frontiers | Frontiers in Microbiology

TYPE Original Research
PUBLISHED 13 April 2023
DOI 10.3389/fmicb.2023.1170785

OPEN ACCESS

EDITED BY
Ettayapuram Ramaprasad Azhagiya Singam,
University of California, Berkeley, United States

REVIEWED BY
Vijaya Sundar Jeyaraj,
University of Illinois at Urbana-Champaign, United States
Liang Cheng,
Harbin Medical University, China
Hao Wu,
School of Software, Shandong University, China

*CORRESPONDENCE
Hasan Zulfiqar
✉ hasanzulfiqar@uestc.edu.cn
Zhao-Yue Zhang
✉ zyzhang@uestc.edu.cn
Fen Liu
✉ nmlf906@163.com

SPECIALTY SECTION
This article was submitted to Evolutionary and Genomic Microbiology, a section of the journal Frontiers in Microbiology

RECEIVED 21 February 2023
ACCEPTED 17 March 2023
PUBLISHED 13 April 2023

CITATION
Zulfiqar H, Ahmed Z, Kissanga Grace-Mercure B, Hassan F, Zhang Z-Y and Liu F (2023) Computational prediction of promotors in *Agrobacterium tumefaciens* strain C58 by using the machine learning technique. *Front. Microbiol.* 14:1170785. doi: 10.3389/fmicb.2023.1170785

Computational prediction of promotors in *Agrobacterium tumefaciens* strain C58 by using the machine learning technique

Hasan Zulfiqar[1,2]*, Zahoor Ahmed[1], Bakanina Kissanga Grace-Mercure[2], Farwa Hassan[2], Zhao-Yue Zhang[2]* and Fen Liu[3]*

[1]Yangtze Delta Region Institute (Huzhou), University of Electronic Science and Technology of China, Huzhou, China, [2]School of Life Science and Technology and Center for Informational Biology, University of Electronic Science and Technology of China, Chengdu, China, [3]Department of Radiation Oncology, Peking University Cancer Hospital (Inner Mongolia Campus), Affiliated Cancer Hospital of Inner Mongolia Medical University, Inner Mongolia Cancer Hospital, Hohhot, China

Promotors are those genomic regions on the upstream of genes, which are bound by RNA polymerase for starting gene transcription. Because it is the most critical element of gene expression, the recognition of promoters is crucial to understand the regulation of gene expression. This study aimed to develop a machine learning-based model to predict promotors in *Agrobacterium tumefaciens* (*A. tumefaciens*) strain C58. In the model, promotor sequences were encoded by three different kinds of feature descriptors, namely, accumulated nucleotide frequency, *k*-mer nucleotide composition, and binary encodings. The obtained features were optimized by using correlation and the mRMR-based algorithm. These optimized features were inputted into a random forest (RF) classifier to discriminate promotor sequences from non-promotor sequences in *A. tumefaciens* strain C58. The examination of 10-fold cross-validation showed that the proposed model could yield an overall accuracy of 0.837. This model will provide help for the study of promoters in *A. tumefaciens* C58 strain.

KEYWORDS

prokaryotic promotors, feature extraction, *agrobacterium tumefaciens* strain C58, feature selection, algorithms

1. Introduction

Agrobacterium belongs to the family of ubiquitous gram-negative soil bacteria. Infectious strains of agrobacterium such as agrobacterium tumefaciens strain C58 cause hairy root and crown gall diseases in plants (Goodner et al., 2001). Promotors are the genomic regions upstream of a gene on DNA where transcription factor and RNA polymerase bind together to initiate gene transcription (Sawadogo and Roeder, 1985; Zhao et al., 2017; Zhang et al., 2018). The biological process of prokaryotic promotors is shown in Figure 1. The study of promoters is the first step to understanding gene expression.

Correct identification of the promotor sequence could produce vital signs for understanding its mechanism of the regulation (Cao et al., 2022; Li et al., 2022b). Currently, numerous tentative techniques, such as mass spectrometry (Flusberg et al., 2010), reduced-representation bisulfite sequencing (Doherty and Couldrey, 2014), and single-molecule real-time sequencing (Boch and Bonas, 2010), have been developed. Though these procedures are quite helpful in the identification of promotors prediction, they are

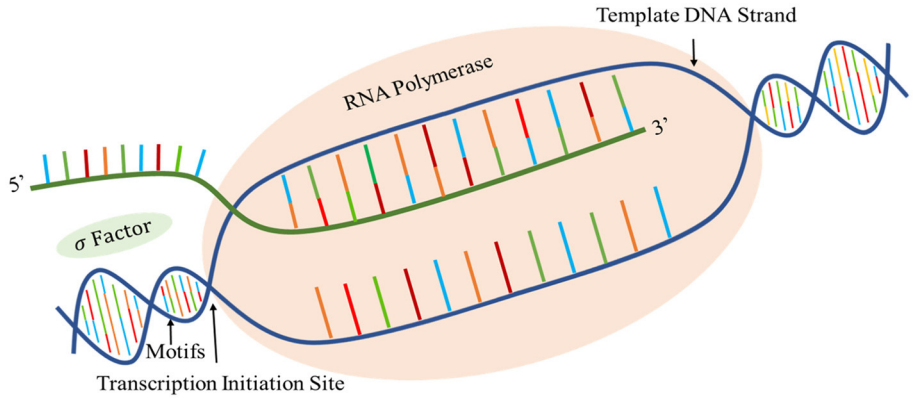

FIGURE 1
Schematic diagram of the prokaryotic promotor structure and its biological processes.

costly when applied to large sequencing data. Thus, a bioinformatics tool to recognize the promotor sequence is urgently needed. At present, some computational tools have been presented to recognize promotors in multiple species, such as PePPer (de Jong et al., 2012) for *Escherichia coli* (*E.coli*) and *Bacillus subtilis* (*B.subtilis*); Promotech for *Bacillus amyloliquefaciens* (*B. amyloliquefaciens*) XH_7 bacterium (Chevez-Guardado and Peña-Castillo, 2021); DeePromotors (Oubounyt et al., 2019) for TATA promotors (Zou et al., 2016) in eukaryotic genomes; iProEP (Lai et al., 2019) for *Homo sapiens* (*H. sapiens*), *Drosophila melanogaster* (*D. melanogaster*), *Caenorhabditis elegans* (*C. elegans*), *B. subtilis*, and *E. coli*; and iPromotor-2L (Liu et al., 2018) for bacterial promotors. However, there is no such model for *A. tumefaciens* C58 strain. To address the above-mentioned problems, we designed an RF-based model to predict promotor sequences in agrobacterium tumefaciens strain C58. Figure 2 illustrates the workflow of the projected model.

Accumulated nucleotide frequency, binary encodings, and *k*-mer nucleotide composition were utilized to convert sequences into numerical features, and then these features were optimized by using correlation and the mRMR-based feature selection algorithm. After this, these optimized features were inputted into a random forest classifier for the identification of promotor sequences on the basis of 10-fold cross-validation. As a result, an ideal model was attained.

2. Materials and methods

A precise and accurate dataset is necessary to establish a prediction model (Liang et al., 2017; Ning et al., 2021a,b; Su et al., 2021). Therefore, we obtained the experimentally verified *Agrobacterium tumefaciens* strain C58 promotors data of 706 sequences from PPD (http://lin-group.cn/database/ppd/index.php) and also collected negative data of 2860 sequences of 81 bp from (http://bioinformatics.hitsz.edu.cn/iPromotor-2L/data). Moreover, we divided the dataset into 80/20 ratios for training and testing the model.

2.1. Feature descriptors

Selecting the feature encodings that are useful and autonomous is a key stage in establishing machine learning-based models (Lv et al., 2021; Zhang D. et al., 2021; Ao et al., 2022a; Li et al., 2022a; Ning et al., 2022; Teng et al., 2022; Wei et al., 2022). Representing the DNA sequences with a mathematical manifestation is very important in functional element identification. Some DNA sequences coding strategies such as accumulated nucleotide frequency, physiochemical properties, binary encodings, nucleotide chemical properties and *k*-tuple nucleotide frequency component, nucleotide pair spectrum encoding, and natural vector have been applied in bioinformatics (Dao et al., 2020; Yang X. et al., 2021; Zhang Y. et al., 2021; Ao et al., 2022b; Ren et al., 2022). The performance of these feature descriptors was good. Here, to extract DNA sequence information as more as possible, accumulated nucleotide frequency, *k*-mer nucleotide composition, and binary encodings were presented to describe the DNA sequences based on their superior performance.

2.1.1. Accumulated nucleotide frequency

The encoding of ANF consists of the distribution and frequency of nucleotides n_i in the sequences. The nucleotide density D_i at any position in the sequence can be calculated as follows:

$$D_i = \frac{1}{|n_i|}\sum_{k=1}^{z} f(n_i), \quad f(g) = \begin{cases} 1 \; if \; n_i = g \\ 0 \; in \; other \; case \end{cases} \tag{1}$$

where z is the sequence length, n_i is the length of the string $\{n_1, n_2, \ldots, n_i\}$ (Li et al., 2022c,d) in the sequence, and $g \in \{A, G, C, T\}$.

2.1.2. k-mer nucleotide composition

k-mer nucleotide composition can reflect short-range nucleotide interaction of sequences (Salimi and Moeini, 2021; Zhang et al., 2022b; Dao et al., 2023). The nucleotide residues can be obtained *via* a sliding window method by setting the window

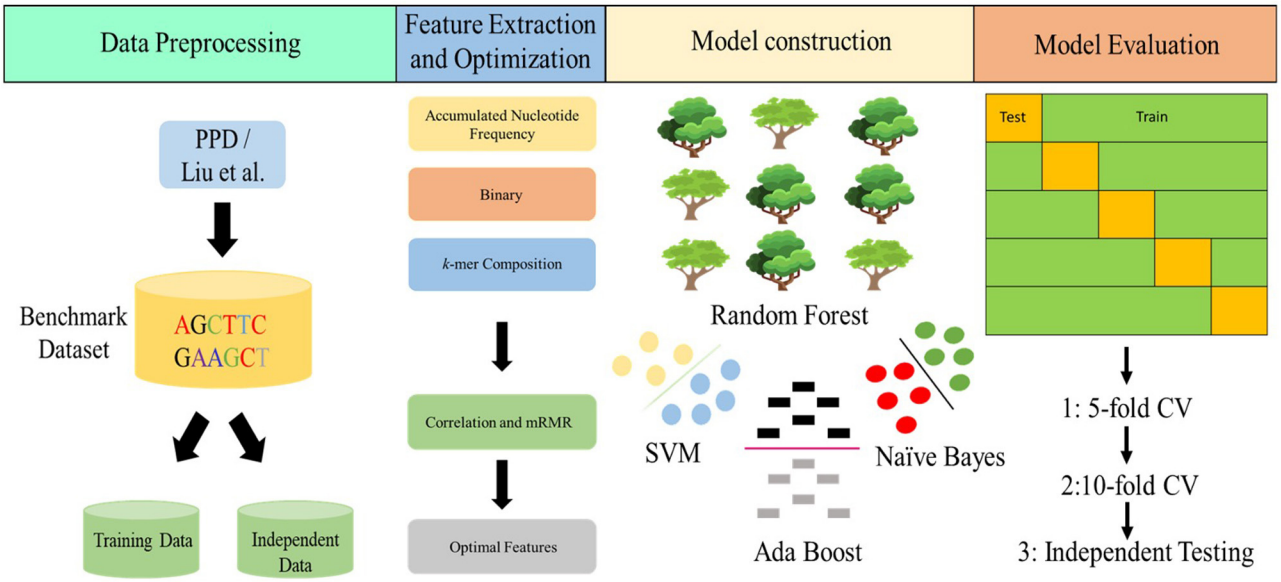

FIGURE 2
Overall workflow of the study.

size of k bp with a step size of 1 bp to examine a sequence with n bp. An arbitrary sample Z with the sequence length of n (where n is 81bp) can be characterized as

$$Z = Q_1\ Q_2\ Q_3 \ldots .. Q_i \ldots .. Q_{(n-1)}\ Q_n \tag{2}$$

where Q_i signifies the nucleotide {A, G, C, T} at the i-th position. The sequences can be transformed into the 4^k D vector using *k-mer* nucleotide composition as follows:

$$Q_k = [p_1^{k-tuple} p_2^{k-tuple} \ldots . p_i^{k-tuple} \ldots .. p_{4^k}^{k-tuple}]^t \tag{3}$$

where t denotes the transposition of the vector, and $p_1^{k-tuple}$ symbolizes the occurrence of the i-th *k-mer* nucleotide composition in the sequence. When $k = 1$, a DNA sample can be decoded into a 4 D vector $Q_1 = [p(A), p(T), p(G), p(C)]^t$. When $k = 2$, the DNA sample can be described by a 16-dimension vector. In this study, the value of k was set as 4 due to the best results. The whole results of k-mer nucleotide composition (k = 1,2,3,4,5,6) on training and independent data are shown in Supplementary Table S1.

2.1.3. Binary encoding

Encoding "0" and "1" can represent any information in the computational work (Zou et al., 2019). Therefore, we can directly convert a DNA sequence into a string of characters, which is consisted of "0" and "1." A = (1,0,0,0), T = (0,1,0,0), G = (0,0,1,0), and C = (0,0,0,1). Thus, a DNA sample of 81 bp length is converted into a 324 (4 × 81) dimension vector in this study.

2.2. Feature selection

2.2.1. Correlation

Feature selection is an important step for improving model performance (Dao et al., 2020). Correlation is a familiar comparison measure between two features. If two features are linearly dependent, then their correlation coefficient will be "±1." If the features are uncorrelated, the correlation coefficient will be "0." There are two comprehensive classes that can be used to measure the correlation between two random variables. One is based on information theory, and the other is classical linear correlation. The most familiar measure is the linear correlation coefficient. The linear correlation coefficient "d" for a pair of (m, n) variables is specified as

$$d = \frac{\sum (m_i - \overline{m}_i)(n_i - \overline{n}_i)}{\sqrt{\sum (m_i - \overline{m}_i)^2}\sqrt{\sum (n_i - \overline{n}_i)^2}} \tag{4}$$

Due to the expansion of the data, the correlation coefficient which is good for a sample may not produce decent outcomes for the whole population. Therefore, it is necessary to determine the significant association between the features, while captivating the whole population. The most commonly used method to examine statistical correlation is the t-test. The procedure used in the projected algorithm is to use the t-test for choosing the most important features from the whole feature set. The formula for calculating the suitable "T" value to test the consequence of a correlation coefficient employs the "T" distribution. The "T" value can be calculated as

$$T = d\sqrt{\frac{i-2}{1-d^2}} \tag{5}$$

where "*i*" is the number of instances and "*d*" is the correlation coefficient for sample data. The significance of the relationship is expressed in probability levels: p (e.g., significant at $p = 0.05$). The degrees of freedom for entering the T-distribution are $i - 2$. If the value of "T" is higher than the threshold value at the 0.05 significant level, then the feature will be significant and selected (Zulfiqar et al., 2022a).

2.2.2. mRMR

mRMR is a very popular feature selection technique, and it has been applied in many bioinformatics and biological applications (He et al., 2020; Zulfiqar et al., 2021b; Su et al., 2023). The compactness functions are described as "*i*" and "*y*," and their corresponding probabilities are $P(i)$ and $P(y)$. The common information between these two functions can be defined as

$$Q_{\min}(f_i, f_y) = \sum_{i \in Q} \sum_{y \in Y} P(fi, fy) \log \frac{P(i,y)}{P(i), P(y)}) \quad (6)$$

If the target is J_i, then calculating the mutual information in relation to the target and can be defined as

$$Q_{\max}(f_i, J_i) = \sum_{fi \in Q} \sum_{Ji \in i} P(fi, Ji) \log \frac{P(fi, Ji)}{P(fi), P(Ji)}) \quad (7)$$

Thus, $mRMR(f_i)$can be calculated as

$$mRMR(f_i) = \frac{Q_{max}(f_i, J_i)}{Q_{min}(f_i, f_y)} \quad (8)$$

2.3. Machine learning classifiers

Naïve Bayes (NB) classifier has been used widely in bioinformatics due to its simplicity (Ye et al., 2021). This classification method totally depends on the Bayes theorems. Ada boost (AB) is another popular machine learning technique. The main idea of AB is to set the classifiers' weights and trained the data in each and every iteration. The support vector machine (SVM) is also very famous and has been used in many bioinformatics and computational biology-related tools (Tao et al., 2020; Ahmed et al., 2022; Manavalan and Patra, 2022; Zou et al., 2022; Bupi et al., 2023; Zulfiqar et al., 2023). It is mostly used to perform binary classification. We implemented these algorithms in Weka version 3. 8.4. by using the default values. RF is a combined knowledge algorithm and is widely used in bioinformatics (Ao et al., 2022c; Zhang et al., 2023). The main idea of this is to combine several weak classifiers and outcomes generated on the basis of voting. The brief description is clearly described by Zulfiqar et al. (2021a). We have used randomized and grid search cross-validations to tune the hyperparameters. We executed this job in the Scikit-learn package version 0.22.2, and its parameters are summarized in Table 1. All experiments were carried out on a Windows operating system with 1.7 GHz intel quad-core i5.

TABLE 1 Best parameters of the proposed model.

Best parameters	
"N-estimators"	80
"Max_depth"	20
"Bootstrap"	True
"Min_samples_leaf"	1
"Min_samples_split"	2

Input: Training data: = H (x_1, x_2, ..., x_k, x_c)
Output: H_{best}
1st Round
1 Start
2 **for** i =1 to k **do**
3 d = calculate correlational coefficient (x_i, x_c)
end
4 let p = 0.05 significant level
5 let ρ = 0 / suppose there is no significant correlation between f_i and f_c
6 **for** i = 1 to k **do**
q = calculate the significance (d, ρ) for x_i / by using the *T-test*
7 **if** T > CV / critical value
8 H_{best} = H_{list}
9 **end**
10 return H_{best}
2nd Round
11 Start
12 By sorting the features
13 **for** each feature *fi* in *Z* **do**
14 By calculating the mutual information in relation to other features as
15 $Q_{\min}(f_i, f_y) = \sum_{i \in Q} \sum_{y \in Y} P(fi, fy) \log \frac{P(i,y)}{P(i), P(y)})$
16 By calculating the mutual information in relation to the target:
17 $Q_{\max}(f_i, J_i) = \sum_{fi \in Q} \sum_{Ji \in i} P(fi, Ji) \log \frac{P(fi, Ji)}{P(fi), P(Ji)})$
18 By calculating the $mRMR(f_i)$ as
19 $mRMR(f_i) = \frac{Q_{\max}(f_i, J_i)}{Q_{\min}(f_i, f_y)}$
20 **end**
21 **for** by sorting the features in descending order
22 By updating the matrix *Z'* with sorted features
23 **end**
24 return *Z'*

Algorithm 1. Correlation and mRMR-based Feature Selection Algorithm.

10.3389/fmicb.2023.1170785

TABLE 2 Performance of models using different classifiers on the training and independent dataset.

	Training dataset							Independent dataset					
Classifier	FS	*k*	Method	Accuracy	Precision	Recall	F1	AUC	Accuracy	Precision	Recall	F1	AUC
AB	256	4	*k*-mer	0.761	0.772	0.761	0.791	0.812	0.775	0.820	0.801	0.798	0.881
	50	4	*k*-mer	0.799	0.802	0.785	0.789	0.856	0.787	0.824	0.799	0.805	0.872
	324		Binary	0.738	0.742	0.756	0.712	0.786	0.700	0.702	0.700	0.730	0.765
	48		Binary	0.745	0.742	0.698	0.789	0.820	0.720	0.732	0.702	0.726	0.789
	82		ANF	0.684	0.645	0.689	0.743	0.731	0.641	0.692	0.688	0.655	0.699
	38		ANF	0.743	0.726	0.775	0.746	0.796	0.696	0.702	0.698	0.710	0.756
	662		Fusion	0.745	0.732	0.785	0.775	0.799	0.720	0.732	0.775	0.745	0.774
	136		Fusion	0.778	0.768	0.792	0.800	0.845	0.738	0.745	0.765	0.725	0.806
SVM	256	4	*k*-mer	0.761	0.802	0.789	0.799	0.865	0.749	0.838	0.761	0.648	0.860
	50	4	*k*-mer	0.796	0.802	0.802	0.812	0.883	0.753	0.748	0.753	0.756	0.832
	324		Binary	0.744	0.747	0.778	0.765	0.792	0.725	0.755	0.760	0.763	0.786
	48		Binary	0.774	0.775	0.732	0.778	0.815	0.748	0.800	0.778	0.769	0.845
	82		ANF	0.666	0.697	0.732	0.705	0.766	0.612	0.623	0.633	0.605	0.699
	38		ANF	0.755	0.768	0.748	0.759	0.820	0.695	0.703	0.713	0.705	0.806
	662		Fusion	0.710	0.722	0.708	0.709	0.745	0.705	0.700	0.700	0.710	0.740
	136		Fusion	0.752	0.759	0.758	0.768	0.801	0.741	0.750	0.770	0.765	0.810
NB	256	4	*k*-mer	0.748	0.780	0.778	0.719	0.823	0.788	0.801	0.799	0.802	0.884
	50	4	*k*-mer	0.802	0.821	0.823	0.827	0.881	0.792	0.778	0.792	0.802	0.878
	324		Binary	0.737	0.775	0.765	0.789	0.794	0.776	0.770	0.778	0.793	0.835
	48		Binary	0.777	0.789	0.759	0.788	0.864	0.782	0.810	0.815	0.816	0.891
	82		ANF	0.675	0.689	0.720	0.696	0.756	0.665	0.685	0.691	0.701	0.741
	38		ANF	0.735	0.741	0.728	0.733	0.770	0.723	0.715	0.705	0.740	0.762
	662		Fusion	0.712	0.754	0.726	0.745	0.768	0.764	0.777	0.756	0.750	0.788
	136		Fusion	0.778	0.802	0.808	0.810	0.880	0.790	0.807	0.803	0.800	0.892
RF	256	4	*k-mer*	0.809	0.830	0.810	0.74	0.861	0.808	0.841	0.811	0.799	0.897
	50	4	*k*-mer	0.837	0.840	0.841	0.801	0.900	0.831	0.842	0.837	0.818	0.900
	324		Binary	0.792	0.632	0.792	0.701	0.842	0.784	0.804	0.808	0.788	0.887
	48		Binary	0.796	0.653	0.801	0.732	0.865	0.806	0.825	0.811	0.806	0.892
	82		ANF	0.791	0.630	0.791	0.702	0.850	0.788	0.803	0.773	0.778	0.878
	38		ANF	0.795	0.642	0.789	0.743	0.866	0.794	0.726	0.792	0.80	0.868
	662		Fusion	0.792	0.630	0.790	0.708	0.822	0.794	0.771	0.790	0.789	0.856
	136		Fusion	0.801	0.786	0.795	0.800	0.881	0.807	0.799	0.820	0.812	0.889

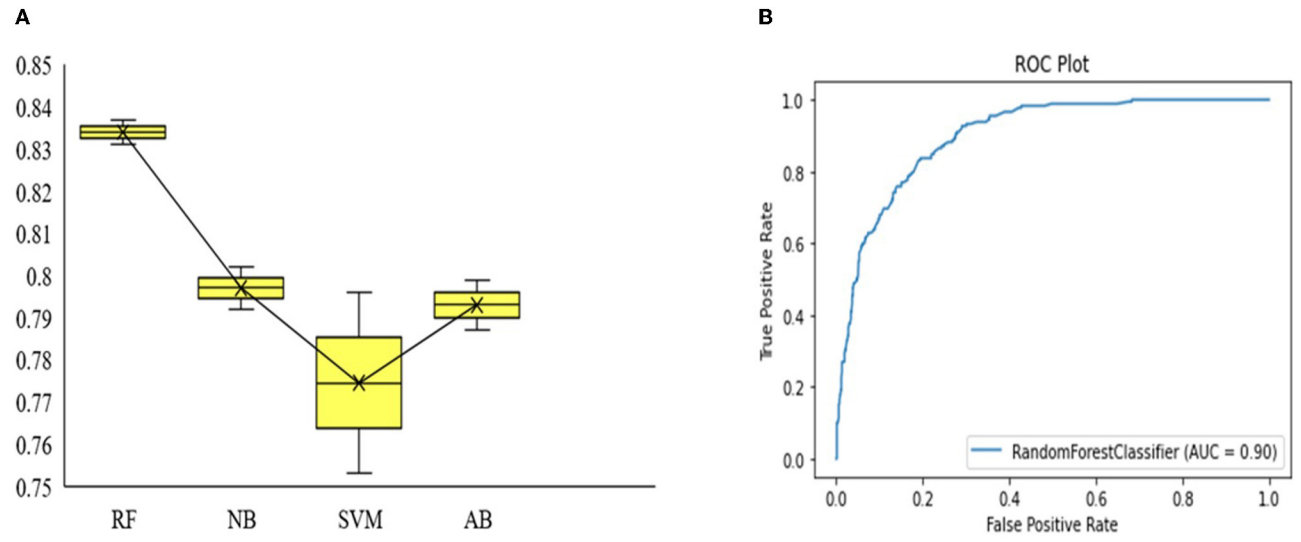

FIGURE 3
Performance comparison on different machine learning classifiers by using training and independent datasets. The higher point represents the training accuracy and the lower point represents the accuracy on independent data **(A).** *AUC* curve of the anticipated model **(B)**.

2.4. Evaluation metrics

Accuracy, precision, recall, and F1 (Hasan et al., 2020; Zhang et al., 2020; Wei et al., 2021b; Shoombuatong et al., 2022; Yang et al., 2022; Zulfiqar et al., 2022b) were employed to assess the performance of the prediction model and are expressed as

$$\begin{cases} Acc = \frac{tp + tn}{tp+fp + tn+fn} \\ Pre = \frac{tp}{tp +fp} \\ Rec = \frac{tp}{tp +fn} \\ F1 = 2 \times \frac{Pre \times Rec}{Pre + Rec} \end{cases} \quad (9)$$

where *tp* symbolizes the correctly predicted promotor sequences and *fp* signifies the non-promotor sequences classified as the promotor sequence. On the other hand, *tn* represents the correctly identified non-promotor sequences, and *fn* demonstrates the promotor sequences, which were classified as the non-promotor sequence.

3. Results and discussion

3.1. Performance evaluation

On the basis of sequence features, we constructed an anticipated model to recognize promotor sequences in *A. tumefaciens* C58 strain. First, the training data were converted into numerical feature vectors using accumulated nucleotide frequency, binary encodings, and *k*-mer nucleotide composition. After this, these features were optimized by using correlation and the mRMR-based algorithm. First, correlation measures and then mRMR were used to select the finest feature subset for the improved prediction outcomes. Afterward, these features were inputted into four machine learning methods. Cross-validation (CV) is a statistical analysis procedure and has been applied in machine learning to evaluate the model's performance (Yang H. et al., 2021; Chen et al., 2022; Liao et al., 2022; Xiao et al., 2022; Zhang et al., 2022a; Yang et al., 2023). In this study, the 10-fold CV test was used to investigate the performance of machine learning methods. In 10-fold CV, the benchmark dataset was randomly separated into ten groups of about equal size. Each group was individually tested by the model which trained with the remaining nine groups. Therefore, the 10-fold CV method was performed 10 times, and the average of the results was the final result (Charoenkwan et al., 2021; Wei et al., 2021a; Hasan et al., 2022). We have trained 32 models on AB, SVM, NB, and RF. At first, we used single encodings and their fusion to train and test the models, and then we optimized the feature encodings and their fusions by using correlation and the mRMR-based algorithm. In this phase, we utilized the *t*-test and picked the significant features by selecting the probability of the significance relation 0.05, and then used mRMR and picked the top features. Moreover, we inputted these features into AB, SVM, NB, and RF and found that the performance of *k*-mer was good as compared to other feature encodings and their fusion. The accuracy of *k*-mer in RF was 3.5%−4.1% higher than the other three classifiers. The *AUC* curve of the anticipated model was 0.900. The *accuracy, precision, recall, and F1* are recorded in Table 2. The performance comparison on different machine learning classifiers by using training and independent datasets and *ROC* plot of the anticipated model is shown in Figures 3A, B.

4. Conclusion

Promotors have a significant role in the transcription process because they are located on upstream of genes where RNA polymerase binds with the transcription factor and initiate the transcription. In this study, an RF model was established to

identify promotors sequences in agrobacterium tumefaciens strain C58. In the proposed model, sequences were encoded using accumulated nucleotide frequency, *k*-mer nucleotide composition, and binary encodings and then optimized with correlation and the mRMR-based algorithm. After this, these optimized features were inputted into the RF-based classifier using the 10-fold CV test and achieved the best model. The estimated outcomes on independent data showed that the projected model provided brilliant performance and oversimplification. We provided the source codes and data freely at https://github.com/linDing-groups/model_promotor. Researchers can yield good results for DNA sequences and recognize their roles by using our freely available source codes. In future, we will further improve the efficiency by using CNN/GNN and release a webserver to make our anticipated model more convenient for users without mathematical and programming knowledge.

Data availability statement

The original contributions presented in the study are included in the article/Supplementary material, further inquiries can be directed to the corresponding authors.

Author contributions

HZ: conceptualization, supervision, methodology, experimentation, visualization, and writing—original draft preparation. ZA and BK: data curation and methodology. FH: data curation. Z-YZ: supervision, methodology, reviewing, and editing. FL: supervision, reviewing, and editing. All authors have read and agreed to the published version of the manuscript.

Funding

This study had been supported by the National Nature Scientific Foundation of China (62102067).

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.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmicb.2023.1170785/full#supplementary-material

References

Ahmed, Z., Zulfiqar, H., Tang, L., and Lin, H. (2022). A statistical analysis of the sequence and structure of thermophilic and non-thermophilic proteins. *Int. J. Mol. Sci.* 23, 10116. doi: 10.3390/ijms231710116

Ao, C., Jiao, S., Wang, Y., Yu, L., and Zou, Q. (2022a). Biological sequence classification: a review on data and general methods. *Research* 2022, 0011. doi: 10.34133/research.0011

Ao, C., Zou, Q., and Yu, L. (2022b). NmRF: identification of multispecies RNA 2'-O-methylation modification sites from RNA sequences. *Brief. Bioinform.* 23, bbab480. doi: 10.1093/bib/bbab480

Ao, C., Zou, Q., and Yu, L. (2022c). RFhy-m2G: Identification of RNA N2-methylguanosine modification sites based on random forest and hybrid features. *Methods* 203, 32–39. doi: 10.1016/j.ymeth.2021.05.016

Boch, J., and Bonas, U. (2010). Xanthomonas AvrBs3 family-type III effectors: discovery and function. *Annu. Rev. Phytopathol.* 48, 419–436. doi: 10.1146/annurev-phyto-080508-081936

Bupi, N., Sangaraju, V. K., Phan, L. T., Lal, A., Vo, T. T. B., Ho, P. T., et al. (2023). An effective integrated machine learning framework for identifying severity of tomato yellow leaf curl virus and their experimental validation. *Research* 6, 0016. doi: 10.34133/research.0016

Cao, C., Wang, J. H., Kwok, D., Cui, F. F., Zhang, Z. L., Zhao, D., et al. (2022). webTWAS: a resource for disease candidate susceptibility genes identified by transcriptome-wide association study. *Nucleic Acids Res.* 50, D1123–D1130. doi: 10.1093/nar/gkab957

Charoenkwan, P., Chiangjong, W., Nantasenamat, C., Hasan, M. M., Manavalan, B., and Shoombuatong, W. (2021). StackIL6: a stacking ensemble model for improving the prediction of IL-6 inducing peptides. *Brief. Bioinform.* 22, bbab172. doi: 10.1093/bib/bbab172

Chen, H., Li, D., Liao, J., Wei, L., and Wei, L. (2022). MultiscaleDTA: a multiscale-based method with a self-attention mechanism for drug-target binding affinity prediction. *Methods* 207, 103–109. doi: 10.1016/j.ymeth.2022.09.006

Chevez-Guardado, R., and Peña-Castillo, L. (2021). Promotech: a general tool for bacterial promoter recognition. *Genome Biol.* 22, 1–16. doi: 10.1186/s13059-021-02514-9

Dao, F.-Y., Lv, H., Yang, Y.-H., Zulfiqar, H., Gao, H., and Lin, H. (2020). Computational identification of N6-methyladenosine sites in multiple tissues of mammals. *Comput. Struct. Biotechnol. J.* 18, 1084–1091. doi: 10.1016/j.csbj.2020.04.015

Dao, F. Y., Liu, M. L., Su, W., Lv, H., Zhang, Z. Y., Lin, H., et al. (2023). AcrPred: A hybrid optimization with enumerated machine learning algorithm to predict Anti-CRISPR proteins. *Int. J. Biol. Macromol.* 228, 706–714. doi: 10.1016/j.ijbiomac.2022.12.250

de Jong, A., Pietersma, H., Cordes, M., Kuipers, O. P., and Kok, J. (2012). PePPER: a webserver for prediction of prokaryote promoter elements and regulons. *BMC Genomics* 13, 1–10. doi: 10.1186/1471-2164-13-299

Doherty, R., and Couldrey, C. (2014). Exploring genome wide bisulfite sequencing for DNA methylation analysis in livestock: a technical assessment. *Front. Genet.* 5, 126. doi: 10.3389/fgene.2014.00126

Flusberg, B. A., Webster, D. R., Lee, J. H., Travers, K. J., Olivares, E. C., Clark, T. A., et al. (2010). Direct detection of DNA methylation during single-molecule, real-time sequencing. *Nat. Methods* 7, 461. doi: 10.1038/nmeth.1459

Goodner, B., Hinkle, G., Gattung, S., Miller, N., Blanchard, M., Qurollo, B., et al. (2001). Genome sequence of the plant pathogen and biotechnology agent

Agrobacterium tumefaciens C58. *Science* 294, 2323–2328. doi: 10.1126/science.1066803

Hasan, M. M., Schaduangrat, N., Basith, S., Lee, G., Shoombuatong, W., and Manavalan, B. (2020). HLPpred-Fuse: improved and robust prediction of hemolytic peptide and its activity by fusing multiple feature representation. *Bioinformatics* 36, 3350–3356. doi: 10.1093/bioinformatics/btaa160

Hasan, M. M., Tsukiyama, S., Cho, J. Y., Kurata, H., Alam, M. A., Liu, X., et al. (2022). Deepm5C: a deep-learning-based hybrid framework for identifying human RNA N5-methylcytosine sites using a stacking strategy. *Mol. Ther.* 30, 2856–2867. doi: 10.1016/j.ymthe.2022.05.001

He, S. D., Guo, F., Zou, Q., and Ding, H. (2020). MRMD2.0: a python tool for machine learning with feature ranking and reduction. *Curr. Bioinform.* 15, 1213–1221. doi: 10.2174/2212392XMTA2bMjko1

Lai, H.-Y., Zhang, Z.-Y., Su, Z.-D., Su, W., Ding, H., Chen, W., et al. (2019). iProEP: a computational predictor for predicting promoter. *Mol. Therapy Nucleic Acids* 17, 337–346. doi: 10.1016/j.omtn.2019.05.028

Li, H., Gong, Y., Liu, Y., Lin, H., and Wang, G. (2022a). Detection of transcription factors binding to methylated DNA by deep recurrent neural network. *Brief. Bioinform.* 23, bbab533. doi: 10.1093/bib/bbab533

Li, H., Shi, L., Gao, W., Zhang, Z., Zhang, L., Zhao, Y., et al. (2022b). dPromoter-XGBoost: detecting promoters and strength by combining multiple descriptors and feature selection using XGBoost. *Methods* 204, 215–222. doi: 10.1016/j.ymeth.2022.01.001

Li, Y., Qiao, G., Gao, X., and Wang, G. (2022c). Supervised graph co-contrastive learning for drug-target interaction prediction. *Bioinformatics* 38, 2847–2854. doi: 10.1093/bioinformatics/btac164

Li, Y., Qiao, G., Wang, K., and Wang, G. (2022d). Drug-target interaction predication via multi-channel graph neural networks. *Brief. Bioinform.* 23, bbab346. doi: 10.1093/bib/bbab346

Liang, Z. Y., Lai, H. Y., Yang, H., Zhang, C. J., Yang, H., Wei, H. H., et al. (2017). Pro54DB: a database for experimentally verified sigma-54 promoters. *Bioinformatics* 33, 467–469. doi: 10.1093/bioinformatics/btw630

Liao, J., Chen, H., Wei, L., and Wei, L. (2022). GSAML-DTA: an interpretable drug-target binding affinity prediction model based on graph neural networks with self-attention mechanism and mutual information. *Comput. Biol. Med.* 150, 106145. doi: 10.1016/j.compbiomed.2022.106145

Liu, B., Yang, F., Huang, D.-S., and Chou, K.-C. (2018). iPromoter-2L: a two-layer predictor for identifying promoters and their types by multi-window-based PseKNC. *Bioinformatics* 34, 33–40. doi: 10.1093/bioinformatics/btx579

Lv, H., Dao, F.-Y., Zulfiqar, H., and Lin, H. (2021). DeepIPs: comprehensive assessment and computational identification of phosphorylation sites of SARS-CoV-2 infection using a deep learning-based approach. *Brief. Bioinform.* 22, bbab244. doi: 10.1093/bib/bbab244

Manavalan, B., and Patra, M. C. (2022). MLCPP 2.0: an updated cell-penetrating peptides and their uptake efficiency predictor. *J. Mol. Biol.* 434, 167604. doi: 10.1016/j.jmb.2022.167604

Ning, L., Abagna, H. B., Jiang, Q., Liu, S., and Huang, J. (2021a). Development and application of therapeutic antibodies against COVID-19. *Int. J. Biol. Sci.* 17, 1486–1496. doi: 10.7150/ijbs.59149

Ning, L., Cui, T., Zheng, B., Wang, N., Luo, J., Yang, B., et al. (2021b). MNDR v3.0: mammal ncRNA-disease repository with increased coverage and annotation. *Nucleic Acids Res.* 49, D160–d164. doi: 10.1093/nar/gkaa707

Ning, L., Liu, M., Gou, Y., Yang, Y., He, B., and Huang, J. (2022). Development and application of ribonucleic acid therapy strategies against COVID-19. *Int. J. Biol. Sci.* 18, 5070–5085. doi: 10.7150/ijbs.72706

Oubounyt, M., Louadi, Z., Tayara, H., and Chong, K. T. (2019). DeePromoter: robust promoter predictor using deep learning. *Front. Genet.* 10, 286. doi: 10.3389/fgene.2019.00286

Ren, L., Xu, Y., Ning, L., Pan, X., Li, Y., Zhao, Q., et al. (2022). TCM2COVID: A resource of anti-COVID-19 traditional Chinese medicine with effects and mechanisms. *iMETA* 1, e42. doi: 10.1002/imt2.42

Salimi, D., and Moeini, A. (2021). Incorporating *K-mers* highly correlated to epigenetic modifications for bayesian inference of gene interactions. *Curr. Bioinform.* 16, 484–492. doi: 10.2174/1574893615999200728193621

Sawadogo, M., and Roeder, R. G. (1985). Interaction of a gene-specific transcription factor with the adenovirus major late promoter upstream of the TATA box region. *Cell* 43, 165–175. doi: 10.1016/0092-8674(85)90021-2

Shoombuatong, W., Basith, S., Pitti, T., Lee, G., and Manavalan, B. (2022). THRONE: a new approach for accurate prediction of human RNA N7-methylguanosine sites. *J. Mol. Biol.* 434, 167549. doi: 10.1016/j.jmb.2022.167549

Su, W., Liu, M. L., Yang, Y. H., Wang, J. S., Li, S. H., Lv, H., et al. (2021). PPD: a manually curated database for experimentally verified prokaryotic promoters. *J. Mol. Biol.* 433, 166860. doi: 10.1016/j.jmb.2021.166860

Su, W., Xie, X. Q., Liu, X. W., Gao, D., Ma, C. Y., Zulfiqar, H., et al. (2023). iRNA-ac4C: a novel computational method for effectively detecting N4-acetylcytidine sites in human mRNA. *Int. J. Biol. Macromol.* 227, 1174–1181. doi: 10.1016/j.ijbiomac.2022.11.299

Tao, Z., Li, Y., Teng, Z., and Zhao, Y. (2020). A method for identifying vesicle transport proteins based on LibSVM and MRMD. *Comput. Math. Methods Med.* 2020, 8926750. doi: 10.1155/2020/8926750

Teng, Z., Zhao, Z., Li, Y., Tian, Z., Guo, M., Lu, Q., et al. (2022). i6mA-vote: cross-species identification of DNA N6-methyladenine sites in plant genomes based on ensemble learning with voting. *Front. Plant Sci.* 13, 845835. doi: 10.3389/fpls.2022.845835

Wei, L., He, W., Malik, A., Su, R., Cui, L., and Manavalan, B. (2021a). Computational prediction and interpretation of cell-specific replication origin sites from multiple eukaryotes by exploiting stacking framework. *Brief. Bioinform.* 22, bbaa275. doi: 10.1093/bib/bbaa275

Wei, L., Ye, X., Sakurai, T., Mu, Z., and Wei, L. (2022). ToxIBTL: prediction of peptide toxicity based on information bottleneck and transfer learning. *Bioinformatics* 38, 1514–1524. doi: 10.1093/bioinformatics/btac006

Wei, L., Ye, X., Xue, Y., Sakurai, T., and Wei, L. (2021b). ATSE: a peptide toxicity predictor by exploiting structural and evolutionary information based on graph neural network and attention mechanism. *Brief. Bioinform.* 22, bbab041. doi: 10.1093/bib/bbab041

Xiao, J., Liu, M., Huang, Q., Sun, Z., Ning, L., Duan, J., et al. (2022). Analysis and modeling of myopia-related factors based on questionnaire survey. *Comput. Biol. Med.* 150, 106162. doi: 10.1016/j.compbiomed.2022.106162

Yang, H., Luo, Y., Ren, X., Wu, M., He, X., Peng, B., et al. (2021). Risk Prediction of Diabetes: Big data mining with fusion of multifarious physical examination indicators. *Inf. Fusion* 75, 140–149. doi: 10.1016/j.inffus.2021.02.015

Yang, K., Li, M., Yu, L., and He, X. (2023). Repositioning linifanib as a potent anti-necroptosis agent for sepsis. *bioRxiv* 9, 57. doi: 10.1038/s41420-023-01351-y

Yang, X., Ye, X., Li, X., and Wei, L. (2021). Idna-mt: identification DNA modification sites in multiple species by using multi-task learning based a neural network tool. *Front. Genet.* 12, 663572. doi: 10.3389/fgene.2021.663572

Yang, Y., Gao, D., Xie, X., Qin, J., Li, J., Lin, H., et al. (2022). DeepIDC: a prediction framework of injectable drug combination based on heterogeneous information and deep learning. *Clin. Pharmacokinet.* 61, 1749–1759. doi: 10.1007/s40262-022-01180-9

Ye, S., Liang, Y., and Zhang, B. (2021). Bayesian functional mixed-effects models with grouped smoothness for analyzing time-course gene expression data. *Curr. Bioinform.* 16, 2–12. doi: 10.2174/1574893615999200520082636

Zhang, D., Chen, H. D., Zulfiqar, H., Yuan, S. S., Huang, Q. L., Zhang, Z. Y., et al. (2021). iBLP: an XGBoost-based predictor for identifying bioluminescent proteins. *Comput. Math. Methods. Med.* 2021, 6664362. doi: 10.1155/2021/6664362

Zhang, S., Wang, Y., Gu, Y., Zhu, J., Ci, C., Guo, Z., et al. (2018). Specific breast cancer prognosis-subtype distinctions based on DNA methylation patterns. *Mol. Oncol.* 12, 1047–1060. doi: 10.1002/1878-0261.12309

Zhang, Y., Liu, T., Hu, X., Wang, M., Wang, J., Zou, B., et al. (2021). CellCall: integrating paired ligand-receptor and transcription factor activities for cell-cell communication. *Nucleic Acids Res.* 49, 8520–8534. doi: 10.1093/nar/gkab638

Zhang, Y.-F., Wang, Y.-H., Gu, Z.-F., Pan, X.-R., Li, J., Ding, H., et al. (2023). Bitter-RF: a random forest machine model for recognizing bitter peptides. *Front. Med.* 10, 1052923. doi: 10.3389/fmed.2023.1052923

Zhang, Z. M., Wang, J. S., Zulfiqar, H., Lv, H., Dao, F. Y., and Lin, H. (2020). Early diagnosis of pancreatic ductal adenocarcinoma by combining relative expression orderings with machine-learning method. *Front. Cell Dev. Biol.* 8, 582864. doi: 10.3389/fcell.2020.582864

Zhang, Z. Y., Ning, L., Ye, X., Yang, Y. H., Futamura, Y., Sakurai, T., et al. (2022a). iLoc-miRNA: extracellular/intracellular miRNA prediction using deep BiLSTM with attention mechanism. *Brief. Bioinform.* 23, bbac395. doi: 10.1093/bib/bbac395

Zhang, Z. Y., Sun, Z.-J., Yang, Y.-H., and Lin, H. (2022b). Towards a better prediction of subcellular location of long non-coding RNA. *Front. Comput. Sci.* 16, 165903. doi: 10.1007/s11704-021-1015-3

Zhao, Y., Wang, F., Chen, S., Wan, J., and Wang, G. (2017). Methods of MicroRNA promoter prediction and transcription factor mediated regulatory network. *Biomed. Res. Int.* 2017, 7049406. doi: 10.1155/2017/7049406

Zou, Q., Wan, S., Ju, Y., Tang, J., and Zeng, X. (2016). Pretata: predicting TATA binding proteins with novel features and dimensionality reduction strategy. *BMC Syst. Biol.* 10, 114. doi: 10.1186/s12918-016-0353-5

Zou, Q., Xing, P. W., Wei, L. Y., and Liu, B. (2019). Gene2vec: gene subsequence embedding for prediction of mammalian N-6-methyladenosine sites from mRNA. *RNA* 25, 205–218. doi: 10.1261/rna.069112.118

Zou, Y., Ding, Y. J., Peng, L., and Zou, Q. (2022). FTWSVM-SR: DNA-binding proteins identification via fuzzy twin support vector machines on self-representation. *Interdisc. Sci. Comput. Life Sci.* 14, 372–384. doi: 10.1007/s12539-021-00489-6

Zulfiqar, H., Guo, Z., Grace-Mercure, B. K., Zhang, Z. Y., Gao, H., Lin, H., et al. (2023). Empirical comparison and recent advances of computational prediction

of hormone binding proteins using machine learning methods. *Comput. Struct. Biotechnol. J.* 21, 2253–2261. doi: 10.1016/j.csbj.2023.03.024

Zulfiqar, H., Huang, Q.-L., Lv, H., Sun, Z.-J., Dao, F.-Y., and Lin, H. (2022a). Deep-4mCGP: a deep learning approach to predict 4mC sites in Geobacter pickeringii by using correlation-based feature selection technique. *Int. J. Mol. Sci.* 23, 1251. doi: 10.3390/ijms23031251

Zulfiqar, H., Khan, R. S., Hassan, F., Hippe, K., Hunt, C., Ding, H., et al. (2021a). Computational identification of N4-methylcytosine sites in the mouse genome with machine-learning method. *Math. Biosci. Eng.* 18, 3348–3363. doi: 10.3934/mbe.2021167

Zulfiqar, H., Sun, Z.-J., Huang, Q.-L., Yuan, S.-S., Lv, H., Dao, F.-Y., et al. (2022b). Deep-4mCW2V: A sequence-based predictor to identify N4-methylcytosine sites in *Escherichia coli*. *Methods* 203, 558–563. doi: 10.1016/j.ymeth.2021.07.011

Zulfiqar, H., Yuan, S.-S., Huang, Q.-L., Sun, Z.-J., Dao, F.-Y., Yu, X.-L., et al. (2021b). Identification of cyclin protein using gradient boost decision tree algorithm. *Comput. Struct. Biotechnol. J.* 19, 4123–4131. doi: 10.1016/j.csbj.2021.07.013