cells

MDPI

Article

DNC4mC-Deep: Identification and Analysis of DNA N4-Methylcytosine Sites Based on Different Encoding Schemes By Using Deep Learning

Abdul Wahab [1,†], Omid Mahmoudi [1,†], Jeehong Kim [2,*] and Kil To Chong [3,4,*]

1 Department of Electronics and Information Engineering, Jeonbuk National University, Jeonju 54896, Korea; me.wahabqayyum@gmail.com (A.W.); omidmahmoudi75@jbnu.ac.kr (O.M.)
2 Department of New & Renewable Energy, VISION College of Jeonju, Jeonju 55069, Korea
3 Department of Electronics Engineering, Jeonbuk National University, Jeonju 54896, Korea
4 Advance Electronics & Information Research Center, Jeonbuk National University, Jeonju 54896, Korea
* Correspondence: jeehong@jvision.ac.kr (J.K.); kitchong@jbnu.ac.kr (K.T.C.)
† These authors equally contributed to this work.

Received: 26 June 2020; Accepted: 17 July 2020; Published: 22 July 2020

Abstract: N4-methylcytosine as one kind of modification of DNA has a critical role which alters genetic performance such as protein interactions, conformation, stability in DNA as well as the regulation of gene expression same cell developmental and genomic imprinting. Some different 4mC site identifiers have been proposed for various species. Herein, we proposed a computational model, DNC4mC-Deep, including six encoding techniques plus a deep learning model to predict 4mC sites in the genome of *F. vesca*, *R. chinensis*, and Cross-species dataset. It was demonstrated by the 10-fold cross-validation test to get superior performance. The DNC4mC-Deep obtained 0.829 and 0.929 of MCC on *F. vesca* and *R. chinensis* training dataset, respectively, and 0.814 on cross-species. This means the proposed method outperforms the state-of-the-art predictors at least 0.284 and 0.265 on *F. vesca* and *R. chinensis* training dataset in turn. Furthermore, the DNC4mC-Deep achieved 0.635 and 0.565 of MCC on *F. vesca* and *R. chinensis* independent dataset, respectively, and 0.562 on cross-species which shows it can achieve the best performance to predict 4mC sites as compared to the state-of-the-art predictor.

Keywords: N4-methylcytosine; rosaceae genome; DNA encoding methods; computational biology; deep learning; bioinformatics

1. Introduction

Dynamic DNA modifications, such as methylation and demethylation have an essential role in the regulation of gene expression. DNA methylation as a heritable epigenetic marker is one type of chemical modification of DNA, which alters genetic performance without altering the DNA sequence [1,2]. Several researches have shown that it has the ability to change DNA protein interactions, DNA conformation, DNA stability, and chromatin structure. Meanwhile, it can regulate some different functions including cell developmental, genomic imprinting, and gene expressions [3,4]. N4-methylcytosine (4mC), 5-Methylcytosine (5mC), and N6-methyladenine (6mA) as three common methylations by specific methyltransferase enzymes occur in both prokaryotes and eukaryotes [5–7].

In prokaryotes, the host DNA from exogenous pathogenic DNA can be identified by 6mA and 4mC [8], and also 4mC regulates DNA replication and its errors [9,10]. Meanwhile, 4mC as part of a restriction-modification (R-M) system prevents restriction enzymes from degrading host DNA [11]. In eukaryotes, 5mC has a crucial role in transposon suppression, gene imprinting, and regulation. By high sensitivity techniques, 6mA and 4mC can only be detected in eukaryotes [12].

The 5mC, as the most well-explored and common type of cytosine methylation plays a significant role in several biological processes [13] and can be caused by cancer, diabetes, and also some neurological diseases [14–16]. The 4mC as effective methylation protects its own DNA from the restriction of enzyme-mediated degradation. It has an important role in controlling some various processes including cell cycle, gene expression levels, differentiating self and non-self-DNA, DNA replication, and correcting DNA replication errors [9,17].

Some extensive experimental studies have been performed to detect 4mC sites in the whole genome such as 4mC-Tet-assisted bisulfite sequencing, methylation-precise PCR, mass spectrometry, and Single-Molecule of Real-Time (SMRT) sequencing [18–21]. The aforementioned experimental approaches are laborious and expensive work when performing genome-wide testing. Therefore, it is necessary to develop a computational method for identifying 4mC sites.

Lately, several 4mC sites identifiers [22,23] have been proposed for different species such as *C. elegans*, *D. melanogaster*, *A. thaliana*, *E. coli*, *G. subterraneus*, *G. pickeringii*. The i4mC-ROSE [24] as the first computational tool for predicting 4mC sites within the *Rosaceae* genomes has been proposed to identify the 4mC sites in the genomes of *F. vesca* [25] and *R. chinensis* [26]. It generated six probabilistic scores by using six encoding methods; random forest (RF), algorithms with k-space spectral nucleotide composition (KSNC), electron-ion interaction pseudopotentials (EIIP), k-mer composition (Kmer), binary encoding (BE), dinucleotide physicochemical properties (DPCP), and trinucleotide physicochemical properties (TPCP) that cover various aspects of DNA sequence information. Then, those scores were combined with a linear regression model for enhanced prediction performance [24]. The 4mcDeep-CBI [27] as a deep learning framework has been proposed to predict the 4mC sites in an expanded dataset of *Caenorhabditis elegans* (*C. elegans*). 3-CNN and BLSTM were used to extract deep information and to obtain advanced features.

In this work, a novel predictor, DNC4mC-Deep, has been established for the identification of 4mC sites in the genome of *F. vesca*, *R. chinensis*, and cross-species which is newly prepared. The overall framework of our work summarized as; Firstly we used the six encoding techniques named 2Kmer [28], 3Kmer [29], binary encoding (BE) [30,31], nucleotide chemical property (NCP) [32], nucleotide chemical property, and nucleotide frequency (NCPNF) [32], and multivariate mutual information (MMI) [33,34]. Then, we made a deep learning model by using the Convolution Neural Network (CNN). We applied a grid search algorithm to obtain the optimal model with tuned hyperparameters. All six encoding schemes were input separately in the optimal selected model and recorded the results of each encoding scheme and used the K-fold cross-validation method with the value of K as 10. To evaluate and analyze the results of the model on each encoding scheme, we used the performance evaluation metrics. We also presented two different applications; the first one is silico mutagenesis [35] representation using heat maps, and the second is distinguishing the most significant portions of a sequence using saliency maps [36]. After getting the results from the model by all six different feature encoding methods, we ended up with that Dinucleotide composition (DNC) is outperforms from all six encoding schemes and the state-of-the-art model. In comparison to the state-of-the-art model, DNC4mC-Deep successfully achieves 0.635, 0.565, and 0.562 of MCC on *F. vesca*, *R. chinensis*, and cross-species independent dataset, respectively.

2. Materials and Methods

2.1. Benchmark Datasets

The benchmark dataset of DNA 4mC obtained from Md. Mehedi Hassan et al. [24]. It contains the *F. vesca* and *R. chinensis* genome. To prepare the high-quality dataset they have applied the sequences with ModQV score greater than 20, whereas the remaining sequences were excluded. To solve the homology bias problem, the CD-HIT-EST [37] software was used to exclude redundant sequences with a cut-off of 0.65. All sequences contain a central cytosine (C) nucleotide with a length of 41 base pairs (bp).

In both datasets, *F. vesca* and *R. chinensis* genome were considered 75% and 25% samples from all data as the training and the independent dataset. The training dataset consists of 4854 and 2337 positive DNA sequences as 4mC samples for *F. vesca*, and *R. chinensis*, severally. The negative DNA sequences, such as non-4mC, consists of 4854 and 2337 samples for *F. vesca*, and *R. chinensis* genome. Furthermore, the independent dataset included 1617 for both positive and negative DNA sequences of *F. vesca* genome whereas for *R. chinensis* positive and negative DNA contains 779 samples.

Moreover, we made the cross-species as a new benchmark dataset from the two above datasets. To avoid the redundancy in the original datasets we used CD-HIT-EST with different threshold values. The recent dataset was also divided into the training and the independent dataset with the same proportion (75% and 25% samples) where we obtained the cross-species dataset with the most attentive threshold at 0.80 containing 7190 and 5874 positive and negative DNA sequences, respectively, on the training dataset. Meanwhile, we assumed 2394 positive and 2234 negative DNA sequences on the independent dataset. The length of each sample is 41nt. Details of the benchmark datasets are shown in Table 1.

Table 1. Benchmark datasets demonstration.

Species	Dataset	Training Dataset	Total	Independent Dataset	Total
F. vesca	4mC samples non-4mC samples	4854 4854	9708	1617 1617	3234
R. chinensis	4mC samples non-4mC samples	2337 2337	4674	779 779	1558
Cross-species	4mC sampels non-4mC samples	7190 5874	13,064	2394 2234	4628

2.2. Feature Encoding Methods

Feature encoding has a vital role in the construction of the model [38]. A DNA sequence is represented as a fixed length of feature vectors which can be classified by deep learning algorithms. In this article, six various types of feature encoding methods, binary encoding [39], DNC (2kmer), TNC (3kmer) [40–43], Multivariate Mutual Information (MMI) [44], Nucleotide Chemical Property (NCP) and Nucleotide Chemical Property and Nucleotide Frequency (NCPNF) [28,29,45–47] were employed to formulate methylcytosine samples.

2.2.1. Binary Encoding (BE)

Binary encoding is a simple and effective feature algorithm converts each nucleotide into a binary vector as follows: A (1, 0, 0, 0, 0), C (0, 1, 0, 0, 0), G (0, 0, 1, 0, 0), T (0, 0, 0, 1, 0) and N (0, 0, 0, 0, 0). A DNA sequence with m nucleotides can be represented into a vector of $5 \times m$ features [30,31].

2.2.2. Kmer

Kmer is a common feature encoding algorithm that has been widely used in various prediction works [28,29,45,48,49]. A DNA sample is expressed as V = N1, N2, N3, ... NL, where L denotes the length of the sequence and Ni is one of the regular nucleotides A, C, G, T, and N. In this work, di-nucleotide composition (DNC) and tri-nucleotide composition (TNC) were considered. In DNC all samples of 41 nt produce 40 components with the equation of $L - k + 1$. The DNC scheme generated a 25 (5^2) dimensional feature. Whereas in TNC samples of 41 nt generated 39 elements with the equation of $L - k + 2$. The TNC form into a 125 (5^3)-dimensional vector. In both equations, the L denotes the length of the sequence and k represents the value of Kmer as an integer.

2.2.3. Nucleotide Chemical Property (NCP)

The four nucleic acids have different chemical properties [50]. In terms of ring structures, A and G each contain two rings, whereas C and T contain only one. Regarding secondary structures, A and T form weak hydrogen bonds, whereas C and G form strong hydrogen bonds. In terms of chemical functionality, A and C can be classified into the amino group, while G and T can be classified into the keto group. The cluster of four nucleotides was shown in Table 2.

Table 2. Cluster of nucleotides based on chemical properties.

Chemical Property	Class	Nucleotides
Ring structure	Two ring	A, G
	One ring	C, T
Hydrogen bond	Strong	C, G
	Weak	A, T
Functional group	Amino	A, C
	Keto	G, T

Three coordinates x, y, and z were used to represent ring structure, the hydrogen bond, the chemical functionality, respectively, and the value of 0 and 1 was assigned to each one. The feature extraction algorithm can be formulated as follows:

$$x_i = \begin{cases} 1 & if s_i \epsilon \{A, G\} \\ 0 & if s_i \epsilon \{C, T\} \end{cases}, y_i = \begin{cases} 1 & if s_i \epsilon \{A, T\} \\ 0 & if s_i \epsilon \{C, G\} \end{cases}, z_i = \begin{cases} 1 & if s_i \epsilon \{A, C\} \\ 0 & if s_i \epsilon \{G, T\} \end{cases}$$

where $n(s_i)$ represents A, C, G, T, and N nucleotide, which can be converted by the coordinates (1, 1, 1), (0, 0, 1), (1, 0, 0), (0, 1, 0), and (0, 0, 0), respectively.

We also tried the nucleotide chemical properties (NCP) with the frequencies of each nucleotide (NF) position in a sample. The method was got from Chen et al. [32] for both encoding schemes. We integrated the NCP and NF to represent a matrix with 41 columns and 5 rows for each sample of the DNA sequence. Each DNA base of the sequence was designated as a column of the matrix and for each column, their initial three components were characterized as the nucleotide chemical property and the last one represented as a nucleotide frequency which we denoted as NCPNF.

2.2.4. Multivariate Mutual Information (MMI)

MMI has been used in many works [33,34,51] to extract features of the nucleotides sequence. We used the MMI based feature encoding algorithm which was proposed by Pan et al. [52]. First of all, they modified a two-tuple and three-tuple nucleotides set as follows:

$$T_2 = \{AA, AC, AG, AT, AN, CC, CG, CT, CN, GG, GT, GN, TT, TN, NN\}$$
$$T_3 = \{AAA, AAC, AAG, AAT, AAN, ACC, ACG, ACT, ACN, AGG, AGT, AGN, ATT, ATN, ANN, CCC, CCG, CCT, CCN, CGG, CGT, CGN, CTT, CTN, CNN, GGG, GGT, GGN, GTT, GTN, TTT, TTN, NNN\}$$

Then, the mutual information for the elements was calculated as a frequency of nucleotides in the sequence with respect to 2-tuple and 3-tuple. We extracted 55 MMI features.

3. The Proposed Deep Learning Model

In this study, an efficient deep learning model based on CNN was proposed for the identification of 4mC sites in the genome of *F. vesca*, *R. chinensis* and cross-species. CNN does not require manually extracted features like a conventional supervised learning processes. The immense advantage of a CNN, it can extract the features by itself automatically for the classification process. Additionally,

a handy crafted feature can also be fed to CNN to build a heterogeneous model. A CNN has a big impact on various fields of natural language processing, image processing [53–56] and computational biology [57,58]. To get an optimum model we applied grid search and during learning the CNN, six hyperparameters were tuned. The ranges within each hyper-parameter was tuned to are listed in Table 3.

Table 3. Hyper-parameters tuning demonstration.

Parameters	Range
Convolution layers	[1, 2, 3, 4, 5]
Filters in convolution Layer	[8, 12, 16, 22, 32, 42, 64, 128]
Filter size	[2, 3, 4, 5, 6, 7, 8, 10, 12, 14]
Pool-size in Maxpooling	[2, 4]
Stride length in Maxpooling	[2, 4]
Dropout values	[0.2, 0.25, 0.3, 0.35, 0.4]

After getting the best model from the grid search, we used six different encoding schemes (DNC, TNC, BE, NCP, NCPNF, MMI) for the input of the CNN model. Each encoding technique converted into vectorization of the input sequence and used the same CNN model for training and testing also verified the robustness from the independent dataset. All the feature encoding approaches had a different impact on a single model.

In the proposed model, initially, two blocks used with the same number of layers but different values of parameters. Each block contains one convolution layer Conv1D (f, k, s) where parameter f is the number of filters, k is the kernel-size, and s represents the stride value are equal to 32, 5 and 1, respectively on both blocks. The convolution layer utilizes its ability to fetch the features by self-regulating for the input sequence of positive and negative 4mC samples. As a parameter of the convolution layer, we used L2 regularization and bias regularization to make sure that the model has no overfitting problem. We set the values for both regularizations with 0.0001 for the two Conv1D of blocks. As an activation function, an exponential linear unit (ELU) is used. Each Conv1D was followed by a group normalization layer (GN) as GroupNormalization (g) where g is a number of groups, to decrease the outcomes of convolution layers. Group normalization divides channels into groups and normalizes the feature within each group. The number of groups was set to 4 on both blocks of GN. To reduce the redundancy of the features after GN layers, we employed a max-pooling layer in each block as MaxPooling1D (l, r) where l denotes pool-size and r is the stride were set as 4 and 2, respectively. The max-pooling layer helps to reduce the dimensionality of the features from former layers. The outputs of the max-pooling layers were passed through dropout layers, Dropout (d) with a probability of 0.25 as a value of d on both blocks for the prevention of overfitting during the training. Dropout helps to switch off the effects of a few hidden nodes by adjusting the output of nodes to zero at training.

After both blocks, to unstack the output, a flatten function was used to squash the feature vectors from the previous layers. Right after a flatten layer, a fully connected (FC) dense layer used as Dense (n) with the number of n neurons which was set as 32 and also used the L2 regularization parameter for bias and weights with the value of 0.0001. ELU activation function used in the FC layer. At last, a FC layer was applied and used sigmoid function for the binary classification. Sigmoid is used to squeezes the values between the range of 0 and 1 to represent the probability of having 4mC and non-4mC sites. Figure 1 shows the complete architecture of the presented model.

The DNC4mC-Deep was carried out on the Keras Framework [59]. In DNC4mC-Deep we used stochastic gradient descent (SGD) optimizer with a momentum of 0.95 and the learning rate is set as 0.005. For the loss function, binary cross-entropy was used. On the fit function, we set the 100 for the epoch and 32 for the batch size. The checkpoint was used on call back function for saving the models and their best weights whereas early stopping was also implemented to halt the prediction accuracy at the time when validation stops improving. The patience level was set to 30 in early stopping.

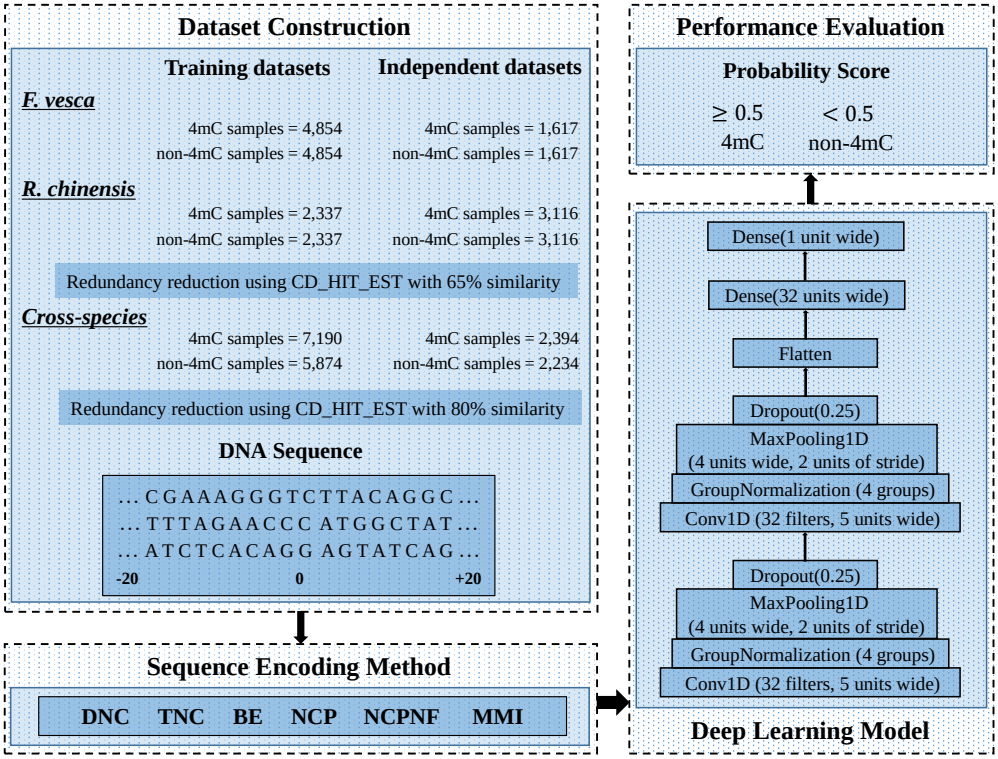


Figure 1. A complete structure of DNC4mC-Deep.

4. Performance Evaluation Metrics

The performance of the prediction model can be measured by using k-fold cross-validation. In DNC4mC-Deep we used 10 fold cross-validation to achieve the foremost prediction calculation. Cross-validation is a resampling technique which provides a precise performance estimation for the predictive model. It intermixes the entire dataset and divides into a k number of clusters, where each cluster contains eight folds for training, one fold for validation, and one for testing. The model was trained and tested k times, recorded performance each time, and concised by taking the mean score for the performance evaluation. The most common criteria which is used to evaluate the performance of the predicted models are four metrics; Mathew's correlation coefficient (MCC), accuracy (ACC), sensitivity (Sn) and specificity (Sp) with the following mathematical formulations [60–62].

$$MCC = \frac{TP * TN - FP * FN}{\sqrt{(TP + FP) * (TP + FN) * (TN + FP) * (TN + FN)}} \quad (1)$$

$$ACC = \frac{TP + TN}{TP + TN + FP + FN} \quad (2)$$

$$SN = \frac{TP}{TP + FN} \quad (3)$$

$$SP = \frac{TN}{TN + FP} \quad (4)$$

where TP and TN as true positive and true negative indicate the correct numbers of predicted samples for 4mCs and non-4mCs, respectively. Meanwhile, FP and FN as false positive and false negative represent the false numbers of predicted samples for 4mCs and non-4mCs, respectively. Besides, the receiver operating characteristics curve (ROC) and area under the ROC curve (AUC) were also used to show the performance of proposed model.

5. Results and Discussion

Six different encoding methods, namely DNC, TNC, NCP, BE, NCPNF, and MMI were used on various feature encodings for identification of the best classifier for the 4mC site prediction.

5.1. Performance Evaluation of Various Feature Methods on the Training Datasets

By comparing the effectiveness of the proposed methods with i4mC-ROSE model which used the same datasets, the DNC scheme yielded MCC, ACC, Sn and Sp of 0.829, 0.914, 0.926, and 0.903, respectively as the best performances for *F. vesca* dataset (Figure 2). Similarly, it achieved 0.828 for MCC, 0.914 for ACC, 0.919 for Sn and 0.910 for Sp as the maximum value on the *R. chinensis* dataset (Figure 3). The detailed performances of DNC as the best encoding method for ten different models on the *R. chinensis* dataset are given in Supplementary File 1. The TNC scheme yielded the highest value for Sp of 0.909 on the *F.vesca* dataset. Table 4, summarized the prediction performances by each six encoding methods and existing state-of-the-art model on *F. vesca* and *R. chinensis* datasets.

Furthermore, the performance evaluation by six different encoding models on the cross-species dataset is shown in Figure 4. The DNC scheme yielded the highest value for all those metrics except Sp which the TNC scheme achieved the highest value of 0.882 shown in Table 5.

Table 4. Performance evaluation of six encoding methods with state-of-the-art model on training benchmark dataset for *F. vesca* and *R. chinensis* species.

Dataset	Method	MCC	ACC	Sn	Sp	AUC
Fragaria Vesca	DNC	0.829	0.914	0.926	0.903	0.96
	TNC	0.825	0.912	0.916	0.909	0.96
	NCP	0.797	0.898	0.922	0.874	0.95
	BE	0.760	0.879	0.905	0.854	094
	NCPNF	0.782	0.891	0.907	0.874	0.95
	MMI	0.659	0.829	0.864	0.794	0.90
	i4mC-ROSE	0.545	0.767	0.635	0.899	0.88
Rosa Chinensis	DNC	0.828	0.914	0.919	0.910	0.96
	TNC	0.811	0.906	0.906	0.906	0.96
	NCP	0.811	0.906	0.901	0.910	0.96
	BE	0.805	0.903	0.891	0.914	0.95
	NCPNF	0.794	0.897	0.892	0.901	0.95
	MMI	0.691	0.846	0.833	0.858	0.92
	i4mC-ROSE	0.563	0.784	0.668	0.900	0.89

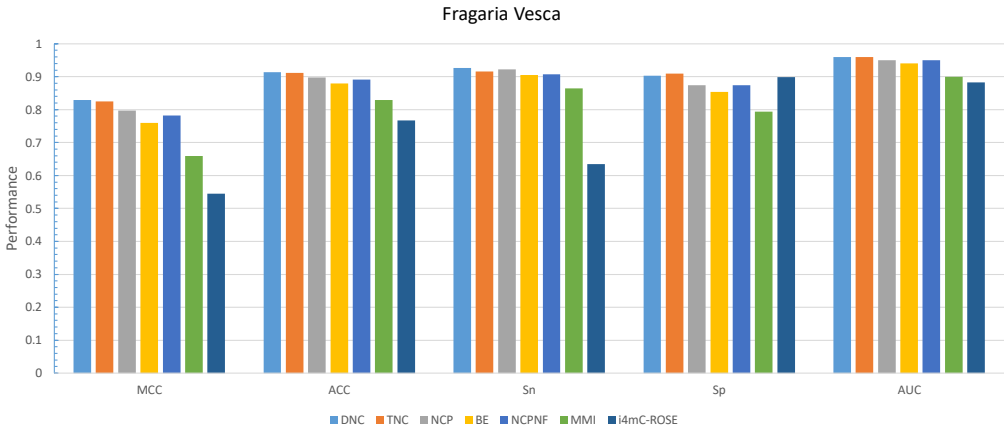


Figure 2. Grapical demonstration of performance comparison between six encoding methods and state-of-the-art model on training *Fragaria Vesca* dataset.

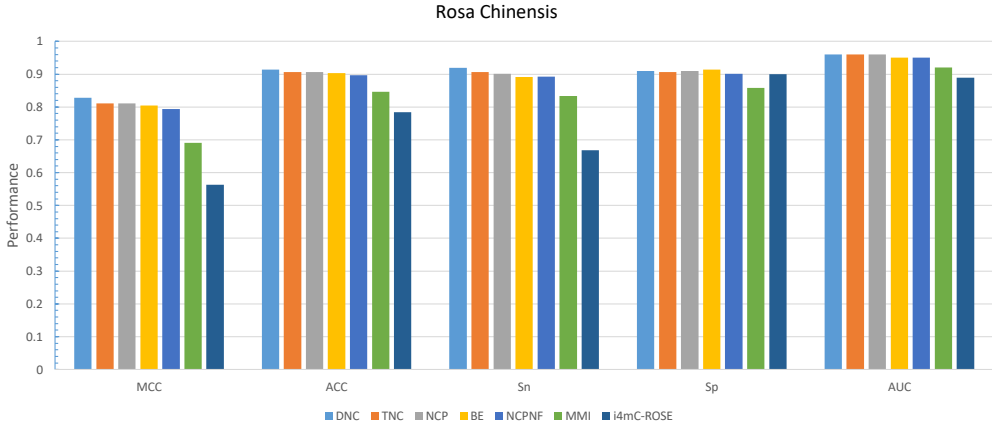


Figure 3. Grapical demonstration of performance comparison between six encoding methods and state-of-the-art model on training *Rosa Chinensis* dateset.

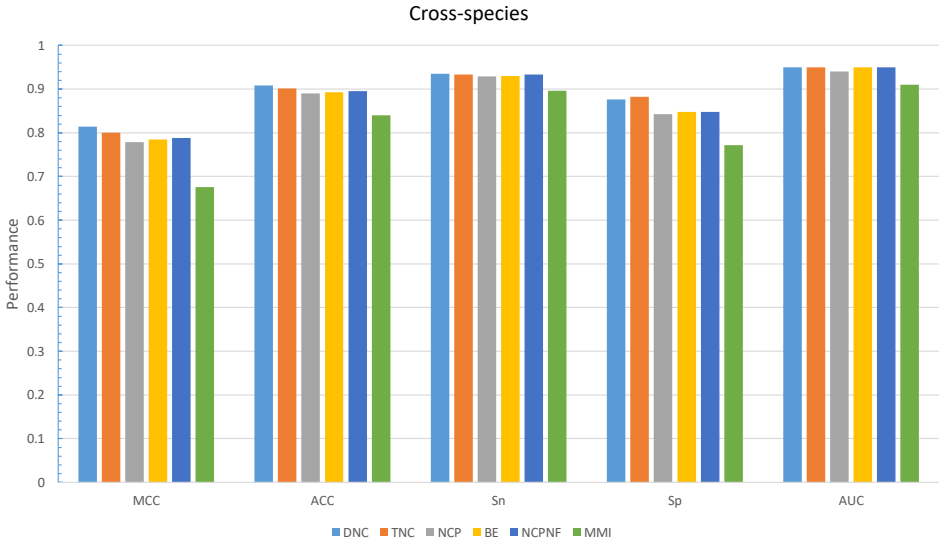
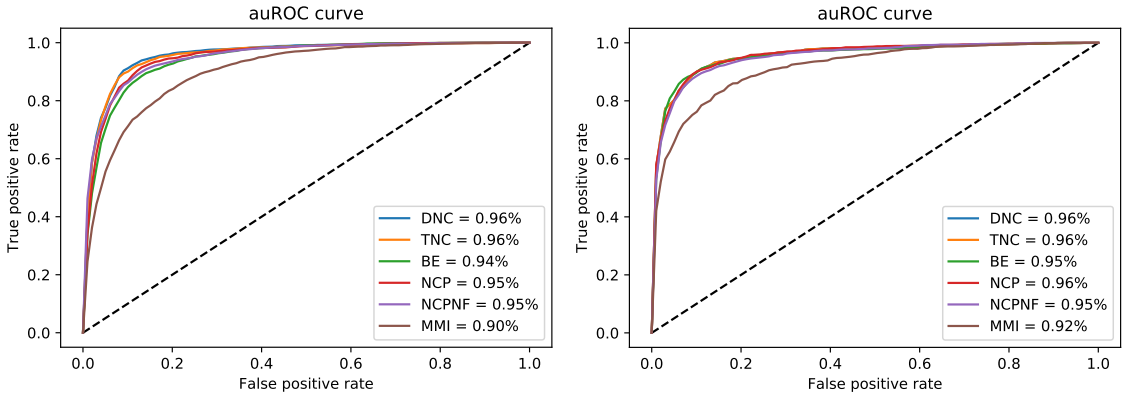


Figure 4. Grapical demonstration of performance comparison between six encoding methods on the training cross-species dataset.

Table 5. Performance evaluation of six encoding methods on training benchmark dataset for cross-species.

Dataset	Method	MCC	ACC	Sn	Sp	AUC
Cross-species	DNC	0.814	0.908	0.935	0.876	0.95
	TNC	0.800	0.901	0.933	0.882	0.95
	NCP	0.779	0.890	0.929	0.843	0.94
	BE	0.785	0.893	0.930	0.848	0.95
	NCPNF	0.788	0.895	0.933	0.848	0.95
	MMI	0.676	0.840	0.896	0.772	0.91

The ROC curve of six encoding models was shown in Figure 5 and compared to the i4mC-ROSE model for both genomes. On the *F. vesca* dataset, the DNC and TNC achieved the best performance with an AUC value of 0.96 followed by NCP, NCPNF, BE, and MMI (Figure 5a). However, the highest AUC value was presented by DNC, TNC, and NCP of 0.96 equally and next BE, NCPNF, and MMI provided 0.95, 0.95, and 0.92, respectively on *R. chinensis* dataset (Figure 5b). Besides, DNC, TNC, BE, NCNF, and NCP all have the highest value of 0.95 on training benchmark dataset cross-species (Figure 6).



(**a**) Fragaria Vesca (**b**) Rosa Chinensis

Figure 5. auRoc curves of six encoding methods for the proposed model on two training benchmark dataset.

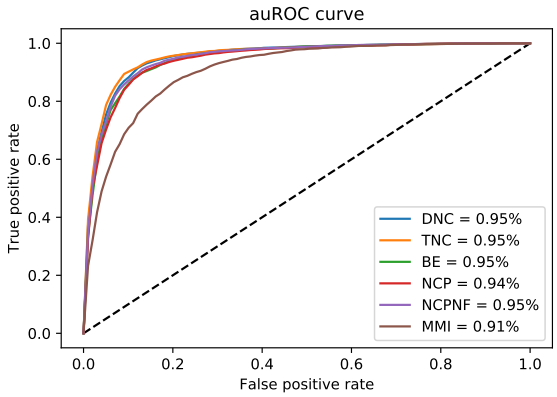


Figure 6. auRoc curves of six encoding methods for the proposed model on new cross-species training benchmark dataset.

5.2. Performance Evaluation of Various Encoding Methods on the Independent Datasets

We considered DNC as an encoder to characterize our proposed model, DNC4mC-Deep, due to its consistent performance on training datasets. It means we used the DNC4mC-Deep term instead of DNC scheme on the independent datasets. As represented in Table 6, the DNC4mC-Deep scheme achieved MCC, ACC, Sn and Sp of 0.635, 0.815, 0.878, and 0.753, respectively on *F. vesca* dataset (Figure 7). However, it yielded 0.565 MCC, 0.783 ACC, 0.801 Sn and 0.765 Sp on *R. chinensis* dataset (Figure 8). It can be seen clearly, comparing with the i4mC-ROSE method, the performance of the proposed predictor outperformed on both datasets. Additionally, as can be seen in Table 7, we compared the performance of six different encoding schemes on the cross-species dataset. The DNC4mC-Deep yielded the highest values for MCC, ACC, Sp, and AUC of 0.562, 0.780, 0.706, and 0.85, respectively. However, the NCPNF provided 0.871 Sn as the highest value Figure 9. Furthermore, we reached to 0.89, 0.87, and 0.85 of ROC for *F. vesca*, *R. chinensis*, and cross-species datasets which are depicted in Figure 10.

Table 6. Performance evaluation between the DNC4mC-Deep and state-of-the-art model on independent benchmark dataset for *F. vesca* and *R. chinensis* species.

Dataset	Method	MCC	ACC	Sn	Sp	AUC
Fragaria Vesca	DNC4mC-Deep	0.635	0.815	0.878	0.753	0.89
	i4mC-ROSE	0.601	0.797	0.721	0.873	0.89
Rosa Chinensis	DNC4mC-Deep	0.565	0.783	0.801	0.765	0.87
	i4mC-ROSE	0.535	0.759	0.636	0.881	0.86

Table 7. Performance evaluation of DNC4mC-Deep and other five encoding methods on independent benchmark dataset for cross-species.

Dataset	Method	MCC	ACC	Sn	Sp	AUC
Cross-species	DNC4mC-Deep	0.562	0.780	0.849	0.706	0.85
	TNC	0.542	0.769	0.854	0.678	0.84
	NCP	0.530	0.764	0.828	0.696	0.84
	BE	0.546	0.770	0.867	0.666	0.84
	NCPNF	0.560	0.777	0.871	0.677	0.85
	MMI	0.512	0.753	0.858	0.640	0.83

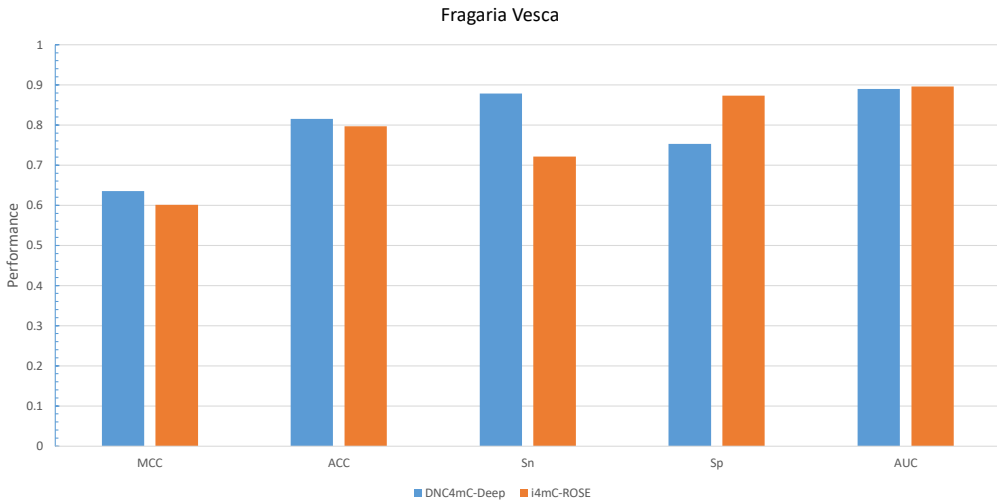


Figure 7. Grapical illustration of performance comparison between DNC4mC-Deep and state-of-the-art model on the independent Fragaria Vesca dataset.

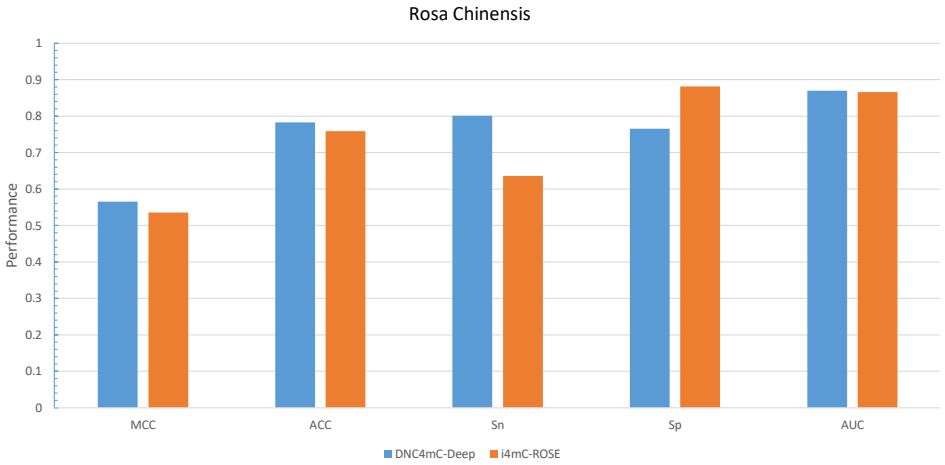


Figure 8. Grapical illustration of performance comparison between DNC4mC-Deep and state-of-the-art model on the independent *Rosa Chinensis* dataset.

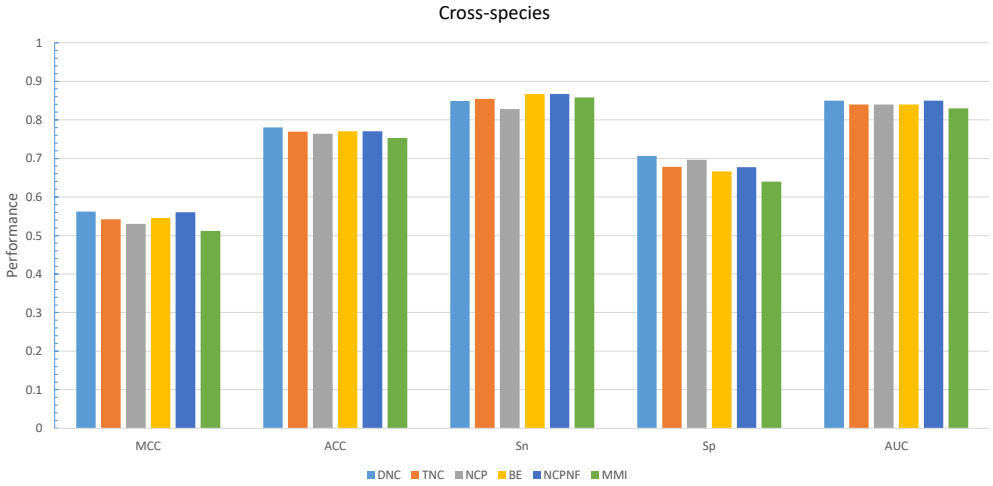


Figure 9. Grapical illustration of performance comparison between six encoding methods on the independent cross-species dataset.

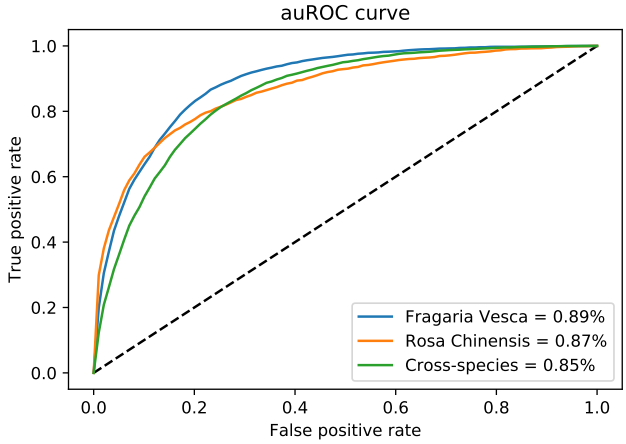


Figure 10. auRoc curves of three independent benchmark dataset on DNC4mC-Deep model.

5.3. Interpreting Applications of Deep Learning Models

Deep learning has an ability to accomplish the state-of-the-art results but it is further challenging to construe the algorithms as a standard statistical model. In the presented work, we demonstrated two applications to understand why those deep learning models perform well apart from others and analyze their prediction by presenting the various visualization methods.

The first most authenticated and reliable method to interpret a CNN model for computational biology is silico mutagenesis which is used in several research works [35,63,64]. We computationally mutated the nucleotides by mutating each nucleotide of a single sequence with a fixed length of five nucleotides A, C, G, T, and N. During this systematic approach, the model recomposes the output of every mutation and stores the output as an absolute difference. Next, the average of mutated predictive results of the whole dataset was taken.

A heat map was used to show the mutated modifications. CNN has the capability to visualize each convolution filter as a heat map or weight matrix. Figure 11, depicts the visualization of the mutation on *F. vesca* dataset as a local feature while learning the model. In the center of the sequences, the impact of mutation is more impactful on the final predictions because of C nucleotide which is representing the methylcytosine modification, the alteration of C nucleotide can lead to different types of gene modification. In contrast, the other sides of the heat map show the low effect of a mutation on the prediction which indicates the alteration of nucleotides cannot affect the outcome of N4-methylcytosine identification.

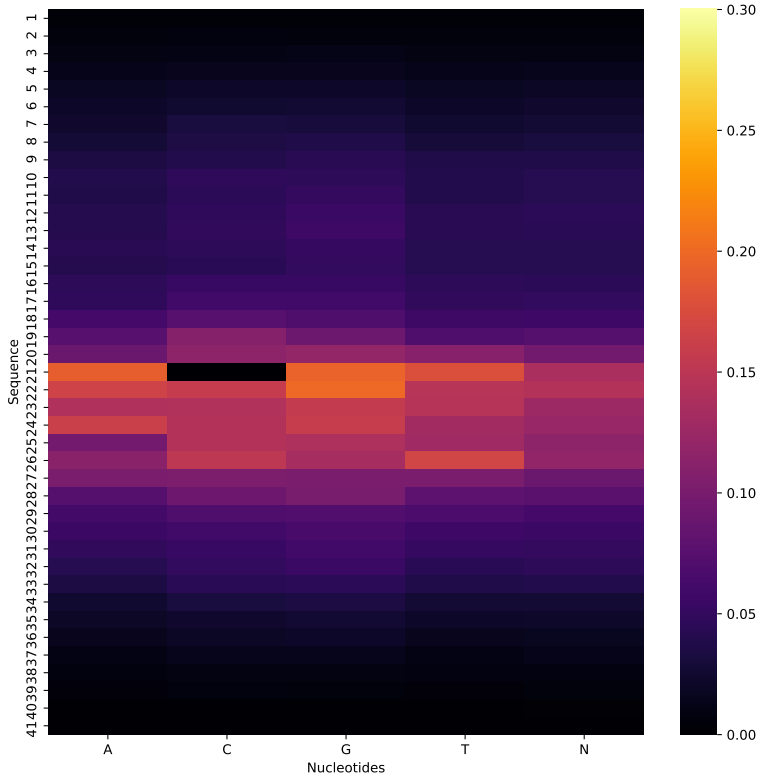


Figure 11. Heatmap visualization of silico mutation, where center of sequence with C nucleotide represents the highest effect on final prediction of Fragaria Vesca dataset.

There is another application to interpret the CNN model for knowing about the important features in the sequence, which help to gradients of the model for the final prediction. Saliency maps are the opted option to know about the most influential parts of the sequences for the classification because many researchers used in their works [36,65,66]. To visualize the effect of each position, we performed a pointwise product of the saliency map with the binary encoded sequence to acquire the derivative values for the original nucleotide letters of the sequences (A, C, G, T, and N). Samples were divided into 2-mer components across all sequences by $L - k + 1$ formulation. In Figure 12, we can see the impact of di-nucleotide characters at each position on the output score of the whole *F. vesca* dataset. At the center of the bases, di-nucleotide motif CC has high magnitude value which represents the most important feature motif in the sequences for the prediction of the CNN model. Motif CC also indicates the N4-methycytosine modification which is our considerable problem for the prediction.

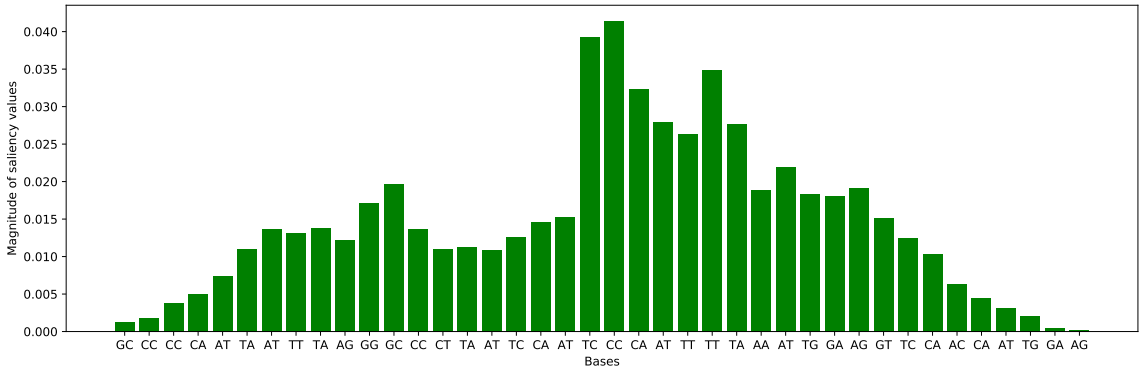


Figure 12. Saliency map of each di-nucleotide influence on model's output in the Fragaria Vesca dataset.

6. Conclusions

In this work, we presented an influential computational model named as DNC4mC-Deep to identify the N4-methylcytosine sites. There are two benchmark datasets related to the Rosaceae genome used Fragaria Vesca and Rosa Chinensis, from those two datasets we constructed a new benchmark dataset: cross-species. We used six different types of feature encoding schemes to input DNA sequence and fed to the CNN model one after another. The CNN based predictor was derived after applying the grid search algorithm. The results obtained from each encoding technique, we concluded that dinucleotide composition (DNC) outperforms and is most imperative for the strong performance of deep learning algorithms to predict 4mC sites. However, to compare with the state-of-the-art models, the CNN model with DNC encoding scheme shows the utmost effective performance and indicates the high capability of prediction. We used different evaluation metrics such as MCC, ACC, Sn, Sp, and AUC, to acquire the efficiency of the proposed predictor. Finally, we interpreted our deep learning model from two techniques: silico mutagenesis and saliency map. DNC4mC-Deep can make a high impact on the biologist to identify the N4-methylcytosine sites and can be used in brain development abnormalities. In the future, we will extend the work to prepare some new datasets and make computational models related to deep learning. Meanwhile, we established the webserver http://home.jbnu.ac.kr/NSCL/DNC4mC-Deep.htm, for users to achieve their desired results easily.

Supplementary Materials: The following are available online at http://www.mdpi.com/2073-4409/9/8/1756/s1. Supplementary File 1: The detailed performances of ten different models on DNC.

Author Contributions: Conceptualization, A.W., O.M., J.K. and K.T.C.; Methodology, A.W. and O.M., J.K.; Software, A.W.; Validation, A.W., O.M., J.K. and K.T.C.; Investigation, A.W., O.M., J.K. and K.T.C.; Writing—original draft preparation: A.W. and O.M.; Writing, review and editing, A.W., O.M., J.K. and K.T.C.; Supervision, K.T.C. All authors have read and agreed to the published version of the manuscript.

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

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]
2. Jeltsch, A.; Jurkowska, R.Z. New concepts in DNA methylation. *Trends Biochem. Sci.* **2014**, *39*, 310–318. [CrossRef]
3. 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]
4. Chatterjee, A.; Eccles, M.R. *DNA Methylation and Epigenomics: New Technologies and Emerging Concepts*; Springer: Berlin, Germany, 2015; Volume 103.
5. 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] [PubMed]
6. 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]
7. 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]
8. Heyn, H.; Esteller, M. An adenine code for DNA: A second life for N6-methyladenine. *Cell* **2015**, *161*, 710–713. [CrossRef]
9. Cheng, X. DNA modification by methyltransferases. *Curr. Opin. Struct. Biol.* **1995**, *5*, 4–10. [CrossRef]

10. 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]
11. Schweizer, H.P. Bacterial genetics: Past achievements, present state of the field, and future challenges. *Biotechniques* **2008**, *44*, 633–641. [CrossRef]
12. Suzuki, M.M.; Bird, A. DNA methylation landscapes: Provocative insights from epigenomics. *Nat. Rev. Genet.* **2008**, *9*, 465–476. [CrossRef]
13. Robertson, K.D. DNA methylation and human disease. *Nat. Rev. Genet.* **2005**, *6*, 597–610. [CrossRef] [PubMed]
14. Jones, P.A. Functions of DNA methylation: Islands, start sites, gene bodies and beyond. *Nat. Rev. Genet.* **2012**, *13*, 484–492. [CrossRef] [PubMed]
15. Yao, B.; Jin, P. Cytosine modifications in neurodevelopment and diseases. *Cell. Mol. Life Sci.* **2014**, *71*, 405–418. [CrossRef] [PubMed]
16. Ling, C.; Groop, L. Epigenetics: A molecular link between environmental factors and type 2 diabetes. *Diabetes* **2009**, *58*, 2718–2725. [CrossRef] [PubMed]
17. Chen, K.; Zhao, B.S.; He, C. Nucleic acid modifications in regulation of gene expression. *Cell Chem. Biol.* **2016**, *23*, 74–85. [CrossRef]
18. Doherty, R.; Couldrey, C. Exploring genome wide bisulfite sequencing for DNA methylation analysis in livestock: A technical assessment. *Front. Genet.* **2014**, *5*, 126. [CrossRef]
19. Flusberg, B.A.; Webster, D.R.; Lee, J.H.; Travers, K.J.; Olivares, E.C.; Clark, T.A.; Korlach, J.; Turner, S.W. Direct detection of DNA methylation during single-molecule, real-time sequencing. *Nat. Methods* **2010**, *7*, 461. [CrossRef]
20. Boch, J.; Bonas, U. Xanthomonas AvrBs3 family-type III effectors: Discovery and function. *Annu. Rev. Phytopathol.* **2010**, *48*, 419–436. [CrossRef]
21. Buryanov, Y.I.; Shevchuk, T. DNA methyltransferases and structural-functional specificity of eukaryotic DNA modification. *Biochemistry* **2005**, *70*, 730–742. [CrossRef]
22. 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**. [CrossRef] [PubMed]
23. 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]
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.* **2019**, *157*, 752–758. [CrossRef] [PubMed]
25. 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]
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. Zeng, F.; Fang, G.; Yao, L. A deep neural network for identifying DNA N4-methylcytosine sites. *Front. Genet.* **2020**, *11*, 209. [CrossRef]
28. Fu, J.; Tang, J.; Wang, Y.; Cui, X.; Yang, Q.; Hong, J.; Li, X.; Li, S.; Chen, Y.; Xue, W.; et al. Discovery of the consistently well-performed analysis chain for SWATH-MS based pharmacoproteomic quantification. *Front. Pharmacol.* **2018**, *9*, 681. [CrossRef]
29. He, W.; Jia, C.; Zou, Q. 4mCPred: Machine learning methods for DNA N4-methylcytosine sites prediction. *Bioinformatics* **2019**, *35*, 593–601. [CrossRef]
30. 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.
31. 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. *Brief. Bioinform.* **2020**, *21*, 1047–1057. [CrossRef]

32. Feng, P.; Yang, H.; Ding, H.; Lin, H.; Chen, W.; Chou, K.C. iDNA6mA-PseKNC: Identifying DNA N6-methyladenosine sites by incorporating nucleotide physicochemical properties into PseKNC. *Genomics* **2019**, *111*, 96–102. [CrossRef] [PubMed]
33. Nanni, L.; Lumini, A. An ensemble of K-local hyperplanes for predicting protein–protein interactions. *Bioinformatics* **2006**, *22*, 1207–1210. [CrossRef] [PubMed]
34. Cao, J.; Xiong, L. Protein sequence classification with improved extreme learning machine algorithms. *BioMed Res. Int.* **2014**, *2014*, 103054 . [CrossRef] [PubMed]
35. Raimondi, D.; Orlando, G.; Tabaro, F.; Lenaerts, T.; Rooman, M.; Moreau, Y.; Vranken, W.F. Large-scale in-silico statistical mutagenesis analysis sheds light on the deleteriousness landscape of the human proteome. *Sci. Rep.* **2018**, *8*, 1–11. [CrossRef] [PubMed]
36. Simonyan, K.; Vedaldi, A.; Zisserman, A. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv* **2013**, arXiv:1312.6034.
37. 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]
38. Wang, Y.; Yang, S.; Zhao, J.; Du, W.; Liang, Y.; Wang, C.; Zhou, F.; Tian, Y.; Ma, Q. Using machine learning to measure relatedness between genes: A multi-features model. *Sci. Rep.* **2019**, *9*, 1–15. [CrossRef]
39. Xu, Z.C.; Feng, P.M.; Yang, H.; Qiu, W.R.; Chen, W.; Lin, H. iRNAD: A computational tool for identifying D modification sites in RNA sequence. *Bioinformatics* **2019**, *35*, 4922–4929. [CrossRef]
40. Liu, B.; Liu, F.; Wang, X.; Chen, J.; Fang, L.; Chou, K.C. Pse-in-One: A web server for generating various modes of pseudo components of DNA, RNA, and protein sequences. *Nucleic Acids Res.* **2015**, *43*, W65–W71. [CrossRef]
41. He, J.; Fang, T.; Zhang, Z.; Huang, B.; Zhu, X.; Xiong, Y. PseUI: Pseudouridine sites identification based on RNA sequence information. *BMC Bioinform.* **2018**, *19*, 306. [CrossRef]
42. Su, Z.D.; Huang, Y.; Zhang, Z.Y.; Zhao, Y.W.; Wang, D.; Chen, W.; Chou, K.C.; Lin, H. iLoc-lncRNA: Predict the subcellular location of lncRNAs by incorporating octamer composition into general PseKNC. *Bioinformatics* **2018**, *34*, 4196–4204. [CrossRef] [PubMed]
43. Zhu, X.; He, J.; Zhao, S.; Tao, W.; Xiong, Y.; Bi, S. A comprehensive comparison and analysis of computational predictors for RNA N6-methyladenosine sites of Saccharomyces cerevisiae. *Brief. Funct. Genom.* **2019**, *18*, 367–376. [CrossRef] [PubMed]
44. Pan, Y.; Wang, Z.; Zhan, W.; Deng, L. Computational identification of binding energy hot spots in protein–rna complexes using an ensemble approach. *Bioinformatics* **2018**, *34*, 1473–1480. [CrossRef] [PubMed]
45. Wei, L.; Chen, H.; Su, R. M6APred-EL: A sequence-based predictor for identifying N6-methyladenosine sites using ensemble learning. *Mol. Ther.-Nucleic Acids* **2018**, *12*, 635–644. [CrossRef]
46. 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]
47. 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]
48. Yang, H.; Yang, W.; Dao, F.Y.; Lv, H.; Ding, H.; Chen, W.; Lin, H. A comparison and assessment of computational method for identifying recombination hotspots in Saccharomyces cerevisiae. *Brief. Bioinform.* **2019**. [CrossRef]
49. Zhou, Y.; Zeng, P.; Li, Y.H.; Zhang, Z.; Cui, Q. SRAMP: Prediction of mammalian N6-methyladenosine (m6A) sites based on sequence-derived features. *Nucleic Acids Res.* **2016**, *44*, e91. [CrossRef]
50. 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]
51. Cerf, N.J.; Adami, C. Information theory of quantum entanglement and measurement. *Phys. D Nonlinear Phenom.* **1998**, *120*, 62–81. [CrossRef]
52. Pan, G.; Jiang, L.; Tang, J.; Guo, F. A novel computational method for detecting DNA methylation sites with DNA sequence information and physicochemical properties. *Int. J. Mol. Sci.* **2018**, *19*, 511. [CrossRef] [PubMed]

53. ur Rehman, M.; Khan, S.H.; Rizvi, S.D.; Abbas, Z.; Zafar, A. Classification of skin lesion by interference of segmentation and convolotion neural network. In Proceedings of the 2018 2nd International Conference on Engineering Innovation (ICEI), Bangkok, Thailand, 5–6 July 2018; pp. 81–85.
54. Khan, S.H.; Abbas, Z.; Rizvi, S.D. Classification of Diabetic Retinopathy Images Based on Customised CNN Architecture. In Proceedings of the 2019 Amity International Conference on Artificial Intelligence (AICAI), Dubai, UAE, 4–6 February 2019; pp. 244–248.
55. Nizami, I.F.; Majid, M.; ur Rehman, M.; Anwar, S.M.; Nasim, A.; Khurshid, K. No-reference image quality assessment using bag-of-features with feature selection. *Multimed. Tools Appl.* **2020**, *79*, 7811–7836. [CrossRef]
56. 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]
57. Wahab, A.; Ali, S.D.; Tayara, H.; Chong, K.T. iIM-CNN: Intelligent identifier of 6mA sites on different species by using convolution neural network. *IEEE Access* **2019**, *7*, 178577–178583. [CrossRef]
58. Mahmoudi, O.; Wahab, A.; Chong, K.T. iMethyl-Deep: N6 Methyladenosine Identification of Yeast Genome with Automatic Feature Extraction Technique by Using Deep Learning Algorithm. *Genes* **2020**, *11*, 529. [CrossRef]
59. Chollet, F. Keras: Deep learning library for theano and tensorflow. *Io/k* **2015**, *7*, T1.
60. Khanal, J.; Tayara, H.; Chong, K.T. Identifying Enhancers and Their Strength by the Integration of Word Embedding and Convolution Neural Network. *IEEE Access* **2020**, *8*, 58369–58376. [CrossRef]
61. Tayara, H.; Chong, K.T. Improving the quantification of DNA sequences using evolutionary information based on deep learning. *Cells* **2019**, *8*, 1635. [CrossRef]
62. Tahir, M.; Tayara, H.; Chong, K.T. Convolutional neural networks for discrimination of RNA pseudouridine sites. *IBRO Rep.* **2019**, *6*, S552. [CrossRef]
63. 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. [CrossRef]
64. McCafferty, C.L.; Sergeev, Y.V. Global computational mutagenesis provides a critical stability framework in protein structures. *PLoS ONE* **2017**, *12*, e0189064. [CrossRef] [PubMed]
65. Lanchantin, J.; Singh, R.; Wang, B.; Qi, Y. Deep motif dashboard: Visualizing and understanding genomic sequences using deep neural networks. In *Pacific Symposium on Biocomputing 2017*; World Scientific: Singapore, 2017; pp. 254–265.
66. Baehrens, D.; Schroeter, T.; Harmeling, S.; Kawanabe, M.; Hansen, K.; Müller, K.R. How to explain individual classification decisions. *J. Mach. Learn. Res.* **2010**, *11*, 1803–1831.