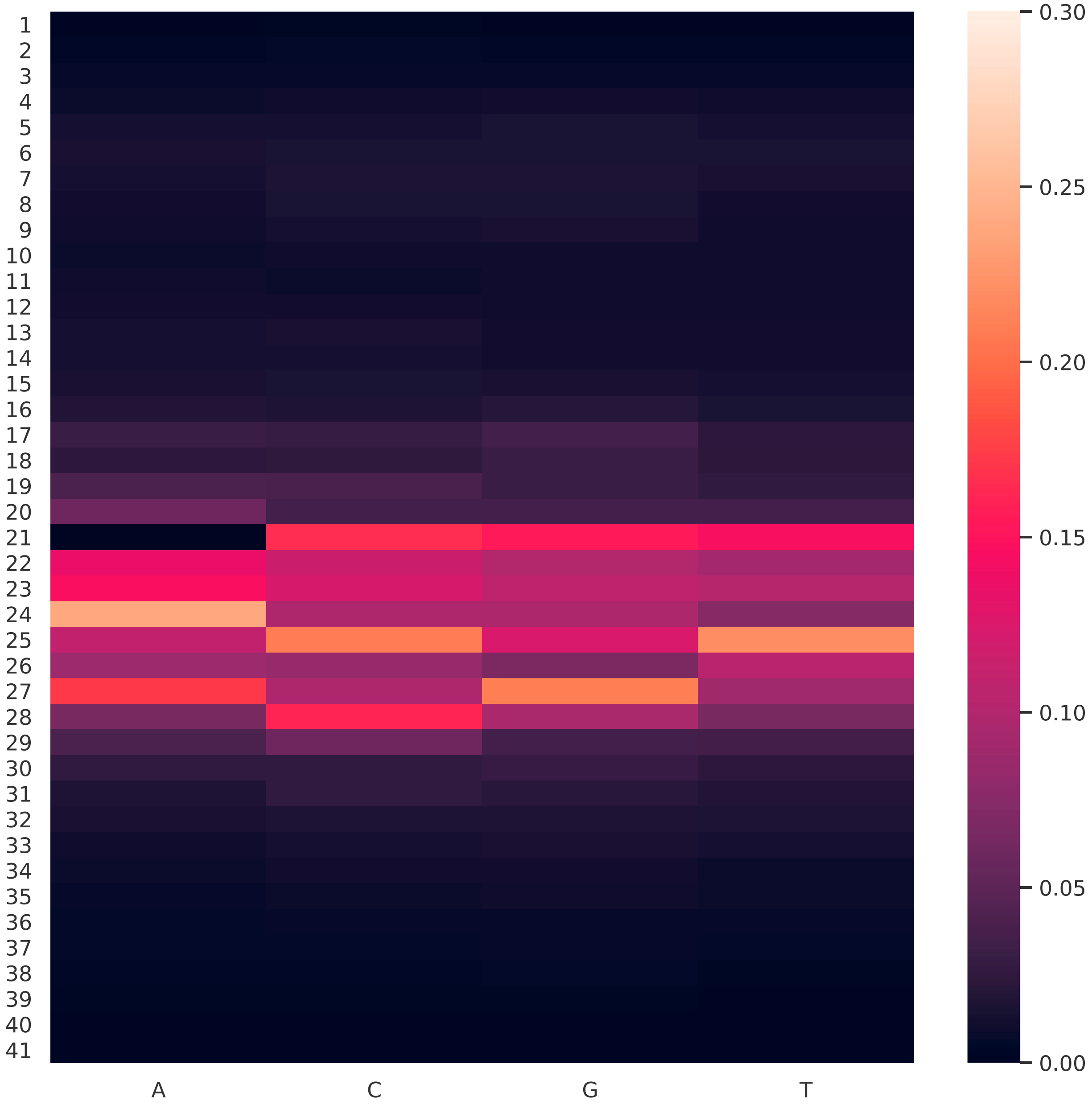


Fig. 5. The heatmap of the effects of the mutations on the prediction results.

i6mA-Pred [16]. This figure shows that our proposed model produced the best result. Fig. 4 shows the visualization of the confusion matrix performance of i6mA-DNC.

For the comprehensive evaluation, we tried to test the model on an independent dataset which is available at (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE103145). In this dataset, we selected the sequences with association modification score (ModQV) with more than 30 as well as they share less than 60% of similarity of the sequences in the benchmark dataset. As a result, we obtained 221 6 mA positive sequences. On the other hand, the 221 negative sequences were chosen with A at the center but not detected by SMRT-seq. The achieved results are 88.64%, 89.09%, 88.18%, 77.28% for Acc, Sn, Sp, and MCC, respectively.

Finally, we study the ability of the proposed model to figure out in-silico mutagenesis. For this, we computationally mutate the nucleotides in the benchmark dataset and we study the effect of the mutation on the final prediction outcomes. For example; for each sequence with length L, we have L x 3 possible mutations. For each mutation, we save the absolute difference. Then, we take an average of the prediction score modifications due to the mutations over the whole sequences in the benchmark dataset. The heat map in Fig. 5 illustrates that the mutations in the center of the sequence affect the final prediction results more than the mutations in both sides of the sequence.

4. Web-server for predictor model

The last step is to develop an easily accessible web-server for the proposed model. Many researchers employed this step in their publications e.g. Ref. [53–58], to enable everyone, especially biologists, to operate a user-friendly computational prediction model. Due to those web oriented computational tools, the impact of bioinformatics on medical science raised dramatically [59]. For more convenience, web-server for i6mA-DNC predictor has also been established and made available at https://home.jbnu.ac.kr/NSCL/i6mA-DNC.htm.

5. Conclusion

Identification of DNA methylation sites is crucial for understanding various biological processes for both drug development and academia. In this study, we designed a novel computational model based on CNN to identify 6 mA sites from DNA samples. We frist converted the raw genomic sequences into groups of dinucleotide components and then fed the pre-processed samples to CNN to identify 6 mA sites. Our proposed model was tested on a benchmark dataset and surpassed the existing state-of-the-art models with regard to every evaluation parameters(sensitivity, specificity, accuracy, Mathew's correlation coefficient, and area under the receiver operating characteristics curve). We also tried the comprehensive evaluation of the proposed model based on independent dataset and achieved better results. To observe the ability of the proposed model, in-silico mutagenesis was tried by computationally mutate the nucleotides in the benchmark dataset. Finally, a user-friendly web server for the proposed model has been constructed and made available at https://home.jbnu.ac.kr/NSCL/i6mA-DNC.htm.

Declaration of competing interest

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

CRediT authorship contribution statement

Sehi Park: Conceptualization, Methodology, Investigation, Writing - original draft, Writing - review & editing, Visualization, Supervision. **Abdul Wahab:** Methodology, Investigation, Writing - review & editing, Visualization, Supervision. **Iman Nazari:** Investigation, Writing - original draft, Visualization. **Ji Hyoung Ryu:** Writing - original draft, Visualization. **Kil To Chong:** Writing - original draft, Visualization.

Acknowledgements

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

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.chemolab.2020.104102.

References

[1] B. Vanyushin, S. Tkacheva, A. Belozersky, Rare bases in animal dna, Nature 225 (5236) (1970) 948.

[2] B. Vanyushin, A. Belozersky, N. Kokurina, D. Kadirova, 5-methylcytosine and 6-methylaminopurine in bacterial dna, Nature 218 (5146) (1968) 1066.

[3] D. Dunn, J. Smith, Occurrence of a new base in the deoxyribonucleic acid of a strain of bacterium coli, Nature 175 (4451) (1955) 336.

[4] Z.K. O'Brown, E.L. Greer, N6-methyladenine: a conserved and dynamic dna mark, in: DNA Methyltransferases-Role and Function, Springer, 2016, pp. 213–246.

[5] J.L. Campbell, N. Kleckner, E. coli, Oric and the dnaa gene promoter are sequestered from dam methyltransferase following the passage of the chromosomal replication fork, Cell 62 (5) (1990) 967–979.

[6] 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 (20) (2005) 7027–7037.

[7] 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 (4) (1983) 571–582.

[8] S.E. Luria, M.L. Human, A nonhereditary, host-induced variation of bacterial viruses, J. Bacteriol. 64 (4) (1952) 557.

[9] M. Meselson, R. Yuan, Dna restriction enzyme from e. coli, Nature 217 (5134) (1968) 1110.

[10] S. Linn, W. Arber, Host specificity of dna produced by escherichia coli, x. in vitro restriction of phage fd replicative form, Proc. Natl. Acad. Sci. U.S.A. 59 (4) (1968) 1300.

[11] E.L. Greer, M.A. Blanco, L. Gu, E. Sendinc, J. Liu, D. Aristizábal-Corrales, C.-H. Hsu, L. Aravind, C. He, Y. Shi, Dna methylation on n6-adenine in c. elegans, Cell 161 (4) (2015) 868–878.

[12] 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 (6) (2010) 461.

[13] K.R. Pomraning, K.M. Smith, M. Freitag, Genome-wide high throughput analysis of dna methylation in eukaryotes, Methods 47 (3) (2009) 142–150.

[14] A.M. Krais, M.G. Cornelius, H.H. Schmeiser, Genomic n6-methyladenine determination by mekc with lif, Electrophoresis 31 (21) (2010) 3548–3551.

[15] C. Zhou, C. Wang, H. Liu, Q. Zhou, Q. Liu, Y. Guo, T. Peng, J. Song, J. Zhang, L. Chen, et al., Identification and analysis of adenine n 6-methylation sites in the rice genome, Nat. Plants 4 (8) (2018) 554.

[16] W. Chen, H. Lv, F. Nie, H. Lin, i6ma-pred: identifying dna n6-methyladenine sites in the rice genome, Bioinformatics 35 (16) (2019) 2796–2800.

[17] N.Q.K. Le, in6-methylat (5-step): identifying dna n 6-methyladenine sites in rice genome using continuous bag of nucleobases via chou's 5-step rule, Mol. Genet. Genom. (2019) 1–10.

[18] M. Tahir, H. Tayara, K.T. Chong, idna6ma (5-step rule): identification of dna n6-methyladenine sites in the rice genome by intelligent computational model via chou's 5-step rule, Chemometr. Intell. Lab. Syst. 189 (2019) 96–101.

[19] K.-C. Chou, Using amphiphilic pseudo amino acid composition to predict enzyme subfamily classes, Bioinformatics 21 (1) (2004) 10–19.

[20] 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.

[21] 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.

[22] 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.

[23] 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.

[24] W. Chen, H. Tang, J. Ye, H. Lin, K.-C. Chou, irna-pseu: identifying rna pseudouridine sites, Mol. Ther. Nucleic Acids 5 (2016) e332.

[25] 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 (3) (2015) 362–369.

[26] B. Liu, R. Long, K.-C. Chou, idhs-el: identifying dnase i hypersensitive sites by fusing three different modes of pseudo nucleotide composition into an ensemble learning framework, Bioinformatics 32 (16) (2016) 2411–2418.

[27] 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.

[28] 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.

[29] B. Linder, A.V. Grozhik, A.O. Olarerin-George, C. Meydan, C.E. Mason, S.R. Jaffrey, Single-nucleotide-resolution mapping of m6a and m6am throughout the transcriptome, Nat. Methods 12 (8) (2015) 767.

[30] 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.

[31] 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–e91.

[32] P. Xing, R. Su, F. Guo, L. Wei, Identifying n 6-methyladenosine sites using multi-interval nucleotide pair position specificity and support vector machine, Sci. Rep. 7 (2017) 46757.

[33] 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.

[34] 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 (2017) 2220–2230.

[35] I. Nazari, H. Tayara, K.T. Chong, Branch point selection in rna splicing using deep learning, IEEE Access 7 (2018) 1800–1807.

[36] 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.

[37] 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.

[38] A. Wahab, S.D. Ali, H. Tayara, K.T. Chong, iim-cnn: intelligent identifier of 6ma sites on different species by using convolution neural network, IEEE Access 7 (2019) 178577–178583.

[39] H. Tayara, K.T. Chong, Improving the quantification of dna sequences using evolutionary information based on deep learning, Cells 8 (12) (2019) 1635.

[40] F. Chollet, et al., Keras: deep learning library for theano and tensorflow (8) (2015) T1, https://keras.io/k 7, 2015, 8, T1.

[41] D.P. Kingma, J. Ba, Adam, A Method for Stochastic Optimization, 2014 arXiv preprint arXiv:1412.6980.

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

[43] K.-C. Chou, Prediction of signal peptides using scaled window, Peptides 22 (12) (2001) 1973–1979.

[44] 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.

[45] Y.-N. Fan, X. Xiao, J.-L. Min, K.-C. Chou, inr-drug, Predicting the interaction of drugs with nuclear receptors in cellular networking, Int. J. Mol. Sci. 15 (3) (2014) 4915–4937.

[46] J. Jia, Z. Liu, X. Xiao, B. Liu, K.-C. Chou, ippbs-opt: a sequence-based ensemble classifier for identifying protein-protein binding sites by optimizing imbalanced training datasets, Molecules 21 (1) (2016) 95.

[47] 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.

[48] L. Cai, W. Yuan, Z. Zhang, L. He, K.-C. Chou, In-depth comparison of somatic point mutation callers based on different tumor next-generation sequencing depth data, Sci. Rep. 6 (2016) 36540.

[49] 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.

[50] J. Khanal, H. Tayara, K.T. Chong, Identifying enhancers and their strength by the integration of word embedding and convolution neural network, IEEE Access 8 (2020) 58369–58376.

[51] O. Mahmoudi, A. Wahab, K.T. Chong, imethyl-deep: N6 methyladenosine identification of yeast genome with automatic feature extraction technique by using deep learning algorithm, Genes 11 (5) (2020).

[52] Z.C. KC, Review: prediction of protein structural classes, Crit. Rev. Biochem. Mol. Biol. 30 (1995) 275–349.

[53] X. Cheng, S.-G. Zhao, W.-Z. Lin, X. Xiao, K.-C. Chou, ploc-manimal: predict subcellular localization of animal proteins with both single and multiple sites, Bioinformatics 33 (22) (2017) 3524–3531.

[54] 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 (2017) 330, 09.

[55] 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.

[56] 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.

[57] J. Wang, B. Yang, J. Revote, A. Leier, T.T. Marquez-Lago, G. Webb, J. Song, K.-C. Chou, T. Lithgow, Possum: a bioinformatics toolkit for generating numerical sequence feature descriptors based on pssm profiles, Bioinformatics 33 (17) (2017) 2756–2758.

[58] Z. Chen, P. Zhao, F. Li, A. Leier, T.T. Marquez-Lago, Y. Wang, G.I. Webb, A.I. Smith, R.J. Daly, K.-C. Chou, et al., ifeature: a python package and web server for features extraction and selection from protein and peptide sequences, Bioinformatics 34 (14) (2018) 2499–2502.

[59] K.-C. Chou, Impacts of bioinformatics to medicinal chemistry, Med. Chem. 11 (3) (2015) 218–234.