A deep learning-based method for the prediction of DNA interacting residues in a protein

Sumeet Patiyal, Anjali Dhall and Gajendra P. S. Raghava [ID]

Corresponding author: Gajendra P. S. Raghava, Department of Computational Biology, Indraprastha Institute of Information Technology, Delhi, A-302 (R&D Block), Okhla Industrial Estate, Phase III (Near Govind Puri Metro Station), New Delhi 110020, India. Tel.: +011-26907444; E-mail: raghava@iiitd.ac.in

Abstract

DNA–protein interaction is one of the most crucial interactions in the biological system, which decides the fate of many processes such as transcription, regulation and splicing of genes. In this study, we trained our models on a training dataset of 646 DNA-binding proteins having 15 636 DNA interacting and 298 503 non-interacting residues. Our trained models were evaluated on an independent dataset of 46 DNA-binding proteins having 965 DNA interacting and 9911 non-interacting residues. All proteins in the independent dataset have less than 30% of sequence similarity with proteins in the training dataset. A wide range of traditional machine learning and deep learning (1D-CNN) techniques-based models have been developed using binary, physicochemical properties and Position-Specific Scoring Matrix (PSSM)/evolutionary profiles. In the case of machine learning technique, eXtreme Gradient Boosting-based model achieved a maximum area under the receiver operating characteristics (AUROC) curve of 0.77 on the independent dataset using PSSM profile. Deep learning-based model achieved the highest AUROC of 0.79 on the independent dataset using a combination of all three profiles. We evaluated the performance of existing methods on the independent dataset and observed that our proposed method outperformed all the existing methods. In order to facilitate scientific community, we developed standalone software and web server, which are accessible from https://webs.iiitd.edu.in/raghava/dbpred.

Keywords: DNA-binding residues, deep learning, 1D-CNN, machine learning, evolutionary profiles

Introduction

In every living organism, life is entirely dependent on molecular interactions, such as DNA–protein, RNA–protein and protein–protein interactions. These interactions perform several biological functions in the cells of living organisms [1]. DNA–protein interactions play a crucial role in a wide range of biological activities that include transcription, gene expression regulation and splicing [2–5]. Several experimental methods are used to confirm the interactions between protein and DNA. The availability of experimental data on 3D structures of protein–DNA complexes supports researchers to reveal the essential knowledge on protein–DNA interactions. These tertiary structures of protein–DNA complexes are essential to understand conformational changes of DNA molecules, the importance of hydrogen bonds, amino acid properties, electrostatic interaction, van der Waals interaction, etc. [6–15]. Due to advancement in sequencing technology, DNA-binding proteins with amino acid sequences are growing with an exponential rate over the years. Unfortunately, due to limitations of structure determination techniques, only a fraction of protein–DNA complex structure has been deposited in

Protein Data Bank (PDB) [16]. AlphaFold is the recent advancement in the field of protein structure prediction, which allows to predict the tertiary structure of a protein with high accuracy [17].

In the last few decades, several attempts have been made for the prediction of DNA-binding residues using computational methods [2, 18–20]. Broadly, these tools can be divided into four major categories, i.e. sequence-based methods [21], structure-based methods [22, 23], evolutionary methods [24] and hybrid methods which used both structure and sequence information [25, 26]. The comprehensive information of all the available methods and tools is provided in Table 1. Most of the earlier methods have been trained on a limited number of protein–DNA complex structures that include BindN [19], BindN+ [27], BindN-RF [28], MetaDBSite, CNNsite [29], DP-Bind [21], SVMnuc and NucBind [30]. Recently, methods have been trained on a large dataset of protein–DNA complexes that include HybridNAP [31], DRNApred [32], ProNA2020 [33], GraphBind [34] and GraphSite [35]. Despite tremendous advancement in the field over the years, the performance of DNA-binding residues prediction methods is far from satisfactory. Thus, there is

Sumeet Patiyal is currently working as PhD in computational biology at the Department of Computational Biology, Indraprastha Institute of Information Technology, New Delhi, India.
Anjali Dhall is currently working as PhD in computational biology at the Department of Computational Biology, Indraprastha Institute of Information Technology, New Delhi, India.
Gajendra P. S. Raghava is currently working as a professor and head of the Department of Computational Biology, Indraprastha Institute of Information Technology, New Delhi, India.

Table 1. List of tools/software developed for the prediction of DNA-interacting residues

Tool	Year	Description (link/standalone)	Dataset	Redundancy	Functional
DBS-Pred [36]	2004	Neural network-based method http://www.netasa.org/dbs-pred/	PDNA-62	25%	NO
DBS-PSSM [37]	2005	PSSM-based prediction method http://www.netasa.org/dbs-pssm/	PDNA-62	25%	NO
Pro-DNA [38]	2005	Structure-based prediction method http://proteomics.bioengr.uic.edu/pro-dna	115 protein-DNA complexes	–	NO
BindN [19]	2006	SVM based DNA/RNA-binding site prediction http://bioinformatics.ksu.edu/bindn/	PDNA-62	25%	NO
DNABindR [39]	2006	Naïve Bayes classifier-based method http://turing.cs.iastate.edu/PredDNA/predict.html	171 DNA-binding proteins	30%	NO
DP-Bind [21]	2006, 2007	PSSM-based prediction method http://lcg.rit.albany.edu/dp-bind/	PDNA-62	25%	YES
BindN-RF [28]	2009	RF-based prediction method http://bioinfo.ggc.org/bindn-rf/	PDNA-62	25%	NO
DBindR [40]	2009	Evolutionary information-based prediction method http://www.cbi.seu.edu.cn/DBindR/DBindR.htm	DBP-374	25%	NO
BindN+ [27]	2010	PSSM-based prediction method http://bioinfo.ggc.org/bindn+/	PDNA-62	25%	NO
MetaDBSite [41]	2011	Integrative tool for the prediction http://projects.biotec.tu-dresden.de/metadbsite/ http://sysbio.zju.edu.cn/metadbsite	PDNA-316	30%	NO
DNABR [42]	2012	RF-based prediction method http://www.cbi.seu.edu.cn/DNABR/	DBP-337	25%	NO
DNABind [25]	2013	Structure-based prediction method http://mleg.cse.sc.edu/DNABind/	DS123	25%	YES
SPOT-Seq (DNA) [43]	2014	Structure-based prediction method http://sparks-lab.org	DB179	35%	YES
PDNAsite [44]	2016	SVM and ensemble learning-based prediction method http://hlt.hitsz.edu.cn:8080/PDNAsite/	PDNA-62 and PDNA-224	25%	NO
CNNsite [29]	2016	Convolutional Neural Network-based method http://hlt.hitsz.edu.cn:8080/CNNsite/	PDNA-62 and PDNA-224	25%	NO
TargetDNA [45]	2017	Evolutionary information-based prediction method http://csbio.njust.edu.cn/bioinf/TargetDNA/	PFNA-543	30%	YES
HybridNAP [31]	2017	DNA-, RNA-, protein-binding residue prediction method http://biomine.cs.vcu.edu/servers/hybridNAP/	817 DNA-binding proteins	–	YES
funDNApred [46]	2018	Fuzzy cognitive map prediction model http://biomine.cs.vcu.edu/servers/funDNApred/	817 DNA-binding proteins	–	YES
iProDNA-CapsNet [47]	2019	Neural network-based prediction method https://github.com/ngphubinh/iProDNA-CapsNet	PDNA-543	30%	YES
DNAPred [48]	2019	Ensembled Hyperplane-Distance-based SVM http://202.119.84.36:3079/dnapred/	PDNA-543 PDNA-335	–	YES
SVMnuc & NucBind [30]	2019	Support vector machine-based ab-initio method https://yanglab.nankai.edu.cn/NucBind/	YFK16_DNA YFK17_DNA	30%	YES
ProNA2020 [33]	2020	Neural network-based prediction method www.predictprotein.org	308 DNA-binding proteins	20%	YES
NCBRPred [49]	2021	Multi-label learning framework method http://bliulab.net/NCBRPred/	YK17 YK16-3.5 YK16–5 MW15	30%	YES
GraphBind [34]	2021	Structure-based hierarchical graph neural network method http://www.csbio.sjtu.edu.cn/bioinf/GraphBind/	573 DNA-binding proteins	30%	YES
GraphSite [35]	2022	AlphaFold2-based prediction using graph transformer method https://biomed.nscc-gz.cn/apps/GraphSite	573 DNA-binding proteins	30%	YES

a need to develop methods for predicting DNA interacting residues in a protein with high precision using sequence information.

In order to facilitate scientific community and complement existing methods, a new method has been proposed for predicting DNA-binding residues with high accuracy.

In this study, we created two datasets: one for training and the other for external or independent validation, which we called training dataset and independent dataset, respectively. In order to provide unbiased evaluation, we remove redundant protein between training and independent datasets. No protein in the independent dataset has more than 30% similarity with any protein in a training dataset. All models have been trained on the training dataset, including optimization of parameters of machine learning and deep learning techniques. We optimized parameters of machine/deep learning techniques using a 5-fold cross-validation technique. The final model has been evaluated on independent dataset, in order to avoid any biasness in evaluation. In other words, the independent dataset has not been used to train or tune model parameters.

Materials and methods
Dataset creation
We have downloaded the dataset from the hybridNAP webserver [31] and recently published article ProNA2020 [33], which consists of 864 and 308 annotated protein sequences. In order to remove redundancy, we implemented CD-HIT software [50] on these datasets. In order to compare with the hybridNAP and ProNA2020, we have used their training dataset to train the model and independent dataset to evaluate the performance of the final model, after removing the redundant protein sequences using CD-HIT. No protein in independent dataset has more than 30% similarity with any protein in the training dataset. Our final training dataset contains 646 protein sequences and independent dataset contains 46 protein sequences. Finally, we got 15 636 DNA-interacting and 298 503 non-interacting residues in the training dataset and 965 interacting and 9911 non-interacting residues in the independent dataset.

Pattern size
The overlapping patterns for each sequence of length 17 are generated using in-house python scripts. The central or ninth residue is taken as the representative of the pattern. The pattern is specified as a positive or interacting pattern if the central residue is DNA-interacting, else pattern is specified as non-interacting or negative pattern. In order to handle the terminal residues, eight counterfeit residues using the formula $(N-1)/2$ (where N represents the pattern length which is 17), as 'X' are added at both sides of the protein sequences, as shown in Figure 1 along with the complete workflow for this study.

Compositional analysis
In order to understand the nature of residues involved in DNA interaction, we have calculated the amino acid composition, residue propensity and physicochemical properties-based composition. The percent amino acid composition was calculated to understand the abundance of residues in DNA interaction. The

residues propensity is computed to understand the preference of particular type of residues in the DNA-binding site. The functionality of residues is based on their sole physicochemical properties, and hence we have determined physicochemical properties-based composition for 25 distinct properties. All composition properties are computed using Pfeature package [51].

Profile of pattern
In order to provide numerical representation of 17 residue patterns, we compute two types of profiles corresponding to patterns: (i) binary profile based on residues and (ii) physicochemical property profile based on residue properties. The profile-based features were calculated by modifying Pfeature [51] scripts. In the case of binary profile, each amino acid is represented by the vector size of length 21; for instance, A is described as 1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0, which comprises 20 natural amino acids and 1 dummy variable, whereas X is denoted as 0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1 [52]. Therefore, each pattern is represented by the vector size of 357 (17 * 21). In physicochemical property profile, each amino acid is designated by the vector of size 25; for instance, A is denoted by 0,0,1,0,1,1,0,0,0,0,1,1,0,0,0,0,1,0, 0,1,0,0,1,1,0, where each position exhibits a particular physicochemical property, and each element denotes the presence (1) or absence (0) of that property. Therefore, the resulting vector for each property is of length 425 (17 * 25), whereas for X, all the elements are 0.

PSSM profile
The third feature that we have used in this study is the evolutionary information. In order to compute evolutionary information, we generate Position-Specific Scoring Matrix (PSSM) profile corresponding to each protein [53]. The PSSM profile was generated by employing PSI-BLAST [54], where each protein is searched against proteins in SwissProt database [55]. The parameters used for running PSI-BLAST were three iterations, with e-value as $1e$-3. Further, the profile was normalized using Equation (1). The final matrix for each sequence is of dimension $N \times 21$, where N is the length of the protein sequence, and each pattern is depicted as the vector of length 357 (17 * 21).

$$\text{PSSM}_N = \frac{1}{1 + e^{-x}} \tag{1}$$

where PSSM_N is the normalized value and x is the PSSM score.

Machine/deep learning-based predictors
We implement python library scikit-learn for developing predictors based on traditional machine learning techniques. In order to develop deep learning (one-dimensional CNN-based classifier) based predictors, we used python library TensorFlow. In the case of machine learning techniques, we have implemented conventional classifiers, such as Random Forest (RF),
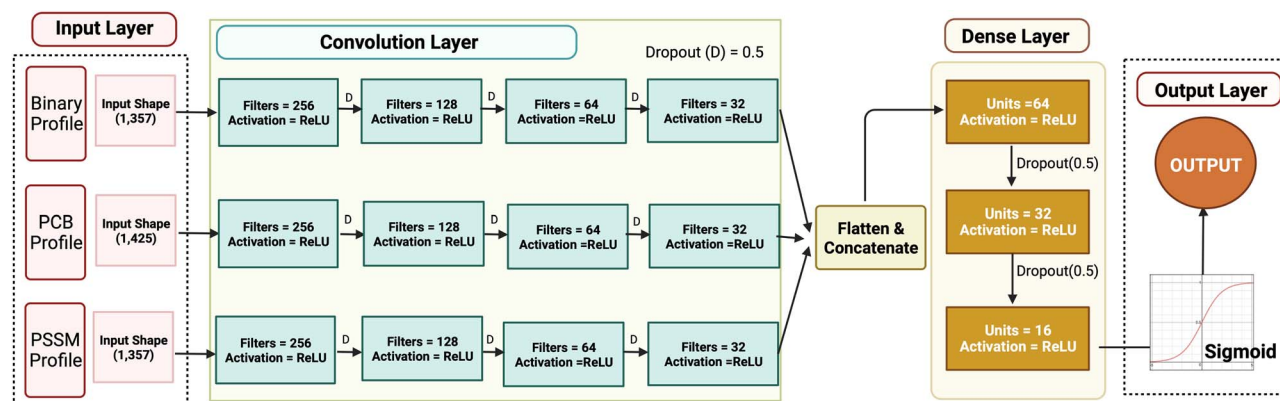
Figure 1. A comprehensive workflow for feature generation (**A**) and model development (**B**). The following steps were taken to generate different profiles from sequence: (a) generation of fixed length patterns from a sequence, (b) binary profile from pattern, (c) generation of physicochemical properties profile and (d) PSSM profile. Overall algorithm for predicting DNA-binding residues is shown in **B**.

Decision Tree (DT), eXtreme Gradient Boosting (XGB), Logistic Regression (LR) and Gaussian Naive Bayes (GNB) to develop the prediction model.

ID-CNN model architecture

In this study, we implemented standard CNN architecture for developing prediction models. It was implemented using Python library Keras, which is based on TensorFlow. Overall architecture of our hybrid model implemented in this study is shown in Figure 2. As shown in Figure 2, each branch have four convolutional layers, first layer used 256 filters. It means input features are represented by 256 filters using first layers; these features are reduced to half in each layer. As shown in Figure 2, final or fourth layer will provide 32 features. Finally, we flattened all these vectors, concatenated them and used them as a feature vector. Instead of passing the entire vector directly for the sake of classification, we passed it through the densely connected neural network layers to capture the importance of each feature for the classification task. We have used the ReLU activation function for each hidden layer because of its simplicity and effectiveness [56, 57]. In the final layer, we have used the sigmoid function to get the values between 0 and 1, which were further employed to find optimal threshold that can provide balanced sensitivity and specificity.

Training of models using 5-fold cross-validation

All our models were trained on the training dataset, where parameters of models have been optimized.

In order to avoid overfitting and biasness, we have implemented the 5-fold cross-validation. In 5-fold cross-validation, dataset is divided into five non-overlapping sets, four out of five sets are used for the training, and the fifth set is kept for testing. The same process is repeated five times so that each set gets the chance to be used for testing. The overall performance would be the mean of the performances of five iterations [58–60]. This five-fold cross-validation technique is performed on the training dataset only to optimize parameters of our models. Best model on training dataset is used for the final evaluation on independent dataset.

We have hyper-tuned the parameters at three levels, such as layers that included the number of filters in each convolution layer, size of the convolution filters, number of dense layers and number of neurons in each layer; functions that included loss function, type of optimizer and activation function for different layers; and rates such as the rate of dropout and learning.

Performance on independent dataset

Though we used a 5-fold cross-validation technique for optimizing the parameters of machine/deep learning-based models, still biasness in performance or overoptimization cannot be ruled out. In order to provide unbiased evaluation of newly developed models, we evaluate the performance of our final models on independent dataset. As independent dataset has no similarity with training dataset, so performance is unbiased. In addition, previously existing methods are also evaluated on the independent dataset.

Figure 2. Overall architecture of hybrid model implemented in this study using 1D-CNN.

Evaluation parameters

In this study, we have calculated various threshold-dependent and threshold-independent parameters in order to evaluate prediction models. Threshold-dependent parameters include sensitivity (sens, equation 2), which signifies the percentage of correctly predicted DNA-interacting residues; specificity (spec, equation 3), which explains the proportion of correctly predicted DNA non-interacting residues; accuracy (acc, Equation 4), which defines the percentage of correctly predicted DNA-interacting and non-interacting residues; and Matthews correlation coefficient (MCC, Equation 5), which exhibits the correlation between observed and predicted values. On the other hand, the threshold-independent parameter includes area under the receiver operating characteristics (AUROC), which is the plot between true positive rate (TPR) and false positive rate (FPR). The module of the R named 'pROC' was used to plot the AUROC curve [61]. The equations for threshold-dependent parameters are as follows:

$$\text{Sensitivity} = \frac{TP}{TP + FN} \text{ X } 100 \qquad (2)$$

$$\text{Specificity} = \frac{TN}{TN + FP} \text{ X } 100 \qquad (3)$$

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \text{ X } 100 \qquad (4)$$

$$\text{MCC} = \frac{(TP * TN) - (FP * FN)}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \qquad (5)$$

Where FP, FN, TP and TN are false positive, false negative, true positive and true negative, respectively.

Results
Compositional analysis

We have analyzed the amino acid composition of DNA-interacting and non-interacting residues in DNA-interacting proteins. As shown in Figure 3, DNA-interacting residues are rich in H, K, N, R, Y, whereas

A, D, I, L, P are sparse. It means residues like H, K, N, R, Y are preferred and residues like A, D, I, L, P are not preferred in DNA interactions. We have calculated the propensity of each residue, which exhibits that K, R, W and Y are most favored in the DNA-binding sites, as shown in Figure 4. We have also analyzed the residues' properties involved in interaction with DNA and found that positively charged, basic, hydrophilic, possessing helix secondary structure, and large are more abundant in DNA-interacting residues shown in Figure 5.

Binary profile-based models

In order to develop the prediction models, we have generated binary profile, as it captures the compositional as well as positional information of each residue. We have generated the binary profile for the training dataset consisting of 15 636 patterns for DNA-interacting and 298 503 non-interacting patterns; and the independent dataset comprises 965 DNA-interacting and 9911 non-interacting patterns. The best result for each classifier on the independent dataset is shown in Table 1, where performances on training dataset are provided in Supplementary Table S1 available online at http://bib.oxfordjournals.org/. As shown in Table 2, the logistic regression-based model and one-dimensional CNN-based classifier (1D-CNN) obtain an AUROC of 0.74 with MCC of 0.21 on the independent dataset. It means that machine learning-based models and 1D-CNN-based models developed using binary profile have similar performance. It is important to note that 1D-CNN achieved a high-performance AUROC of 0.83 on the training dataset (Supplementary Table S1 available online at http://bib.oxfordjournals.org/). This shows that the performance of 1D-CNN is highly overoptimized on the training dataset, despite we used 5-fold cross-validation. Thus, it is important to evaluate these models on the independent dataset, which is not used for optimizing parameters.

Physicochemical property profile-based models

We have also used the binary profiles based on physicochemical properties for the first time in the literature to
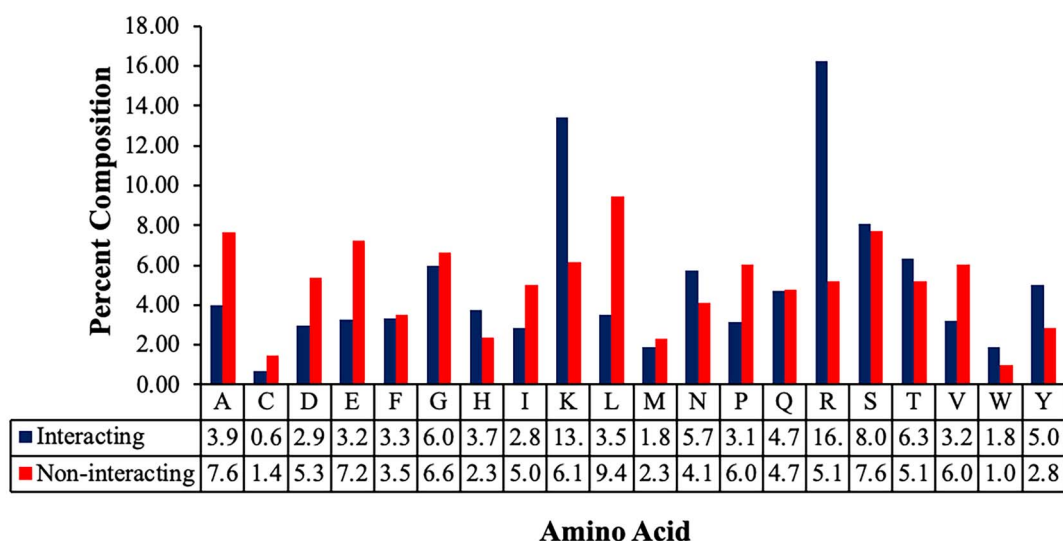
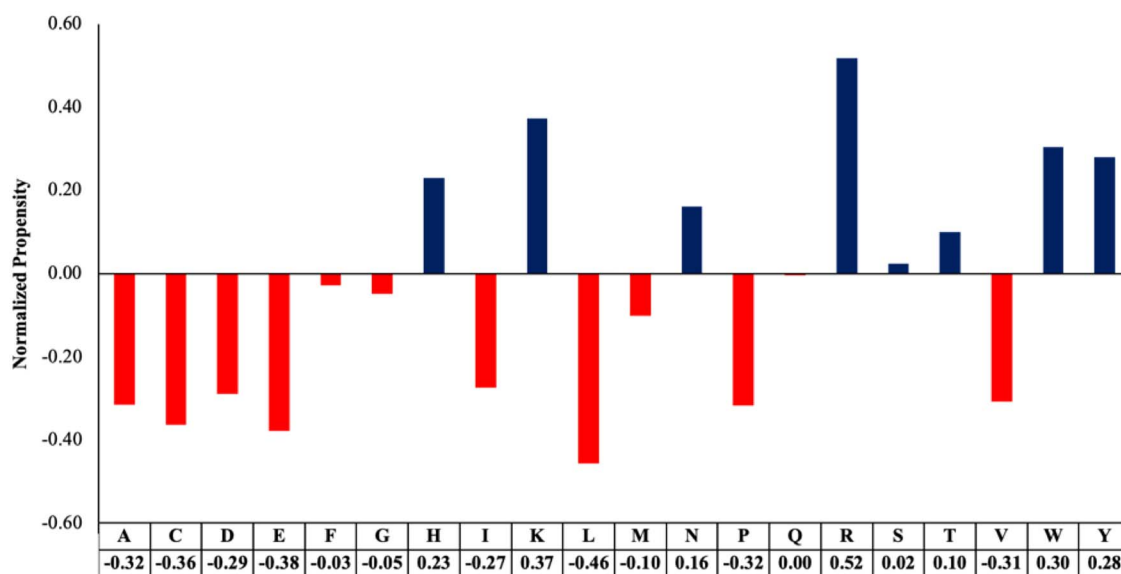**Figure 3.** Percent composition of DNA-interacting and non-interacting residues.

Figure 4. Normalized propensity scores for DNA-interacting and non-interacting residues.

Table 2. The performance of various classifiers on independent dataset developed using binary profile

Classifier	Sensitivity	Specificity	Accuracy	AUROC	MCC
DT	12.62	92.81	85.61	0.53	0.06
RF	67.05	65.29	65.45	0.72	0.19
LR	68.19	66.59	66.73	0.74	0.21
XGB	67.15	68.17	68.08	0.73	0.21
GNB	66.22	63.19	63.46	0.70	0.17
1D-CNN	70.67	66.54	66.00	0.74	0.21

DT, decision tree; RF, random forest; LR, logistic regression; XGB, eXtreme Gradient Boosting; GNB, Gaussian Naive Bayes; 1D-CNN, one-dimensional convolutional neural network; AUROC, Area under the receiver operating characteristic curve; MCC, Matthews correlation coefficient.

develop the prediction models. As shown in Table 3, 1D-CNN and logistic regression-based models acquire nearly similar performance (AUROC 0.73) on independent dataset. Performances of each classifier on the training dataset are reported in Supplementary Table S2 available online at http://bib.oxfordjournals.org/. The performance of models based on physicochemical properties profiles acquires nearly same performance as we got in case of binary profile-based models. We found similar trend here also, 1D-CNN achieved high performance
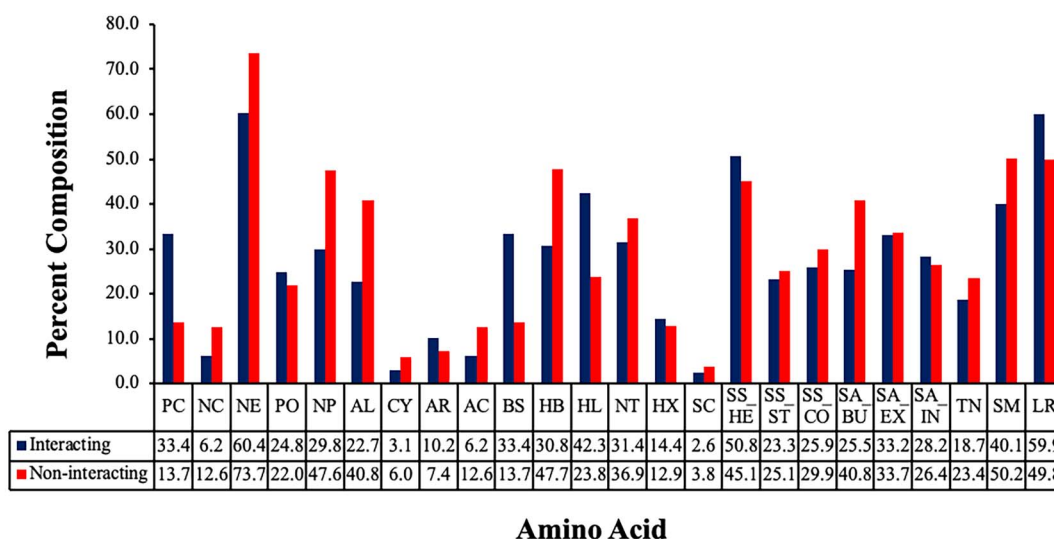
Figure 5. Percent composition of physicochemical properties in DNA-interacting and non-interacting residues.

Table 3. The performance of various classifiers on independent dataset developed using physicochemical properties profile

| Classifier	Sensitivity	Specificity	Accuracy	AUROC	MCC
DT	09.32	94.82	87.24	0.52	0.05
RF	63.11	63.67	63.62	0.69	0.16
LR	68.39	66.50	66.67	0.73	0.21
XGB	63.32	68.98	68.48	0.72	0.19
GNB	67.46	58.87	59.64	0.68	0.15
1D-CNN	67.08	67.86	67.79	0.73	0.20

DT, decision tree; RF, random forest; LR, logistic regression; XGB, eXtreme Gradient Boosting; GNB, Gaussian Naive Bayes; 1D-CNN, one-dimensional convolutional neural network; AUROC, area under the receiver operating characteristic curve; MCC, Matthews correlation coefficient.

(AUROC 0.86) in comparison with machine learning techniques on the training dataset (Supplementary Table S2 available online at http://bib.oxfordjournals.org/).

PSSM profile-based models

It has been shown in number of studies that evolutionary information of a protein provides more information than amino acid sequence. In order to capture evolutionary information of proteins, we generate PSSM profile for each protein. We have developed various prediction models by using normalized PSSM profile as the input feature, and the performance of each classifier on the independent dataset is exhibited in Table 4. Supplementary Table S3 available online at http://bib.oxfordjournals.org/ comprises different evaluation parameters evaluated on the training dataset for each classifier. As shown in Table 3, our machine learning-based (XGB) achieved a maximum AUROC of 0.77 on the independent dataset, which is better than models developed using 1D-CNN-based classifier exceeded other classifiers' performance with AUROC of 0.74 and MCC of 0.21 for the independent dataset.

Performance based on combined features

The combined features were generated by concatenating the amino acid binary profile, physicochemical property-based binary profile and PSSM profile in the column-wise manner for each pattern, which generated a vector of length 1175. A wide range of machine learning-based classifiers have been implemented to develop prediction methods. As shown in Table 5, we got maximum AUCROC 0.77 using LR, which is same as we got in case of PSSM only. It means our machine learning-based classifiers unable to capture more information from additional information. It is interesting to note that our 1D-CNN-based classifier achieved a maximum AUROC of 0.79 on the independent dataset. It means 1D-CNN able to capture additional information. Performances for each classifier for the training dataset are provided in Supplementary Table S4 available online at http://bib.oxfordjournals.org/. It is clear from these results that deep learning-based classifiers perform better than machine learning-based classifiers in case of additional features. In order to check the performance of the model at protein level, we have predicted the interaction on each protein of independent dataset and calculated the performance measures. Finally, we have computed the average ± standard deviation for each measure and we obtained sensitivity of 69.45 ± 2.62, specificity of 74.87 ± 4.61, accuracy of 73.25 ± 4.32, AUROC of 0.78 ± 0.021 and MCC of 0.34 ± 0.026 on the independent dataset.

Table 4. The performance of various classifiers developed using PSSM profile on independent dataset

Classifier	Sensitivity	Specificity	Accuracy	AUROC	MCC
DT	13.26	94.47	87.27	0.54	0.09
RF	73.06	62.46	63.41	0.74	0.21
LR	69.33	67.98	68.10	0.75	0.22
XGB	72.12	67.51	67.92	0.77	0.24
GNB	64.87	56.24	57.91	0.63	0.12
1D-CNN	64.89	69.97	69.43	0.74	0.21

DT, decision tree; RF, random forest; LR, logistic regression; XGB, eXtreme Gradient Boosting; GNB, Gaussian Naive Bayes; 1D-CNN, one-dimensional convolutional neural network; AUROC, area under the receiver operating characteristic curve; MCC, Matthews correlation coefficient.

Table 5. The performance of various classifiers on independent dataset developed using combined features

Classifier	Sensitivity	Specificity	Accuracy	AUROC	MCC
DT	15.34	94.91	87.84	0.55	0.18
RF	70.98	63.02	63.73	0.75	0.20
LR	70.88	69.18	69.33	0.77	0.29
XGB	69.95	62.62	63.27	0.72	0.19
GNB	66.84	62.70	63.06	0.70	0.17
1D-CNN	70.78	78.40	77.72	0.79	0.32

DT, decision tree; RF, random forest; LR, logistic regression; XGB, eXtreme Gradient Boosting; GNB, Gaussian Naive Bayes; 1D-CNN, one-dimensional convolutional neural network; AUROC, area under the receiver operating characteristic curve; MCC, Matthews correlation coefficient.

Comparison with the existing methods

In order to concede the newly developed method, its comparison with the existing methods is of uttermost importance. The comparison conveys the merits and demerits of the newly developed method. Since there are many existing methods for predicting DNA-binding residues in a protein [27, 28, 31, 33], a comprehensive comparison is must to understand the benefits of the newly developed method 'DBPred'. In order to provide an unbiased comparison, we evaluated the performance of existing methods and the proposed method on independent dataset of 46 proteins used in this study. The performances of all methods are reported in terms of sensitivity, specificity, AUROC, accuracy and MCC in Table 6. Among existing methods, DRNAPred [32] achieved a maximum AUROC of 0.75 and MCC of 0.22, whereas SVMnuc, NucBind [30], DNAPred [48], DNAbindR [39] and ProNA2020 [33] achieved equivalent performance in terms of MCC and AUROC. As shown in Table 6, our method DBPred outperformed the existing methods with an AUROC of 0.79 and MCC of 0.32 on the independent dataset. The comparison between the AUROC of the existing methods has been shown in ROC curve (Figure 6). We are unable to compare our methods with few methods, which are either non-function methods or whose webserver/standalone software is not available.

Facilities to scientific community

In order to serve the scientific community, we have developed a webserver DBPred, to predict the DNA-interacting residues in a protein using its primary structure information. The facilities provided by the webserver are available in various modules (as shown in Figure 7) such as 'Sequence', 'PSSM profile', 'Hybrid' and 'Standalone'. Sequence module allows the user to predict the DNA-interacting residues using the binary
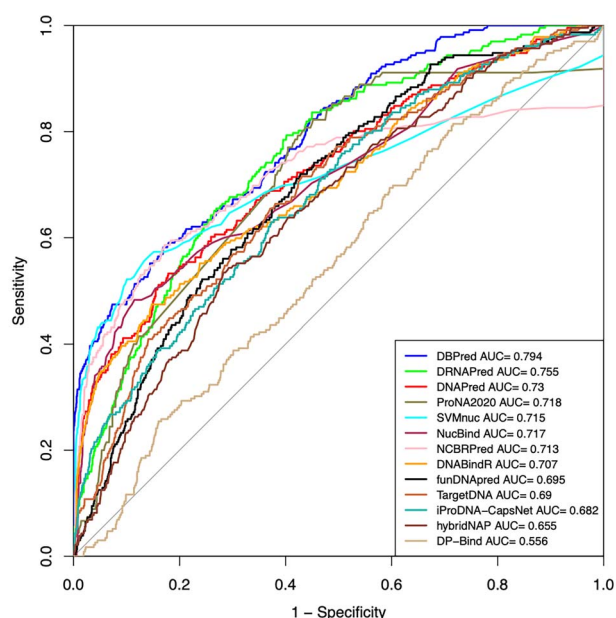
Figure 6. AUROC plots obtained for existing methods using independent dataset.

and physicochemical profiles. The PSSM module allowed to predict DNA-interacting residues in a protein using evolutionary information. Hybrid module is based on hybrid features (binary, physicochemical properties, PSSM profile) for predicting the DNA-interacting residues. In addition, a standalone software package is also available to run on local machine of users. This web server DBPred is compatible with smart devices such as iPhone, iPad, laptops, and android mobile phones.

Discussion and Conclusion

Biological interactions between proteins and DNA are very crucial to understand several aspects of biological

Table 6. The performance of existing methods and proposed method on independent dataset

Method	Year	Sensitivity	Specificity	AUROC	Accuracy	MCC
DNABindR [39]	2006	52.16	78.09	0.71	75.80%	0.20
DP-Bind [20]	2007	47.41	71.14	0.56	69.06%	0.11
DRNAPred [32]	2017	67.67	69.19	0.75	69.06%	0.22
TargetDNA [41]	2017	48.71	77.52	0.69	74.98%	0.17
HybridNAP [28]	2017	38.79	79.58	0.66	75.99%	0.13
funDNApred [42]	2018	62.93	63.70	0.69	63.70%	0.16
DNAPred [48]	2019	67.10	65.50	0.73	65.64%	0.19
SVMnuc [30]	2019	66.81	66.57	0.72	66.59%	0.20
NucBind [30]	2019	62.50	64.86	0.72	64.66%	0.16
iProDNA-CapsNet [43]	2019	63.79	61.28	0.68	61.28%	0.14
ProNA2020[a] [30]	2020	42.22	76.28	0.72	74.31%	0.22
NCBRPred[a] [49]	2021	67.67	67.44	0.71	67.46%	0.21
DBPred[a]	2022	70.78	78.40	0.79	77.72%	0.32

[a]Standalone is also available.
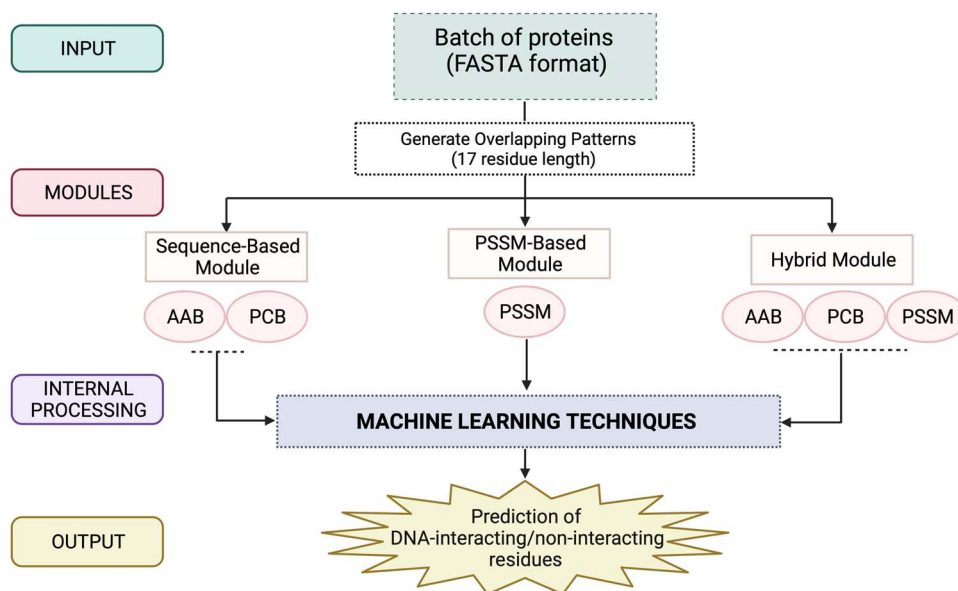
Figure 7. Flowchart representing the processing of input data using three different modules of DBPred server for the prediction of DNA-interacting residues.

processes such as transcription, translation and gene regulation [62]. The comprehensive understanding of interacting residues is an important aspect in the designing of novel drugs [63, 64]. Of note, the interacting residues can only be extracted through the three-dimensional information of protein. PDB database reports an ample of experimentally verified protein structures identified using X-ray crystallography and NMR technology [65, 66]. Studies have reported that the 3D information of protein-binding residues are helpful in structure-based drug designing [67], where one can understand the interaction of drug molecules with the DNA-interacting residues [68–70]. Therefore, in the past few decades, a number of researchers have been working very hard to understand the physical interaction between DNA and protein molecules. Several computational methods have also been developed by researchers to predict the DNA-interacting sites on protein, which can be classified into three classes such as sequence-based, structure-based and hybrid approaches [24, 71].

However, the major limitation of the structure-based or hybrid methods is their dependency on the protein structural information, which limits their application, as determination of the protein structure is a costly, time-consuming and very complex process [52]. On the other hand, sequence information in various databases is growing exponentially, enhancing the application of sequence-based methods with reliable performance.

In the last few years, a number of computational methods have been developed for the prediction of DNA-interacting residues. However, the datasets used in most of the studies are very small and performance on independent dataset is poor. Therefore, this is a need of the hour to develop a new method using the largest dataset for the prediction of DNA-interacting residues using protein sequences. In the current study, we have made a systematic attempt to develop a prediction method using the latest benchmark dataset of ProNA2020 and hybridNAP. We have a total of 15 636 DNA-interacting and 298 503 non-interacting residues in the training dataset

and 965 interacting and 9911 non-interacting residues in the independent dataset. It was observed that certain residues like lysine, arginine and tyrosine are more frequent in DNA interaction. Most of the DNA-interacting residues possess positively charged, basic, hydrophilic residues and helix secondary structure properties.

In this study, we developed prediction models using a wide range of machine learning techniques. As shown in Results section, models based on evolutionary information perform better than the binary and physicochemical property-based models. Our PSSM-based models got a maximum AUROC of 0.77 with MCC of 0.24 (Table 4). This is expected and agrees with the previous studies, where it has been demonstrated that evolutionary information provide more information than single sequence. In order to improve the performance of our model, we developed models using hybrid features, which combine all three types of features (binary, physicochemical and PSSM). The maximum performance of hybrid feature-based model was AUROC 0.77, which is same as we achieved in the case of PSSM-based model. It means that we are unable to combine different type of feature effectively. In this study, we used 1D-CNN-based model using different types of features and got a maximum AUROC of 0.74 using PSSM features. The performance of our 1D-CNN-based model increases significantly from AUROC 0.74 to 0.79, when we used hybrid features. It means our 1D-CNN-based model predicts DNA interacting residues successfully. It is important to understand the reason of the success behind 1D-CNN prediction of DNA-binding residues. It is a known fact that CNN (basically 2D-CNN) is the most successful method in image classification as it identifies local objects. In prediction of DNA interaction, we are using string of 17 characters (17 residues) for prediction. Thus, in this study, we used 1D-CNN, which identifies motifs in pattern and the effect of these patterns on central residue. One of the challenges in machine learning methods is a combination of different types of features, as these features have different magnitude and characteristics. As shown in Figure 2, 1D-CNN applies filters and generates a fixed number of features for each type of features. Finally, convolutional layers generate 32 features of each type of filters; these features are combined to develop models. It means that 1D-CNN is suitable for combining different types of features in the prediction of DNA interactions. We anticipate that this method can be an efficient tool for correctly predicting DNA-interacting residues in a protein sequence. To serve the scientific community, we have provided a standalone package and web server 'DBPred' to assist biologists in the finding of DNA-interacting residues for the sake of annotation and functional analysis. DBPred is freely available and accessible on https://webs.iiitd.edu.in/raghava/dbpred/ and python-based/docker-based standalone package is available at https://webs.iiitd.edu.in/raghava/dbpred/stand.html.

Key Points

- DNA–protein interactions play vital roles in numerous biological processes.
- Understanding of interacting residues is a crucial aspect of drug designing.
- *In silico* model was developed using deep learning algorithm to predict DNA-interacting residues.
- Binary, physicochemical properties and PSSM profiles were used as input features.
- Available as web server, python- and Perl-based standalone package and Docker container.

Supplementary data

Supplementary data are available online at https://academic.oup.com/bib.

Authors' contributions

S.P. and G.P.S.R. collected and processed the datasets. S.P. and G.P.S.R. implemented the algorithms and developed the prediction models. S.P., A.D. and G.P.S.R. analyzed the results. S.P. created the back end of the web server and A.D. created the front-end user interface. S.P., A.D. and G.P.S.R. penned the manuscript. G.P.S.R. conceived and coordinated the project. All authors have read and approved the final manuscript.

Data Availability Statement

The dataset used in this study is available at "DBPred" web server, at https://webs.iiitd.edu.in/raghava/dbpred/download.php.

Ethics approval

Not applicable.

Consent to participate

Not applicable.

References

1. Emamjomeh A, Choobineh D, Hajieghrari B, *et al.* DNA-protein interaction: identification, prediction and data analysis. *Mol Biol Rep* 2019;**46**(3):3571–96.

2. Si J, Zhao R, Wu R. An overview of the prediction of protein DNA-binding sites. *Int J Mol Sci* 2015;**16**(12):5194–215.

3. Aeling KA, Steffen NR, Johnson M, *et al.* DNA deformation energy as an indirect recognition mechanism in protein-DNA interactions. *IEEE/ACM Trans Comput Biol Bioinform* 2007;**4**(1):117–25.

4. Wong KC, Li Y, Peng C, *et al.* A comparison study for DNA motif modeling on protein binding microarray. *IEEE/ACM Trans Comput Biol Bioinform* 2016;**13**(2):261–71.

5. Choi S, Han K. Prediction of RNA-binding amino acids from protein and RNA sequences. *BMC Bioinformatics* 2011;**12**(Suppl 13):S7.

6. Collas P. The current state of chromatin immunoprecipitation. *Mol Biotechnol* 2010;**45**(1):87–100.

7. Berger MF, Philippakis AA, Qureshi AM, *et al.* Compact, universal DNA microarrays to comprehensively determine transcription-factor binding site specificities. *Nat Biotechnol* 2006;**24**(11):1429–35.

8. Furlan-Magaril M, Rincon-Arano H, Recillas-Targa F. Sequential chromatin immunoprecipitation protocol: ChIP-reChIP. *Methods Mol Biol* 2009;**543**:253–66.

9. Ponting CP, Schultz J, Milpetz F, *et al.* SMART: identification and annotation of domains from signalling and extracellular protein sequences. *Nucleic Acids Res* 1999;**27**(1):229–32.

10. Jones S, van Heyningen P, Berman HM, *et al.* Protein-DNA interactions: a structural analysis. *J Mol Biol* 1999;**287**(5):877–96.

11. Ho SW, Jona G, Chen CT, *et al.* Linking DNA-binding proteins to their recognition sequences by using protein microarrays. *Proc Natl Acad Sci U S A* 2006;**103**(26):9940–5.

12. Jayaram B, McConnell K, Dixit SB, *et al.* Free-energy component analysis of 40 protein-DNA complexes: a consensus view on the thermodynamics of binding at the molecular level. *J Comput Chem* 2002;**23**(1):1–14.

13. Lejeune D, Delsaux N, Charloteaux B, *et al.* Protein-nucleic acid recognition: statistical analysis of atomic interactions and influence of DNA structure. *Proteins* 2005;**61**(2):258–71.

14. Nadassy K, Wodak SJ, Janin J. Structural features of protein-nucleic acid recognition sites. *Biochemistry* 1999;**38**(7):1999–2017.

15. Nagarajan R, Ahmad S, Gromiha MM. Novel approach for selecting the best predictor for identifying the binding sites in DNA binding proteins. *Nucleic Acids Res* 2013;**41**:7606–14.

16. Rose PW, Prlic A, Bi C, *et al.* The RCSB Protein Data Bank: views of structural biology for basic and applied research and education. *Nucleic Acids Res* 2015;**43**:D345–56.

17. Jumper J, Evans R, Pritzel A, *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* 2021;**596**(7873):583–9.

18. Schmidtke P, Barril X. Understanding and predicting druggability. A high-throughput method for detection of drug binding sites. *J Med Chem* 2010;**53**(15):5858–67.

19. Wang L, Brown SJ. BindN: a web-based tool for efficient prediction of DNA and RNA binding sites in amino acid sequences. *Nucleic Acids Res* 2006;**34**(Web Server):W243–8.

20. Miao Z, Westhof E. A large-scale assessment of nucleic acids binding site prediction programs. *PLoS Comput Biol* 2015;**11**:e1004639.

21. Hwang S, Gou Z, Kuznetsov IB. DP-Bind: a web server for sequence-based prediction of DNA-binding residues in DNA-binding proteins. *Bioinformatics* 2007;**23**:634–6.

22. Jones S, Barker JA, Nobeli I, *et al.* Using structural motif templates to identify proteins with DNA binding function. *Nucleic Acids Res* 2003;**31**(11):2811–23.

23. Tjong H, Zhou HX. DISPLAR: an accurate method for predicting DNA-binding sites on protein surfaces. *Nucleic Acids Res* 2007;**35**(5):1465–77.

24. Chowdhury SY, Shatabda S, Dehzangi A. iDNAProt-ES: Identification of DNA-binding Proteins using Evolutionary and Structural Features. *Sci Rep* 2017;**7**(1):14938.

25. Liu R, Hu J. DNABind: a hybrid algorithm for structure-based prediction of DNA-binding residues by combining machine learning- and template-based approaches. *Proteins* 2013;**81**(11):1885–99.

26. Li BQ, Feng KY, Ding J, *et al.* Predicting DNA-binding sites of proteins based on sequential and 3D structural information. *Mol Gen Genomics* 2014;**289**(3):489–99.

27. Wang L, Huang C, Yang MQ, *et al.* BindN+ for accurate prediction of DNA and RNA-binding residues from protein sequence features. *BMC Syst Biol* 2010;**4**(Suppl 1):S3.

28. Wang L, Yang MQ, Yang JY. Prediction of DNA-binding residues from protein sequence information using random forests. *BMC Genomics* 2009;**10**(Suppl 1):S1.

29. Zhou J, Lu Q, Xu R, Gui L, Wang H. CNNsite: Prediction of DNA-binding residues in proteins using Convolutional Neural Network with sequence features. *IEEE International Conference on Bioinformatics and Biomedicine (BIBM)* 2016;78–85.

30. Su H, Liu M, Sun S, *et al.* Improving the prediction of protein-nucleic acids binding residues via multiple sequence profiles and the consensus of complementary methods. *Bioinformatics* 2019;**35**(6):930–6.

31. Zhang J, Ma Z, Kurgan L. Comprehensive review and empirical analysis of hallmarks of DNA-, RNA- and protein-binding residues in protein chains. *Brief Bioinform* 2019;**20**(4):1250–68.

32. Yan J, Kurgan L. DRNApred, fast sequence-based method that accurately predicts and discriminates DNA- and RNA-binding residues. *Nucleic Acids Res* 2017;**45**:e84.

33. Qiu J, Bernhofer M, Heinzinger M, *et al.* ProNA2020 predicts protein-DNA, protein-RNA, and protein-protein binding proteins and residues from sequence. *J Mol Biol* 2020;**432**(7):2428–43.

34. Xia Y, Xia CQ, Pan X, *et al.* GraphBind: protein structural context embedded rules learned by hierarchical graph neural networks for recognizing nucleic-acid-binding residues. *Nucleic Acids Res* 2021;**49**(9):e51.

35. Yuan Q, Chen S, Rao J, *et al.* AlphaFold2-aware protein-DNA binding site prediction using graph transformer. *Brief Bioinform* 2022;**23**.

36. Ahmad S, Gromiha MM, Sarai A. Analysis and prediction of DNA-binding proteins and their binding residues based on composition, sequence and structural information. *Bioinformatics* 2004;**20**(4):477–86.

37. Ahmad S, Sarai A. PSSM-based prediction of DNA binding sites in proteins. *BMC Bioinformatics* 2005;**6**(1):33.

38. Bhardwaj N, Langlois R, Zhao G, *et al.* Structure based prediction of binding residues on DNA-binding proteins. *Conf Proc IEEE Eng Med Biol Soc* 2005;**2005**:2611–4.

39. Yan C, Terribilini M, Wu F, *et al.* Predicting DNA-binding sites of proteins from amino acid sequence. *BMC Bioinformatics* 2006;**7**(1):262.

40. Wu J, Liu H, Duan X, *et al.* Prediction of DNA-binding residues in proteins from amino acid sequences using a random forest model with a hybrid feature. *Bioinformatics* 2009;**25**(1):30–5.

41. Si J, Zhang Z, Lin B, *et al*. MetaDBSite: a meta approach to improve protein DNA-binding sites prediction. *BMC Syst Biol* 2011;**5**(Suppl 1):S7.

42. Ma X, Guo J, Liu HD, *et al*. Sequence-based prediction of DNA-binding residues in proteins with conservation and correlation information. *IEEE/ACM Trans Comput Biol Bioinform* 2012;**9**(6): 1766–75.

43. Zhao H, Wang J, Zhou Y, *et al*. Predicting DNA-binding proteins and binding residues by complex structure prediction and application to human proteome. *PLoS One* 2014;**9**(5):e96694.

44. Zhou J, Xu R, He Y, *et al*. PDNAsite: identification of DNA-binding site from protein sequence by incorporating spatial and sequence context. *Sci Rep* 2016;**6**(1):27653.

45. Hu J, Li Y, Zhang M, *et al*. Predicting protein-DNA binding residues by weightedly combining sequence-based features and boosting multiple SVMs. *IEEE/ACM Trans Comput Biol Bioinform* 2017;**14**(6):1389–98.

46. Amirkhani A, Kolahdoozi M, Wang C, *et al*. Prediction of DNA-binding residues in local segments of protein sequences with fuzzy cognitive maps. *IEEE/ACM Trans Comput Biol Bioinform* 2020;**17**(4):1372–82.

47. Nguyen BP, Nguyen QH, Doan-Ngoc GN, *et al*. iProDNA-CapsNet: identifying protein-DNA binding residues using capsule neural networks. *BMC Bioinformatics* 2019;**20**(S23):634.

48. Zhu YH, Hu J, Song XN, *et al*. DNAPred: accurate identification of DNA-binding sites from protein sequence by ensembled hyperplane-distance-based support vector machines. *J Chem Inf Model* 2019;**59**(6):3057–71.

49. Zhang J, Chen Q, Liu B. NCBRPred: predicting nucleic acid binding residues in proteins based on multilabel learning. *Brief Bioinform* 2021;**22**.

50. Huang Y, Niu B, Gao Y, *et al*. CD-HIT Suite: a web server for clustering and comparing biological sequences. *Bioinformatics* 2010;**26**(5):680–2.

51. Pande A, Patiyal S, Lathwal A, *et al*. Computing wide range of protein/peptide features from their sequence and structure. *BioRxiv* 2019;599126.

52. Patiyal S, Agrawal P, Kumar V, *et al*. NAGbinder: an approach for identifying N-acetylglucosamine interacting residues of a protein from its primary sequence. *Protein Sci* 2020;**29**(1): 201–10.

53. Chen K, Mizianty MJ, Kurgan L. Prediction and analysis of nucleotide-binding residues using sequence and sequence-derived structural descriptors. *Bioinformatics* 2012;**28**(3): 331–41.

54. Altschul SF, Madden TL, Schaffer AA, *et al*. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* 1997;**25**:3389–402.

55. Bairoch A, Apweiler R. The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Res* 2000;**28**(1):45–8.

56. Agarap AFM. Deep Learning using Rectified Linear Units (ReLU). *arXiv* 2018;1803.08375v2.

57. Gühring I, Kutyniok G, and Petersen P. Error bounds for approximations with deep ReLU neural networks in $W^{s,p}$ norm. *Anal Appl* 2020;**18**(5):803–59.

58. Dhall A, Patiyal S, Sharma N, *et al*. Computer-aided prediction and design of IL-6 inducing peptides: IL-6 plays a crucial role in COVID-19. *Brief Bioinform* 2021;**22**(2):936–45.

59. Sharma N, Patiyal S, Dhall A, *et al*. AlgPred 2.0: an improved method for predicting allergenic proteins and mapping of IgE epitopes. *Brief Bioinform* 2020;**22**(4):bbaa294.

60. Dhall A, Patiyal S, Sharma N, *et al*. Computer-aided prediction of inhibitors against STAT3 for managing COVID-19 associate cytokine storm. *Computers in biology and medicine* 2021;**137**:104780.

61. Sachs MC. plotROC: a tool for plotting ROC curves. *J Stat Softw* 2017;**79**.

62. Ofran Y, Mysore V, Rost B. Prediction of DNA-binding residues from sequence. *Bioinformatics* 2007;**23**(13):i347–53.

63. Hopkins AL. Network pharmacology: the next paradigm in drug discovery. *Nat Chem Biol* 2008;**4**(11):682–90.

64. Csermely P, Korcsmaros T, Kiss HJ, *et al*. Structure and dynamics of molecular networks: a novel paradigm of drug discovery: a comprehensive review. *Pharmacol Ther* 2013;**138**:333–408.

65. Burley SK, Bhikadiya C, Bi C, *et al*. RCSB Protein Data Bank: powerful new tools for exploring 3D structures of biological macromolecules for basic and applied research and education in fundamental biology, biomedicine, biotechnology, bioengineering and energy sciences. *Nucleic Acids Res* 2021;**49**(D1): D437–51.

66. Berman HM, Westbrook J, Feng Z, *et al*. The Protein Data Bank. *Nucleic Acids Res* 2000;**28**(1):235–42.

67. Anderson AC. The process of structure-based drug design. *Chem Biol* 2003;**10**(9):787–97.

68. Goodwin KD, Long EC, Georgiadis MM. A host-guest approach for determining drug-DNA interactions: an example using netropsin. *Nucleic Acids Res* 2005;**33**(13):4106–16.

69. Pradhan S, Das P, Mattaparthi VSK. Characterizing the binding interactions between DNA-binding proteins XPA and XPE: a molecular dynamics approach. *ACS Omega* 2018;**3**(11):15442–54.

70. Moravek Z, Neidle S, Schneider B. Protein and drug interactions in the minor groove of DNA. *Nucleic Acids Res* 2002;**30**:1182–91.

71. Mishra A, Pokhrel P, Hoque MT. StackDPPred: a stacking based prediction of DNA-binding protein from sequence. *Bioinformatics* 2019;**35**(3):433–41.