symmetry

MDPI

Article

iRG-4mC: Neural Network Based Tool for Identification of DNA 4mC Sites in Rosaceae Genome

Dae Yeong Lim [1,†], Mobeen Ur Rehman [1,2,†] and Kil To Chong [1,3,*]

[1] Department of Electronics and Information Engineering, Jeonbuk National University, Jeonju 54896, Korea; ldy@jbnu.ac.kr (D.Y.L.); cmobeenrahman@jbnu.ac.kr (M.U.R.)
[2] Institute of Avionics and Aeronautics (IAA), Air University, Islamabad 44000, Pakistan
[3] Advances Electronics and Information Research Center, Jeonbuk National University, Jeonju 54896, Korea
* Correspondence: kitchong@jbnu.ac.kr
† These authors equally contributed to this work.

Abstract: DNA N4-Methylcytosine is a genetic modification process which has an essential role in changing different biological processes such as DNA conformation, DNA replication, DNA stability, cell development and structural alteration in DNA. Due to its negative effects, it is important to identify the modified 4mC sites. Further, methylcytosine may develop anywhere at cytosine residue, however, clonal gene expression patterns are most likely transmitted just for cytosine residues in strand-symmetrical sequences. For this reason many different experiments are introduced but they proved not to be viable choice due to time limitation and high expenses. Therefore, to date there is still need for an efficient computational method to deal with 4mC sites identification. Keeping it in mind, in this research we have proposed an efficient model for Fragaria vesca (*F. vesca*) and Rosa chinensis (*R. chinensis*) genome. The proposed iRG-4mC tool is developed based on neural network architecture with two encoding schemes to identify the 4mC sites. The iRG-4mC predictor outperformed the existing state-of-the-art computational model by an accuracy difference of 9.95% on *F. vesca* (training dataset), 8.7% on *R. chinesis* (training dataset), 6.2% on *F. vesca* (independent dataset) and 10.6% on *R. chinesis* (independent dataset). We have also established a webserver which is freely accessible for the research community.

Keywords: Convolution Neural Network (CNN); bioinformatics; Long Short-Term Memory (LSTM); N4-methylcytosine; computational biology

Citation: Lim, D.Y.; Rehman, M.U.; Chong, K.T. iRG-4mC: Neural Network Based Tool for Identification of DNA 4mC Sites in Rosaceae Genome. *Symmetry* **2021**, *13*, 899. https://doi.org/10.3390/sym13050899

Academic Editors: Jeffrey A. Thompson, Filip Jagodzinski and Ellen Palmer

Received: 13 April 2021
Accepted: 15 May 2021
Published: 19 May 2021

1. Introduction

DNA modification consisting of methylation and demethylation plays a crucial role in gene regulation. DNA methylation being a heritable epigenetic marker is a kind of chemical modification of DNA that changes the genetic functionality without disrupting the DNA sequence [1,2]. Research demonstrates that DNA modification exhibits the property to modify DNA protein interactions, chromatin structure, stability and conformation [3,4]. It also has a role in the regulation of a few activities such as cell development, chromosome stability and genomic imprinting [5–7].

In prokaryotes and eukaryotes most commonly observed methylations are N4-methylcytosine (4mC) [8], 5-methylcytosine (5mC) [9], and N6-methyladenine (6mA) [10]. Each methylation process has its specific altering site, i.e., 4mC, 5mC and 6mC occurs at 4th, 5th and 6th position of the cytosine respectively. Host DNA present in exogenous pathogenic DNA of prokaryotes can be detected by 6mA and 4mC [11], where 4mC perform error correction and regulation of DNA replication [12,13]. Additionally, 4mC belongs to a restriction-modification system that resists the degradation of host DNA caused by restriction enzymes [14]. 5mC plays an essential role in various activities of eukaryotes, such as gene imprinting, regulation, and transposon suppression. In eukaryotes, 6mA and 4mC can only be identified using high sensitive techniques [15].

The 5mC has been extensively studied and previous studies have proved that 5mC is responsible for various biological processes [16], such as diabetes, few neurological defects and cancer [17–19]. The 4mC exhibits the ability to resist enzyme-mediated degradation to safeguard its own DNA. Furthermore, it can manipulate different activities entailing gene expression levels, DNA replication, cell cycle, discriminating self and non-self-DNA and amendment in DNA replication abnormalities [12,20]. Even after extensive research, the accurate mechanism of 4mC epigenetic modification remains unrevealed. This makes the identification of 4mC sites to be an important task, as its identification can give a better understanding of pathological and physiological mechanisms.

Several experimental studies have been conducted to find 4mC sites throughout the genome, some of them are 4mC-Tet-assisted bisulphite sequencing, Single Molecule of Real Time (SMRT) sequencing, methylation-precise PCR and mass spectrometry [21,22]. These methodologies are time exhaustive, laborious, and financially expensive when utilized for genome-wide testing. Henceforth, it is crucial to follow a computational approach for finding 4mC sites.

Recently, various 4mC sites identifiers have been suggested for several species entailing *A. thaliana*, *C. elegans*, *D. melanogaster*, *G. pickeringii*, *E. coli* and *G. subterraneus* [23,24]. Further a first computational tool for 4mC identification in Rosaceae genomes is recently introduced which is, i4mC-ROSE [25]. This tool has been suggested to predict the 4mC sites within the genome of *Rosa chinensis* (*R. chinensis*) [26] and *Fragaria vesca* (*F. vesca*) [27]. It produced six probabilistic scores by utilizing six encoding schemes; k-space spectral nucleotide composition (KSNC), k-mer composition (Kmer), dinucleotide physicochemical properties (DPCP), electron-ion interaction pseudopotentials (EIIPs), trinucleotide physicochemical properties (TPCPs), and binary encoding (BE) that works on different characteristics of DNA sequence information. These encoded sequences are used to train random forest classifier to identify the 4mC modification. The scores acquired are concatenated with a linear regression model for improved prediction results. Studies have suggested that in plant methylation DNA is mostly found in cytosines belonging to symmetrical sequences, which makes the classification problem further difficult [28].

In recent years, neural networks have acquired great importance due to their high performance in different fields like medical imaging [29–31], agriculture [32–37], image quality assessment and others. Moreover, neural networks have exhibited great performance in the identification of 6mA sites [38,39], m6A sites [40,41], 4mC sites [23,24,42], functional piRNAs [43], N4-acetylcytidine sites [44], promoters classification [45] and others. Inspired from the high performance given by neural network tools for modification identification in different sites, we have proposed a neural network based tool for 4mC identification in *Rosaceae* genomes.

The proposed iRG-4mC tool encodes the input sequence using two techniques which are one-hot encoding and nucleotide chemical properties (NCP). The outputs of the two encoding scheme are combined and given to the Convolution Neural Network (CNN) model. The CNN model extracts the features using convolution layers and then gives optimal representation to these features using Long Short Term Memory (LSTM). The optimized features are used for the prediction of 4mC sites. The proposed architecture has obtained high performance. Performance analysis is carried out by using K-fold cross-validation on the training dataset and by using an independent dataset. While in comparison with the existing state-of-the-art tool which is i4mC-Rose, the proposed iRG-4mC tool have achieved higher performance for training as well as an independent dataset.

2. Benchmark Dataset

The benchmark dataset used in this study is acquired from MDR database [46]. The dataset contains two genomes which are *F. vesca* and *R. chinensis*. All of the sequences present in the dataset have a length of 41 base pairs (bp) and centered with cytosine nucleotide. For maintaining the high quality of the dataset, the positive sequences are verified with the help of modification score (ModQV). The cytosine nucleotide is considered to

be modified if the modification score is greater than 19. Further, CD-HIT software [47] is utilized to encounter the homology bias problem, where sequences with more than 65% similarity are removed.

A large number of negative 4mC sites were gathered too and CD-HIT was applied to remove the the sequences that were having more than 65%. Out of this big dataset of negative samples, we have randomly selected the same numbers of sites as available for positive 4mC, so that there would be no class imbalance problem. The total sites collected were 6471 and 3116 for *F. vesca* and *R. chinensis* respectively for each class. Each dataset is further divided into two sets where 75% is kept for training and remaining 25% is kept as an independent dataset. Table 1 shows the summary of the database.

Table 1. Summary of the databases utilized in this study.

Specie	Sequence Class	Number of Sequences of Seq	Sequence Length
F. vesca (training)	4mC	4854	41 bp
	non-4mC	4854	
R. chinensis (training)	4mC	2337	41 bp
	non-4mC	2337	
F. vesca (Independent)	4mC	1617	41 bp
	non-4mC	1617	
R. chinensis (Independent)	4mC	779	41 bp
	non-4mC	779	

3. Methodology

The DNA sequences were in string form, therefore, an encoding scheme was required before feeding the data to the model. Previous methods in site identification have used different encoding schemes. One-hot encoding and NCP [48] encoding schemes are the most commonly used schemes. In this study, we used both of these schemes, by combining them both. Table 2 shows the summary of the encoding schemes. One-hot encoding assigned 4 bits for each nucleotide, while NCP allocated 3 bits for each nucleotide. The fusion of both encoding schemes resulted in 7 bits for each nucleotide. Each input DNA sequence was 41bp long, the encoding mechanism converted the sequence into a matrix of size 41×7.

Table 2. Summary of encoding scheme.

Nucleotide	One-Hot	NCP	Fusion
A	1,0,0,0	1,1,1	1,0,0,0,1,1,1
C	0,1,0,0	0,0,1	0,1,0,0,0,0,1
G	0,0,1,0	1,0,0	0,0,1,0,1,0,0
T	0,0,0,1	0,1,0	0,0,0,1,0,1,0

Figure 1 shows the proposed iRG-4mC architecture. The final encoded sequence was given to a CNN model that contained multiple layers in a stacked way. The proposed CNN model had two main feature extraction blocks. Each feature extraction block contained a convolution layer followed by a batch normalization layer, max-pooling layer and a dropout layer. The convolution layer fetched the features from the input encoded sequence by a self-regulating mechanism. To get the optimized model, different layers had different parameter settings. In the first feature extraction stage, the convolution layer used 32 filters of size 9 with strides equal to 1 with the dropout layer being set with a ratio of 0.1, while the convolution layer of the second feature extractor used eight filters of size 5 with a single stride and the dropout ratio in this block was set to 0.25. The max-pooling layer of both

stages used a pool-size of 4 and the stride value was kept at 1. Batch normalization was used in both blocks to achieve model stability by normalizing the extracted features.

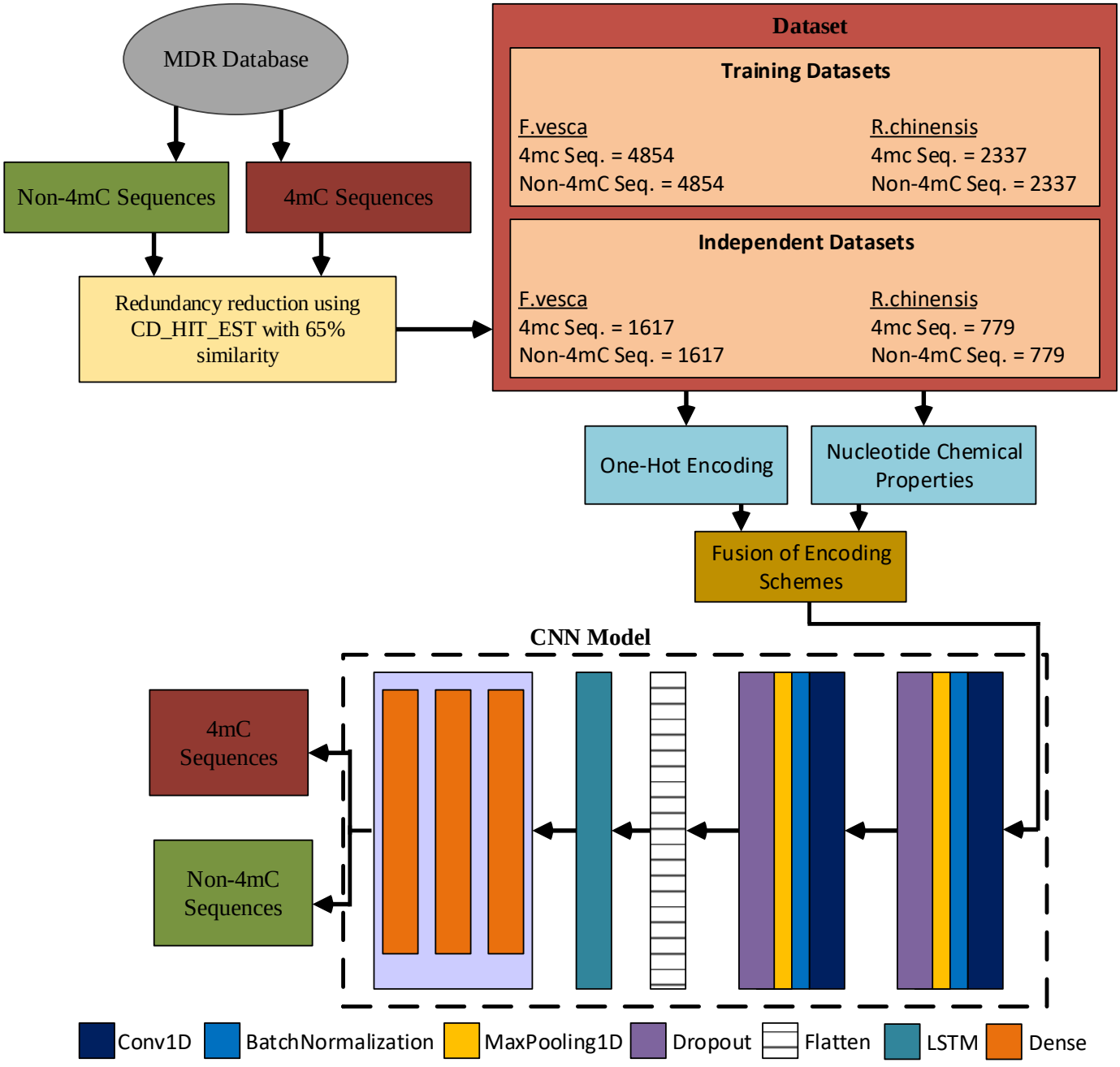

Figure 1. iRG-4mC Architecture for Identification of 4mC sites.

Feature extractor blocks were followed by an LSTM layer which gives an optimal interpretation to the extracted features. It helped the architecture to learn the internal representation of the input sequence. Further LSTM also solved the vanishing gradient problem by adding extra interactions. The output of the LSTM layer was unstacked using the flatten layer. The final feature vector was then given to the three fully connected layers to get the classification of 4mC and non-4mC sites. Table 3 shows the neural network architecture details for iRG-4mC.

Table 3. Neural Network Architecture of iRG-4mC.

Layer	Output Shape	Number of Parameters
Input	(41,7)	-
Conv1D (32,9,1)	(33,32)	2048
Batch Normalization	(33,32)	128
Max Pooling (4,2)	(15,32)	0
Dropout (0.1)	(15,32)	0
Conv1D (8,5,1)	(11,8)	1288
Batch Normalization	(11,8)	32
Max Pooling (4,2)	(4,8)	0
Dropout (0.25)	(4,8)	0
LSTM (16)	(4,16)	0
Flatten	64	0
Dense (64)	64	4160
Dense (32)	32	2080
Dense (1)	1	33

To avoid the over-fitting problem, we used L2 regularization mechanism for weights and bias of the filters. The regularization rates were set to 0.0001. The ReLU activation function was used in convolution, LSTM and first two dense layers. The numerical representation of ReLU is,

$$F(x) = max(0, x) \tag{1}$$

The third dense layer is have a single neuron, so it uses a sigmoid activation function which is represented as,

$$Sigmoid(x) = \frac{1}{1 + e^{-x}} \tag{2}$$

The loss function used in this study was binary cross entropy and stochastic gradient descent (SGD) was used as an optimizer with momentum set at 0.7 and learning rate set at 0.005. The SGD was used as it achieved faster iterations to reduce the system complexity. All filter sizes, number of filters, number of convolution layers, pool sizes, stride length and dropout values were optimized using hyper-parameter tuning.

4. Figure of Merits

The iRG-4mC tool for training datasets was evaluated using k-fold cross-validation, with the value of k set to 10 for a fair comparison with the previous methodology. The whole dataset under consideration was divided into k subsets, where a single subset was kept for testing while the other subsets were used for training purpose. This method was iteratively repeated till the time all subsets were considered for testing. For getting the final performance evaluation of the model, an average was taken of k-trials. For selecting the figure of merits, we followed the previous related work so that we could compare the similar figure of merits including sensitivity (*Sn*), specificity (*Sp*), accuracy (*ACC*) and Matthews correlation coefficient (*MCC*). The metrics can be defined mathematically using the equations:

$$Sn = \frac{TP}{TP + FN} \tag{3}$$

$$Sp = \frac{TN}{TN + FP} \tag{4}$$

$$ACC = \frac{TP + TN}{TP + TN + FP + FN} \tag{5}$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \quad (6)$$

where,

True Positive (*TP*) = 4mC correctly classified as 4mC
False Positive (*FP*) = Non 4mC incorrectly classified as 4mC
True Negative *(TN)* = Non 4mC correctly classified as Non 4mC
False Negative (*FN*) = 4mC incorrectly classified as Non 4mC

Besides these matrices, we used the Receiver Operating Characteristic (ROC) curve to evaluate the performance of the proposed model.

5. Results and Discussion

We evaluated the iRG-4mC tool on four datasets, where two datasets were for *F. vesca* genome and the other two are for *R. chinensis*. For each genome one training dataset and one independent dataset was taken into consideration. The performance evaluation of the training dataset was done using k-fold cross-validation where the value of k was kept at 10. The 10-fold cross-validation was adopted because i4mC-ROSE have also used the same technique and keeping a similar experimental setup allowed having better comparative analysis between both techniques. For the performance evaluation of independent datasets, the training was performed on its respective training dataset and tested on the independent dataset.

Table 4 shows the performance comparison between the proposed model and the i4mC-ROSE model for all datasets taken into account. As can be seen, the proposed iRG-4mC tool showed a noticeable improvement in performance than i4mC-ROSE. In the case of *F. vesca (training/independent)* dataset, the proposed model showed better results on all evaluation matrices, while in the case of *R. chinensis (training/independent)* an improvement was seen in all of the matrices except specificity. The specificity of i4mC-ROSE was slightly higher than the proposed model. As can be observed however, that i4mC-ROSE had a high difference between sensitivity and specificity for all the datasets. This difference depicted that the i4mC-ROSE tool was biased towards one class and that led to a biased decision from the tool. However, in the case of iRG-4mC, the difference between sensitivity and specificity was minimum, which also led to an improvement in MCC. The iRG-4mC tool achieved high accuracy in all the datasets. The achieved accuracies for *F. vesca (training)*, *F. vesca (independent)*, *R. chinensis (training)* and *R. chinensis (independent)* were 0.867, 0.859, 0.871 and 0.865 respectively.

Table 4. Performance comparison of proposed iRG-4mC tool with existing i4mC-ROSE model (The highest values for different performance parameters are shown in bold).

Dataset	Tool	Sensitivity	Specificity	Accuracy	MCC
F. vesca (training)	i4mC-ROSE	0.635	0.899	0.767	0.545
	iRG-4mC	**0.825**	**0.908**	**0.8665**	**0.732**
R. chinensis (training)	i4mC-ROSE	0.668	**0.9**	0.784	0.563
	iRG-4mC	**0.869**	0.864	**0.871**	**0.739**
F. vesca (independent)	i4mC-ROSE	0.721	0.873	0.797	0.601
	iRG-4mC	**0.835**	**0.882**	**0.859**	**0.706**
R. chinensis (independent)	i4mC-ROSE	0.636	**0.881**	0.759	0.535
	iRG-4mC	**0.854**	0.875	**0.865**	**0.714**

Figure 2 shows the graphical performance comparison between the two techniques. Figure 3 shows the ROC curve of 10 folds for *F. vesca* dataset. An ROC curve is the plot between sensitivity and specificity at different classification thresholds, while the area

under the ROC curve (AUC) is the accumulated performance measure over all possible thresholds. The mean AUC achieved was 0.903 with a standard deviation of 0.03.

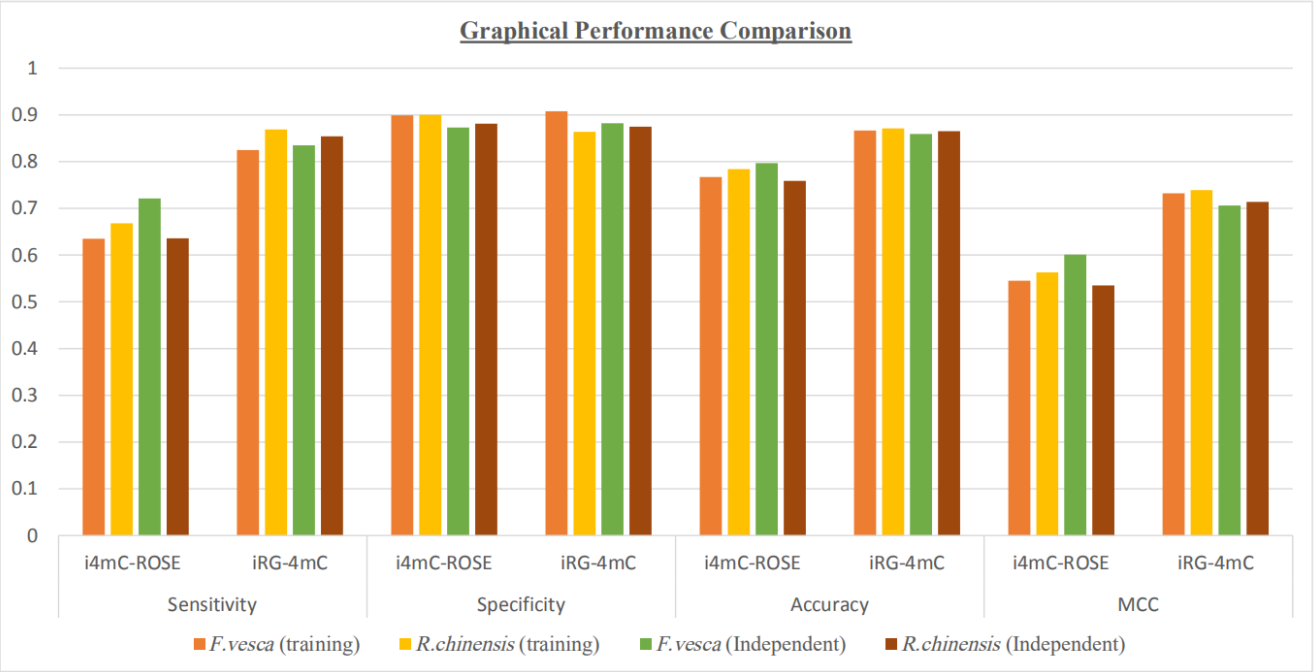

Figure 2. Visual performance comparison between state-of-the-art i4mC-ROSE and proposed iRG-4mC.

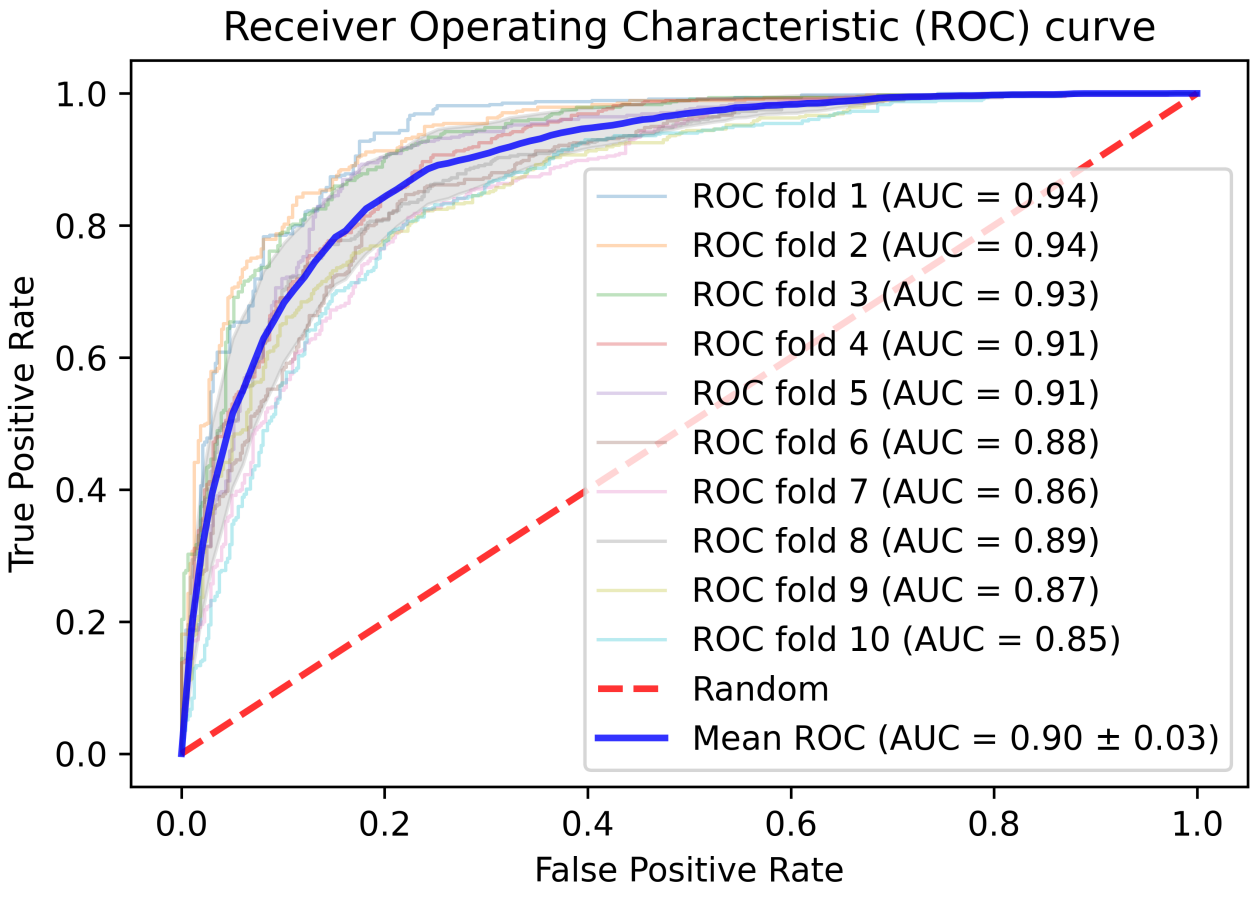

Figure 3. ROC curve.

6. Conclusions

Getting inspired from the importance of N4-methylcytosine sites identification, we have proposed an identification computational tool named as iRG-4mC. The proposed identification tool is for the *Rosaceae* genome and that is why two different species *F. vesca* and *R. chinensis* are considered in this study. In proposed tool the input sequences are encoded using combination of one-hot encoding and nucleotide chemical properties (NCP). The final attained sequence is given to the neural network architecture for classification between 4mC and non-4mC sites. The neural network architecture contains two blocks for feature extraction followed by LSTM for feature optimization and three fully connected layers for final prediction. The architecture is optimized by hyper parameter tuning. Different figure of merits are taken into account to have comparison with existing method. The achieved results have illustrated great improvement in performance by the iRG-4mC tool. This computational tool can be of great importance for the researchers from the field of biology and bio-informatics. A user friendly web-server is made for the researcher's convenience. The webserver is freely available at: http://nsclbio.jbnu.ac.kr/tools/iRG-4mC/.

Author Contributions: Conceptualization, D.Y.L., M.U.R. and K.T.C.; methodology, D.Y.L., M.U.R.; software, D.Y.L., M.U.R.; validation, D.Y.L., M.U.R. and K.T.C.; investigation, D.Y.L., M.U.R. and K.T.C.; writing—original draft preparation: D.Y.L., M.U.R.; writing—review and editing, D.Y.L., M.U.R. and K.T.C.; supervision, K.T.C. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported in part by the selection of research-oriented professor of Jeonbuk National University in 2020 and in part by National Research Foundation of Korea (NRF) grant funded by the Korea government(MSIT) (No. 2020R1A2C2005612) and in part by Basic Science Research Program through the National Research Foundation of Korea funded by the Ministry of Education (No. 2019R1A6A3A01094685).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

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. Biol. Sci.* **2018**, *373*, 20170078. [CrossRef]
2. Jeltsch, A.; Jurkowska, R.Z. New concepts in DNA methylation. *Trends Biochem. Sci.* **2014**, *39*, 310–318. [CrossRef]
3. Jin, Z.; Liu, Y. DNA methylation in human diseases. *Genes Dis.* **2018**, *5*, 1–8. [CrossRef]
4. Zhang, H.; Lang, Z.; Zhu, J.-K. Dynamics and function of DNA methylation in plants. *Nat. Rev. Mol. Cell Biol.* **2018**, *8*, 489–506.
5. Liang, Z.; Shen, L.; Cui, X.; Bao, S.; Geng, Y.; Yu, G.; Liang, F.; Xie, S.; Lu, T.; Gu, X.; et al. DNA N6-adenine methylation in Arabidopsis thaliana. *Dev. Cell* **2018**, *45*, 406–416. [CrossRef] [PubMed]
6. Law, J.A.; Jacobsen, S.E. Establishing, maintaining and modifying DNA methylation patterns in plants and animals. *Nat. Rev. Genet.* **2010**, *11*, 204–220. [CrossRef] [PubMed]
7. Chatterjee, A.; Eccles, M.R. DNA methylation and epigenomics: New technologies and emerging concepts. *Genome Biol.* **2015**, *16*. [CrossRef]
8. Fu, Y.; Luo, G.Z.; Chen, K.; Deng, X.; Yu, M.; Han, D.; Hao, Z.; Liu, J.; Lu, X.; Doré, L.C.; et al. N6-methyldeoxyadenosine marks active transcription start sites in Chlamydomonas. *Cell* **2015**, *161*, 879–892. [CrossRef]
9. Blow, M.J.; Clark, T.A.; Daum, C.G.; Deutschbauer, A.M.; Fomenkov, A.; Fries, R.; Froula, J.; Kang, D.D.; Malmstrom, R.R.; Morgan, R.D.; et al. The epigenomic landscape of prokaryotes. *PLoS Genet.* **2016**, *12*, e1005854. [CrossRef] [PubMed]
10. Chen, W.; Yang, H.; Feng, P.; Ding, H.; Lin, H. iDNA4mC: Identifying DNA N4-methylcytosine sites based on nucleotide chemical properties. *Bioinformatics* **2017**, *33*, 3518–3523. [CrossRef]
11. Heyn, H.; Esteller, M. An adenine code for DNA: A second life for N6-methyladenine. *Cell* **2015**, *161*, 710–713. [CrossRef]
12. Cheng, X. DNA modification by methyltransferases. *Curr. Opin. Struct. Biol.* **1995**, *5*, 4–10. [CrossRef]
13. 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]

14. Schweizer, H.P. Bacterial genetics: Past achievements, present state of the field, and future challenges. *Biotechniques* **2008**, *44*, 633–641. [CrossRef] [PubMed]
15. Suzuki, M.M.; Bird, A. DNA methylation landscapes: Provocative insights from epigenomics. *Nat. Rev. Genet.* **2008**, *9*, 465–476. [CrossRef]
16. Robertson, K.D. DNA methylation and human disease. *Nat. Rev. Genet.* **2005**, *6*, 597–610. [CrossRef] [PubMed]
17. Jones, P.A. Functions of DNA methylation: Islands, start sites, gene bodies and beyond. *Nat. Rev. Genet.* **2012**, *13*, 484–492. [CrossRef] [PubMed]
18. Yao, B.; Jin, P. Cytosine modifications in neurodevelopment and diseases. *Cell. Mol. Life Sci.* **2014**, *71*, 405–418. [CrossRef]
19. Ling, C.; Groop, L. Epigenetics: A molecular link between environmental factors and type 2 diabetes. *Diabetes* **2009**, *58*, 2718–2725. [CrossRef]
20. Chen, K.; Zhao, B.S.; He, C. Nucleic acid modifications in regulation of gene expression. *Cell Chem. Biol.* **2016**, *23*, 74–85. [CrossRef]
21. 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]
22. Buryanov, Y.I.; Shevchuk, T. DNA methyltransferases and structural-functional specificity of eukaryotic DNA modification. *Biochemistry* **2005**, *70*, 730–742. [CrossRef]
23. Liu, Q.; Chen, J.; Wang, Y.; Li, S.; Jia, C.; Song, J.; Li, F. DeepTorrent: A deep learning-based approach for predicting DNA N4-methylcytosine sites. *Brief. Bioinform.* **2020**, 1–14. [CrossRef] [PubMed]
24. Khanal, J.; Nazari, I.; Tayara, H.; Chong, K.T. 4mCCNN: Identification of N4-methylcytosine sites in prokaryotes using convolutional neural network. *IEEE Access* **2019**, *7*, 145455–145461. [CrossRef]
25. 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]
26. Raymond, O.; Gouzy, J.; Just, J.; Badouin, H.; Verdenaud, M.; Lemainque, A.; Vergne, P.; Moja, S.; Choisne, N.; Pont, C.; et al. The Rosa genome provides new insights into the domestication of modern roses. *Nat. Genet.* **2018**, *50*, 772–777. [CrossRef] [PubMed]
27. Edger, P.P.; VanBuren, R.; Colle, M.; Poorten, T.J.; Wai, C.M.; Niederhuth, C.E.; Alger, E.I.; Ou, S.; Acharya, C.B.; Wang, J.; et al. Single-molecule sequencing and optical mapping yields an improved genome of woodland strawberry (*Fragaria vesca*) with chromosome-scale contiguity. *Gigascience* **2018**, *7*, gix124. [CrossRef]
28. Gruenbaum, Y.; Naveh-Many, T.; Cedar, H.; Razin, A. Sequence specificity of methylation in higher plant DNA. *Nature* **1981**, *292*, 860–862. [CrossRef]
29. Rehman, M.U.; Cho, S.; Kim, J.H.; Chong, K.T. BU-Net: Brain Tumor Segmentation Using Modified U-Net Architecture. *Electronics* **2020**, *9*, 2203. [CrossRef]
30. Rehman, M.U.; Cho, S.; Kim, J.; Chong, K.T. BrainSeg-Net: Brain Tumor MR Image Segmentation via Enhanced Encoder-Decoder Network. *Diagnostics* **2021**, *11*, 169. [CrossRef]
31. Rehman, M.U.; Abbas, Z.; Khan, S.H.; Ghani, S.H. Diabetic retinopathy fundus image classification using discrete wavelet transform. In Proceedings of the 2018 IEEE 2nd International Conference on Engineering Innovation (ICEI), Bangkok, Thailand, 5–6 July 2018; pp. 75–80.
32. Ilyas, T.; Khan, A.; Umraiz, M.; Kim, H. Seek: A framework of superpixel learning with cnn features for unsupervised segmentation. *Electronics* **2020**, *9*, 383. [CrossRef]
33. Ilyas, T.; Umraiz, M.; Khan, A.; Kim, H. DAM: Hierarchical Adaptive Feature Selection Using Convolution Encoder Decoder Network for Strawberry Segmentation. *Front. Plant Sci.* **2021**, *12*, 189. [CrossRef]
34. Okinda, C.; Nyalala, I.; Korohou, T.; Okinda, C.; Wang, J.; Achieng, T.; Wamalwa, P.; Mang, T.; Shen, M. A review on computer vision systems in monitoring of poultry: A welfare perspective. *Artif. Intell. Agric.* **2020**, *4*, 184–208. [CrossRef]
35. Heinrich, F.; Wutke, M.; Das, P.P.; Kamp, M.; Gültas, M.; Link, W.; Schmitt, A.O. Identification of regulatory SNPs associated with vicine and convicine content of Vicia faba based on genotyping by sequencing data using deep learning. *Genes* **2020**, *11*, 614. [CrossRef] [PubMed]
36. Yik, S.; Benjamin, M.; Lavagnino, M.; Morris, D. DIAT (Depth-Infrared Image Annotation Transfer) for Training a Depth-Based Pig-Pose Detector. In Proceedings of the 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Las Vegas, NV, USA, 25–29 October 2020. [CrossRef]
37. Wutke, M.; Schmitt, A.O.; Traulsen, I.; Gültas, M. Investigation of Pig Activity Based on Video Data and Semi-Supervised Neural Networks. *AgriEngineering* **2020**, *2*, 39. [CrossRef]
38. Rehman, M.U.; Chong, K.T. DNA6mA-MINT: DNA-6mA modification identification neural tool. *Genes* **2020**, *11*, 898. [CrossRef] [PubMed]
39. 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]
40. 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]
41. 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]

42. Abbas, Z.; Tayara, H.; Chong, K.T. 4mCPred-CNN—Prediction of DNA N4-Methylcytosine in the Mouse Genome Using a Convolutional Neural Network. *Genes* **2021**, *12*, 296. [CrossRef]
43. Ali, S.D.; Alam, W.; Tayara, H.; Chong, K. Identification of functional piRNAs using a convolutional neural network. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **2020**. [CrossRef] [PubMed]
44. Alam, W.; Tayara, H.; Chong, K.T. XG-ac4C: Identification of N4-acetylcytidine (ac4C) in mRNA using eXtreme gradient boosting with electron-ion interaction pseudopotentials. *Sci. Rep.* **2020**, *10*, 20942. [CrossRef] [PubMed]
45. Shujaat, M.; Wahab, A.; Tayara, H.; Chong, K.T. pcPromoter-CNN: A CNN-Based Prediction and Classification of Promoters. *Genes* **2020**, *11*, 1529. [CrossRef] [PubMed]
46. Liu, Z.Y.; Xing, J.F.; Chen, W.; Luan, M.W.; Xie, R.; Huang, J.; Xie, S.Q.; Xiao, C.L. MDR: An integrative DNA N6-methyladenine and N4-methylcytosine modification database for Rosaceae. *Hortic. Res.* **2019**, *6*, 1–7. [CrossRef]
47. Fu, L.; Niu, B.; Zhu, Z.; Wu, S.; Li, W. CD-HIT: Accelerated for clustering the next-generation sequencing data. *Bioinformatics* **2012**, *28*, 3150–3152. [CrossRef]
48. Jeong, B.S.; Bari, A.G.; Reaz, M.R.; Jeon, S.; Lim, C.G.; Choi, H.J. Codon-based encoding for DNA sequence analysis. *Methods* **2014**, *67*, 373–379. [CrossRef] [PubMed]