International Journal of
Molecular Sciences

MDPI

Communication

Deep-4mCGP: A Deep Learning Approach to Predict 4mC Sites in *Geobacter pickeringii* by Using Correlation-Based Feature Selection Technique

Hasan Zulfiqar, Qin-Lai Huang, Hao Lv, Zi-Jie Sun, Fu-Ying Dao and Hao Lin *

School of Life Science and Technology, Center for Informational Biology, University of Electronic Science and Technology of China, Chengdu 610054, China; hasanzulfiqar66@gmail.com (H.Z.); huangqinlai2016@outlook.com (Q.-L.H.); hao.lyu@uzh.ch (H.L.); 15848174320@163.com (Z.-J.S.); N2107277C@e.ntu.edu.sg (F.-Y.D.)
* Correspondence: Correspondence: hlin@uestc.edu.cn

Abstract: 4mC is a type of DNA alteration that has the ability to synchronize multiple biological movements, for example, DNA replication, gene expressions, and transcriptional regulations. Accurate prediction of 4mC sites can provide exact information to their hereditary functions. The purpose of this study was to establish a robust deep learning model to recognize 4mC sites in Geobacter pickeringii. In the anticipated model, two kinds of feature descriptors, namely, binary and *k*-mer composition were used to encode the DNA sequences of *Geobacter pickeringii*. The obtained features from their fusion were optimized by using correlation and gradient-boosting decision tree (GBDT)-based algorithm with incremental feature selection (IFS) method. Then, these optimized features were inserted into 1D convolutional neural network (CNN) to classify 4mC sites from non-4mC sites in *Geobacter pickeringii*. The performance of the anticipated model on independent data exhibited an accuracy of 0.868, which was 4.2% higher than the existing model.

Keywords: deep learning; alteration; features vector; genomics; algorithm

Citation: Zulfiqar, H.; Huang, Q.-L.; Lv, H.; Sun, Z.-J.; Dao, F.-Y.; Lin, H. Deep-4mCGP: A Deep Learning Approach to Predict 4mC Sites in *Geobacter pickeringii* by Using Correlation-Based Feature Selection Technique. *Int. J. Mol. Sci.* **2022**, *23*, 1251. https://doi.org/10.3390/ijms23031251

Academic Editors: Jung Hun Oh and Mingon Kang

Received: 24 December 2021
Accepted: 20 January 2022
Published: 23 January 2022

1. Introduction

Alterations in DNA play a significant role in gene expression and regulation, DNA replication, and transcriptional regulation. Methylcytosine is a key epigenetic trait at 5′-cytosine-phosphate-guanine-3′ site. Methylcytosine is precisely correlated with cell growth and chromosomal protection [1,2]. 5-Hydroxymethylcytosine (5hmC), 5-methylcytosine (5mC), and 4-methylcytosine (4mC) are the familiar cytosine methylations in multiple genomes of prokaryotes and eukaryotes [3,4]. 5mC is a frequent type of methylcytosine and responsible for many neurodegenerative and cancerous diseases [5]. 4mC is a significant alteration that protects genomic knowledge from weakening by restriction enzymes [6].

Precise identification of 4mC sites can give important signs to understand the method of gene regulation. At present, there are several techniques to recognize 4mC sites, for example, single-molecule real-time sequencing [7], mass spectrometry [8], and bisulfite sequencing [9], but these techniques are time-consuming and expensive when utilized on next-generation sequencing data. Hence, a computational model to identify 4mC sites is needed on an urgent basis. Currently, a few computational and mathematical methods have been introduced to predict 4mC sites in multiple species. In 2017, Chen at al. [10] introduced the first computational model to predict 4mC sites in multiple species on the basis of confirmed 4mC dataset. Subsequently, Wei at al. [11] designed the novel iterative feature illustrative algorithm for the prediction of 4mC sites. Tang et al. [12] introduced the new linear integration method by merging the existing models for the identification of 4mC sites. Afterwards, Manavalan et al. [13] established the new tool Meta-4mCpred to recognize 4mC sites in six different species. Khanal et al. [14] introduced the first deep

learning model 4mCCNN by utilizing numerous feature combinations [15–17] for the prediction of 4mC sites in multiple genomes [18]. Although the prediction model 4mCCNN can yield good outcomes, there is still space for more improvement.

To tackle these hitches, we constructed a 1D CNN model to recognize 4mC sites in *Geobacter pickeringii*. Figure 1 illustrates the flowchart of the whole study. Binary and *k*-mer nucleotide composition descriptors were used to encode DNA sequences of *Geobacter pickeringii* into feature vectors and then these features were optimized by using a correlation and gradient-boosting decision tree (GBDT)-based algorithm with incremental feature selection (IFS) method. After this, these optimized features were inserted into 1D CNN-based classifier using 10-fold cross-validation and we attained the finest model to classify 4mC from non-4mC.

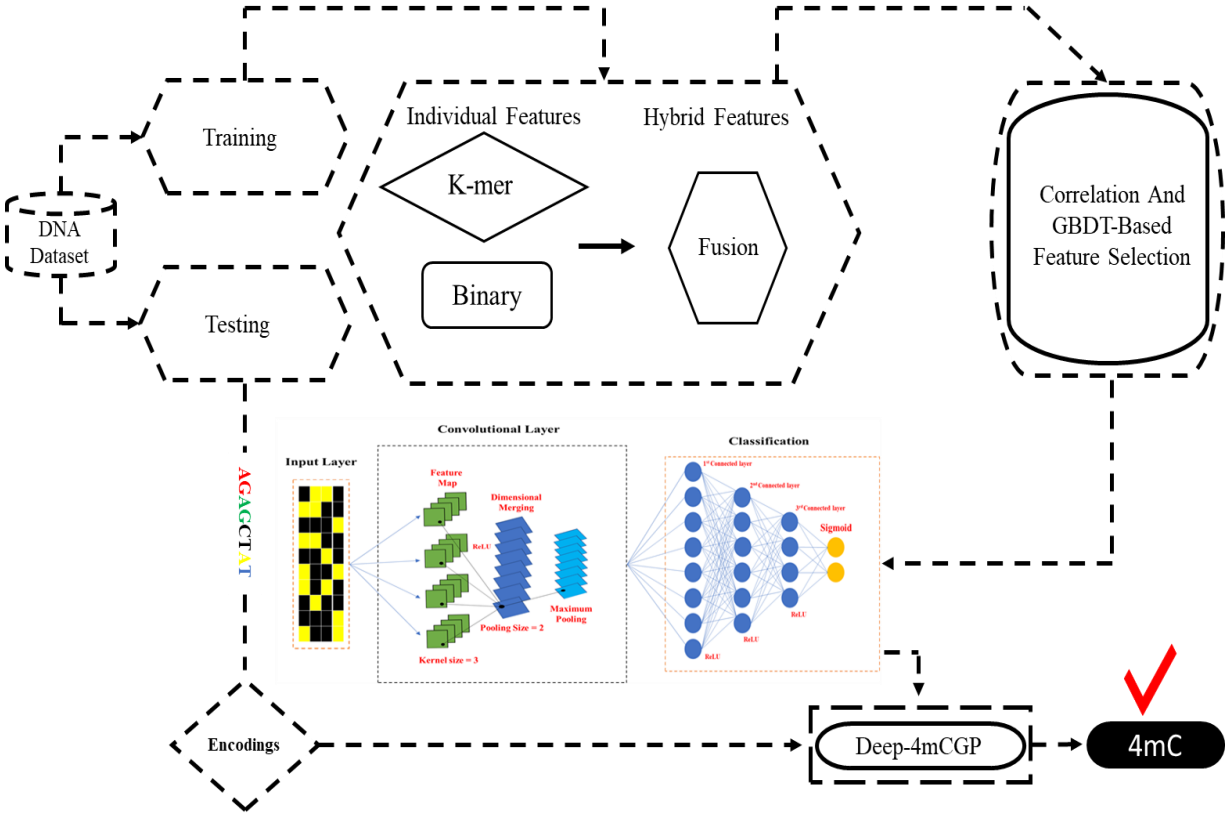

Figure 1. Flowchart of the whole study.

2. Results and Discussion

2.1. Performance Evaluation

We constructed a 1D CNN-based model named Deep-4mCGP for the identification of 4mC sites in *Geobacter pickeringii.* In the first step, we converted the sequence data in to feature vectors by using *k*-mer nucleotide composition and binary encodings. Subsequently, these feature vectors were improved by means of correlation and GBDT-based algorithm with IFS method. Initially, correlation and then GBDT with IFS were utilized to pick the finest features. Figure 2A,B displays the IFS curve of top features. Afterward, these finest features were inserted into 1D CNN by using 10-fold cross-validation to classify 4mC sites from non-4mC sites in *Geobacter pickeringii.* In this work, 10-fold cross-validation was employed to examine the efficiency of the model. The data were arbitrarily divided into 10 segments of equal proportion. Each segment was independently tested by the model, which was trained on the outstanding nine segments. Thus, 10-fold cross-validation technique was executed 10 times, and the average of the outcomes was the ultimate result. *AUROC* of the anticipated model was 0.986, which was 6.5% higher than the existing model.

The *accuracy*, *precision*, *recall*, and *F1* are shown in Table 1, and the *ROC* curve is shown in Figure 2C.

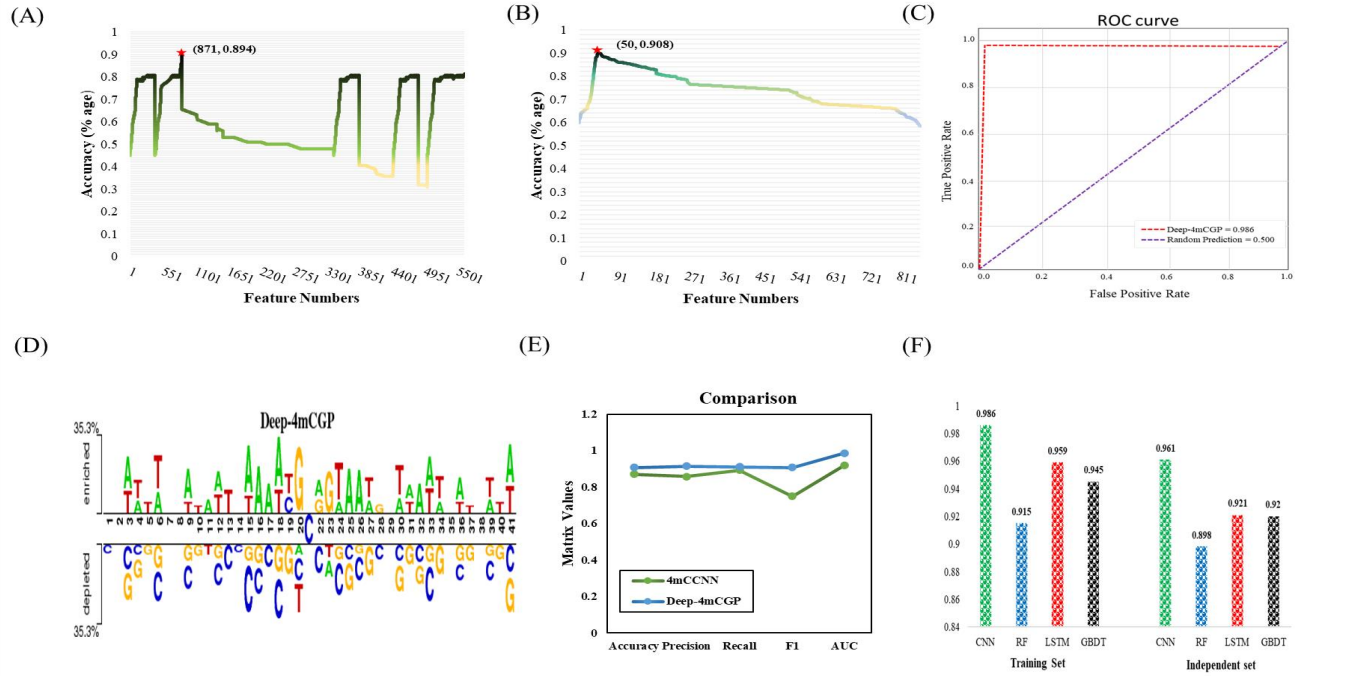

Figure 2. (**A**,**B**) The IFS technique for recognizing 4mC sites. Initially, 871 best features were picked from an overall 5624 by correlation measures (**A**). A total of 50 more optimized features were also attained from 871 best features by the using of GBDT on 10-fold CV. The Acc increases from 0.894 to 0.908 (**B**). Plot showing the *AUROC* curve of Deep-4mCGP on 10-fold CV (**C**). Nucleotides allocation along the alteration site (**D**). Performance comparison of Deep-4mCGP with 4mCCNN on 10-fold cross-validation (**E**). *AUROC* of predictors on training and independent data (**F**).

Table 1. Outcomes of single encodings and their fusion based-models on training and independent data by using different classification algorithms. Bold is used to highlight the best results.

			Training Data					Independent Data				
Algorithm	**FS**	**Method**	***Accuracy***	***Precision***	***Recall***	***F1***	***AUROC***	***Accuracy***	***Precision***	***Recall***	***F1***	***AUROC***
LSTM	5460	*k*-mer	0.861	0.872	0.861	0.811	0.943	0.825	0.820	0.812	0.819	0.882
	164	Binary	0.834	0.828	0.837	0.838	0.875	0.801	0.804	0.798	0.801	0.872
	5624	Fusion	0.868	0.865	0.859	0.862	0.937	0.810	0.814	0.808	0.813	0.902
	871	Fusion	0.859	0.857	0.847	0.857	0.925	0.808	0.801	0.807	0.800	0.876
	50	Fusion	0.884	0.878	0.881	0.879	0.959	0.841	0.842	0.839	0.842	0.921
RF	5460	*k*-mer	0.831	0.862	0.758	0.664	0.936	0.809	0.838	0.761	0.648	0.909
	164	Binary	0.772	0.763	0.755	0.770	0.863	0.753	0.748	0.753	0.756	0.832
	5624	Fusion	0.844	0.847	0.839	0.845	0.891	0.795	0.788	0.783	0.794	0.887
	871	Fusion	0.847	0.849	0.851	0.846	0.897	0.801	0.800	0.800	0.798	0.878
	50	Fusion	0.866	0.858	0.861	0.854	0.915	0.812	0.808	0.814	0.812	0.898
GBDT	5460	*k*-mer	0.848	0.881	0.776	0.676	0.962	0.828	0.861	0.770	0.669	0.931
	164	Binary	0.827	0.821	0.823	0.827	0.895	0.782	0.778	0.779	0.781	0.862
	5624	Fusion	0.835	0.832	0.830	0.832	0.893	0.786	0.780	0.786	0.786	0.882
	871	Fusion	0.851	0.853	0.848	0.854	0.901	0.814	0.810	0.815	0.810	0.893
	50	Fusion	0.875	0.874	0.868	0.860	0.945	0.836	0.835	0.830	0.841	0.920
CNN	5460	*k*-mer	0.880	0.879	0.887	0.880	0.949	0.848	0.844	0.841	0.845	0.927
	164	Binary	0.868	0.836	0.834	0.832	0.928	0.798	0.802	0.807	0.790	0.881
	5624	Fusion	0.868	0.865	0.859	0.862	0.937	0.810	0.814	0.808	0.813	0.903
	871	Fusion	0.894	0.877	0.897	0.889	0.955	0.846	0.845	0.841	0.838	0.920
	50	Fusion	**0.908**	**0.914**	**0.910**	**0.908**	**0.986**	**0.868**	**0.876**	**0.773**	**0.859**	**0.961**

2.2. Sequence Composition Analysis

The pattern of sequence along the alteration site is a crucial phase to recognize and understand the definition of genomic disparities [19]. In this work, we utilized Two Sample Logo [20] to inspect the dispersal of nucleotides along the 4mC site. Figure 2D illustrates the dispersal of nucleotides. Nucleotides 'A' and 'T' were separately rich at the upstream and downstream of the positive sequences, e.g., five consecutive 'A' nucleotides (30–34) and four successive 'A' (15–18, 24–27) originated in positive sequences. Nucleotides 'C' and 'G' were abundant at the upstream and downstream of the negative sequences, e.g., five repeated 'G' nucleotides (30–34) and four repeated 'G' nucleotides (3–6, 24–27) and four consecutive 'C' nucleotides (15–18) were noticed in negative sequences. Figure 2D shows that there was a significant variance amongst 4mC sequences and non-4mC sequences. The consequences proposed that the dispersal of nucleotides in diverse places are supportive for the precise identification of 4mC.

2.3. Comparison on the Basis of Independent Data

Features fusion were inserted into LSTM [21], GBDT [22], and RF [23,24] to compare with the CNN-based model [25]. Ultimately, on the basis of *AUROC*, we achieved a perfect model for each predictor, which is shown in Table 1 and Figure 2F. Comparison of anticipated model with 4mCCNN by using 10-fold cross-validation is shown in Figure 2E. On the independent data (200 Pos. seq and 200 Neg. seq) the efficiency of Deep-4mCGP was checked and then compared with the existing 4mCCNN. The *accuracy, precision, recall, F1,* and *AUROC* of the 4mCCNN were 0.826, 0.818, 0.823, 0.825, and 0.920, respectively. The *accuracy, precision, recall, F1,* and *AUROC* of Deep-4mCGP were 0.868, 0.876, 0.773, 0.859, and 0.961, respectively. The performance of the anticipated Deep-4mCGP on independent data exhibited the accuracy of 0.868, which was 4.2% higher than the 4mCCNN. The performance comparison is shown in Table 2.

Table 2. Performance comparison of Deep-4mCGP with 4mCCNN.

Predictor	*CV*	*Accuracy*	*Precision*	*Recall*	*F1*	*AUROC*	Reference
4mcCNN	10 (folds)	0.871	0.857	0.893	0.750	0.921	[14]
Deep-4mCGP	10 (folds)	0.908	0.914	0.910	0.908	0.986	Deep-4mCGP
4mcCNN	Test (Ind)	0.826	0.818	0.823	0.825	0.920	[14]
Deep-4mCGP	Test (Ind)	0.868	0.876	0.773	0.859	0.961	Deep-4mCGP

3. Materials and Methods

Authentic data are a significant requirement for the construction of a machine learning-based model [26,27]. Thus, we acquired the data of 1138 (569 Pos. seq and 569 Neg. seq) sequences of *Geobacter pickeringii* from the work of Chen et al. [10] for training and testing the model. Moreover, we attained the data of 400 sequences (200 Pos. seq and 200 Neg. seq) from the work of Manavalan et al. [13] for the sake of independent testing.

3.1. Feature Descriptors

Selecting useful and ideal features is an important step in developing machine learning models [4,28–37]. Converting the DNA sequences into numerical feature vectors is key in the recognition of functional elements, e.g., physiochemical properties, natural vectors, binary composition, and *k*-mer nucleotide compositions, which have been utilized in computational biology and bioinformatics [38,39]. In this study, binary and *k*-mer composition were used to encode DNA sequences of *Geobacter pickeringii.*

3.1.1. *k-mer*

k-mer composition has the ability to show interactions between nucleotides of DNA sequences [40]. The residues of nucleotides can be attained by setting the size of window and steps. A random sample F with n sequence length can be designated as

$$F = S_1\ S_2\ S_3 \ldots..S_i \ldots..S_{(n-1)}\ S_n \tag{1}$$

where S_i indicates the i-th nucleotide of the DNA sequences and can be converted in to 4^k D features vector with the help of *k-mer.*

$$F_k = \left[d_1^{k-tuple} d_2^{k-tuple} \ldots\ldots\ldots d_i^{k-tuple} \ldots..d_{4^k}^{k-tuple}\right]^T \tag{2}$$

where $d_1{}^{k\text{-}tuple}$ denotes the incidence of i-th *k-mer* and T represents the transposition. If the value of k is equal to 1, then DNA sequence will be decoded in to 4D features vector, and if the value of k is equal to 2, then DNA sequence will be 16D features vector. In this work, k was set as 1, 2, 3, 4, 5, 6. Consequently, DNA sequences were converted into ($4^1 + 4^2 + 4^3 + 4^4 + 4^5 + 4^6$ = 5460D) formulated as

$$F = F_1\ \cup\ F_2\ \cup\ F_3\ \cup\ F_4\ \cup\ F_5\ \cup\ F_6 \tag{3}$$

3.1.2. Binary

Binary encodings such as 0s and 1s have the ability to illustrate any information. Therefore, we can transform DNA sequence in the form of 0s and 1s. In this work, DNA sequences of *Geobacter pickeringii* with length of 41bp was encoded into the (4×41 = 164D) features vector.

3.2. Feature Selection

3.2.1. Correlation

Correlation is a familiar comparison amongst two different features, e.g., if the features are un-correlated, then the correlation will be zero; otherwise, it will be ± 1. Two complete modules named classical linear correlation and correlation on the basis of information theory were implemented to compute the correlation amongst the two unique variables. Linear correlation coefficient is the most acquainted and utilizable. The linear correlation coefficient 'r' for a pair of (p, q) variables is specified as

$$r = \frac{\sum (p_i - \overline{p}_i)(q_i - \overline{q}_i)}{\sqrt{\sum\ (p_i - \overline{p}_i)^2}\ \sqrt{\sum\ (q_i - \overline{q}_i)^2}} \tag{4}$$

Correlation generates good results in smaller datasets, but the performance of correlation coefficient is not up to the mark on gigantic amounts of data. Therefore, it is necessary to determine the substantial relationship amongst the features. Thus, we utilized the t-test to investigate the statistical correlation between the features and picked the significant features. The value of 't' can be computed as

$$t = r\sqrt{\frac{n-2}{1-r^2}} \tag{5}$$

where 'r' signifies the coefficient of correlation and 'n' represents the occurrences. '$n-2$' denotes the degree of freedom. Probability of the significance relation is 0.05. If 't' is greater than the probability of the significance relation 0.05, then the feature will be selected.

3.2.2. GBDT with IFS

GBDT is a popular machine learning-based classifier that has been utilized in various mathematical, cheminformatics, and bioinformatics tools [41,42]. It has the ability to

establish a scalable and reliable prediction model by utilizing non-linear joints of weak learners [43].

$$\{(x_1, y_1) \ldots (x_n, y_n)\} \ (\because \ x_i \epsilon \ x \ \subseteq \ S_n, \ and \ y_i \epsilon \ y \ \subseteq \ S) \ q_k \ (x) := \sum_{k=1}^{k} D \ (x; \theta_k) \tag{6}$$

where θ_k is minimal risk of the decision tree and $D_k(x; \theta_k)$ is the decision tree.

$$\hat{\theta_k} = argmin \sum_{i=1}^{n} P \ (y_i, \ q_{k-1}(x) + D \ (x; \theta_k) \ ) (\because P \text{ is the loss function}) \tag{7}$$

GBDT also computes the concluding evaluations in an advancing mode.

$$q_k(x) = q_{k-1}(x) + D \ (x; \theta_k) \tag{8}$$

Negative gradient loss function q_{k-1} is applied for residual computation.

$$S_{ki} = -\left[\frac{\partial P(y_i, q(x_i))}{\partial q(y_i)}\right]_{q(x)=q_{k-1}(x)} \ (\because \ i = 1, 2, 3 \ldots . n) \tag{9}$$

Hence, we trained the anticipated model through S_{ki} to compute the minimal risk θ_k. This kind of trees rationally represents the relations between variables, e.g., plotting the input X into J fragments $S_1 \ \ldots \ S_J$, and output is Z_J for area S_J.

$$D(x; \theta) = \sum_{j=1}^{J} z_j I\left(x_j \ \epsilon \ S_j \ \right) \tag{10}$$

The IFS [44,45] method was implemented in this work to pick the finest feature. IFS estimates the performance of the best q-ranked features repetitively for $q \ \epsilon \ (1, 2, 3, \ldots \ n)$, where '$n$' is the overall number of the features. IFS frequently stops at the first scrutiny of performance. In IFS, features were picked incrementally from a randomly taken initial feature and the finest result from several randomly re-instated IFS processes were outputted. A brief explanation of the IFS technique can be found in [46].

Algorithms 1: Correlation and GBDT-based Feature Selection Algorithm

Input: Training Data: = $Q \ (L_1, L_2, \ldots \ \ldots \ , L_k, L_c)$
Output: Q_{best}

1st Round

Begin
for $i = 1$ *to* k **do**
r = calculate correlational coefficient (L_i, L_c)
end
$p = 0.05$
$\rho = 0$ ($\because$ if there is no correlation among the F_i and F_c)
for $i = 1$ *to* k **do**
t = to calculate the significance (r, ρ) for L_i ($\because$ by utilizing the t-test value from Equation (5))
if $t >$ critical value
$Q_{best} = Q$ *list*
end
return Q_{best}

Algorithms 1: *Cont.*

2nd Round
Input: Q_{best}: = $(x_i, y_i)_{i=1}^{n}$
Where, (x_i = *data and* y_i = *label*)
LF: = *P* (y_i, *q* (*x*))
By initializing the model
$q_{\circ}(x)$: = argument minimum $\sum_{i=1}^{n} P(y_i, z)$
for I = {1, 2, 3, 4, 5 . . . , n} **do**
for k = {1, 2, 3, 4, 5 . . . , K} **do**
Pseudo residual error calculations: $S_{ki} = -\left[\frac{\partial P(y_i, q(x_i))}{\partial q(y_i)}\right]_{q(x)=q_{k-1}(x)}$
end
end
On the basis of S_{ki} , θ_k= $\{S_{kj} j = [1, 2, 3 \ldots J]\}$, we built a decision tree $D_k(x; \theta_k)$
for j = {1, 2, 3, 4, 5 . . . ., J} **do**
z_{kj} = argument minimum $\sum_{x_i \in S_{kj}}^{n} P(y_i, q_{k-1}(x) + z)$
end
Updating the model $q_k(x) = q_{k-1}(x) + \sum_{j=1}^{j} z_{kj} I\left(x \hat{I} S_{kj}\right)$
q (*x*) = $\sum_{k=1}^{K} \sum_{j=1}^{J} z_{kj} I\left(x \hat{I} S_{kj}\right)$
Output: The decision tree function *q* (*x*)

3.3. Convolutional Neural Network

LeCun at al. [47] introduced convolutional neural network, and now it has been roughly utilized in many biological and bioinformatics advances [48–50]. The fundamental principle of CNN is to create abundant filters that have the ability to produce hidden topological features from data by executing pooling procedures and layer-wise convolutions. The performance of CNN on 2D data of images and matrices is exceptional [51]. Subsequently, 1D CNN has been used to tackle the difficulties of biomedical sequence data identification and the research associated with natural language processing [41,52]. In this work, we implemented 1D CNN to identify 4mC sites in *Geobacter pickeringii.* We employed Keras 2.3.1 [53], TensorFlow 2.1.0, and Python 3.5.4 to perform this experiment. The best tuning parameters are recorded in Table 3.

Table 3. Program in TensorFlow 2.1.0 with employed parameters.

Classifier	Parameters
RF	N-estimators = 100, Learning-rate = 0.001, Mean absolute error = 0.143, Mean square error = 0.220
GBDT	N-estimators = 120, Learning-rate = 0.01, Mean absolute error = 0.117, Mean square error = 0.212
LSTM	nn.LSTM(input_size = feature_size, hidden_size = 128) nn.Linear(int_features = 128, out_features = 1) nn.Sigmoid() learning-rate = 0.001, Epoch = 100, Batch-size = 32
CNN	nn. Conv1d (in_channels = feature size, out_channels = 32, padding = valid, strides = 1, kernel_size = 2) nn.ReLU() nn.MaxPool 1d (padding = valid, strides = 2, pool_size = 2) nn. Dropout (p = 0.5) nn.Sigmoid() Learning-rate = 0.01, epoch = 80, batch-size = 32

3.4. Metrics Evaluation

Precision, accuracy, recall, and F1 [54–56] were employed to examine the effectiveness of the anticipated prediction model and formulated as

$$\begin{cases} Precision = \frac{TP}{TP+FP} \\ Recall = \frac{TP}{TP+FN} \\ Accuracy = \frac{TP+TN}{TP+FP+TN+FN} \\ F1 = 2 \times \frac{Precision \times Recall}{Precision+Recall} \end{cases} \quad (11)$$

where '*TP*' symbolizes the accurately predicted 4mC sequences, '*TN*' represents the perfectly predicted non-4mC sequences, '*FP*' indicates the non-4mC sequences predicted as 4mC sequences, and '*FN*' indicates the 4mC sequences predicted as non-4mC sequences.

4. Conclusions

4mC is a type of DNA alteration that has the ability to synchronize multiple biological movements for example DNA replication, gene expressions, and transcriptional regulations. Accurate prediction of 4mC sites can provide exact information to their hereditary functions. Currently, several machine learning models have been used to predict 4mC sites in multiple genomes [10,12,13,57–60]. However, there is only one deep learning-based model, 4mCCNN [14], that exists for *Geobacter pickeringii*. In this work, a deep learning model was constructed to recognize 4mC sites in *Geobacter pickeringii*. In the anticipated model, two kinds of feature descriptors, namely, binary and *k*-mer composition were used to encode the DNA sequences of *Geobacter pickeringii*. The obtained features from their fusion were optimized by using correlation and GBDT-based algorithm with IFS method. Then, these optimized features were inserted into a 1D CNN-based classifier using 10-fold cross-validation, and we attained the finest model to classify 4mC from non-4mC. The performance of the anticipated Deep-4mCGP on independent data exhibited an accuracy of 0.868, which was 4.2% higher than the 4mCCNN. The source code and data are available at GitHub: https://github.com/linDing-groups/Deep-4mCGP (accessed on 19 January 2022). In future work, we have a plan to release a web-based application to make our anticipated model more convenient for the users without programming and statistical knowledge.

Author Contributions: H.Z.: methodology, coding, data curation, visualization, writing—original draft preparation. Q.-L.H.: data curation, methodology. H.L. (Hao Lv): data curation, methodology, visualization. Z.-J.S.: data curation, methodology. F.-Y.D.: data curation, methodology, visualization. H.L. (Hao Lin): conceptualization, supervision, reviewing, editing. All authors have read and agreed to the published version of the manuscript.

Funding: This work has been supported by the Sichuan Provincial Science Fund for Distinguished Young Scholars (2020JDJQ0012) and National Nature Scientific Foundation of China (62172078).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: All the data are available at https://github.com/linDing-groups/Deep-4mCGP (accessed on 19 January 2022).

Acknowledgments: We are very thankful to Hui Ding (Center for Informational Biology, University of Electronic Science and Technology of China) for their constructive suggestions and support on this work.

Conflicts of Interest: All the authors are in agreement and declare that there is no conflict of interest.

References

1. Schübeler, D. Function and information content of DNA methylation. *Nature* **2015**, *517*, 321–326. [CrossRef]
2. Ao, C.; Yu, L.; Zou, Q. Prediction of bio-sequence modifications and the associations with diseases. *Brief. Funct. Genom.* **2021**, *20*, 1–18. [CrossRef] [PubMed]

3. Pataillot-Meakin, T.; Pillay, N.; Beck, S. 3-methylcytosine in cancer: An underappreciated methyl lesion? *Epigenomics* **2016**, *8*, 451–454. [CrossRef] [PubMed]
4. Yalcin, D.; Otu, H.H. An Unbiased Predictive Model to Detect DNA Methylation Propensity of CpG Islands in the Human Genome. *Curr. Bioinform.* **2021**, *16*, 179–196. [CrossRef]
5. Robertson, K.D. DNA methylation and human disease. *Nat. Rev. Genet.* **2005**, *6*, 597–610. [CrossRef] [PubMed]
6. Iyer, L.M.; Abhiman, S.; Aravind, L. Natural history of eukaryotic DNA methylation systems. *Prog. Mol. Biol. Transl. Sci.* **2011**, *101*, 25–104. [PubMed]
7. 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] [PubMed]
8. Doherty, R.; Couldrey, C. Exploring genome wide bisulfite sequencing for DNA methylation analysis in livestock: A technical assessment. *Front. Genet.* **2014**, *5*, 126. [CrossRef]
9. Boch, J.; Bonas, U. Xanthomonas AvrBs3 family-type III effectors: Discovery and function. *Annu. Rev. Phytopathol.* **2010**, *48*, 419–436. [CrossRef]
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. Wei, L.; Su, R.; Luan, S.; Liao, Z.; Manavalan, B.; Zou, Q.; Shi, X. Iterative feature representations improve N4-methylcytosine site prediction. *Bioinformatics* **2019**, *35*, 4930–4937. [CrossRef]
12. Tang, Q.; Kang, J.; Yuan, J.; Tang, H.; Li, X.; Lin, H.; Huang, J.; Chen, W. DNA4mC-LIP: A linear integration method to identify N4-methylcytosine site in multiple species. *Bioinformatics* **2020**, *36*, 3327–3335. [CrossRef] [PubMed]
13. Manavalan, B.; Basith, S.; Shin, T.H.; Wei, L.; Lee, G. Meta-4mCpred: A Sequence-Based Meta-Predictor for Accurate DNA 4mC Site Prediction Using Effective Feature Representation. *Mol. Ther.-Nucleic Acids* **2019**, *16*, 733–744. [CrossRef] [PubMed]
14. 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]
15. 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] [PubMed]
16. 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. [CrossRef] [PubMed]
17. Zulfiqar, H.; Khan, R.S.; Hassan, F.; Hippe, K.; Hunt, C.; Ding, H.; Song, X.-M.; Cao, R. Computational identification of N4-methylcytosine sites in the mouse genome with machine-learning method. *Math. Biosci. Eng.* **2021**, *18*, 3348–3363. [CrossRef]
18. 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. [CrossRef]
19. Smith, Z.D.; Meissner, A. DNA methylation: Roles in mammalian development. *Nat. Rev. Genet.* **2013**, *14*, 204–220. [CrossRef]
20. Vacic, V.; Iakoucheva, L.M.; Radivojac, P. Two Sample Logo: A graphical representation of the differences between two sets of sequence alignments. *Bioinformatics* **2006**, *22*, 1536–1537. [CrossRef]
21. Gers, F.A.; Schmidhuber, J.; Cummins, F. Learning to forget: Continual prediction with LSTM. *Neural Comput.* **2000**, *12*, 2451–2471. [CrossRef] [PubMed]
22. Ye, J.; Chow, J.-H.; Chen, J.; Zheng, Z. Stochastic gradient boosted distributed decision trees. In Proceedings of the 18th ACM Conference on Information and Knowledge Management, Hong Kong, China, 2–6 November 2009; pp. 2061–2064.
23. Qi, Y. Random forest for bioinformatics. In *Ensemble Machine Learning*; Springer: Berlin/Heidelberg, Germany, 2012; pp. 307–323.
24. Ahmed, F.F.; Khatun, M.S.; Mosharaf, M.P.; Mollah, M.N.H. Prediction of Protein-protein Interactions in Arabidopsis thaliana Using Partial Training Samples in a Machine Learning Framework. *Curr. Bioinform.* **2021**, *16*, 865–879. [CrossRef]
25. Zhang, Y.; Li, Y.; Wang, R.; Lu, J.; Ma, X.; Qiu, M. PSAC: Proactive Sequence-aware Content Caching via Deep Learning at the Network Edge. *IEEE Trans. Netw. Sci. Eng.* **2020**, *7*, 2145–2154. [CrossRef]
26. Su, W.; Liu, M.L.; Yang, Y.H.; Wang, J.S.; Li, S.H.; Lv, H.; Dao, F.Y.; Yang, H.; Lin, H. PPD: A Manually Curated Database for Experimentally Verified Prokaryotic Promoters. *J. Mol. Biol.* **2021**, *433*, 166860. [CrossRef] [PubMed]
27. Sharma, A.K.; Srivastava, R. Protein Secondary Structure Prediction Using Character bi-gram Embedding and Bi-LSTM. *Curr. Bioinform.* **2021**, *16*, 333–338. [CrossRef]
28. Hasan, M.M.; Alam, M.A.; Shoombuatong, W.; Deng, H.W.; Manavalan, B.; Kurata, H. NeuroPred-FRL: An interpretable prediction model for identifying neuropeptide using feature representation learning. *Brief. Bioinform.* **2021**, *22*, bbab167. [CrossRef] [PubMed]
29. Charoenkwan, P.; Chiangjong, W.; Nantasenamat, C.; Hasan, M.M.; Manavalan, B.; Shoombuatong, W. StackIL6: A stacking ensemble model for improving the prediction of IL-6 inducing peptides. *Brief. Bioinform.* **2021**, *22*, bbab172. [CrossRef] [PubMed]
30. Zulfiqar, H.; Sun, Z.J.; Huang, Q.L.; Yuan, S.S.; Lv, H.; Dao, F.Y.; Lin, H.; Li, Y.W. Deep-4mCW2V: A sequence-based predictor to identify N4-methylcytosine sites in Escherichia coli. *Methods* **2021**, *in press*. [CrossRef]
31. Ju, Z.; Wang, S.-Y. Prediction of Neddylation Sites Using the Composition of k-spaced Amino Acid Pairs and Fuzzy SVM. *Curr. Bioinform.* **2020**, *15*, 725–731. [CrossRef]
32. Zhang, D.; Chen, H.-D.; Zulfiqar, H.; Yuan, S.-S.; Huang, Q.-L.; Zhang, Z.-Y.; Deng, K.-J. iBLP: An XGBoost-based predictor for identifying bioluminescent proteins. *Comput. Math. Methods Med.* **2021**, *2021*, 6664362. [CrossRef]

33. Lv, H.; Dao, F.-Y.; Zulfiqar, H.; Lin, H. DeepIPs: Comprehensive assessment and computational identification of phosphorylation sites of SARS-CoV-2 infection using a deep learning-based approach. *Brief. Bioinform.* **2021**, *22*, bbab244. [CrossRef] [PubMed]
34. Zhang, L.; Huang, Z.; Kong, L. CSBPI Site: Multi-Information Sources of Features to RNA Binding Sites Prediction. *Curr. Bioinform.* **2021**, *16*, 691–699. [CrossRef]
35. Lv, H.; Shi, L.; Berkenpas, J.W.; Dao, F.-Y.; Zulfiqar, H.; Ding, H.; Zhang, Y.; Yang, L.; Cao, R. Application of artificial intelligence and machine learning for COVID-19 drug discovery and vaccine design. *Brief. Bioinform.* **2021**, *22*, bbab320. [CrossRef]
36. Zulfiqar, H.; Masoud, M.S.; Yang, H.; Han, S.G.; Wu, C.Y.; Lin, H. Screening of prospective plant compounds as H1R and CL1R inhibitors and its antiallergic efficacy through molecular docking approach. *Comput. Math. Methods Med.* **2021**, *2021*, 6683407. [CrossRef]
37. Hasan, M.M.; Schaduangrat, N.; Basith, S.; Lee, G.; Shoombuatong, W.; Manavalan, B. HLPpred-Fuse: Improved and robust prediction of hemolytic peptide and its activity by fusing multiple feature representation. *Bioinformatics* **2020**, *36*, 3350–3356. [CrossRef]
38. Govindaraj, R.G.; Subramaniyam, S.; Manavalan, B. Extremely-randomized-tree-based Prediction of N(6)-Methyladenosine Sites in Saccharomyces cerevisiae. *Curr. Genom.* **2020**, *21*, 26–33. [CrossRef]
39. Li, Q.; Yu, J.; Yan, Y.; Chen, Y.; Tan, S. PsePSSM-based Prediction for the Protein-ATP Binding Sites. *Curr. Bioinform.* **2021**, *16*, 576–582.
40. Dao, F.-Y.; Lv, H.; Zulfiqar, H.; Yang, H.; Su, W.; Gao, H.; Ding, H.; Lin, H. A computational platform to identify origins of replication sites in eukaryotes. *Brief. Bioinform.* **2021**, *22*, 1940–1950. [CrossRef]
41. Lv, H.; Dao, F.-Y.; Zulfiqar, H.; Su, W.; Ding, H.; Liu, L.; Lin, H. A sequence-based deep learning approach to predict CTCF-mediated chromatin loop. *Brief. Bioinform.* **2021**, *22*, 1–13. [CrossRef]
42. Zulfiqar, H.; Yuan, S.-S.; Huang, Q.-L.; Sun, Z.-J.; Dao, F.-Y.; Yu, X.-L.; Lin, H. Identification of cyclin protein using gradient boost decision tree algorithm. *Comput. Struct. Biotechnol. J.* **2021**, *19*, 4123–4131. [CrossRef]
43. Ke, G.; Meng, Q.; Finley, T.; Wang, T.; Chen, W.; Ma, W.; Ye, Q.; Liu, T.-Y. Lightgbm: A highly efficient gradient boosting decision tree. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 3146–3154.
44. Yang, W.; Zhu, X.-J.; Huang, J.; Ding, H.; Lin, H. A Brief Survey of Machine Learning Methods in Protein Sub-Golgi Localization. *Curr. Bioinform.* **2019**, *14*, 234–240. [CrossRef]
45. Tan, J.-X.; Li, S.-H.; Zhang, Z.-M.; Chen, C.-X.; Chen, W.; Tang, H.; Lin, H. Identification of hormone binding proteins based on machine learning methods. *Math. Biosci. Eng.* **2019**, *16*, 2466–2480. [CrossRef]
46. Alim, A.; Rafay, A.; Naseem, I. PoGB-pred: Prediction of Antifreeze Proteins Sequences Using Amino Acid Composition with Feature Selection Followed by a Sequential-based Ensemble Approach. *Curr. Bioinform.* **2021**, *16*, 446–456. [CrossRef]
47. LeCun, Y.; Bottou, L.; Bengio, Y.; Haffner, P. Gradient-based learning applied to document recognition. *Proc. IEEE* **1998**, *86*, 2278–2324. [CrossRef]
48. Niu, M.; Lin, Y.; Zou, Q. sgRNACNN: Identifying sgRNA on-target activity in four crops using ensembles of convolutional neural networks. *Plant Mol. Biol.* **2021**, *105*, 483–495. [CrossRef] [PubMed]
49. Zhang, Y.; Yan, J.; Chen, S.; Gong, M.; Gao, D.; Zhu, M.; Gan, W. Review of the Applications of Deep Learning in Bioinformatics. *Curr. Bioinform.* **2020**, *15*, 898–911. [CrossRef]
50. Bukhari, S.A.S.; Razzaq, A.; Jabeen, J.; Khan, S.; Khan, Z. Deep-BSC: Predicting Raw DNA Binding Pattern in Arabidopsis thaliana. *Curr. Bioinform.* **2021**, *16*, 457–465. [CrossRef]
51. Kwon, Y.-H.; Shin, S.-B.; Kim, S.-D. Electroencephalography based fusion two-dimensional (2D)-convolution neural networks (CNN) model for emotion recognition system. *Sensors* **2018**, *18*, 1383. [CrossRef]
52. Mo, F.; Luo, Y.; Fan, D.A.; Zeng, H.; Zhao, Y.N.; Luo, M.; Liu, X.B.; Ma, X.L. Integrated Analysis of mRNA-seq and miRNA-seq to identify c-MYC, YAP1 and miR-3960 as Major Players in the Anticancer Effects of Caffeic Acid Phenethyl Ester in Human Small Cell Lung Cancer Cell Line. *Curr. Gene Ther.* **2020**, *20*, 15–24. [CrossRef]
53. Chollet, F. Keras: Deep learning library for theano and tensorflow. *Keras* **2015**, *7*, T1. Available online: https://Keras.Io/ (accessed on 19 January 2022).
54. Cao, R.; Freitas, C.; Chan, L.; Sun, M.; Jiang, H.; Chen, Z. ProLanGO: Protein function prediction using neural machine translation based on a recurrent neural network. *Molecules* **2017**, *22*, 1732. [CrossRef] [PubMed]
55. Gai, D.; Shen, X.; Chen, H. Effective Classification of Melting Curve in Real-time PCR Based on Dynamic Filter-based Convolutional Neural Network. *Curr. Bioinform.* **2021**, *16*, 820–828. [CrossRef]
56. Ao, C.; Zou, Q.; Yu, L. RFhy-m2G: Identification of RNA N2-methylguanosine modification sites based on random forest and hybrid features. *Methods* **2021**, *in press*. [CrossRef] [PubMed]
57. He, W.; Jia, C.; Zou, Q. 4mCPred: Machine learning methods for DNA N4-methylcytosine sites prediction. *Bioinformatics* **2019**, *35*, 593–601. [CrossRef]
58. Lv, H.; Dao, F.-Y.; Zhang, D.; Guan, Z.-X.; Yang, H.; Su, W.; Liu, M.-L.; Ding, H.; Chen, W.; Lin, H. iDNA-MS: An integrated computational tool for detecting DNA modification sites in multiple genomes. *Iscience* **2020**, *23*, 100991. [CrossRef]
59. Zulfiqar, H.; Dao, F.Y.; Lv, H.; Yang, H.; Zhou, P.; Chen, W.; Lin, H. Identification of Potential Inhibitors Against SARS-CoV-2 Using Computational Drug Repurposing Study. *Curr. Bioinform.* **2021**, *16*, 1320–1327. [CrossRef]
60. 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.* **2021**, *22*, bbaa124. [CrossRef]