Molecular Therapy
Nucleic Acids
Original Article

Predicting Preference of Transcription Factors for Methylated DNA Using Sequence Information

Meng-Lu Liu,[1] Wei Su,[1] Jia-Shu Wang,[1] Yu-He Yang,[1] Hui Yang,[1] and Hao Lin[1]

[1]Center for Informational Biology, University of Electronic Science and Technology of China, Chengdu 610054, China

Transcription factors play key roles in cell-fate decisions by regulating 3D genome conformation and gene expression. The traditional view is that methylation of DNA hinders transcription factors binding to them, but recent research has shown that many transcription factors prefer to bind to methylated DNA. Therefore, identifying such transcription factors and understanding their functions is a stepping-stone for studying methylation-mediated biological processes. In this paper, a two-step discriminated method was proposed to recognize transcription factors and their preference for methylated DNA based only on sequences information. In the first step, the proposed model was used to discriminate transcription factors from non-transcription factors. The areas under the curve (AUCs) are 0.9183 and 0.9116, respectively, for the 5-fold cross-validation test and independent dataset test. Subsequently, for the classification of transcription factors that prefer methylated DNA and transcription factors that prefer non-methylated DNA, our model could produce the AUCs of 0.7744 and 0.7356, respectively, for the 5-fold cross-validation test and independent dataset test. Based on the proposed model, a user-friendly web server called TFPred was built, which can be freely accessed at http://lin-group.cn/server/TFPred/.

INTRODUCTION

Transcription factors (TFs) have prominent roles in 3D genome organization and transcriptional regulation. During the formation of 3D genome, TFs mediate long-range interactions of DNA sequences and induce TAD and loop formation, A–B compartment switching, and nuclear repositioning.[1] The 3D genome conformation further affects transcriptional regulation. Transcriptional regulation of gene expression plays a critical role in many cellular processes, including cancer development[2] and plant yield.[3] Identification of TFs is essential for understanding 3D genome functions and gene regulatory mechanisms.[4] The traditional view thought that TFs usually bind to non-methylated DNA motifs, where high-level methylation of CpG dinucleotides can prohibit the recruitment of TFs.[5] However, hundreds of TFs were identified to prefer methylated sequences by experimental methods. For example, KLF4,[6] TET,[7] CEBPA,[8] and ZFP57[9] were identified to bind to methylated DNA with high affinity. The binding of methylated DNA to TFs is usually through hydrophobic interactions.[10] Moreover, the binding motif in methylated DNA may be different from that in non-methylated DNA.[10,11] Study has shown that the methylated DNA-bound TFs are primarily involved in embryonic and organismal development in the human body.[5,10,12,13] As a result, the interaction between methylated DNA and TFs can activate or repress gene expression, initiate transcription, and regulate RNA splicing. Furthermore, abnormal interactions may cause diseases.[11,14,15] However, the detailed function of the interaction between methylated DNA and TFs is still unknown. Identification of such TFs and elucidation of their roles becomes an important step in understanding methylation-mediated biological processes and related human diseases.

There are some high-throughput experimental methods[5] to identify the TFs that prefer bind to methylated DNA, such as tandem mass spectrometry, functional protein microarray, DNA microarray, ChIP-BS-seq (chromatin immunoprecipitation followed by bisulfite sequencing) and systematic evolution of ligands by exponential enrichment (HT-SELEX). However, in the post-genome era, the number of protein sequences exploded. It is costly and time-consuming to annotate new proteins by these experimental methods. The gap between unknown and annotated proteins is becoming wider and wider. Therefore, it is necessary to develop computational methods[16–19] to recognize the TFs that prefer to bind to methylated DNA. However, to the best of our knowledge, there is still not such a tool for the identification of methylated DNA-bound TFs.

In this paper, we first collected the datasets of TFs and non-TFs from literatures and database and then built a two-step classifier based on these datasets. As shown in Figure 1, the first step is to identify TFs based on support vector machine (SVM), and the second step is to further predict the TFs that prefer to bind to methylated DNA by applying XGBoost. Sequence-based feature extraction methods were used to represent protein sequences, and the analysis of variance (ANOVA) combined with incremental feature selection (IFS) technique was applied to obtain the optimal feature set. Finally, for the convenience of experimental scientists, a user-friendly web server was built according to the proposed method.

Received 27 May 2020; accepted 28 July 2020;
https://doi.org/10.1016/j.omtn.2020.07.035.

Correspondence: Hao Lin, Center for Informational Biology, University of Electronic Science and Technology of China, Chengdu 610054, China.
E-mail: hlin@uestc.edu.cn

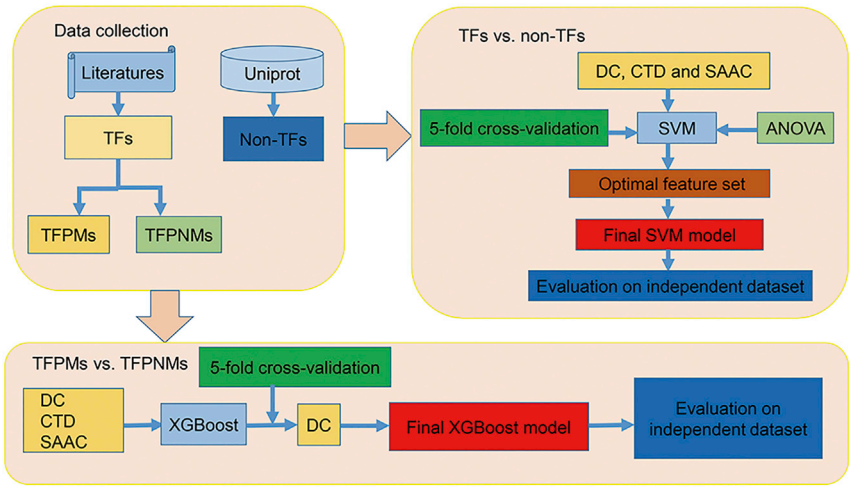

Figure 1. The Workflow of This Work

RESULTS AND DISCUSSION

Amino Acid Composition (AAC) of TFs and Non-TFs

The code arrangement from the primary sequence is the most essential for the formation of protein unique functions. Therefore, the AAC was used to analyze the preference of different sequences for specific amino acids. Figure 2 plotted the average AAC of positive and negative samples in the two datasets, respectively. It can be seen from Figure 2A that the frequency of A, Q, H, G, P, and S in TFs are significantly higher than that in non-TF (t test, $p < 0.01$). In contrast, the frequency of D, C, E, I, N, L, K, M, F, W, Y, and V in non-TFs is higher than that in TFs (t test, $p < 0.01$). These amino acids may be closely related to the function of TFs that further studies need to focus on. As can be seen from Figure 2B, the AAC of the two type of TFs is very similar, so it is difficult to distinguish them by only AAC.

Model Construction by Applying Sequence-Based Features

Identification of TFs by Applying ANOVA

The ensemble features of composition/transition/distribution (CTD), split AAC (SAAC), and dipeptide composition (DC) were used to train the TFs prediction model, so each sample was converted to a 528-dimension vector. SVM was used as the machine learning algorithm. To get rid of the redundant features, we used ANOVA combined with IFS was used for feature selection. The accuracy of 5-fold cross-validation test was used to choose the optimal feature set. Figure 3 showed the corresponding IFS curve. It is obvious that the optimal feature set including the top ranked 201 features could produce the maximum accuracy of 86.54%. Therefore, the 201 optimal features were used to construct the final classification model. Of these 201 features, 28 are from SAAC, 128 are from DC, and 45 are from CTD. All three kinds of features have a positive effect on the classification of sequences. We further investigated the top ten features and found they were all derived from CTD, which are related to the charge, hydrophobicity, van der Waals volume, secondary structure, and polarizability of the protein sequence. This indicates that physicochemical properties are very important for the recognition of TFs, which may be because the function of a TF is closely related to its physicochemical properties.

Prediction of TFs that Prefer Methylated DNA by Comparing Different Features

In order to balance the number of positive and negative samples, 146 positive samples in the benchmark dataset were randomly selected for model training. XGBoost was applied to build a classifier that predict TFs that prefer methylated DNA. In XGBoost algorithm, there are many hyper-parameters like the maximum tree depth for base learners, the learning rate, the booster method, the tree method, and so on. In order to quickly and efficiently obtain a model with high classification ability, we focused on four parameters, including the maximum depth of the tree, the number of boosting iterations, the subsample ratio, and the learning rate. Grid search was used to get the best classification accuracy. The corresponding parameters were listed in Table 1.

In order to compare the impact of different sequence representation methods on classification performance, we used the three feature extraction methods mentioned in the previous section, respectively, to encode the sequences, and obtained the corresponding results of 5-fold cross-validation test, which were recorded in Table 2. It was found that although the ensemble features contained more features, they didn't always perform better. The single feature DC achieved the highest value on all of the evaluation indicators except *Sp*. Its sensitivity (*Sn*), specificity (*Sp*), accuracy (*Acc*), Matthew's correlation coefficient (*MCC*), and area under the curve (*AUC*) were 73.29%, 69.86%, 71.58%, 0.4318, and 0.7744, respectively. Therefore, DC was used to predict TFs that prefer methylated DNA.

It can be seen that the prediction performance in the second step is lower than it in the first step. The reason might be related to the scale of problems. TFs are significantly different in function from other proteins, which is easily observed from the AAC analysis. Therefore, it is relatively easy to distinguish between TFs and non-TFs. However, no matter what kind of TFs, they all have a similar function that could interact with DNA sequences. Their sequence difference is relatively small, and thus it is difficult to identify TFs that prefer methylated DNA.

Comparative Performance among Different Machine Learning Methods

In addition to SVM and XGBoost, many machine learning algorithms have achieved excellent performance on sequence classification

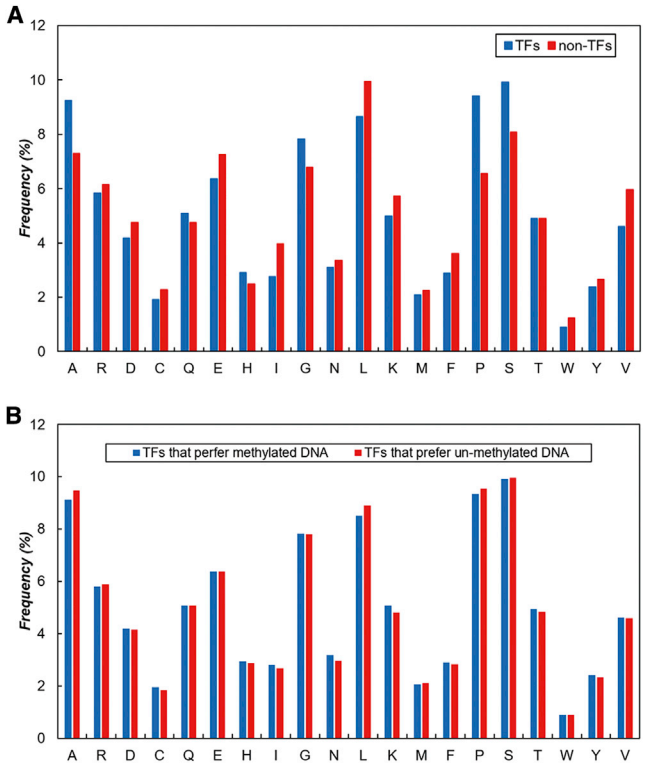

Figure 2. The 20 Amino Acid Composition in the Two Kinds of Data
(A) TFs versus non-TFs. (B) TFs that prefer methylated DNA versus TFs that prefer non-methylated DNA.

problems. We compared the performance obtained in the previous section with the performance of other machine learning algorithms, including linear classification algorithms, such as logistic regression (LR), and non-linear classification algorithms, such as random forest (RF) and k-nearest neighbor (KNN). In LR, KNN, and SVM, ANOVA was used for feature selection, whereas no extra feature selection technique was applied in XGBoost and RF because they have their own feature selection mechanisms. The results of 5-fold cross-validation test were plotted in Figure 4. From Figure 4A, for predicting TFs, the performances of the five classifiers are similar. Among them, SVM achieves the best performance and its results on all evaluation indicators are the highest. On the contrary, these algorithms display quite different performances for distinguishing between two types of TFs. As shown in Figure 4B, the XGBoost achieves the best prediction performance. Above results suggest that different classification algorithms are suitable for different classification problems.

Model Evaluation on the Independent Dataset

The independent dataset was used to evaluate the generalization ability of the models. The corresponding results were recorded in Table 3. TFPM and TFPNM in the table refer to TFs that prefer to bind to methylated DNA and those that prefer to bind to non-methylated DNA. It can be seen from the table that the accuracies of models are 83.02% and 68.87%, respectively, in the identification of TFs

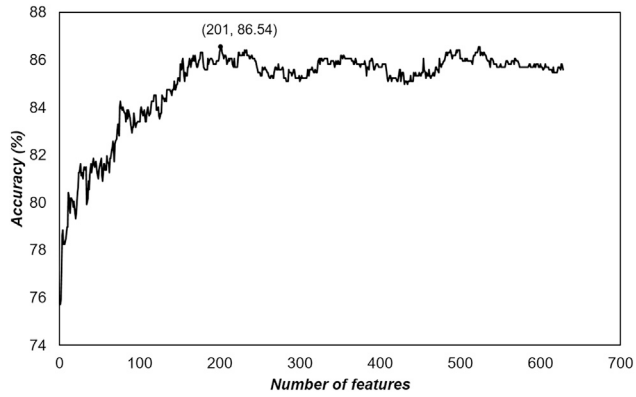

Figure 3. The IFS Curve for Identifying TFs
It reaches a peak of 86.54% when abscissa is 201.

and those that prefer methylated DNA, which are almost equal to the results of the 5-fold cross-validation test. This demonstrates that the model possesses high stability. In order to show the prediction ability of the model more vividly, we draw the ROC curves of 5-fold cross-validation test and independent dataset test in Figure 5. It shows that our models are powerful. For predicting TFs that prefer methylated DNA, 124 samples were not used in the training dataset. We used the model to predict these samples, and 86 were identified, which further indicates that the model is robust. In order to know whether our model is also effective for mouse TFs, 129 TFs that prefer to bind to methylated DNA[5] were used to examine our model. However, only 66 TFs were identified correctly, suggesting the species specificity of TFs' sequences.

Web Server

Based on our proposed models, a user-friendly web server called TFPred was constructed for the convenience of experimental scientists. A step-by-step guide is given as follows:

Step 1: enter the website http://lin-group.cn/server/TFPred in the browser. You can see the home page as shown in Figure 6.

Step 2: click the "Web Server" button and type or paste query protein sequences in the format of FASTA into the input box.

Step 3: select the protein type to be predicted.

Step 4: click on the "submit" button and see the predicted result.

Table 1. The Parameters in XGBoost and the Corresponding Value Ranges

Parameter	Range	Number
Maximum depth of the tree	from 1 to 6	6
Number of boosting iterations	from 1 to 30	30
Subsample ratio	from 0.5 to 1 with step 0.1	5
Learning rate	from 0.1 to 1 with step 0.1	10

Table 2. The Effects of Different Feature Extraction Methods on Prediction of TFs that Prefer Methylated DNA

Feature	*Sn* (%)	*Sp* (%)	*Acc* (%)	*MCC*	*AUC*
(1) CTD	63.01	71.23	67.12	0.3436	0.7447
(2) DC	73.29	69.86	71.58	0.4318	0.7744
(3) SAAC	71.92	69.18	70.55	0.4111	0.7459
(1)+(2)	69.86	68.49	69.18	0.3836	0.7565
(1)+(3)	65.75	72.60	69.18	0.3845	0.7245
(2)+(3)	71.23	70.55	70.89	0.4178	0.7666
(1)+(2)+(3)	69.86	70.55	70.21	0.4041	0.7401

Besides the web server prediction, a stand-alone software package is also available on the download page, so one can download it and perform the prediction locally.

Conclusions

Recently, researchers have found that many TFs prefer to bind to methylated DNA, which is contrary to our traditional view. However, the detailed functions of these TFs are still unknown. Identifying such TFs and understanding their function is essential for studying methylation-mediated biological processes.

This paper constructed a two-step classifier to identify TFs and their preference for methylated DNA. It achieved an encouraging performance on both training dataset and independent dataset. This indicates that the protein sequence contains important information for binding to the target DNA. For the convenience of experimental scientists, we built an online free web server TFPred. We hope the tool could provide important clues for further study of methylated DNA-bound TFs. In the future, we will collect more annotated TFs and deal with TFs of different species.

MATERIALS AND METHODS

Data Collection and Processing

The TFs that prefer methylated DNA were obtained from the work of Wang et al.[5] They manually collected 601 human and 129 mouse TFs from the literature. In our study, the human TFs were used to train models. The TFs that prefer non-methylated DNA were obtained from the work of Yin et al.[10] In their study, they systematically investigated the effects of DNA methylation on TFs and DNA binding by experimental methods. A total of 286 TFs were confirmed to prefer non-methylated DNA. In order to build high-quality datasets, three procedures were performed to further process these sequences. First, sequences that contain non-standard amino acid residues, such as "B," "X," or "Z," were excluded because they have multiple possible meanings and the specific meaning is uncertain. Second, CD-HIT[20] with the cut-off threshold of 25% was used to get rid of redundant samples.[21] Third, sequences with less than 50 amino acids were removed. As a result, 339 TFs that prefer mCpG DNA sequences were extracted as positive samples and 183 TFs that prefer non-methylated DNA were extracted as negative samples. The

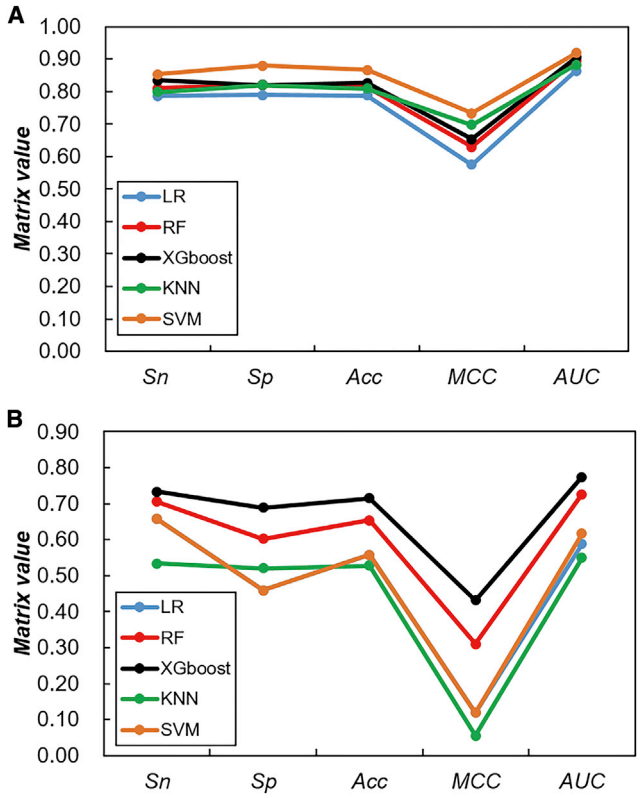

Figure 4. The Performance Comparison of Different Machine Learning Algorithms for Two Classification Problems

(A) TFs versus non-TFs. (B) TFs that prefer methylated DNA versus TFs that prefer non-methylated DNA.

benchmark dataset S_{TF} for the classification of two kinds of TFs can be formulated by:

$$S_{TF} = S_{TFPM} \cup S_{TFPNM}, \quad \text{(Equation 1)}$$

where S_{TFPM} and S_{TFPNM} denotes the TFs that prefer methylated DNA and TFs that prefer non-methylated DNA, respectively.

The TF dataset consists of the two above-mentioned TFs with different methylated DNA preferences. The non-TFs were obtained from the Uniprot database release 2019_11 according to the following criteria: (1) proteins should be reviewed; (2) existence of proteins was proven by "evidence at protein level"; (3) proteins should be full length and not fragmented; (4) proteins should have

Table 3. The Performance of the Models on an Independent Dataset

Data	*Sn* (%)	*Sp* (%)	*Acc* (%)	*MCC*	*AUC*
TF versus non-TF	80.19	85.85	83.02	0.6614	0.9116
TFPM versus TFPNM	71.01	64.86	68.87	0.3471	0.7356

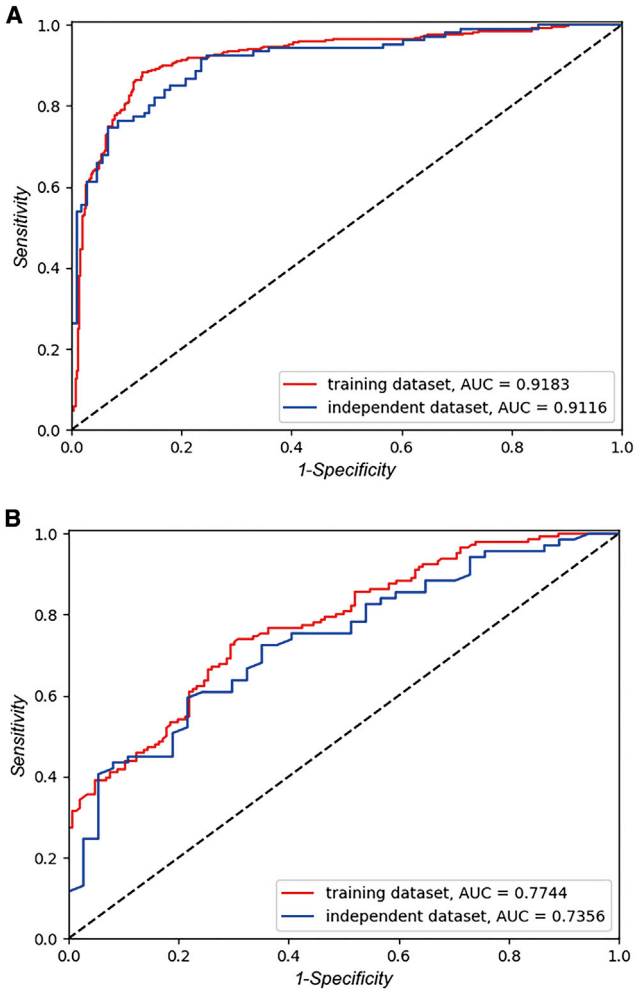

Figure 5. The ROC Curves of 5-Fold Cross-Validation Test and Independent Dataset Test for the Two Problems

(A) The classification of TFs and non-TFs. (B) The classification of TFs that prefer methylated DNA and non-methylated DNA.

more than 50 amino acids; (5) proteins should not have the DNA-binding TF activity; and (6) proteins should come from *Homo sapiens*. CD-HIT with cut-off 25% was also performed to remove redundant sequences. Finally, 522 non-TFs were randomly selected. The dataset for the discrimination between TFs and non-TFs can be formulated by:

$$S = S_{TF} \cup S_{non-TF}, \quad (2)$$

where S_{TF} and S_{non-TF} denotes the TFs and non-TFs, respectively.

To evaluate the performances of predictors, we further divided each dataset into a training dataset and an independent dataset according to a ratio of 8:2. The number of sequences in each dataset was shown in Table 4. All datasets are available at http://lin-group.cn/server/TFPred/.

Feature Description

The varying length of protein sequences prevents machine learning from directly handling such information. They need to be represented as fixed-length vectors. In order to effectively represent protein sequences and preserve the differences between positive and negative samples, many sequence representation methods have been proposed. They are roughly divided into 3 categories,[22] which are sequence-based methods, such as AAC,[23–28] DC,[29–33] tripeptide composition (TC),[34] evolutionary-information-based methods, such as position-specific scoring matrix (PSSM), and annotation-based methods, like Gene Ontology (GO).[35–37] In this study, we used three sequence-based methods to represent protein samples, including SAAC, DC,[29] and CTD.[34]

SAAC

In SAAC, each protein sequence was divided into three parts. The 20 amino acid frequencies of each part were calculated respectively, so the dimension of the final vector was 60 for each sequence.[35,38] We tried 5, 10, 15, 20, and 25 N-terminal and C-terminal amino acid residues and found that 25 and 20 terminal amino acid residues performed best in the first and second step predictions, respectively. Therefore, these two forms were used for model construction.

DC

DC used a sliding window mode with size 2 and step 1 to calculate the frequencies of all kinds of amino acid pairs in a protein sequence.[29–33] Hence, it describes each protein sequence in form of 400-dimension vector.

CTD

CTD was proposed by Dubchak et al.[39] in 1995. The 20 standard amino acids were first divided into 3 types according to their physicochemical properties. For each property, three descriptors including composition (C), transition (T), and distribution (D) were used to describe the global composition in a protein, the property changes along the protein and the distribution pattern along the sequence, respectively. In this study, eight kinds of physicochemical properties were used including hydrophobicity, normalized van der Waals volume, polarity, polarizability, charge, second structure, solvent accessibility, and surface tension. Please refer to Tan et al.[34] and Dubchak et al.[39] for more details about how to calculate CTD features.

Feature Selection

In order to remove redundant features and improve the accuracy of the classifier, many dimensionality reduction methods have been used in bioinformatics, such as ANOVA,[40] minimal redundancy maximal relevance (mRMR),[41] and maximum relevance maximum distance (MRMD).[42,43] Among them, ANOVA usually performs well in the field of protein prediction and its calculation speed is fast, so we used it for feature selection. The main idea of ANOVA is to divide the variance of the data into intra-class variance and inter-class variance, then to calculate the ratio of each feature, and final

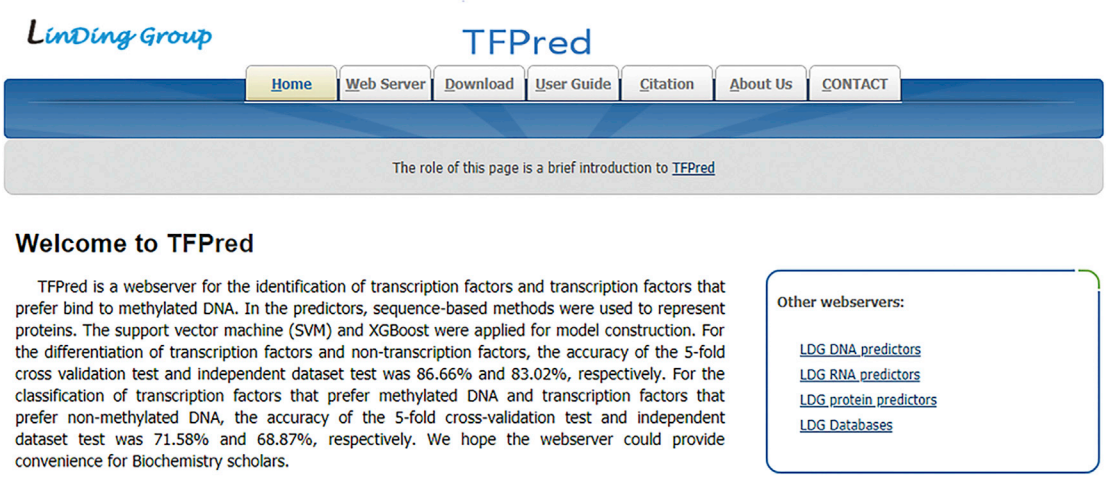

Figure 6. A Semi-Screen Shot for the Top Page of the TFPred Web Server

to sort them in descending order. By combining with IFS,[44] an approximate optimal feature set can be obtained.

Machine Learning

In recent years, machine learning has been widely used for sequence classification problems.[45–54] Among all machine learning algorithms, SVM and XGBoost are widely used because of their good performance. SVM is a single learning method and has been widely used in bioinformatics.[55–63] It maps samples to a new space by kernel function and finds a hyperplane to maximize the interval between positive and negative samples. XGBoost is an ensemble learning method based on boosting tree. Its main idea is to continuously add trees and perform feature splitting to grow a tree. Each time a tree is added, it is actually learning a new function to fit the residuals of the last prediction. The impact of XGBoost also has been widely recognized in a number of machine learning and data mining challenges.[64]

Performance Evaluation

At present, there are two main evaluation methods for machine learning models including k-fold cross-validation test and independent dataset test.[65,66] In this article, both 5-fold cross-validation test and independent dataset test were used to fully evaluate the performance of the classifier.

Four standard statistical indicators were used to measure the performance of each model namely *Acc*, *Sn*, *Sp*, and *MCC*, which are expressed as[67,68]

Table 4. The Number of Samples in Each Dataset

	Training Dataset		Independent Dataset	
Sample	Positive	Negative	Positive	Negative
TF versus non-TF	416	416	106	106
TFPM versus TFPNM	270	146	69	37

$$Acc = \frac{TP + TN}{TP + TN + FP + FN} \quad (3)$$

$$Sn = \frac{TP}{TP + FN} \quad (4)$$

$$Sp = \frac{TN}{TN + FP} \quad (5)$$

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

where *TP* represents true positive, *TN* represents true negative, *FP* represents false positive, and *FN* represents false negative.

The ROC curve is one of the most important indicators to measure the performance of a classifier. It can be drawn with 1-*Sp* as abscissa and *Sn* as ordinate by adjusting the threshold of the classifier, The ROC curve can be characterized by the *AUC*, and the higher the *AUC* value, the better the performance of the classifier. Therefore, we also used the ROC curve and *AUC* to measure the performance of a classifier.

AUTHOR CONTRIBUTIONS

H.L. conceived and designed the study. M.-L.L. conducted the experiments. M.-L.L., W.S., J.-S.W., Y.-H.Y., and H.Y. implemented the algorithms. M.-L.L. and H.Y. established the web server. M.-L.L. and H.L. performed the analysis and wrote the paper. All authors read and approved the final manuscript.

CONFLICTS OF INTEREST

The authors declare no competing interests.

ACKNOWLEDGMENTS

This work has been supported by the National Natural Science Foundation of China (61772119), Sichuan Provincial Science Fund for

Distinguished Young Scholars (2020JDJQ0012), and the Science Strength Promotion Programme of UESTC.

REFERENCES

1. Stadhouders, R., Filion, G.J., and Graf, T. (2019). Transcription factors and 3D genome conformation in cell-fate decisions. Nature *569*, 345–354.
2. Bradner, J.E., Hnisz, D., and Young, R.A. (2017). Transcriptional Addiction in Cancer. Cell *168*, 629–643.
3. Shen, Z., Lin, Y., and Zou, Q. (2019). Transcription factors-DNA interactions in rice: identification and verification. Brief. Bioinform. *21*, 946–956.
4. Wang, Z., Civelek, M., Miller, C.L., Sheffield, N.C., Guertin, M.J., and Zang, C. (2018). BART: a transcription factor prediction tool with query gene sets or epigenomic profiles. Bioinformatics *34*, 2867–2869.
5. Wang, G., Luo, X., Wang, J., Wan, J., Xia, S., Zhu, H., Qian, J., and Wang, Y. (2018). MeDReaders: a database for transcription factors that bind to methylated DNA. Nucleic Acids Res. *46* (D1), D146–D151.
6. Hu, S., Wan, J., Su, Y., Song, Q., Zeng, Y., Nguyen, H.N., Shin, J., Cox, E., Rho, H.S., Woodard, C., et al. (2013). DNA methylation presents distinct binding sites for human transcription factors. eLife *2*, e00726.
7. Liu, D., Li, G., and Zuo, Y. (2019). Function determinants of TET proteins: the arrangements of sequence motifs with specific codes. Brief. Bioinform. *20*, 1826–1835.
8. Mann, I.K., Chatterjee, R., Zhao, J., He, X., Weirauch, M.T., Hughes, T.R., and Vinson, C. (2013). CG methylated microarrays identify a novel methylated sequence bound by the CEBPB|ATF4 heterodimer that is active in vivo. Genome Res. *23*, 988–997.
9. Quenneville, S., Verde, G., Corsinotti, A., Kapopoulou, A., Jakobsson, J., Offner, S., Baglivo, I., Pedone, P.V., Grimaldi, G., Riccio, A., and Trono, D. (2011). In embryonic stem cells, ZFP57/KAP1 recognize a methylated hexanucleotide to affect chromatin and DNA methylation of imprinting control regions. Mol. Cell *44*, 361–372.
10. Yin, Y., Morgunova, E., Jolma, A., Kaasinen, E., Sahu, B., Khund-Sayeed, S., Das, P.K., Kivioja, T., Dave, K., Zhong, F., et al. (2017). Impact of cytosine methylation on DNA binding specificities of human transcription factors. Science *356*, eaaj2239.
11. Zhu, H., Wang, G., and Qian, J. (2016). Transcription factors as readers and effectors of DNA methylation. Nat. Rev. Genet. *17*, 551–565.
12. Li, H., Ta, N., Long, C., Zhang, Q., Li, S., Liu, S., Yang, L., and Zuo, Y. (2019). The spatial binding model of the pioneer factor Oct4 with its target genes during cell reprogramming. Comput. Struct. Biotechnol. J. *17*, 1226–1233.
13. Li, H., Song, M., Yang, W., Cao, P., Zheng, L., and Zuo, Y. (2020). A Comparative Analysis of Single-Cell Transcriptome Identifies Reprogramming Driver Factors for Efficiency Improvement. Mol. Ther. Nucleic Acids *19*, 1053–1064.
14. Yu, L., Yao, S., Gao, L., and Zha, Y. (2019). Conserved Disease Modules Extracted From Multilayer Heterogeneous Disease and Gene Networks for Understanding Disease Mechanisms and Predicting Disease Treatments. Front. Genet. *9*, 745.
15. Yu, L., and Gao, L. (2019). Human Pathway-Based Disease Network. IEEE/ACM Trans. Comput. Biol. Bioinformatics *16*, 1240–1249.
16. Qu, K., Wei, L., and Zou, Q. (2019). A Review of DNA-binding Proteins Prediction Methods. Curr. Bioinform. *14*, 246–254.
17. Liang, P., Yang, W., Chen, X., Long, C., Zheng, L., Li, H., and Zuo, Y. (2020). Machine Learning of Single-Cell Transcriptome Highly Identifies mRNA Signature by Comparing F-Score Selection with DGE Analysis. Mol. Ther. Nucleic Acids *20*, 155–163.
18. Ta, N., Li, H., Liu, S., and Zuo, Y. (2020). Mining Key Regulators of Cell Reprogramming and Prediction Research Based on Deep Learning Neural Networks. IEEE Access *8*, 23179–23185.
19. Wang, G., Wang, Y., Feng, W., Wang, X., Yang, J.Y., Zhao, Y., Wang, Y., and Liu, Y. (2008). Transcription factor and microRNA regulation in androgen-dependent and -independent prostate cancer cells. BMC Genomics *9*, S22.
20. Huang, Y., Niu, B., Gao, Y., Fu, L., and Li, W. (2010). CD-HIT Suite: a web server for clustering and comparing biological sequences. Bioinformatics *26*, 680–682.
21. Zou, Q., Lin, G., Jiang, X., Liu, X., and Zeng, X. (2018). Sequence clustering in bioinformatics: an empirical study. Brief. Bioinform. *21*, 1–10.
22. Li, X., Wu, X., and Wu, G. (2014). Robust feature generation for protein subchloroplast location prediction with a weighted GO transfer model. J. Theor. Biol. *347*, 84–94.
23. Zuo, Y., Li, Y., Chen, Y., Li, G., Yan, Z., and Yang, L. (2017). PseKRAAC: a flexible web server for generating pseudo K-tuple reduced amino acids composition. Bioinformatics *33*, 122–124.
24. Manavalan, B., Basith, S., Shin, T.H., Wei, L., and Lee, G. (2019). AtbPpred: A Robust Sequence-Based Prediction of Anti-Tubercular Peptides Using Extremely Randomized Trees. Comput. Struct. Biotechnol. J. *17*, 972–981.
25. Wei, L., Su, R., Luan, S., Liao, Z., Manavalan, B., Zou, Q., and Shi, X. (2019). Iterative feature representations improve N4-methylcytosine site prediction. Bioinformatics *35*, 4930–4937.
26. 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.
27. Zhao, X., Jiao, Q., Li, H., Wu, Y., Wang, H., Huang, S., and Wang, G. (2020). ECFS-DEA: an ensemble classifier-based feature selection for differential expression analysis on expression profiles. BMC Bioinformatics *21*, 43.
28. Cheng, L., Wang, P., Tian, R., Wang, S., Guo, Q., Luo, M., Zhou, W., Liu, G., Jiang, H., and Jiang, Q. (2019). LncRNA2Target v2.0: a comprehensive database for target genes of lncRNAs in human and mouse. Nucleic Acids Res. *47* (D1), D140–D144.
29. Ding, H., and Li, D. (2015). Identification of mitochondrial proteins of malaria parasite using analysis of variance. Amino Acids *47*, 329–333.
30. Charoenkwan, P., Shoombuatong, W., Lee, H.-C., Chaijaruwanich, J., Huang, H.-L., and Ho, S.-Y. (2013). SCMCRYS: predicting protein crystallization using an ensemble scoring card method with estimating propensity scores of P-collocated amino acid pairs. PLoS ONE *8*, e72368.
31. Pratiwi, R., Malik, A.A., Schaduangrat, N., Prachayasittikul, V., Wikberg, J.E., Nantasenamat, C., et al. (2017). CryoProtect: A Web Server for Classifying Antifreeze Proteins from Nonantifreeze Proteins. J. Chem. *8*, 1–15.
32. Win, T.S., Malik, A.A., Prachayasittikul, V., S Wikberg, J.E., Nantasenamat, C., and Shoombuatong, W. (2017). HemoPred: a web server for predicting the hemolytic activity of peptides. Future Med. Chem. *9*, 275–291.
33. Win, T.S., Schaduangrat, N., Prachayasittikul, V., Nantasenamat, C., and Shoombuatong, W. (2018). PAAP: a web server for predicting antihypertensive activity of peptides. Future Med. Chem. *10*, 1749–1767.
34. Tan, J.X., Li, S.H., Zhang, Z.M., Chen, C.X., Chen, W., Tang, H., and Lin, H. (2019). Identification of hormone binding proteins based on machine learning methods. Math. Biosci. Eng. *16*, 2466–2480.
35. Zuo, Y.C., Peng, Y., Liu, L., Chen, W., Yang, L., and Fan, G.L. (2014). Predicting peroxidase subcellular location by hybridizing different descriptors of Chou' pseudo amino acid patterns. Anal. Biochem. *458*, 14–19.
36. Zhang, Z., Zhang, J., Fan, C., Tang, Y., and Deng, L. (2019). KATZLGO: Large-Scale Prediction of LncRNA Functions by Using the KATZ Measure Based on Multiple Networks. IEEE/ACM Trans. Comput. Biol. Bioinformatics *16*, 407–416.
37. Deng, L., Wang, J., and Zhang, J. (2019). Predicting Gene Ontology Function of Human MicroRNAs by Integrating Multiple Networks. Front. Genet. *10*, 3.
38. Kumar, R., Kumari, B., and Kumar, M. (2017). Prediction of endoplasmic reticulum resident proteins using fragmented amino acid composition and support vector machine. PeerJ *5*, e3561.
39. Dubchak, I., Muchnik, I., Holbrook, S.R., and Kim, S.H. (1995). Prediction of protein folding class using global description of amino acid sequence. Proc. Natl. Acad. Sci. USA *92*, 8700–8704.
40. Deng, L., Guan, J., Wei, X., Yi, Y., Zhang, Q.C., and Zhou, S. (2013). Boosting prediction performance of protein-protein interaction hot spots by using structural neighborhood properties. J. Comput. Biol. *20*, 878–891.
41. Zhang, Z.M., Tan, J.X., Wang, F., Dao, F.Y., Zhang, Z.Y., and Lin, H. (2020). Early Diagnosis of Hepatocellular Carcinoma Using Machine Learning Method. Front. Bioeng. Biotechnol. *8*, 254.

42. 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.

43. 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* (*Suppl 4*), 114.

44. Li, S.H., Zhang, J., Zhao, Y.W., Dao, F.Y., Ding, H., Chen, W., et al. (2019). iPhoPred: a predictor for identifying phosphorylation sites in human protein. IEEE Access *7*, 177517–177528.

45. Liao, Z.J., Wan, S.X., He, Y., and Zou, Q. (2018). Classification of Small GTPases with Hybrid Protein Features and Advanced Machine Learning Techniques. Curr. Bioinform. *13*, 492–500.

46. Liao, Z.J., Li, D.P., Wang, X.R., Li, L.S., and Zou, Q. (2018). Cancer Diagnosis Through IsomiR Expression with Machine Learning Method. Curr. Bioinform. *13*, 57–63.

47. Ru, X., Cao, P., Li, L., and Zou, Q. (2019). Selecting Essential MicroRNAs Using a Novel Voting Method. Mol. Ther. Nucleic Acids *18*, 16–23.

48. Basith, S., Manavalan, B., Shin, T.H., and Lee, G. (2019). SDM6A: A Web-Based Integrative Machine-Learning Framework for Predicting 6mA Sites in the Rice Genome. Mol. Ther. Nucleic Acids *18*, 131–141.

49. Manavalan, B., Basith, S., Shin, T.H., Lee, D.Y., Wei, L., and Lee, G. (2019). 4mCpred-EL: An Ensemble Learning Framework for Identification of DNA N^4-methylcytosine Sites in the Mouse Genome. Cells *8*, 1332.

50. Manavalan, B., Basith, S., Shin, T.H., Wei, L., and Lee, G. (2019). Meta-4mCpred: A Sequence-Based Meta-Predictor for Accurate DNA 4mC Site Prediction Using Effective Feature Representation. Mol. Ther. Nucleic Acids *16*, 733–744.

51. Yu, L., Zhao, J., and Gao, L. (2017). Drug repositioning based on triangularly balanced structure for tissue-specific diseases in incomplete interactome. Artif. Intell. Med. *77*, 53–63.

52. Tang, H., Cao, R.Z., Wang, W., Liu, T.S., Wang, L.M., and He, C.M. (2017). A two-step discriminated method to identify thermophilic proteins. Int. J. Biomath. *10*, 1750050.

53. Stephenson, N., Shane, E., Chase, J., Rowland, J., Ries, D., Justice, N., Zhang, J., Chan, L., and Cao, R. (2019). Survey of Machine Learning Techniques in Drug Discovery. Curr. Drug Metab. *20*, 185–193.

54. Charoenkwan, P., Schaduangrat, N., Nantasenamat, C., Piacham, T., and Shoombuatong, W. (2020). Correction: Shoombuatong, W., et al. iQSP: A Sequence-Based Tool for the Prediction and Analysis of Quorum Sensing Peptides via Chou's 5-Steps Rule and Informative Physicochemical Properties. *Int. J. Mol. Sci.* 2020, *21*, 75. Int. J. Mol. Sci. *21*, 75.

55. Chao, L., Wei, L., and Zou, Q. (2019). SecProMTB: A SVM-based Classifier for Secretory Proteins of Mycobacterium tuberculosis with Imbalanced Data Set. Proteomics *19*, e1900007.

56. Zhang, N., Yu, S., Guo, Y., Wang, L., Wang, P., and Feng, Y. (2018). Discriminating Ramos and Jurkat Cells with Image Textures from Diffraction Imaging Flow Cytometry Based on a Support Vector Machine. Curr. Bioinform. *13*, 50–56.

57. Wang, Y., Shi, F.Q., Cao, L.Y., Dey, N., Wu, Q., Ashour, A.S., et al. (2019). Morphological Segmentation Analysis and Texture-based Support Vector Machines Classification on Mice Liver Fibrosis Microscopic Images. Curr. Bioinform. *14*, 282–294.

58. Yuan, J., Zhang, Y., Liu, H., Tian, Z., Li, X., Zheng, Y., Gao, Q., Song, L., Xiao, X., Sun, J., et al. (2018). Clinical Observation of Patients with Leber's Hereditary Optic Neuropathy Before Gene Therapy. Curr. Gene Ther. *18*, 386–392.

59. Yu, L., Su, R., Wang, B., Zhang, L., Zou, Y., Zhang, J., and Gao, L. (2017). Prediction of Novel Drugs for Hepatocellular Carcinoma Based on Multi-Source Random Walk. IEEE/ACM Trans. Comput. Biol. Bioinformatics *14*, 966–977.

60. Cao, R., Wang, Z., Wang, Y., and Cheng, J. (2014). SMOQ: a tool for predicting the absolute residue-specific quality of a single protein model with support vector machines. BMC Bioinformatics *15*, 120.

61. Jiang, Q., Wang, G., Jin, S., Li, Y., and Wang, Y. (2013). Predicting human microRNA-disease associations based on support vector machine. Int. J. Data Min. Bioinform. *8*, 282–293.

62. Zhao, Y., Wang, F., and Juan, L. (2015). MicroRNA Promoter Identification in Arabidopsis Using Multiple Histone Markers. BioMed Res. Int. *2015*, 861402.

63. 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.

64. Zhong, J., Sun, Y., Peng, W., Xie, M., Yang, J., and Tang, X. (2018). XGBFEMF: An XGBoost-Based Framework for Essential Protein Prediction. IEEE Trans. Nanobioscience *17*, 243–250.

65. Basith, S., Manavalan, B., Hwan Shin, T., and Lee, G. (2020). Machine intelligence in peptide therapeutics: A next-generation tool for rapid disease screening. Med. Res. Rev. Published online January 10, 2020. https://doi.org/10.1002/med.21658.

66. Wang, G., Wang, Y., Teng, M., Zhang, D., Li, L., and Liu, Y. (2010). Signal transducers and activators of transcription-1 (STAT1) regulates microRNA transcription in interferon gamma-stimulated HeLa cells. PLoS ONE *5*, e11794.

67. Chen, W., Feng, P., Liu, T., and Jin, D. (2019). Recent Advances in Machine Learning Methods for Predicting Heat Shock Proteins. Curr. Drug Metab. *20*, 224–228.

68. Liu, B., Han, L., Liu, X., Wu, J., and Ma, Q. (2019). Computational Prediction of Sigma-54 Promoters in Bacterial Genomes by Integrating Motif Finding and Machine Learning Strategies. IEEE/ACM Trans. Comput. Biol. Bioinformatics *16*, 1211–1218.