genes

MDPI

Article

4mCPred-CNN—Prediction of DNA N4-Methylcytosine in the Mouse Genome Using a Convolutional Neural Network

Zeeshan Abbas [1,2], Hilal Tayara [3,*] and Kil To Chong [1,4,*]

1 Department of Electronics and Information Engineering, Jeonbuk National University, Jeonju 54896, Korea; zabbas@jbnu.ac.kr
2 Institute of Avionics and Aeronautics (IAA), Air University, Islamabad 44000, Pakistan
3 School of International Engineering and Science, Jeonbuk National University, Jeonju 54896, Korea
4 Advanced Electronics and Information Research Center, Jeonbuk National University, Jeonju 54896, Korea
* Correspondence: hilaltayara@jbnu.ac.kr (H.T.); kitchong@jbnu.ac.kr (K.T.C.)

Abstract: Among DNA modifications, N4-methylcytosine (4mC) is one of the most significant ones, and it is linked to the development of cell proliferation and gene expression. To know different its biological functions, the accurate detection of 4mC sites is required. Although we have several techniques for the prediction of 4mC sites in different genomes based on both machine learning (ML) and convolutional neural networks (CNNs), there is no CNN-based tool for the identification of 4mC sites in the mouse genome. In this article, a CNN-based model named 4mCPred-CNN was developed to classify 4mC locations in the mouse genome. Until now, we had only two ML-based models for this purpose; they utilized several feature encoding schemes, and thus still had a lot of space available to improve the prediction accuracy. Utilizing only a single feature encoding scheme—one-hot encoding—we outperformed both of the previous ML-based techniques. In a ten-fold validation test, the proposed model, 4mCPred-CNN, achieved an accuracy of 85.71% and Matthews correlation coefficient (MCC) of 0.717. On an independent dataset, the achieved accuracy was 87.50% with an MCC value of 0.750. The attained results exhibit that the proposed model can be of great use for researchers in the fields of biology and bioinformatics.

Keywords: N4-methylcytosine; computational biology; neural networks; epigenetics

Citation: Abbas, Z.; Tayara, H.; Chongx, K.T. 4mCPred-CNN—Prediction of DNA N4-Methylcytosine in the Mouse Genome Using a Convolutional Neural Network. *Genes* **2021**, *12*, 296. https://doi.org/10.3390/genes12020296

Academic Editor: Federico Divina
Received: 24 January 2021
Accepted: 17 February 2021
Published: 20 February 2021

1. Introduction

In prokaryotes and eukaryotes, alterations of DNA, such as 4-methylcytosine (4mC), 5-Methylcytosine (5mC), and N6-methyladenine (6mA), play key roles in the regulation of gene expression [1–3]. N6-methyladenine (6mA) has recently been identified as one of the common modifications in prokaryotes; it has epigenetic functions in the regulation of chromatin organization and retrotransposons [4–9]. The modification of 5-Methylcytosine (5mC) is another popular and quite well-explored type of DNA alteration that plays a significant role in biological advancements associated with diseases like diabetes and cancer, along with several neurological disorders [10–12]. 4-methylcytosine (4mC) is also recognized as an effective epigenetic alteration that safeguards the self-DNA from enzyme-mediated degradation. Although 4mC has been investigated less than 5mC, it has various tasks, including DNA replication control, DNA replication error correction, cell cycle functions, and self- and non-self DNA differentiation [13,14].

To recognize the epigenetic 4mC sites, until now, several methodologies have been used, such as methylation-specific polymerase chain reaction (PCR) [15], mass spectrometry [16], whole-genome bisulfite sequencing [17], and single-molecule real-time (SMRT) sequencing [18]. These experimental methods are very costly as well as labor-intensive, and methods like SMRT frequently overestimate 4mC in prokaryotic and eukaryotic DNA [19]; therefore, cost-effective and systematic computational tools are necessary for the identification of 4mC sites in different genomes.

SMRTseq [20] was used to detect 4mC in *Mus musculus* (0.00008%), *Drosophila melanogaster* (0.904%), *Saccharomyces cerevisiae* (0.046%), and *Arabidopsis thaliana* (1.366%). However, 4mC was not detected using ultra-high-performance liquid chromatography coupled with mass spectrometry (UHPLC-ms/ms) [19] in any of these species, as the limit of the detection was set lower than 0.00005%. Using a recently created database called MethSMRT [20], certain computational tools were suggested for the prediction of 4mC sites in different species, such as *Caenorhabditis elegans*, *Escherichia coli*, *Arabidopsis thaliana*, *Drosophila melanogaster*, *Geobacter pickeringii*, *Rosaceae genome*, and *Geoalkalibacter subterraneus* [21–24].

To the best of our knowledge, there are only two tools available for the prediction of 4mC sites in the mouse genome—namely, 4mCpred-EL [21] and i4mC-Mouse [25]. Generally, the mouse is a well-recognized experimental animal because it has nearly the same collection of genes as humans, and it is used to replicate the effects of epigenetic changes involved in the development of mammalian diseases including humans [26,27]. Both of the above-mentioned computational tools are machine learning (ML)-based; they employ different feature encoding techniques, such as electron–ion interaction pseudopotentials (EIIP), binary profile (BPF), dinucleotide binary encoding (DPE), ring-function hydrogen chemical properties (RFHC), Kmer, and trinucleotide physio-chemical properties (TPCP), but they still have space available for improvement of the prediction accuracy, and until now, no studies have used neural networks (NNs) for the prediction of 4mC sites in the mouse genome.

In conventional machine learning, features need to be extracted by a data scientist in order to minimize the sophistication of the data and make the patterns easily apparent for the learning algorithms. On the other hand, neural networks work in a systematic way to extract the high-level features themselves from the data [28–30]. This removes the need for domain knowledge and the extraction of hard-core features. Therefore, we have proposed a CNN-based model for the first time for this specific mouse dataset to achieve higher accuracy for the prediction of 4mC sites. Unlike the conventional ML-based techniques, it only uses one-hot encoding and nucleotide chemical properties (NCPs) for feature extraction, and it learns high-level abstract features using the neural network architecture.

2. Datasets

A high-quality dataset is required to establish a sequence-based predictor for the identification of 4mC sites. We used the same dataset that was used for 4mCpred-EL [21] and i4mC-Mouse [25]. The positive samples were extracted using the MethSMRT database [20], and they contained cytosine (C) in the center. The length was fixed to 41 bp (base pairs). For the creation of an accurate model and to have a fair comparison with the previous model, 4mCPred-EL [21], we also applied the same 70% CD-HIT and omitted the sequences that displayed more than 70% similarity. Following this screening method, the positive sequences (that had 4mC sites) of the benchmark dataset were eventually collected. The same number of negative sequences (that did not have 4mC sites) were extracted randomly from the sequences that were not detected as 4mCs, and hence, a balanced dataset was acquired. After acquiring the balanced dataset, we divided it into a ratio that could be used to find the training and independent sets. Therefore, the final training set contained 746 positives (4mCs) and 746 negatives (non-4mCs), and the independent set contained 160 positives (4mCs) and 160 negatives (non-4mCs).

3. Proposed Methodology

Based on the benchmark dataset for mice, we developed a CNN-based model called 4mCPred-CNN. The DNA sequences in the benchmark dataset were represented in string form, such as with "AGACT…CTAAT", with each having a length of 41 bp. Since neural networks only recognize numerical data, the strings should be transformed into numerical format before introducing them as input to the model. Both of the earlier approaches, 4mCpred-EL and i4mC-Mouse, used handcrafted features, like Kmer, MBE, EIIP, KSNC, DBE, and DPC to represent the string-like sequences in a numerical format. The handcrafted

feature extraction needs a significant amount of background knowledge; therefore, rather than using all of these, we only used the one-hot encoding and NCP methods. Both of the encoding methods are explained briefly in the following subsections.

3.1. One-Hot Encoding

One-hot encoding is a straightforward and reliable encoding scheme that is also known as binary encoding. Using the binary representation of nucleotides, this encoder establishes sequence characteristics [31]. Nucleotides are translated into the following formats by the one-hot encoding algorithm:

$$\begin{matrix} A : 1,0,0,0 \\ T : 0,1,0,0 \\ C : 0,0,1,0 \\ G : 0,0,0,1 \end{matrix}.$$

It is then possible to transform any DNA sequence of m nucleotides into a vector of $4 \times m$ features [32,33]. The nucleotide representation is not specific, and the A, T, C, and G representations are exchangeable.

3.2. Nucleotide Chemical Properties

Four groups of nucleotides make up DNA: namely, adenine (A), guanine (G), cytosine (C), and guanine (G). There are various properties of DNA, such as functional groups, ring structures, and hydrogen bonds [34–36]. A and G each hold two rings, while there is only one in C and T. A and T form weak hydrogen bonds with regard to secondary structures, while strong hydrogen bonds are formed by C and G. A and C make up the amino group with respect to functional groups, while G and T make up the keto group. The method for feature extraction can be described as follows:

$$a = \begin{cases} 1, & n \in \{\text{A , G}\} \\ 0, & \text{otherwise} \end{cases}$$

$$b = \begin{cases} 1, & n \in \{\text{A , T}\} \\ 0, & \text{otherwise} \end{cases}$$

$$c = \begin{cases} 1, & n \in \{\text{A , C}\} \\ 0, & \text{otherwise.} \end{cases}$$

Here, n denotes a nucleotide that can be translated into the following format:

$$A : 1,1,1$$

$$T : 0,1,0$$

$$C : 0,0,1$$

$$G : 1,0,0$$

For example, using this technique, a DNA sequence, "ATTCGGT", can be translated into a vector as (1,1,1,0,1,0,0,1,0,0,0,1,1,0,0,1,0,0,0,1,0). The NCPs have similar characteristics to those of one-hot encoding, all of which can be assumed to yield distinct nucleotide representations.

Each sequence is translated into a matrix with 41 rows and four columns using one-hot encoding, where each column constitutes a particular DNA base of the sequence. Similarly, NCP encodes the sequences into a matrix with 41 rows and three columns. In short, only the basic composition derived using one-hot encoding and NCP comprises the data supplied to our model. Figure 1 shows the proposed model's block diagram.

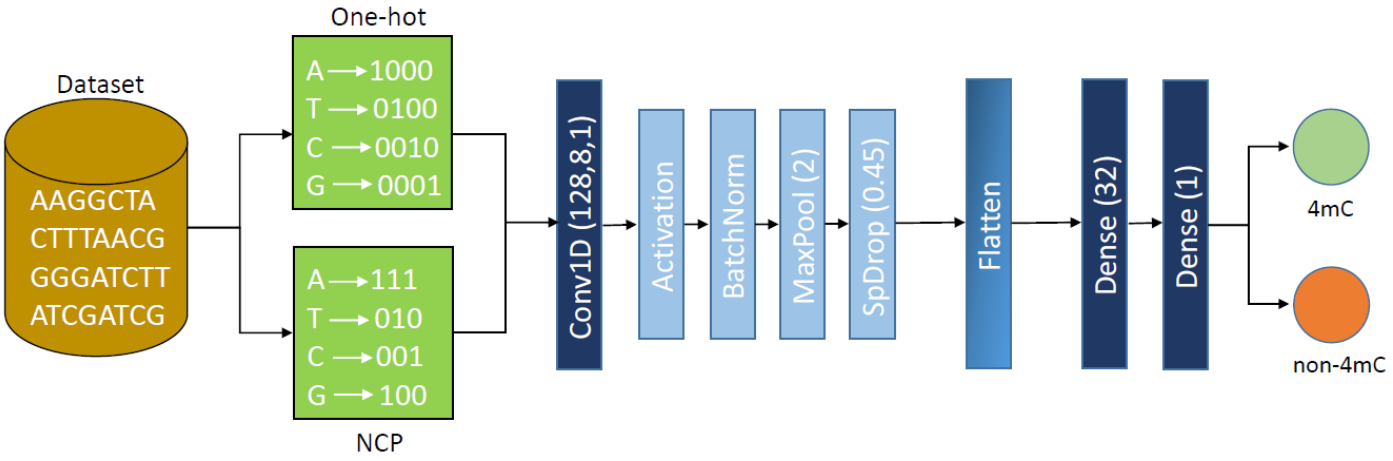


Figure 1. The architecture of 4mCPred-CNN.

Our method uses a standard convolutional neural network consisting of a single 1D convolutional layer (Conv1D), followed by batch-normalization, pooling, and dropout layers. The Conv1D layer consists of 128 filters with a kernel size of 8. It is then followed by a linear activation function. The output of the activation function is then normalized using batch normalization (BatchNorm) to decrease the association of the results that each filter produces, which is then followed by max pooling (MaxPool) with a pool size of 2. Before using the flattening layer, we use the SpatialDropout1D (SpDrop) layer to minimize the number of parameters, with a dropout value of 0.45. The SpatialDropout1D plays the same role as the normal dropout, but rather than dropping individual nodes, it drops full 1D feature maps. If adjacent frames are closely correlated within function maps, the activations will not be regularized by normal dropouts, and will otherwise only result in an overall reduction in the learning rate. Since SpatialDropout1D can help to promote independence among function maps, we preferred to use it. After the spatial dropout, we flattened the output and fed it as input into the first dense layer, which contained 32 hidden units. We kept the same linear activation function as in the first dense layer. This dense layer was then followed by the final dense output layer, which had a single node that used a sigmoid as the activation function, which would help to classify the sequence as 4mC or non-4mC. The sigmoid activation function generated a probability score between 0 and 1. If the generated score was above 0.5, the sequence would be classified as positive or 4mC, and if the score was below 0.5, then the sequence would be classified as non-4mC or negative. Mathematically, this can be represented as:

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

The selected hyper-parameters were noted by using the well-known grid search algorithm.

As an optimizer, stochastic gradient descent (SGD) with a learning rate of 0.003 and a momentum of 0.8 was used in our model. The whole model was based on Keras 2.3.1 https://keras.io/ (accessed on 24 January 2021).

4. Evaluation Metrics

To evaluate the model's performance and to make a fair comparison with previous methodologies, we used the same five metrics, including the accuracy, sensitivity, specificity, Matthews correlation coefficient (MCC), and area under the curve (AUC). These can be defined as:

$$Accuracy = Acc = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

$$Sensitivity = Sn = \frac{TP}{TP + FN} \quad (2)$$

$$Specificity = Sp = \frac{TN}{TN + FP} \quad (3)$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}, \quad (4)$$

where TP is for true positives, TN is for true negatives, FP is for false positives, and FN denotes false negatives.

5. Results

The proposed model was evaluated using the benchmark dataset along with an independent dataset. First, we compared our own results achieved using one-hot encoding and the NCPs on the independent dataset; then, the best approach was compared with previous methodologies.

5.1. One-Hot vs. NCPs

We carried out experiments using the two encoding schemes, one-hot encoding and NCPs, and compared their results, as shown in Table 1. It can be clearly seen that one-hot encoding provided a better result compared to that of the NCPs. So, we will carry out the rest of our paper using the results achieved using one-hot encoding to compare with the previous methodologies.

Table 1. Performance comparison between one-hot encoding and nucleotide chemical properties (NCPs) on an independent dataset.

Encoding Technique	Acc (%)	Sn (%)	Sp (%)	MCC	AUC
One-hot	87.50	88.75	86.25	0.75	0.95
NCPs	83.71	81.72	85.68	0.73	0.93

5.2. Comparison of 4mCPred-CNN with Previous Models on the Benchmark Dataset

To prove that our neural network based technique is superior to the previous machine-learning-based methodologies, 4mCpred-EL [21] and i4mC-Mouse [25], using the benchmark dataset, we used the same ten-fold cross-validation to get a fair comparison of the results. The results clearly depict the better performance of our model, 4mCPred-CNN, compared to the previous state-of-the-art methodologies. We used the same five evaluation metrics mentioned in Section 4 to remain consistent with the criteria of measurement seen in these studies. The outputs of 4mCpred-EL [21] and i4mC-Mouse [25] were explicitly quoted from the existing analyses. We observed that both 4mCpred-EL and i4mC-Mouse were outperformed by 4mCPred-CNN with respect to the five assessment metrics. Table 2 shows a comparison of the aforementioned existing methods with the proposed model, which outperformed 4mCpred-EL and i4mC-Mouse by 6.22 and 6.42 percentage points (p.p) for accuracy, 0.08 and 12.01 p.p for sensitivity, 12.42 and 0.92 p.p for specificity, 12.6 and 6.6 p.p for the MCC, and 3.6 and 0.6 p.p for AUC, respectively. Figure 2 provides a graphical representation of the achieved results.

Table 2. Performance comparison between 4mCPred-CNN and previous machine learning (ML)-based methods on the benchmark dataset.

Methods	Acc (%)	Sn (%)	Sp (%)	MCC	AUC
4mCpred-EL	79.50	80.40	78.70	0.591	0.874
i4mC-Mouse	79.30	68.31	90.20	0.651	0.904
4mCPred-CNN	85.72	80.32	91.12	0.717	0.910

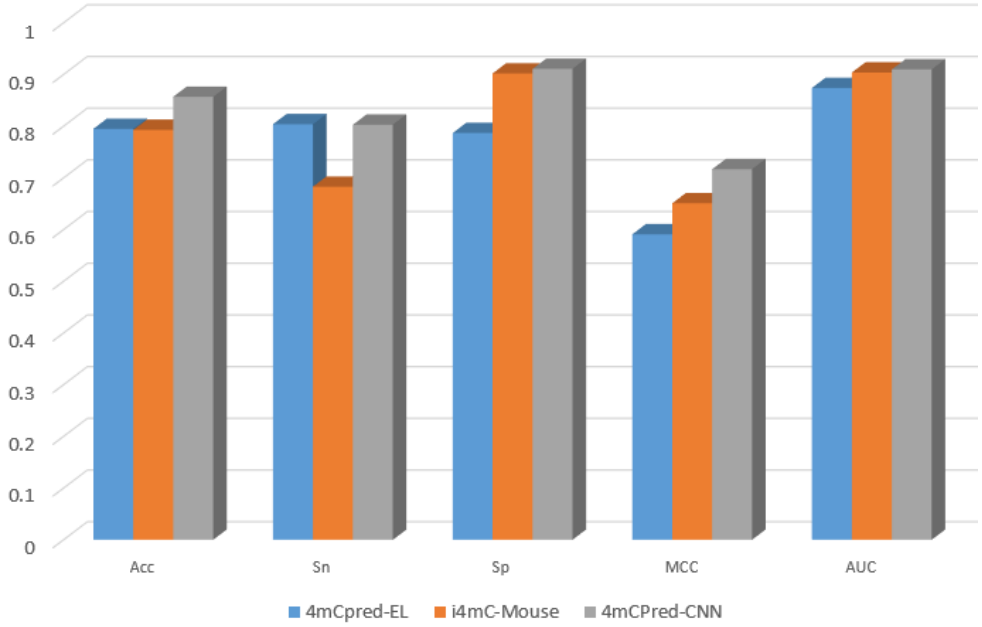


Figure 2. Performance comparison between 4mCPred-CNN and previous ML-based methods on the benchmark dataset—graphical representation.

5.3. Comparison of 4mCPred-CNN with Previous Models on the Independent Dataset

Using the same independent dataset used for 4mCpred-EL and i4mC-Mouse, which was comprising of 160 4mCs and the same number of non-4mCs, we compared 4mCPred-CNN with these two state-of-the-art techniques and illustrated the results in Table 3. Similarly to what was found above, the outputs of 4mCpred-EL and i4mC-Mouse on the independent dataset were directly quoted from i4mC-Mouse [25], which was yielded by submitting the dataset directly to the server. The proposed model, 4mCPred-CNN, outperformed 4mCpred-EL and i4mC-Mouse by 8.4 and 5.89 p.p for accuracy, 13.03 and 8.04 p.p for sensitivity, 3.74 and 3.73 p.p for specificity, 16.6 and 11.7 p.p for the MCC, and 6.9 and 3 p.p for AUC, respectively, on the independent dataset. A graphical representation of the achieved results can be seen in Figure 3.

Our system achieved an accuracy of 87.50%, and we believe that room for improvement is still available, but the achieved result is far better than the results achieved by the techniques presented in the literature. In order to further improve the performance, we need to collect additional experimentally validated data to train a more robust model.

To study the effectiveness of the model in the case of a change in DNA motifs, we applied in silico mutagenesis. We mutated the nucleotides of the available sequences so that we could predict the effect of the mutation. One after another, every nucleotide in the sequence was mutated, and the prediction was made using the proposed model. At the final stage, the average of all of the predicted scores attained for a single sequence was used to compute a heat map. Figure 4 shows the generated heat map, and it can be seen that the nucleotides at the center of the sequence had a much higher impact on the prediction than that of the nucleotides present at the edges of the sequence.

Table 3. Performance comparison between 4mCPred-CNN and previous ML-based methods on an independent dataset.

Methods	Acc (%)	Sn (%)	Sp (%)	MCC	AUC
4mCpred-EL	79.10	75.72	82.51	0.584	0.881
i4mC-Mouse	81.61	80.71	82.52	0.633	0.920
4mCPred-CNN	87.50	88.75	86.25	0.750	0.950

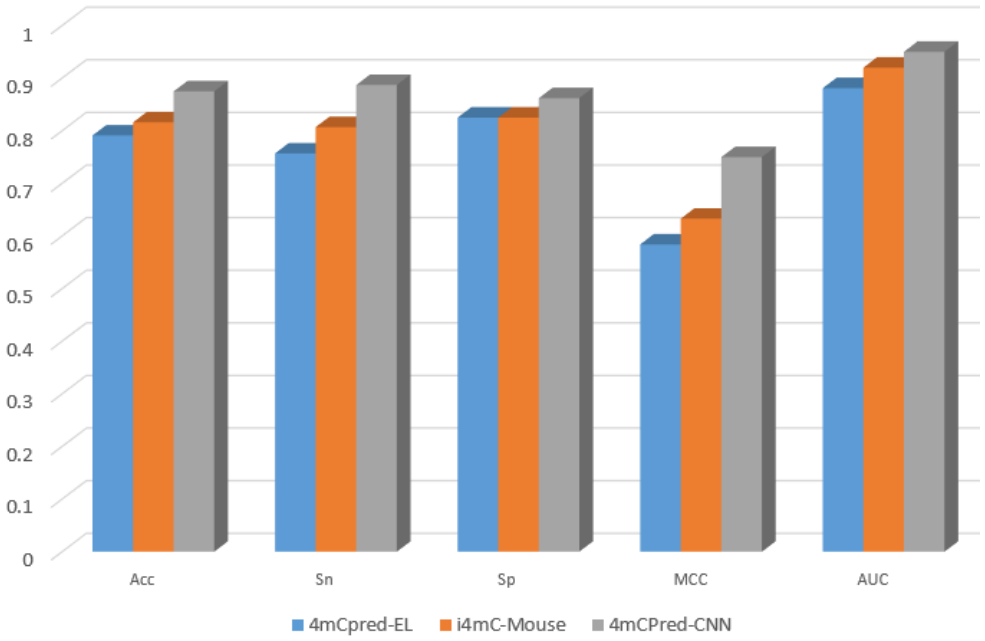


Figure 3. Performance comparison between 4mCPred-CNN and previous ML-based methods on an independent dataset—graphical representation.

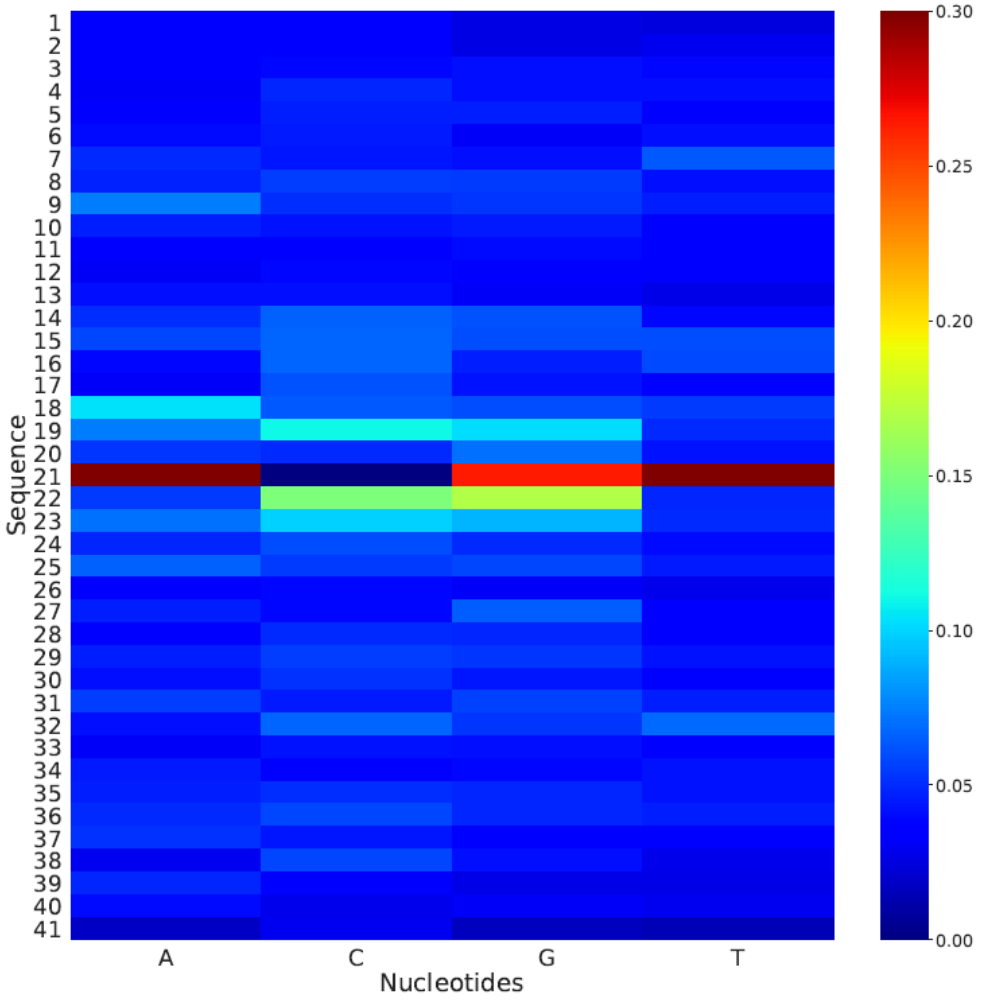
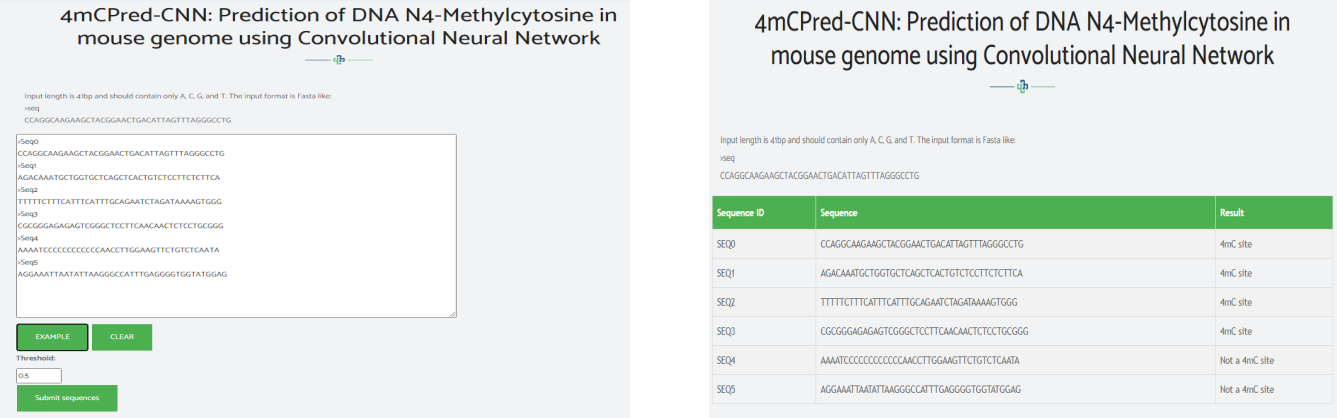


Figure 4. Heat map for analyzing the impacts of mutations in the model assessment.

6. Web Server

To provide the research community with quick access to the proposed tool, a web server was created and made publicly available at http://nsclbio.jbnu.ac.kr/tools/4mCPred-CNN/ (accessed on 24 January 2021). Figure 5 shows a snippet from the web server; an example of adding prediction sequences is shown in Figure 5a, and the performance of the predictor is shown in Figure 5b.



(a) Insertion of prediction sequences **(b)** Predictor output

Figure 5. 4mCPred-CNN: web server snippet.

7. Conclusions

In DNA modifications, 4mC plays a significant role, and it is actively engaged in controlling cell replication and levels of gene expression. Accurate detection of these sites is, therefore, an important step in identifying the particular biological processes. In order to classify 4mC sites among DNA sequences of different species, many computational models have been created by different researchers using both ML and NNs [21,23,37–39], but for the mouse genome, only two ML-based models are available, and no studies have used NNs for this particular species. Using NNs, we proposed a new state-of-the-art model called 4mCPred-CNN to enhance the accuracy of prediction of 4mC sites in the mouse genome. 4mCPred-CNN surpassed 4mCpred-EL and i4mC-Mouse by 8.4 and 5.89 p.p for accuracy, 13.03 and 8.04 p.p for sensitivity, 3.74 and 3.73 p.p for specificity, 16.6 and 11.7 p.p for the MCC, and 6.9 and 3 p.p for AUC, respectively, when tested on an independent dataset.

Author Contributions: Conceptualization, Z.A., H.T., and K.T.C.; methodology, Z.A.; software, Z.A. and H.T.; validation, Z.A., H.T., and K.T.C.; investigation, Z.A., H.T., and K.T.C.; writing—original draft preparation, Z.A.; writing—review and editing, Z.A., H.T., and K.T.C.; supervision, H.T. and K.T.C. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported in part by a National Research Foundation of Korea (NRF) grant funded by the Korean government (MSIT) (No. 2020R1A2C2005612) and in part by the Brain Research Program of the National Research Foundation (NRF) funded by the Korean government (MSIT) (No. NRF-2017M3C7A1044816).

Institutional Review Board Statement: Not applicable as the publically available dataset is used.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data sharing not applicable. No new data were created or analyzed in this study.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

References

1. Rathi, P.; Maurer, S.; Summerer, D. Selective recognition of N 4-methylcytosine in DNA by engineered transcription-activator-like effectors. *Philos. Trans. R. Soc. B: Biol. Sci.* **2018**, *373*, 20170078. [CrossRef] [PubMed]
2. Jeltsch, A.; Jurkowska, R.Z. New concepts in DNA methylation. *Trends Biochem. Sci.* **2014**, *39*, 310–318. [CrossRef] [PubMed]
3. Alam, W.; Ali, S.D.; Tayara, H.; to Chong, K. A CNN-based RNA n6-methyladenosine site predictor for multiple species using heterogeneous features representation. *IEEE Access* **2020**, *8*, 138203–138209. [CrossRef]
4. Wu, T.P.; Wang, T.; Seetin, M.G.; Lai, Y.; Zhu, S.; Lin, K.; Liu, Y.; Byrum, S.D.; Mackintosh, S.G.; Zhong, M.; et al. DNA methylation on N 6-adenine in mammalian embryonic stem cells. *Nature* **2016**, *532*, 329–333. [CrossRef]
5. Ma, C.; Niu, R.; Huang, T.; Shao, L.W.; Peng, Y.; Ding, W.; Wang, Y.; Jia, G.; He, C.; Li, C.Y.; et al. N6-methyldeoxyadenine is a transgenerational epigenetic signal for mitochondrial stress adaptation. *Nat. Cell Biol.* **2019**, *21*, 319–327. [CrossRef]
6. Liu, J.; Zhu, Y.; Luo, G.Z.; Wang, X.; Yue, Y.; Wang, X.; Zong, X.; Chen, K.; Yin, H.; Fu, Y.; et al. Abundant DNA 6mA methylation during early embryogenesis of zebrafish and pig. *Nat. Commun.* **2016**, *7*, 1–7. [CrossRef]
7. Abbas, Z.; Tayara, H.; to Chong, K. SpineNet-6mA: A Novel Deep Learning Tool for Predicting DNA N6-Methyladenine Sites in Genomes. *IEEE Access* **2020**, *8*, 201450–201457. [CrossRef]
8. Rehman, M.U.; Chong, K.T. DNA6mA-MINT: DNA-6mA modification identification neural tool. *Genes* **2020**, *11*, 898. [CrossRef]
9. Rehman, M.U.; Hong, K.J.; Tayara, H.; to Chong, K. m6A-NeuralTool: Convolution Neural Tool for RNA N6-Methyladenosine Site Identification in Different Species. *IEEE Access* **2021**, *9*, 17779–17786. [CrossRef]
10. Jones, P.A. Functions of DNA methylation: Islands, start sites, gene bodies and beyond. *Nat. Rev. Genet.* **2012**, *13*, 484–492. [CrossRef]
11. Ling, C.; Groop, L. Epigenetics: A molecular link between environmental factors and type 2 diabetes. *Diabetes* **2009**, *58*, 2718–2725. [CrossRef]
12. Yao, B.; Jin, P. Cytosine modifications in neurodevelopment and diseases. *Cell. Mol. Life Sci.* **2014**, *71*, 405–418. [CrossRef] [PubMed]
13. Cheng, X. DNA modification by methyltransferases. *Curr. Opin. Struct. Biol.* **1995**, *5*, 4–10. [CrossRef]
14. Chen, K.; Zhao, B.S.; He, C. Nucleic acid modifications in regulation of gene expression. *Cell Chem. Biol.* **2016**, *23*, 74–85. [CrossRef]
15. Ku, J.L.; Jeon, Y.K.; Park, J.G. Methylation-specific PCR. In *Epigenetics Protocols*; Springer: Berlin/Heidelberg, Germany, 2011; pp. 23–32.
16. Domon, B.; Aebersold, R. Mass spectrometry and protein analysis. *Science* **2006**, *312*, 212–217. [CrossRef]
17. Doherty, R.; Couldrey, C. Exploring genome wide bisulfite sequencing for DNA methylation analysis in livestock: A technical assessment. *Front. Genet.* **2014**, *5*, 126. [CrossRef] [PubMed]
18. Ardui, S.; Ameur, A.; Vermeesch, J.R.; Hestand, M.S. Single molecule real-time (SMRT) sequencing comes of age: applications and utilities for medical diagnostics. *Nucleic Acids Res.* **2018**, *46*, 2159–2168. [CrossRef] [PubMed]
19. O'Brown, Z.K.; Boulias, K.; Wang, J.; Wang, S.Y.; O'Brown, N.M.; Hao, Z.; Shibuya, H.; Fady, P.E.; Shi, Y.; He, C.; et al. Sources of artifact in measurements of 6mA and 4mC abundance in eukaryotic genomic DNA. *BMC Genom.* **2019**, *20*, 1–15. [CrossRef]
20. Ye, P.; Luan, Y.; Chen, K.; Liu, Y.; Xiao, C.; Xie, Z. MethSMRT: An integrative database for DNA N6-methyladenine and N4-methylcytosine generated by single-molecular real-time sequencing. *Nucleic Acids Res.* **2016**, *45*, D85–D89.
21. Manavalan, B.; Basith, S.; Shin, T.H.; Lee, D.Y.; Wei, L.; Lee, G. 4mCpred-EL: An ensemble learning framework for identification of DNA N4-methylcytosine sites in the mouse genome. *Cells* **2019**, *8*, 1332. [CrossRef]
22. He, W.; Jia, C.; Zou, Q. 4mCPred: Machine learning methods for DNA N4-methylcytosine sites prediction. *Bioinformatics* **2019**, *35*, 593–601. [CrossRef]
23. Wei, L.; Luan, S.; Nagai, L.A.E.; Su, R.; Zou, Q. Exploring sequence-based features for the improved prediction of DNA N4-methylcytosine sites in multiple species. *Bioinformatics* **2019**, *35*, 1326–1333. [CrossRef] [PubMed]
24. Hasan, M.M.; Manavalan, B.; Khatun, M.S.; Kurata, H. i4mC-ROSE, a bioinformatics tool for the identification of DNA N4-methylcytosine sites in the Rosaceae genome. *Int. J. Biol. Macromol.* **2020**, *157*, 752–758. [CrossRef] [PubMed]
25. Hasan, M.M.; Manavalan, B.; Shoombuatong, W.; Khatun, M.S.; Kurata, H. i4mC-Mouse: Improved identification of DNA N4-methylcytosine sites in the mouse genome using multiple encoding schemes. *Comput. Struct. Biotechnol. J.* **2020**, *18*, 906–912.
26. Espada, J.; Esteller, M. Mouse models in epigenetics: Insights in development and disease. *Briefings Funct. Genom.* **2013**, *12*, 279–287. [CrossRef] [PubMed]
27. Uhl, E.W.; Warner, N.J. Mouse models as predictors of human responses: Evolutionary medicine. *Curr. Pathobiol. Rep.* **2015**, *3*, 219–223. [CrossRef]
28. Zhang, L.; Tan, J.; Han, D.; Zhu, H. From machine learning to deep learning: Progress in machine intelligence for rational drug discovery. *Drug Discov. Today* **2017**, *22*, 1680–1685. [CrossRef]
29. Shen, C.; Ding, J.; Wang, Z.; Cao, D.; Ding, X.; Hou, T. From machine learning to deep learning: Advances in scoring functions for protein–ligand docking. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **2020**, *10*, e1429.
30. Ongsulee, P. Artificial intelligence, machine learning and deep learning. In Proceedings of the 2017 15th International Conference on ICT and Knowledge Engineering (ICT&KE), Bangkok, Thailand, 22–24 November 2017; pp. 1–6.
31. Zou, J.; Huss, M.; Abid, A.; Mohammadi, P.; Torkamani, A.; Telenti, A. A primer on deep learning in genomics. *Nat. Genet.* **2019**, *51*, 12–18.

32. Hao, L.; Dao, F.Y.; Guan, Z.X.; Zhang, D.; Tan, J.X.; Zhang, Y.; Chen, W.; Lin, H. iDNA6mA-Rice: A computational tool for detecting N6-methyladenine sites in rice. *Front. Genet.* **2019**, *10*, 793.
33. Chen, Z.; Zhao, P.; Li, F.; Marquez-Lago, T.T.; Leier, A.; Revote, J.; Zhu, Y.; Powell, D.R.; Akutsu, T.; Webb, G.I.; et al. iLearn: An integrated platform and meta-learner for feature engineering, machine-learning analysis and modeling of DNA, RNA and protein sequence data. *Briefings Bioinform.* **2020**, *21*, 1047–1057. [CrossRef]
34. Tan, J.X.; Lv, H.; Wang, F.; Dao, F.Y.; Chen, W.; Ding, H. A survey for predicting enzyme family classes using machine learning methods. *Curr. Drug Targets* **2019**, *20*, 540–550. [CrossRef]
35. Xue, W.; Yang, F.; Wang, P.; Zheng, G.; Chen, Y.; Yao, X.; Zhu, F. What contributes to serotonin–norepinephrine reuptake inhibitors' dual-targeting mechanism? The key role of transmembrane domain 6 in human serotonin and norepinephrine transporters revealed by molecular dynamics simulation. *ACS Chem. Neurosci.* **2018**, *9*, 1128–1140. [CrossRef] [PubMed]
36. He, S.; Zhang, G.; Wang, J.; Gao, Y.; Sun, R.; Cao, Z.; Chen, Z.; Zheng, X.; Yuan, J.; Luo, Y.; et al. 6mA-DNA-binding factor Jumu controls maternal-to-zygotic transition upstream of Zelda. *Nat. Commun.* **2019**, *10*, 1–14. [CrossRef] [PubMed]
37. Wahab, A.; Mahmoudi, O.; Kim, J.; Chong, K.T. DNC4mC-Deep: Identification and analysis of DNA N4-methylcytosine sites based on different encoding schemes by using deep learning. *Cells* **2020**, *9*, 1756. [CrossRef]
38. Yang, J.; Lang, K.; Zhang, G.; Fan, X.; Chen, Y.; Pian, C. SOMM4mC: A second-order Markov model for DNA N4-methylcytosine site prediction in six species. *Bioinformatics* **2020**, *36*, 4103–4105. [CrossRef]
39. Xu, H.; Jia, P.; Zhao, Z. Deep4mC: Systematic assessment and computational prediction for DNA N4-methylcytosine sites by deep learning. *Brief. Bioinform.* **2020**, bbaa099. [CrossRef]