frontiers in Cell and Developmental Biology

ORIGINAL RESEARCH
published: 28 July 2020
doi: 10.3389/fcell.2020.00614

iPromoter-5mC: A Novel Fusion Decision Predictor for the Identification of 5-Methylcytosine Sites in Genome-Wide DNA Promoters

Lei Zhang, Xuan Xiao and Zhao-Chun Xu**

Computer Department, Jing-De-Zhen Ceramic Institute, Jingdezhen, China

OPEN ACCESS

Edited by:
Yu Xue,
Huazhong University of Science and Technology, China

Reviewed by:
Leyi Wei,
Shandong University, China
Wei Chen,
North China University of Science and Technology, China
Jianbo Pan,
Johns Hopkins Medicine, United States

***Correspondence:**
Xuan Xiao
jdzxiaoxuan@163.com
Zhao-Chun Xu
jdzxuzhaochun@163.com

Specialty section:
This article was submitted to Cellular Biochemistry, a section of the journal Frontiers in Cell and Developmental Biology

Received: *08 May 2020*
Accepted: *22 June 2020*
Published: *28 July 2020*

Citation:
Zhang L, Xiao X and Xu Z-C (2020) iPromoter-5mC: A Novel Fusion Decision Predictor for the Identification of 5-Methylcytosine Sites in Genome-Wide DNA Promoters. Front. Cell Dev. Biol. 8:614. doi: 10.3389/fcell.2020.00614

The hypomethylation of the whole cancer genome and the hypermethylation of the promoter of specific tumor suppressor genes are the important reasons for the rapid proliferation of cancer cells. Therefore, obtaining the distribution of 5-methylcytosine (5mC) in promoters is a key step to further understand the relationship between promoter methylation and mRNA gene expression regulation. Large-scale detection of DNA 5mC through wet experiments is still time-consuming and laborious. Therefore, it is urgent to design a method for identifying the 5mC site of genome-wide DNA promoters. Based on promoter methylation data of the small cell lung cancer (SCLC) from the database named cancer cell line Encyclopedia (CCLE), we built a fusion decision predictor called iPromoter-5mC for identifying methylation modification sites in promoters using deep neural network (DNN). One-Hot Encoding (One-hot) was used to encode the promoter samples for the classification. The method achieves average AUC of 0.957 on the independent testing dataset, indicating that our predictor is robust and reliable. A user-friendly web-server called iPromoter-5mC could be freely accessible at http://www.jci-bioinfo.cn/iPromoter-5mC, which will provide simple and effective means for users to study promoter 5mC modification. The source code of the proposed methods is freely available for academic research at https://github.com/zlwuxi/iPromoter-5mC.

Keywords: promoter, 5-methylcytosine, fusion decision, predictor, web-server, deep neural network

INTRODUCTION

DNA methylation dominates any cell processes, and plays a particularly important role in regulating expression of gene (Bird, 2007; Deichmann, 2016; Nicoglou and Merlin, 2017). DNA methylation at promoters and enhancers has been associated with cell differentiation, developmental processes, cancer development, and regulation of the immune system (Muller et al., 2019). At present, N6-methyladenine (6mA), N4-methylcytosine (4mC) and 5-methylcytosine (5mC) are the three most well-studied types of DNA methylation (Wei et al., 2019). 5mC is a covalent addition between the methyl group and the 5-carbon of the cytosine ring. In somatic cells, 5mC occurs almost exclusively in the context of paired symmetrical methylation of a CpG site.

Recent study (Michalak et al., 2019) suggests that aberrant levels of 5mC at CpG islands in promoter regions is associated with inactivation of various tumor suppressor genes (TSGs). In

young normal cells, 5mC is low in the promoter regions but high in the genic and intergenic regions. However, in aging and in cancer, a limited number of genomic loci acquire 5mC, especially at the CpG islands in promoter regions of tumor suppressor and Polycomb-repressed gene, resulting in gene silencing and loss of function. In normal tissue, heterochromatin contains repeating elements and is highly methylated. The aberrant promoter methylation can lead to cancer initiation and progression, which has been confirmed in CpG island methylator phenotype (CIMP) cancers (Gessler, 1999; Kang et al., 2002; Mansour, 2014). Thus, promoter methylation can be used as a potential biomarker for cancer diagnosis and for helping determine prognosis, indicating that identification of 5mC modification in promoter regions by analyzing CpG islands in cell systems of a specific cancer could provide a reference for cancer early diagnosis and precise treatment.

Among cancers worldwide, both the incidence and death rate of lung cancer are in the first place. Small cell lung cancer (SCLC) poses approximately 15% of newly increasing clinical cases with lung cancer each year (Siegel et al., 2018). Its pattern is significantly different from other lung cancer, and is closely related to the high expression of E2F target and EZH2 gene of histone methyltransferase. Furthermore, SCLC is famous for its dense cluster of high-level methylation in CpG islands of discrete promoter. Therefore, in this study, we are concentrating on improving the ability to access the methylation status of promoters for a large number of genes or the entire genome in SCLC.

One of the most usual methods for identifying DNA methylation is distinguishing the cytosine-5 methylation within the CpG dinucleotides (Bianchi and Zangi, 2015; Muller et al., 2019). The popular sequencing technology for identifying 5mC sites includes Methylated DNA immunoprecipitation sequencing (MeDIP-seq), Methyl-Binding Domain sequencing (MBD-seq) and DNA methylome profiling at single-base resolution through bisulfite sequencing (MB-seq) (Down et al., 2008). However, these wet-lab methods are expensive and time-consuming. Therefore, it is urgent to develop a number of methods or tools for the accurate detection of DNA 5mC modification sites.

Over the past decade, computational methods have been proposed to identify 5mC modification sites. Bhasin et al. (2005) developed a SVM-based model called "Methylator," for the prediction of 5mC modification sites using the methylated and non-methylated CpG dinucleotide sequences from various sources ranging from plants to humans in MethDB database (Amoreira, 2003). Fang et al. (2006) developed a SVM-based classifier called "MethCGI" using nucleotide sequence contents and transcription factor binding sites as features. Compared with the previous two, the predictor "iDNA-Methyl" (Liu et al., 2015) constructed by using the trinucleotide composition and pseudo amino acid components achieved higher success rates. Recently, a novel computational tool called NanoMod (Liu et al., 2018) was designed to improve the performance of detecting candidate positions with DNA modifications. Based on deep neural networks, a computational approach called DeepCpG (Angermueller et al., 2017) was developed to predict methylation states in single cells.

Though the research about the recognition of DNA 5mC modification sites have had a significant advance in recent years, but still exist shortness. Compared to increasing massive high-throughput data, previous studies are of small sample size. Furthermore, among above-mentioned methods, there are three webservers developed by the researchers: Methylator, MethCGI, and iDNA-Methyl, however, only the latter is available but slow, causing much inconvenience to scholars. Most importantly, there is still no computation tool to identify DNA 5mC modification sites in promoters to detect the biomarkers of a specific cancer. Therefore, in the current study, we are devoted to solve these problems and to develop a tool or software for quickly and precisely identifying DNA 5mC modification sites in promoters.

MATERIALS AND METHODS

Benchmark Datasets

The construction of the high-quality data sets is an essential step in the process of establishing the classification model. In the current study, all the sequence samples were collected from the database named cancer cell line Encyclopedia (CCLE) (Barretina et al., 2012; Li et al., 2019), which provided the location information of gene promoter regions and 5mC modification sites experimented by reduced representation bisulfite sequencing (RRBS) (Ghandi et al., 2019) in cell lines of various cancers. Due to the high incidence rate and mortality rate of lung cancer, here we focused on the small cell lung cancer (SCLC) to reveal the distribution of 5mC modification in promoters.

In accordance with the forward/reverse (±) chain and 5mC modification sites' positions in promoters, we collected the sequence samples from the most recent human assembly GRCh37/hg19 on UCSC Genome Browser. It is noteworthy that the sample sequence containing 5mC modification site described as the base G (guanine) in the reverse chain should convert to the reverse complementary sequence, compatible with the principle that the DNA 5mC methylation tends to occur at cytosine (C). Generally, we considered the base C with the methylation level greater than zero as the true 5mC modification site, otherwise, as the false 5mC modification site.

In order to more succinctly describe the promoter sequence fragment potentially containing 5mC modification site, the sample sequence can be expressed as

$$E_\delta(\mathbb{C}) = E_{-\delta}E_{-(\delta-1)}\cdots E_{-2}E_{-1}\mathbb{C}E_{+1}E_{+2}\cdots E_{+(\delta-1)}E_{+\delta} \quad (1)$$

where the double letter $\mathbb{C}$ represents the cytosine; the subscript δ is an integer, indicating the location of the base in the sequence; $E_{-\delta}$ is the δ -th base upstream from the center and $E_{+\delta}$ is the δ -th base downstream from the center.

The sample sequence thus obtained can be divided into two categories:

$$E_\delta(\mathbb{C}) \in \begin{cases} E_\delta^-(\mathbb{C}) \\ E_\delta^+(\mathbb{C}) \end{cases} \quad (2)$$

where $E_\delta^-(\mathbb{C})$ represents a false 5mC modification segment with $\mathbb{C}$ at its center, $E_\delta^+(\mathbb{C})$ denotes a true 5mC modification segment

TABLE 1 | Distribution of experimental data sets.

Attribute	Total	Training data	Testing data
Positive	69,750	55,800	13,950
Negative	823,576	658,861	164,715

with $\mathbb{C}$ at its center, and the symbol $\in$ denotes "a member of" in the set theory.

Therefore, the benchmark dataset can be formulated by

$$S_\delta = S_\delta^- \cup S_\delta^+ \tag{3}$$

where S_δ^- denotes the negative subset containing the false 5mC modification site samples; S_δ^+, the positive subset containing the true 5mC modification site samples; and symbol $\cup$ represents union in the set theory.

Unbalanced data between the true 5mC modification site samples and the false 5mC modification site samples could more objectively reflect the distribution of 5mC modification in promoters. Therefore, the proportion of positive samples and negative samples was set to about 1:11 in this study. In order to reduce the adverse effects of redundancy and homologous bias, sequences with more than 80% sequence similarity were removed using CD-HIT software.

Finally, we obtained the benchmark dataset S_δ composed of 893,326 methylation sample sequences in promoter regions, of which 69,750 sample sequences belong to the positive sample dataset S_δ^+ and 823,576 sample sequences belong to the negative sample dataset S_δ^-. To investigate the stability and robustness of the prediction model, we randomly selected 80% data in S_δ^+ and S_δ^-, respectively, as training set S_1 for constructing and training the prediction model, and remained 20% as independent testing dataset S_2 to test the constructed model (**Table 1**). These datasets can be downloaded from the website http://www.jci-bioinfo.cn/iPromoter-5mC/download.

Extract Features From DNA Sequences

Feature extraction, fusion and selection are the important steps in machine learning process. Many feature extraction methods for protein, RNA and DNA sequences, including PseAAC, PseKNC, PCPs, PCM, PS(k-mer)NP (Zou et al., 2016), have been proposed to overcome the prediction problem of modification sites. In the current study, we employed two effective feature extraction methods (one-hot and DPF) to extract feature directly form DNA sample sequences.

One-Hot Encoding Method (One-Hot)

One-hot is a simple but effective feature extraction method, especially for deep learning model. The nucleotides A, C, G and T are denoted as one of the four one-hot vectors [1,0,0,0], [0,1,0,0], [0,0,1,0], and [0,0,0,1] (**Figure 1**).

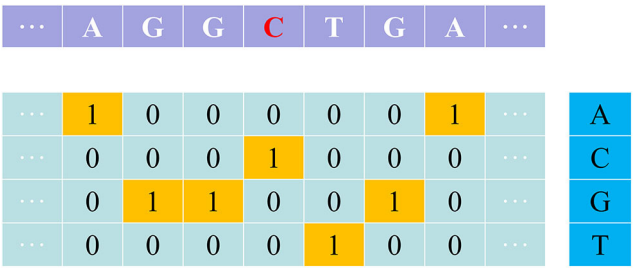

FIGURE 1 | An illustration showing the one-hot encoding method.

The Deoxynucleotide Property and Frequency (DPF)

Deoxynucleotides are the basic structural and functional units of DNA, and the sequence generated by deoxynucleotides determines biological diversity. Therefore, their chemical properties can influence the inherited characteristics of the DNA sequence to a certain extent. Similar to the encoding method of RNA sequences used in identifying 4mC sites, the deoxynucleotide property and frequency (DPF) (Xia et al., 2019; Xu et al., 2019) is an effective sequence encoding scheme for computationally identifying 5mC modification sites.

Each of the four deoxynucleotides has a different chemical property. Given the sample sequence Q represented by Equation (1), the k-th deoxynucleotide in Equation (1) can be converted into a three-dimensional vector, as shown in the Equation (4). Considering that purines have two rings between them and pyrimidines have only one ring, we added the feature of ring structure to feature extraction. Since there is an amino group between A and C, but A keto group between G and T, we added functional group features to feature extraction. In terms of the strength of the hydrogen bond between the base pair, the hydrogen bond between C and G is stronger than the hydrogen bond between A and T, because A is always paired with T by two hydrogen bonds, but C is bound to G by three hydrogen bonds. So we added hydrogen bond features to Q, as shown in the following expression.

$$Q_k = (x_k, y_k, z_k) \tag{4}$$

where x_k represents the "ring structure"; y_k, the "functional group"; z_k, the "hydrogen bond."

x_k, y_k and z_k can be formulated by Equation (5):

$$\begin{aligned} x_k &= \begin{cases} 1 \; if \; Q_k \in \{A \; , \; G\} \\ 0 \; if \; Q_k \in \{C \; , \; T\} \end{cases} \\ y_k &= \begin{cases} 1 \; if \; Q_k \in \{A \; , \; C\} \\ 0 \; if \; Q_k \in \{G \; , \; T\} \end{cases} \\ z_k &= \begin{cases} 1 \; if \; Q_k \in \{A \; , \; T\} \\ 0 \; if \; Q_k \in \{C \; , \; G\} \end{cases} \end{aligned} \tag{5}$$

In order to extract the sequence position information as much as possible (Chen et al., 2017), the cumulative frequency characteristics of deoxynucleotides were adopted:

$$\lambda_k = \frac{\sum_{j=1}^{k} \mathcal{F}(M_j)}{k} \left(1 \le k \le 2\delta + 1\right) \tag{6}$$

where k is the length of the sample sequence, λ_k is the density of the deoxynucleotide Q_k along the subsequence from position 1 to

position k in the sample sequence, and $\mathcal{F}(M_j)$ can be expressed as below.

$$\mathcal{F}(M_j) = \begin{cases} 1 \ if \ M_j = Q_k \\ 0 \ otherwise \end{cases} \quad (7)$$

Then we obtained a feature vector $\vec{v}$ to represent the k-th deoxynucleotide in the sample sequence, as shown in the following formula,

$$\vec{v} = (x_k, y_k, z_k, \lambda_k) \quad (8)$$

The chemical properties of deoxynucleotides reveal the intrinsic relationship between the four different deoxy nucleotides in the sequence and represent the sequence information as discrete vectors by means of 0–1 coding. Therefore, by this method, we represented the sequence with $4 \times L$-D (dimensional) feature vector $\mathcal{W}$ to represent the sample sequence formulated by Equation (1),

$$\mathcal{W} = \left[x_1 y_1 z_1 \lambda_1 \cdots x_{2\delta+1} \ y_{2\delta+1} \ z_{2\delta+1} \ \lambda_{2\delta+1}\right]^T \quad (9)$$

where the symbol T is the transpose operator.

Feature Fusion

Feature fusion usually joins several kinds of different feature vectors into an integrated one, which could express the local and global sequence order information of the given sample sequence. Therefore, in this study, we not only employed one-hot and DPF methods, but also took into account their combination. According with this method, we represented the sequence with $2 \times 4 \times L$-D (dimensional) feature vector.

Framework of the Integrated Predictor

For imbalance problems existing in positive samples and negative samples, the down-sampling method was adopted in the current study. We randomly divided the negative samples from the training dataset S_1 into 11 groups of equal size, one of which can form the balance training subset by combining with the positive samples in the same amount. And then, we could obtain 11 sub-models. After converting into a numeric vector by one-hot, DPF or their combination, a query sequence with the base C in its center, can be input into 11 sub-models for prediction. The 11 prediction results thus obtained can be used to generate the final decision whether the base C is methylated or not by some judging methods, just like a simple majority vote or weighted voting method (**Figure 2**). The integrated predictor obtained by above-mentioned method was named as iPromoter-5mC, which can be used to identify the 5mC modification sites in promoter sequences.

In this study, a simple deep neural network (DNN) framework (Islam et al., 2018) was employed to consturcted the prediction model. The generated feature matrix was fed into the fully connected neural network for training. The fully connected layer of DNN contained 64, 128, 256, 128, 64 neurons in turn, and the activation function was ReLU (Zhuang et al., 2019). For binary problem, the last layer contained two neurons, and sigmoid

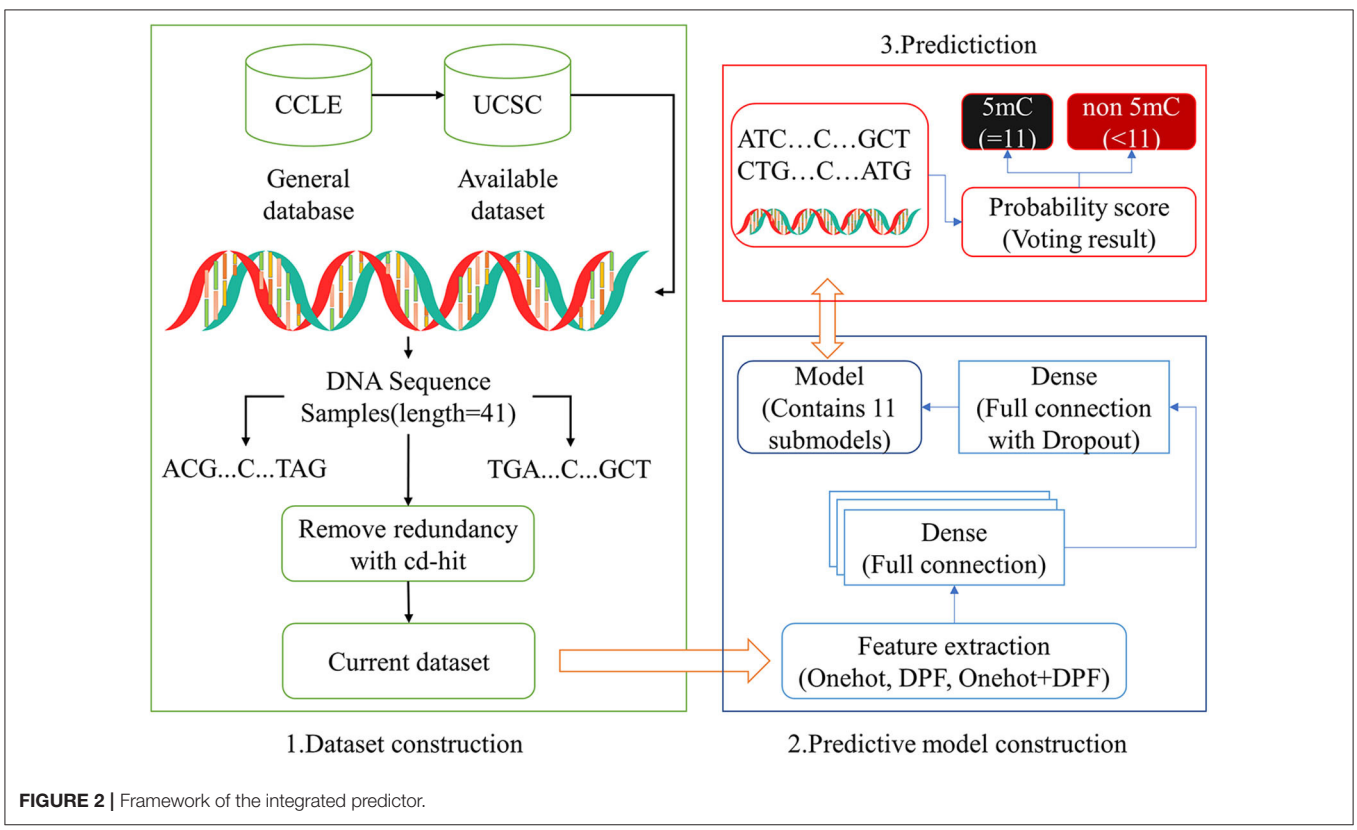

FIGURE 2 | Framework of the integrated predictor.

was selected as the activation function. To prevent overfitting and improve model generalization, a dropout layer was added before the last full connection layer, with a value of 0.3. Five-fold cross validation was conducted to validate the reliability of each sub-model.

Evaluation Metrics

K-fold cross-validation method could effectively utilize limited data, and the evaluation results are as close as possible to the model's performance on the test set. Therefore, we used this method to evaluate the model's performance (Wei et al., 2018; Chen et al., 2019a,b; Dao et al., 2019). For single label system, there are several common evaluation indexes to measure the predictive performance of the predictor, including Sensitivity (Sn), Accuracy (Acc), Specificity (Sp) and Matthew's correlation coefficient (MCC), which can be defined as following,

$$\begin{cases} Sn = 1 - \frac{N_-^+}{N_+}, 0 \le Sn \le 1 \\ Sp = 1 - \frac{N_+^-}{N}, 0 \le Sp \le 1 \\ Acc = 1 - \frac{N_-^+ + N_+^-}{N^+ + N}, 0 \le Acc \le 1 \\ MCC = \frac{1 - \left(\frac{N_-^+ + N_+^-}{N^+ + N}\right)}{\sqrt{\left(1 + \frac{N_+^- N^+}{N^+}\right)\left(1 + \frac{N_-^+ N_+^-}{N}\right)}}, 0 \le MCC \le 1 \end{cases} \quad (10)$$

where N^+ is the total number of 5mC sites actually containing in the sample sequences, i.e., the sum of the quantities of true positive; while N^- denotes the total number of non-5mC site sequences, i.e., the sum of the quantities of true negative; N_-^+ represents the number of true 5mC sites predicted incorrectly as non-5mC sites; N_+^- represents the number of non-5mC sites predicted incorrectly as true 5mC sites.

In addition, we used the Receiver Operating characteristic curve (ROC curve) to exam the performance of the entire integrated predictor model. The true positive rate (Sn) and false positive rate (1-Sp) were set to x-axis and y-axis to plot the ROC curve, respectively. The area under the ROC curve, also known as AUC, was used to quantify the performance of the model.

RESULTS AND CONCLUSIONS

Window Size Analysis

Considering the position specific deoxynucleotide bias, it is necessary to determine the optimal window size δ of sample sequences for identifying 5mC modification sites. Generally, if δ is too small, the residues around the 5mC modification sites cannot carry enough information, leading to poor prediction effect (Xu et al., 2019). Thus, we analyzed the trend of the precision rate of the constructed model with different window size δ. As shown in **Figure 3**, the search step size for δ here was 1nt, with a range of 10–20.

According to the intuitive observation in sub-graphs (A), (B), and (D), when $\delta = 20$, the prediction results generated by the three different methods were the best. In order to distinguish the optimal model obtained by using one-hot, DPF and onehot-DPF, we compared the most important metrics Acc and MCC values, and found that the feature method with the best effect was one-hot, as illustrated in sub-graphs (A) and (D). Therefore,

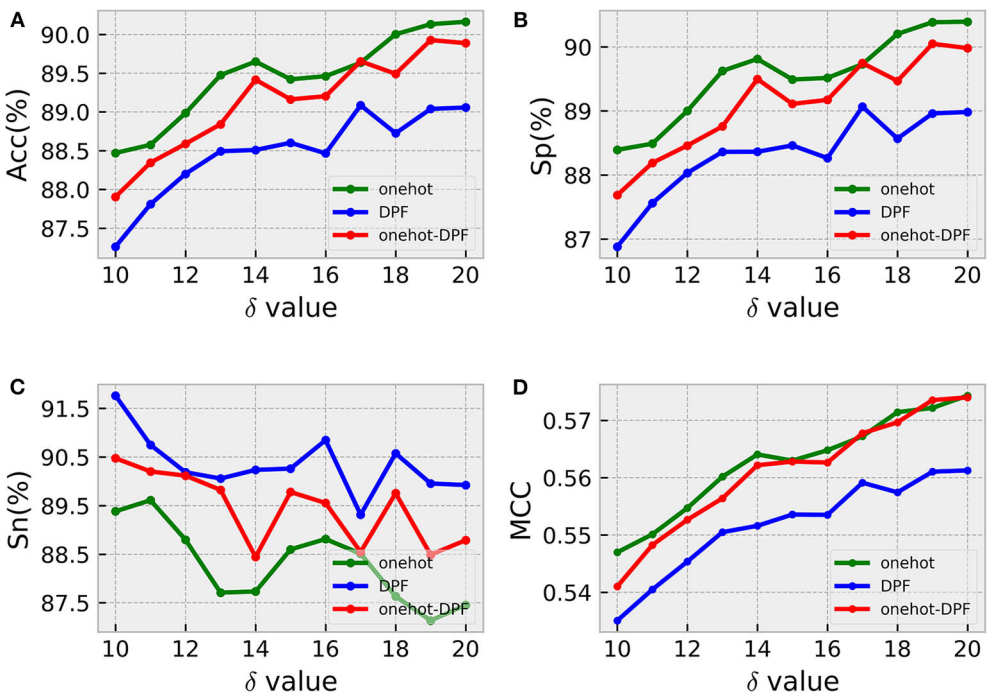

FIGURE 3 | Windows size analysis. Sub-graphs from **(A–D)** represent the ACC, Sp, Sn, MCC values generated by three different feature coding methods under different sliding window sizes, respectively.

the following analysis and calculation were based on δ with 20, indicating the length of the sample sequence formulated by Equation (1) was 41nt.

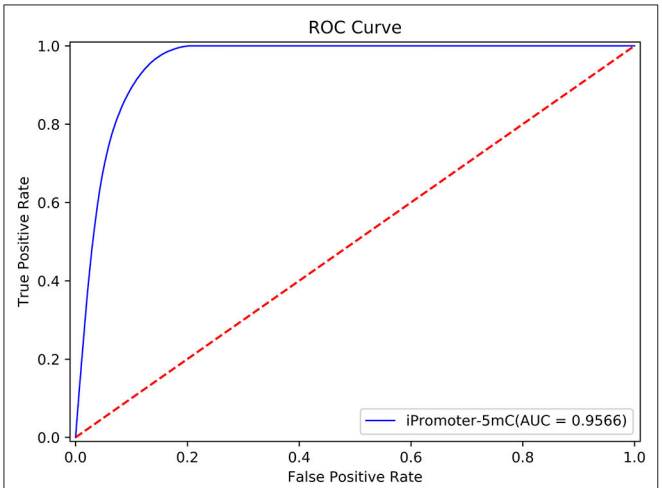

FIGURE 4 | The ROC curve of the S1 dataset on our model.

TABLE 2 | The results obtained by 5-fold cross validation on the training dataset S_1.

Method	Sn (%)	Sp (%)	Acc (%)	MCC
iPromoter-5mC	87.46	90.39	90.16	0.5743

Performance of DNN Models

According to the description in section "Framework of the integrated predictor," we can construct the 11 sub-models based on the training dataset S_1 using one-hot feature extraction method. A simple majority vote strategy was used to integrate all the decisions originated from the 11 sub-models into a final classification result. In the current study, we adopted the strict discriminating standard for identifying 5mC modification sites. If only all the sub-models consider that the potential 5mC sites is a true 5mC modification site, the iPromoter-5mC model could infer the center of this query sequence is a 5mC modification site. After 30 repeated experiments with 5-fold cross validation, we obtained the average values of each metric as the final results of the iPromoter-5mC model, as shown in **Table 2**. The results of the iPromoter-5mC model indicated that the performance of our models was promising, supported by the metric values, such as Sn, 87.46%; Sp, 90.39%; Acc, 90.16%; MCC, 0.5743. To more directly illustrate the performance of the predictor, a ROC curve was plotted using the training dataset S_1, and its corresponding AUC value was calculated. The high AUC value (0.9566) indicates that our predictor iPromoter-5mC has excellent performance and stable performance in predicting the 5mC site (**Figure 4**).

In order to validate the stability of the DNN algorithm model, we compared the performance of the DNN models constructed by one-hot, DPF, and their combination. All the results were displayed as a histogram directly on **Figure 5**. Small discrepancies of every metric value obtained by the three different methods indicated the superior stability of the DNN algorithm model to identify the 5mC modification sites.

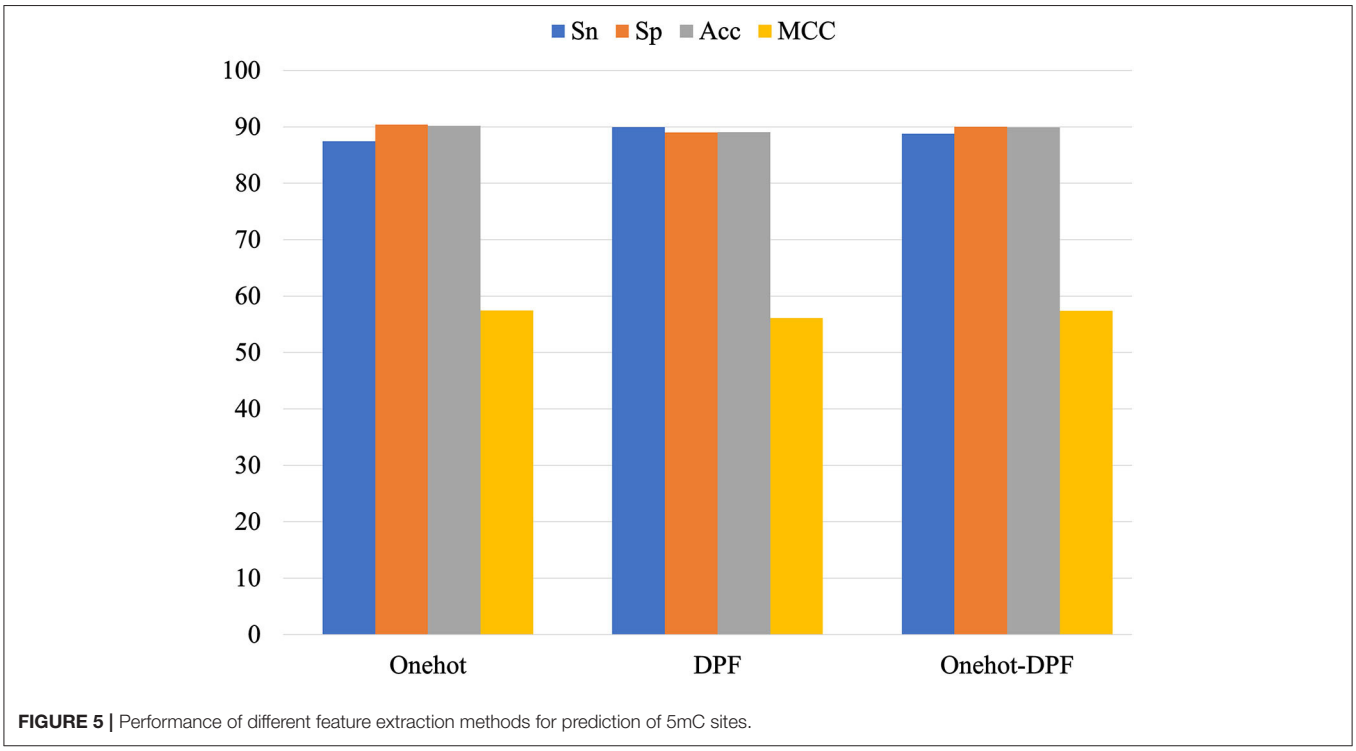

FIGURE 5 | Performance of different feature extraction methods for prediction of 5mC sites.

TABLE 3 | The performance of iPromoter-5mC based on the independent dataset.

Model number	Sn (%)	Sp (%)	Acc (%)	MCC	AUC
1	94.48	86.53	87.15	0.5455	0.9543
2	98.32	83.19	84.37	0.5183	0.9542
3	95.88	85.77	86.56	0.5417	0.9545
4	96.97	84.71	85.66	0.5319	0.9533
5	95.49	85.97	86.72	0.5425	0.9539
6	95.59	85.88	86.64	0.5417	0.9542
7	97.84	83.84	84.93	0.5244	0.9531
8	97.94	83.75	84.86	0.5238	0.9535
9	94.24	86.71	87.29	0.5469	0.9539
10	95.98	85.69	86.49	0.5409	0.9542
11	97.53	84.04	85.09	0.5256	0.9545
iPromoter-5mC	87.77	90.42	90.22	0.5771	0.9570

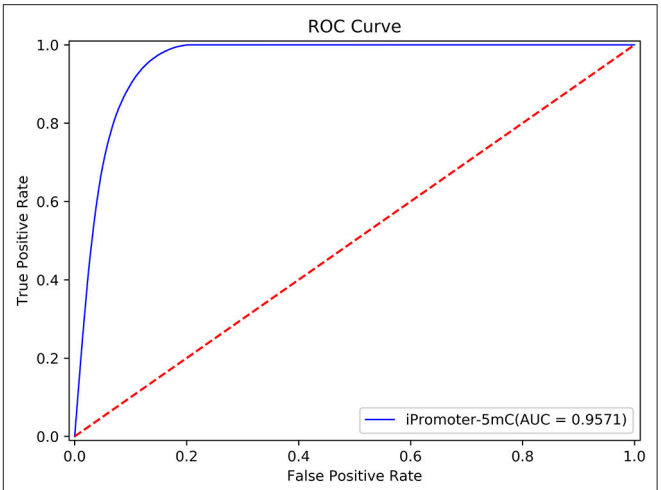

FIGURE 6 | The performance generated by the same prediction framework was still reliable and stable after the expansion of the training data.

The Robustness and Reliability Analysis

Independent test is an effective approach to check the performance of the constructed classification model. Compared with the cross-validation method, it can better verify the robustness and reliability of the prediction models. In the section "Benchmark datasets" in this study, we established the training dataset S_1 and independent testing dataset S_2. Here, we used the independent testing dataset S_2 to further test the performance of the predictor iPromoter-5mC. The results were listed in **Table 3**.

The predictive results of the 11 sub-models using the 5-fold cross-validation method on the independent test dataset S_2 were very stable at about 95, 83, 85%, 0.52 and 0.95 in Sn, Sp, Acc, MCC, and AUC, respectively, indicating that the constructed sub-models are very robust for identifying 5mC modification sites on new data. After integrating all the decisions originated from these sub-models, the independent test performance of this final model were 87.77, 90.42, 90.22%, 0.5771 and 0.9570 in Sn, Sp, Acc, MCC and AUC, respectively. The performance of the predictor iPromoter-5mC was improved, mainly seen in the metrics Acc and MCC. This implied that our designed framework for 5mC modification site prediction is reasonable and efficient, indicating that this method can be extended to realize synthetic problems on accurate prediction of other DNA/RNA modification sites.

To further validate the robustness and reliability of the prediction framework, we implemented 5-fold cross validation on the benchmark dataset S_δ including the training dataset S_1 and the independent test dataset S_2. The results of the ROC curve shown in **Figure 6** showed that the performance generated by the same prediction framework was still reliable and stable after the expansion of the training data, which have laid a solid foundation for establishment of online predictor.

We are also concerned with whether our models are applicable to the data from other cell line or tissues. To do so, we firstly constructed the benchmark dataset according to the 5mC site information in promoter regions of human hepatocarcinoma cell lines (HUH7_LIVER) from database CCLE. This dataset also was divided into the training dataset and the independent test dataset, which were also released on the GitHub and on our online server. And then, we constructed the DNN model using the same method proposed in this study. The results listed in **Table 4** were also promising, indicating that the method using in this study can also be applied to the prediction of 5mC sites in other cancer cell lines.

TABLE 4 | The 5-fold cross validation results on the training set and the independent test set of human hepatocarcinoma cell lines.

Method	Sn (%)	Sp (%)	Acc (%)	MCC	AUC
iPromoter-5mC (training)	80.53	95.79	93.73	0.7408	0.9736
iPromoter-5mC (independent test)	81.22	95.79	93.81	0.7459	0.9735

Comparison With Existing Predictor

Compared with the two early predictors Methylator and MethCGI, the predictor iDNA-Methyl has better prediction performance, which has been demonstrated in the study (Liu et al., 2015). And iDNA-Methyl has own webserver for identifying DNA 5mC sites. Therefore, we compared the performance of iPromoter-5mC with those of iDNA-Methyl. For convenience of comparison, the scores of the four indexes defined in Equation 10 obtained by these two predictors based on the independent test dataset S_2 were listed in **Table 5**. It can be observed from the table that the overall accuracy (Acc) score obtained by the current iPromoter-5mC is significantly higher than that of the existing predictors, as are the other three indicators.

We analyzed its causes and presently summarized as follows: (1) There is the biggest difference between iDNA-Methl and iPromoter-5mC. From the view of the function, iDNA-Methl detected the genome-wide methylation while iPromoter-5mC identified the methylation sites in promoters. (2) Most

TABLE 5 | Comparison of predictors' performance on the independent testing dataset S_2 and sample data from iDNA-Methyl by 5-fold cross validation, respectively.

Success rates	Dataset S2		Sample data from iDNA-Methyl	
	iPromoter-5mC	iDNA-Methyl	iPromoter-5mC	iDNA-Methyl
Sn (%)	87.77	30.62	83.48	61.25
Sp (%)	90.42	90.30	88.04	90.33
Acc (%)	90.22	85.90	86.56	77.49
MCC	0.5771	0.1730	0.7013	0.5471

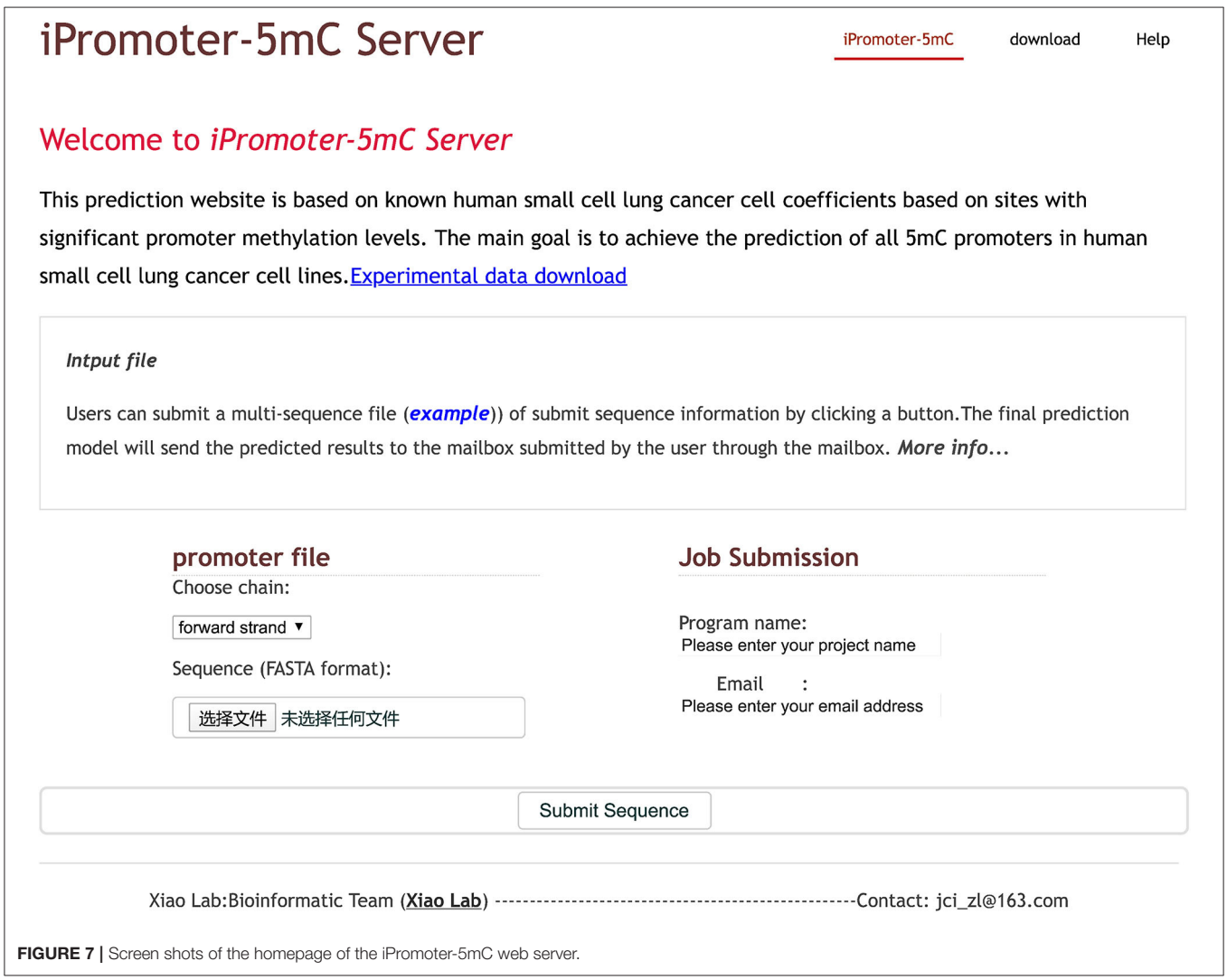

FIGURE 7 | Screen shots of the homepage of the iPromoter-5mC web server.

importantly, the sizes of their benchmark dataset are significantly different. The sample size of iPromoter-5mC is far greater than iDNA-Methl's, which enables our model to obtain better correlation between sequences, causing the phenomenon that the server iPromoter-5mC can identify the 5mC sites of the benchmark dataset from iDNA-Methl effectively while iDNA-Methl cannot. (3) The other reason is that the non-equilibrium degree of the benchmark datasets from these two predictors is significantly different. The unbalance ratio of the positive samples and negative samples from iDNA-Methyl is about 1:2, however, that of the iPromoter-5mC approximately up to 1:11.

In order to further analyze the performance of these two predictors, we implemented experiments to obtain the result by iPromoter-5mC using the sample data from iDNA-Methyl. And we found that the performance of iPromoter-5mC was better than that of iDNA-Methyl (**Table 5**), which also benefits from a large amount of data during our training.

In conclusion, these results indicated that deep learning was better suited for identify 5mC sites on a large dataset, compared to SVM. In fact, parameter optimization of SVM is extremely time-consuming, especially in the case of large amount of data. The predictor iPromoter-5mC can be an outstanding supplemental tool for identifying 5mC sites since the predictor iDNA-Methyl.

Web-Server

A user-friendly web server could provide ease of use for broad scholars to get their desired predictive results without following the complex mathematical calculations. To achieve this, we had developed an online predictor called iPromoter-5mC to identify the 5mC modification sites in promoters, following the principle described below.

For a given promoter sequence, a 41 bp scan window was used to segment the sequence into equal-size sequences. If a DNA query sequence containing potential 5mC modifications sites is in a forward strand, the base C in this DNA sequence will be selected and considered as the fixed length sequence with 41, otherwise, the base G will be found to construct the input sequence, and then be converted to the reverse complementary sequence. After that, users can follow the detailed guide to try out online experience of our web server iPromoter-5mC.

Step 1. Click the link http://www.jci-bioinfo.cn/iPromoter-5mC and then the top page of iPromoter-5mC will be shown in **Figure 7**.

Step 2. Select the strand where the sequence is located from the drop-down list box (the default value is the forward strand).

Step 3. Users can submit the file containing multiple sequences in FASTA format by clicking the submit button.

Step 4. Enter the project name and your e-mail address. The running results will be sent to you by email after finishing the work.

CONCLUSIONS

In this study, we designed a fast and effective DNN model, named iPromoter-5mC, to identify 5mC modification sites in DNA promoter region in cell lines of the small cell lung cancer. The robustness and good performance of the model were verified by feature analysis and various experiments. More importantly, Due to build an easy to use web server can provide users with more convenient, we set up an online web server to identify 5mC modification sites, which can bring great convenience to scholars' research work. The model mentioned in this paper only targets cell lines of lung small cell carcinoma, but the basic method and analysis flow can also be applied to the prediction of 5mC sites of other cancer cell lines.

Although the model in this study achieved higher predictive performance, the future is going to be one that presents many challenges. We are going to continue to study the predictive problem about DNA 5mC methylation. Firstly, with the development of single cell sequencing technology, we will try to accurately predict single-cell DNA 5mC methylation states using deep learning based on single-cell methylation data. Secondly, we plan to design a scheme to achieve accurate classification of DNA 5mC methylation level. Finally, we will construct machine learning models based on other data in cell lines of other cancers to better detect the biomarkers of those cancers.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: http://www.jci-bioinfo.cn/iPromoter-5mC/download.

AUTHOR CONTRIBUTIONS

XX designed the experiments. LZ constructed the predictor and established the online server. Z-CX wrote the manuscript. All authors read and approved the manuscript. In additional, thank Ang Sun for collecting the data information.

FUNDING

This work was partially supported by the National Nature Science Foundation of China (Nos. 31860312, 31760315, 61300139, 61761023), Natural Science Foundation of Jiangxi Province, China (Nos. 20171ACB20023, 20171BAB202020), the Department of Education of Jiangxi Province (GJJ160866, GJJ180733, GJJ180703), China Postdoctoral Science Foundation Funded Project (Project No. 2017M612949), Jingdezhen technology office program (20192GYZD008-04), Jiangxi province graduate student innovation special fund (YC2019-S388). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

REFERENCES

Amoreira, C. (2003). An improved version of the DNA methylation database (MethDB). *Nucl. Acids Res.* 31, 75–77. doi: 10.1093/nar/gkg093

Angermueller, C., Lee, H. J., Reik, W., and Stegle, O. (2017). DeepCpG: accurate prediction of single-cell DNA methylation states using deep learning. *Genome Biol.* 18:67. doi: 10.1186/s13059-017-1233-z

Barretina, J., Caponigro, G., Stransky, N., Venkatesan, K., Margolin, A. A., Kim, S., et al. (2012). The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature* 483, 603–607. doi: 10.1038/nature11003

Bhasin, M., Zhang, H., Reinherz, E. L., and Reche, P. A. (2005). Prediction of methylated CpGs in DNA sequences using a support vector machine. *FEBS Lett.* 579, 4302–4308. doi: 10.1016/j.febslet.2005.07.002

Bianchi, C., and Zangi, R. (2015). Molecular dynamics study of the recognition of dimethylated CpG sites by MBD1 protein. *J. Chem. Inf. Model.* 55, 636–644. doi: 10.1021/ci500657d

Bird, A. (2007). Perceptions of epigenetics. *Nature* 447, 396–398. doi: 10.1038/nature05913

Chen, W., Feng, P., Song, X., Lv, H., and Lin, H. (2019a). iRNA-m7G: identifying N(7)-methylguanosine sites by fusing multiple features. *Mol. Ther. Nucl. Acids* 18, 269–274. doi: 10.1016/j.omtn.2019.08.022

Chen, W., Song, X., Lv, H., and Lin, H. (2019b). iRNA-m2G: identifying N(2)-methylguanosine sites based on sequence-derived information. *Mol. Ther. Nucl. Acids* 18, 253–258. doi: 10.1016/j.omtn.2019.08.023

Chen, W., Yang, H., Feng, P., Ding, H., and Lin, H. (2017). iDNA4mC: identifying DNA N4-methylcytosine sites based on nucleotide chemical properties. *Bioinformatics* 33, 3518–3523. doi: 10.1093/bioinformatics/btx479

Dao, F. Y., Lv, H., Wang, F., Feng, C. Q., Ding, H., Chen, W., et al. (2019). Identify origin of replication in *Saccharomyces cerevisiae* using two-step feature selection technique. *Bioinformatics* 35, 2075–2083. doi: 10.1093/bioinformatics/bty943

Deichmann, U. (2016). Epigenetics: the origins and evolution of a fashionable topic. *Dev. Biol.* 416, 249–254. doi: 10.1016/j.ydbio.2016.06.005

Down, T. A., Rakyan, V. K., Turner, D. J., Flicek, P., Li, H., Kulesha, E., et al. (2008). A Bayesian deconvolution strategy for immunoprecipitation-based DNA methylome analysis. *Nat. Biotechnol.* 26, 779–785. doi: 10.1038/nbt1414

Fang, F., Fan, S., Zhang, X., and Zhang, M. Q. (2006). Predicting methylation status of CpG islands in the human brain. *Bioinformatics* 22, 2204–2209. doi: 10.1093/bioinformatics/btl377

Gessler, M. (1999). WT1 (Wilms' tumor suppressor gene). *Atlas Genet. Cytogenet. Oncol. Haematol.* 3, 177–178. doi: 10.4267/2042/37552

Ghandi, M., Huang, F. W., Jane-Valbuena, J., Kryukov, G. V., Lo, C. C., Mcdonald, E. R. III., et al. (2019). Next-generation characterization of the Cancer Cell Line Encyclopedia. *Nature* 569, 503–508. doi: 10.1038/s41586-019-1186-3

Islam, M. M., Tian, Y., Cheng, Y., Wang, Y., and Hu, P. (2018). A deep neural network based regression model for triglyceride concentrations prediction using epigenome-wide DNA methylation profiles. *BMC Proc.* 12:21. doi: 10.1186/s12919-018-0121-1

Kang, Y. H., Lee, H. S., and Kim, W. H. (2002). Promoter methylation and silencing of PTEN in gastric carcinoma. *Lab. Invest.* 82, 285–291. doi: 10.1038/labinvest.3780422

Li, H., Ning, S., Ghandi, M., Kryukov, G. V., Gopal, S., Deik, A., et al. (2019). The landscape of cancer cell line metabolism. *Nat. Med.* 25, 850–860. doi: 10.1038/s41591-019-0404-8

Liu, Q., Georgieva, D. C., Egli, D., and Wang, K. (2018). NanoMod: a computational tool to detect DNA modifications using nanopore long-read sequencing data. *BMC Genomics* 20(Suppl. 1):78. doi: 10.1101/277178

Liu, Z., Xiao, X., Qiu, W. R., and Chou, K. C. (2015). iDNA-Methyl: identifying DNA methylation sites via pseudo trinucleotide composition. *Anal. Biochem.* 474, 69–77. doi: 10.1016/j.ab.2014.12.009

Mansour, H. (2014). Cell-free nucleic acids as noninvasive biomarkers for colorectal cancer detection. *Front. Genet.* 5:182. doi: 10.3389/fgene.2014.00182

Michalak, E. M., Burr, M. L., Bannister, A. J., and Dawson, M. A. (2019). The roles of DNA, RNA and histone methylation in ageing and cancer. *Nat. Rev. Mol. Cell Biol.* 20, 573–589. doi: 10.1038/s41580-019-0143-1

Muller, F., Scherer, M., Assenov, Y., Lutsik, P., Walter, J., Lengauer, T., et al. (2019). RnBeads 2.0: comprehensive analysis of DNA methylation data. *Genome Biol.* 20:55. doi: 10.1186/s13059-019-1664-9

Nicoglou, A., and Merlin, F. (2017). Epigenetics: a way to bridge the gap between biological fields. *Stud. Hist. Philos. Biol. Biomed. Sci.* 66, 73–82. doi: 10.1016/j.shpsc.2017.10.002

Siegel, R. L., Miller, K. D., and Jemal, A. (2018). Cancer statistics, 2018. *CA Cancer J. Clin.* 68, 7–30. doi: 10.3322/caac.21442

Wei, L., Chen, H., and Su, R. (2018). M6APred-EL: a sequence-based predictor for identifying N6-methyladenosine sites using ensemble learning. *Mol. Ther. Nucl. Acids* 12, 635–644. doi: 10.1016/j.omtn.2018.07.004

Wei, L., Su, R., Luan, S., Liao, Z., Manavalan, B., Zou, Q., et al. (2019). Iterative feature representations improve N4-methylcytosine site prediction. *Bioinformatics* 35, 4930–4937. doi: 10.1093/bioinformatics/btz408

Xia, C., Xiao, Y., Wu, J., Zhao, X., and Li, H. (2019). "A convolutional neural network based ensemble method for cancer prediction using dna methylation data," in *Proceedings of the 2019 11th International Conference on Machine Learning and Computing - ICMLC '19* (Zhuhai), 191–196. doi: 10.1145/3318299.3318372

Xu, Z. C., Feng, P. M., Yang, H., Qiu, W. R., Chen, W., and Lin, H. (2019). iRNAD: a computational tool for identifying D modification sites in RNA sequence. *Bioinformatics* 35, 4922–4929. doi: 10.1093/bioinformatics/btz358

Zhuang, Z., Shen, X., and Pan, W. (2019). A simple convolutional neural network for prediction of enhancer-promoter interactions with DNA sequence data. *Bioinformatics* 35, 2899–2906. doi: 10.1093/bioinformatics/bty1050

Zou, Q., Zeng, J., Cao, L., and Ji, R. (2016). A novel features ranking metric with application to scalable visual and bioinformatics data classification. *Neurocomputing* 173, 346–354. doi: 10.1016/j.neucom.2014.12.123

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.