Deng *et al. BMC Bioinformatics* 2018, **19**(Suppl 19):522
https://doi.org/10.1186/s12859-018-2527-1

BMC Bioinformatics

RESEARCH Open Access

PDRLGB: precise DNA-binding residue prediction using a light gradient boosting machine

Lei Deng[1], Juan Pan[1], Xiaojie Xu[1], Wenyi Yang[1], Chuyao Liu[1] and Hui Liu[2*]

From 29th International Conference on Genome Informatics
Yunnan, China. 3-5 December 2018

Abstract

Background: Identifying specific residues for protein-DNA interactions are of considerable importance to better recognize the binding mechanism of protein-DNA complexes. Despite the fact that many computational DNA-binding residue prediction approaches have been developed, there is still significant room for improvement concerning overall performance and availability.

Results: Here, we present an efficient approach termed PDRLGB that uses a light gradient boosting machine (LightGBM) to predict binding residues in protein-DNA complexes. Initially, we extract a wide variety of 913 sequence and structure features with a sliding window of 11. Then, we apply the random forest algorithm to sort the features in descending order of importance and obtain the optimal subset of features using incremental feature selection. Based on the selected feature set, we use a light gradient boosting machine to build the prediction model for DNA-binding residues. Our PDRLGB method shows better overall predictive accuracy and relatively less training time than other widely used machine learning (ML) methods such as random forest (RF), Adaboost and support vector machine (SVM). We further compare PDRLGB with various existing approaches on the independent test datasets and show improvement in results over the existing state-of-the-art approaches.

Conclusions: PDRLGB is an efficient approach to predict specific residues for protein-DNA interactions.

Keywords: DNA-binding residue, Light gradient boosting, Random forest, Incremental feature selection

Introduction

Th protein-DNA interaction is one of the central issues in molecular biology and widely exists in various biological activities in living organisms, such as DNA replication, repair, and modification processes. To understand the recognition mechanism of protein-DNA complexes, researchers often focus on protein-DNA binding sites especially the interface residues that bind DNA. Experimental approach such as electrophoretic mobility shift assays (EMSAs) [1, 2], conventional chromatin immunoprecipitation (ChIP) [3], X-ray crystallography [4], PNA (peptide nucleic acid)-assisted identification of RNA binding proteins (PAIR) [5], and NMR spectroscopy [6] have been applied to expose the DNA binding amino acids. However, these laboratory methods are expensive and time-consuming. Alternatively, low-cost and efficient computational methods are particularly important in discovering specific interface residues of protein-DNA complexes.

A number of computational approaches have been focused on applying machine learning algorithms to build prediction models based on sequence and structural information. Wei [7] proposed novel evolutionary features for DNA-binding proteins prediction. Jones and his coworkers [8] proposed a simple method to identify DNA-binding residues using the positive electrostatic patches

*Correspondence: hliu@cczu.edu.cn
[2]Lab of Information Management, Changzhou University, 213164, Changzhou, China
Full list of author information is available at the end of the article

on the protein surface. Ahmad et al. [9] developed a neural network classifier to predict DNA-binding residues using a variety of composition, sequence and structural information. Wang et al. [10] built SVM-based models to predict DNA-binding residues by using data examples represented with three sequence characteristics. Ferrer-Costa et al. [11] implemented an effective linear predictor to determine the DNA-binding sites in protein sequences. Yan and his coworkers [12] trained a Naive Bayes classifier to predict whether a given amino acid is a DNA-binding site based on its characteristics and the features of its sequence neighbors. Wang and Yang [13] developed a random forest (RF) classifier according to the evolutionary information to detect the DNA-binding sites. Song et al. [14] employed imbalanced classification techniques for this problem. Carson et al. [15] combined the C4.5 algorithm with bootstrap aggregation and cost-sensitive learning to identify binding residues in protein-RNA complexes. Zou et al. [16] focused on the feature selection techniques and improved the performance. Ozbek et al. [17] presented a prediction method based on residue variations in high frequency forms using the Gaussian network. Other protein-DNA binding residue prediction tools such as DR_bind [18] and PreDNA [19] have also been developed.

Although a lot of studies has been performed, the problem of accurately identifying protein-DNA binding sites still has huge room for improvement. Firstly, effective features to detect DNA-binding interface residues from non-binding amino acids are not fully exploited. Secondly, the imbalanced problem exists since the numbers of DNA-binding and non-binding amino acids in proteins are extremely unbalanced, and will cause over-fitting and poor performance in the prediction of DNA-binding amino acids.

In this work, we develop a innovative computational pipeline, named PDRLGB, for predicting interface residues in protein-DNA complexes. We extract many sequence and structure features and use the random forest to select a subset of optimal features. Based on the selected characteristics, we train the DNA-binding residue prediction models using a new implementation of Gradient boosting decision tree (LightGBM) [20]. Our experiments show that PDRLGB significantly outperforms other state-of-the-art DNA-binding residue prediction approaches.

Materials and methods

Datasets

To access the performance of the PDRLGB method and other existing approaches, two benchmarking datasets (PDNA-62 and PDNA-224) and two independent datasets (TS-72 and TS-61) are used. PDNA-62 was built by Ahmad et al. [9]. It consists of 67 sequences obtained from 62 protein-DNA complexes in the Protein Data Bank (PDB) [21] and the sequence identity between any two sequences is $\leq$ 25%. PDNA-224 was generated by Li et al. [19], which contains 224 proteins and the redundant sequences was removed by using the sequence identity cutoff of 25%. The independent test dataset called TS-72 was extracted by Ma et al. [22]. It contains 72 protein chains. TS-61 was constructed by Zhou et al. [23]. Redundant proteins are removed by using the CD-HIT [24], and the remaining 61 non-redundant DNA-binding protein sequences have $\leq$30% sequence identity with the protein sequences in PDNA-62, PDNA-224, and TS-72.

Similar to previous researches [10, 18], a residue of a protein is defined as a binding amino acid if the closest distance between atoms of the protein and its binding DNA is $\leq$3.5Å. The whole positive samples and negative samples of the four datasets are summarized in Table 1.

Performance measures

To evaluate the performance, we use several typical measures, including accuracy (ACC), sensitivity (SN/Recall), specificity (SP), strength (ST), precision (PRE), F1-score (F1), and Matthews Correlation Coefficient (MCC) score. These measurements are defined as:

$$ACC = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

$$SN = TP/(TP + FN) \tag{2}$$

$$SP = TN/(TN + FP) \tag{3}$$

$$ST = (SN + SP)/2 \tag{4}$$

$$PRE = TP/(TP + FP) \tag{5}$$

$$F1 = \frac{2 \times SN \times PRE}{SN + PRE} \tag{6}$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{7}$$

In these equations, the TP, FP, TN, and FN represent the number of true positives, the number of false positives, the number of true negatives, and the number of

Table 1 Number of positive samples (binding sites) and negative samples (non-binding sites) of the four datasets

Dataset	Positive samples	Negative samples
PDNA-62	1215	6948
PDNA-224	3778	53,570
TS-72	1040	13,226
TS-61	1078	13,175

false negatives, respectively. Because of the imbalanced problem in the data sets, the strength (ST) is the average score of sensitivity and specificity which is used to obtain a fair measure of the model. Additionally, there are two broadly employed measurement to estimate prediction performance including the receiver operating characteristic (ROC) [25] and the area under ROC curve (AUC) [26]. The ROC curve is plotted with the false positive rate against the true positive rate. When AUC takes the maximum value of 1, it represents a perfect predictor, and the values of AUC of random guessing is usually close to 0.5.

The prediction pipeline

The pipeline of PDRLGB is showed in Fig. 1. It is made up of several steps: A) feature extraction: a total of 83 sequence and structure features are extracted, and the feature vectors are generated using a sliding window of *w*; B) feature selection: the features are sorted with random forest and the optimal feature set is selected using the incremental feature selection approach; C) building prediction classifiers: the DNA-binding residue prediction models are built using the light gradient boosting machine. These processes are described in details in the following subsections.

Feature extraction

We extract a variety of features including position-specific scoring matrices (PSSMs) (20 features), physicochemical properties (10 features), disordered features (3 features), side-chain environment (pKa) (2 features), identity vector (20 features), net charge (1 feature), the information from DSSP (15 features), the information from NACCESS (10 features), H-bonds (1 feature) and B-factor(1 feature). These features can be grouped into two categories: sequence and structure features.

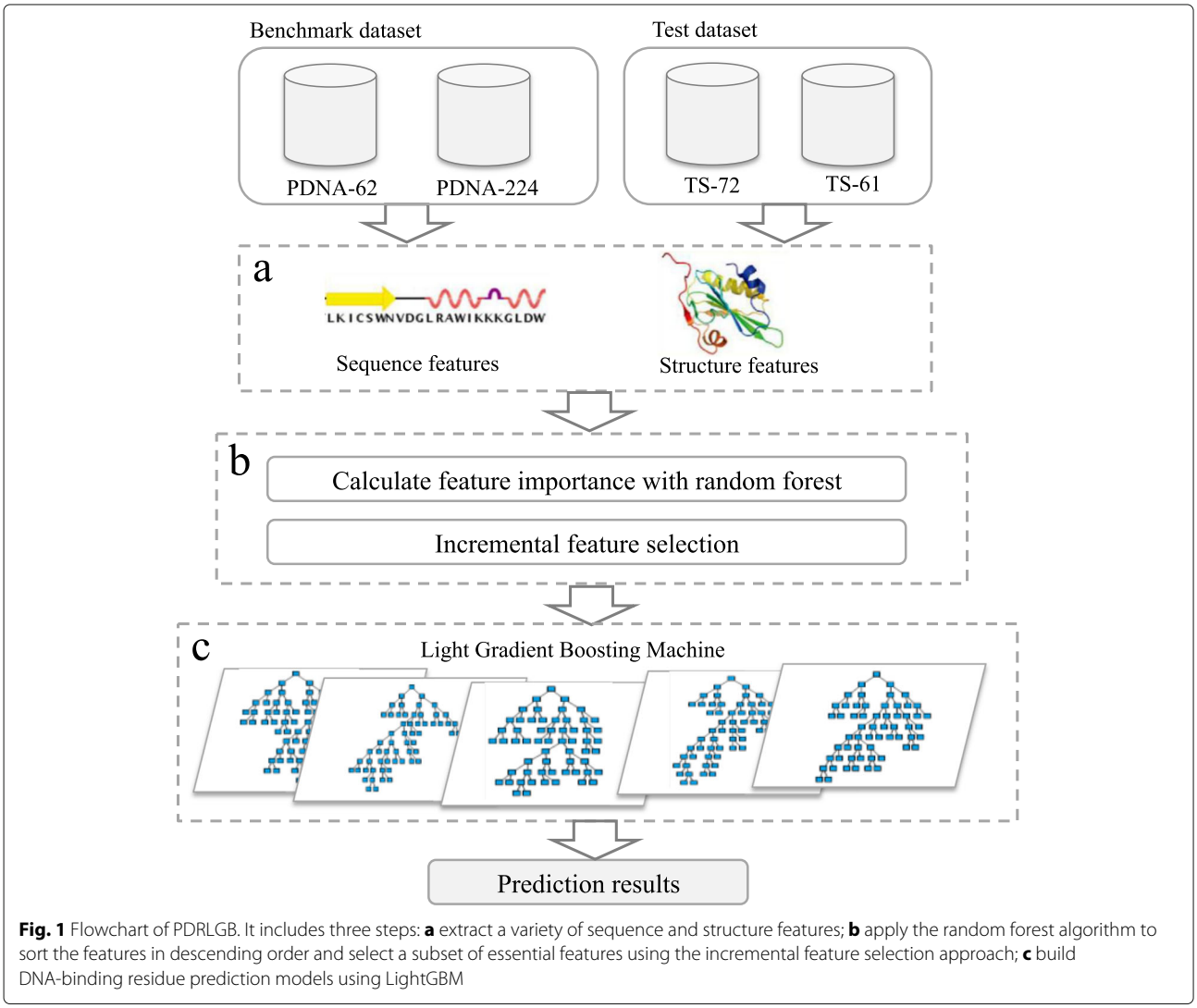

Fig. 1 Flowchart of PDRLGB. It includes three steps: **a** extract a variety of sequence and structure features; **b** apply the random forest algorithm to sort the features in descending order and select a subset of essential features using the incremental feature selection approach; **c** build DNA-binding residue prediction models using LightGBM

Sequence features:

1) Position-specific scoring matrices (PSSMs): PSSM based evolutionary information is obtained from multiple sequence alignment calculated by PSI-BLAST [27] searching against the NCBI non-redundant (NR) database, with iteration number as 3 and e-value as 0.001.
2) Physicochemical properties: The physicochemical properties of a residue include atom numbers, electrostatic charge numbers, potential hydrogen bonds, molecular mass (Mmass), hydrophobicity, hydrophilicity, polarity, polarizability, propensities and average accessible surface area [28]. The original values of the ten physicochemical attributes for each residue are obtained from the AAindex database [29].
3) Disordered regions: Predicted disordered regions within a protein is also a significant property. Avoiding possibly disordered fragments in protein expression constructs can enhance expression, foldability, and stability of the protein. DisEMBL [30] is a useful tool for identifying disordered regions, which is needed for many biochemical studies, particularly structural biology, and structural genomics projects. In this study, DisEMBL is used to indentify dynamically disordered regions of the protein sequence.
4) Side-chain environment (pKa): The value of pKa is an effective metric in determining environmental features of a protein. The side-chain pKa rates are collected from Nelson and Cox [31] representing protein side-chain environmental factors and are broadly used by previous studies.
5) Identity vector: There is a 20-feature vector with 1 when the residue type occurs at the corresponding position and 0 for the remaining amino acid types.
6) Net charge of a residue: Twenty amino acids can be divided into non-polar amino acids, polar charged amino acids, polar uncharged amino acids. The DNA backbone is negatively charged, so the sequence of polar positively charged amino acids is thought to be characteristic of DNA binding. A charge of +1 is assigned to Arg and Lys and -1 to Asp and Glu. His is assigned a charge of +0.5 and all other residues are regarded as neutral.

Structure features:

1) Features from DSSP: we use DSSP [32] to obtain the secondary structures, including solvent accessible surface area (ASA), hydrogen bonds, atom coordinates and backbone torsion angles.
2) Features from NACCESS: We use NACCESS [33] to compute the absolute and relative ASA of all atoms, total side chain, main chain, non-polar side chain and allpolar side chain, respectively. ASA related features has been shown to be a important feature in identifying protein functional sites [34–37].
3) Number of H-bonds: The number of Hydrogen bonds (Hbond) is computed by HBPLUS [38].
4) B-factor of a residue: The B-factor [39] of protein crystal structures, including the B-factor of the C_α and that of the C_β of the amino acids in the sequence, was adopted.

Features selection

We encode the features with a sliding window of w and generate a large feature vector. To eliminate uninformative variables and obtain more cost-effective models, a reliable feature selection approach was applied. Firstly, we use

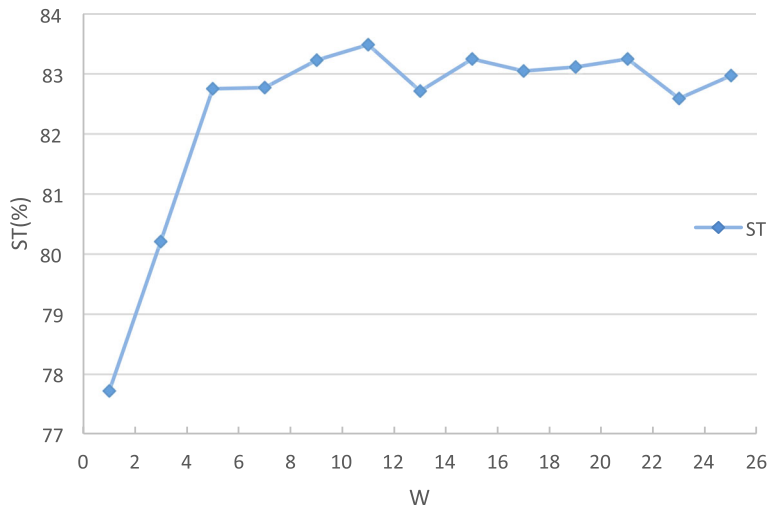

Fig. 2 The effect of window size w on performance

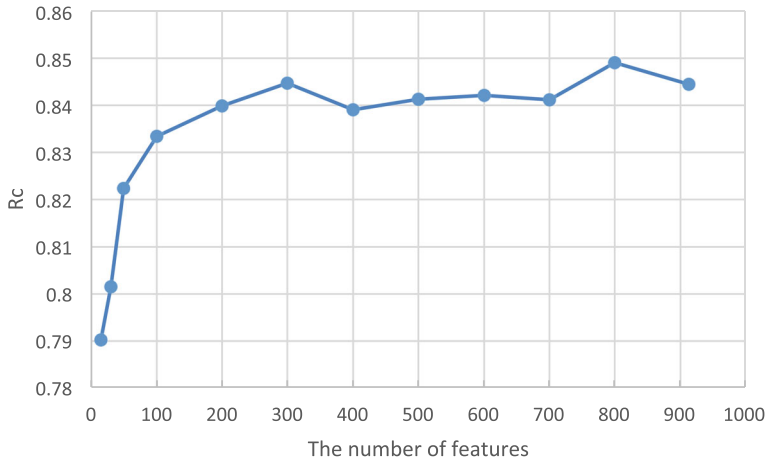

Fig. 3 The R_c values of $top-k$ feature sets obtained by using the LightGBM algorithm

the random forest algorithm [40] to sort the features by using the mean decrease Gini index (MDGI) Z-Score [41]. MDGI Z-Score measures the importance of individual features. Features with higher MDGI Z-Scores are more sensitive to random shuffling of their values, and thus are more important for correctly classifying a residue into DNA-binding site and non-DNA binding site. After ranking the features in descending order of MDGI Z-Score, we utilize the incremental feature selection approach to select the top-k features. We construct the feature subset by incremental adding the features in the ranked list to the subset, and evaluate the performance of the top-k subset using the LightGBM classifier with 5-fold cross-validation. We use a comprehensive evaluation score (R_c) to measure the performance of the feature subset. The R_c score is defined as follows:

$$R_c = \frac{1}{n}\sum_{i=1}^{n}\{ACC_i + SN_i + SP_i + AUC_i\}, \quad (8)$$

where n is the repeat times of the 5-fold cross-validation.

Building prediction classifiers

Gradient boosting decision tree (GBDT) [20] is a widely used and useful algorithm that can be used for both classi fication and regression problems [42–47]. Recently, Ke et al. proposed a novel GBDT algorithm named LightGBM [48], which utilize two novel techniques: Gradient-based One-Side Sampling (GOSS) along with Exclusive Feature Bundling (EFB) to deal with the huge number of data samples along with massive amount of features respectively. GOSS keeps all the examples with large gradients and conducts random sampling on the examples with small gradients. EFB algorithm can bundle many exclusive characteristics to the much fewer dense characteristics, which can dramatically avoid unnecessary calculation for zero feature values. Here we apply LightGBM to build the DNA-binding residue prediction models. The detailed steps of the LightGBM algorithm is shown in Algorithm 1.

Results

Parameter selection

The sliding window describes the target residue's sequence neighborhood, and the window size w should be

Table 2 Performance comparison of LightGBM with other machine learning methods

Dataset	Methods	ACC	SN	SP	ST	PRE	F1	MCC	AUC
PDNA-62	SVM	0.817	0.745	0.829	0.787	0.433	0.547	0.468	0.873
	Adaboost	0.814	0.791	0.818	0.804	0.431	0.558	0.485	0.881
	RF	0.817	0.782	0.822	0.802	0.435	0.559	0.486	0.883
	LightGBM	0.815	0.863	0.806	0.835	0.438	0.581	0.523	0.912
PDNA-224	SVM	0.786	0.765	0.776	0.771	0.194	0.310	0.306	0.847
	Adaboost	0.773	0.761	0.774	0.768	0.192	0.307	0.320	0.851
	RF	0.814	0.750	0.819	0.784	0.226	0.347	0.351	0.864
	LightGBM	0.800	0.833	0.797	0.815	0.224	0.353	0.383	0.896

Algorithm 1 The LightGBM algorithm

Input:

Training data: $D = \{(\chi_1, y_1), (\chi_2, y_2), ..., (\chi_N, y_N)\}$, $\chi_i \epsilon \chi$, $\chi \subseteq R$, $y_i \epsilon \{-1, +1\}$; loss function: $L(y, \Theta(\chi))$; iterations: M; sampling ratio of large gradient data: a; sampling ratio of small gradient data: b;

1: Merge mutually exclusive features(i.e. features never take nonzero values simultaneously) of $\chi_i, i = \{1, ..., N\}$ by exclusive feature bundling(EFB) method;

2: Initialize $\Theta_0(\chi) = arg\ min_c \sum_i^N L(y_i, c)$;

3: **for** $m = 1$ to M **do**

4: Compute absolute values of gradients:
$r_i = \left| \frac{\partial L(y_i, \Theta(\chi_i))}{\partial \Theta(\chi_i)} \right|_{\Theta(\chi) = \Theta_{m-1}(\chi)}, i = \{1, ..., N\}$

5: Resampled dataset by Gradient-based One-Side Sampling (GOSS) method:
$topN = a \times len(D)$; $randN = b \times len(D)$;
$sorted = GetSortedIndices(abs(r))$;
$A = sorted[\,1 : topN]$; $B = RandomPick(sorted[\,topN : len(D)]\,, randN)$; $D' = A + B$;

6: Compute information gains:
$V_j(d) = \frac{1}{n} \left(\frac{(\sum_{\chi_i \epsilon A_l} r_i + \frac{1-a}{b} \sum_{\chi_i \epsilon B_l} r_i)^2}{n_l^j(d)} + \frac{(\sum_{\chi_i \epsilon A_r} r_i + \frac{1-a}{b} \sum_{\chi_i \epsilon B_r} r_i)^2}{n_r^j(d)} \right)$

7: Get a new decision tree $\Theta_m(\chi)'$ on set D'.

8: Update $\Theta_m(\chi) = \Theta_{m-1}(\chi) + \Theta_m(\chi)'$

9: **end for**

10: **return** $\tilde{\Theta}(\chi) = \Theta_M(\chi)$

selected properly. The predictive performance of a variety of different local window sizes (1, 3, ..., 25) is evaluated. As shown in Fig. 2, the ST score increases when the window size increases from 1 to 11, and the highest ST score is achieved when the window size is 11. So we select the optimal window size as 11 in the proposed PDRLGB method.

The number of features (k) is another important parameter. We build LightGBM classifiers for each $top-k$ subset and calculate the performance of 5-fold cross-validation.

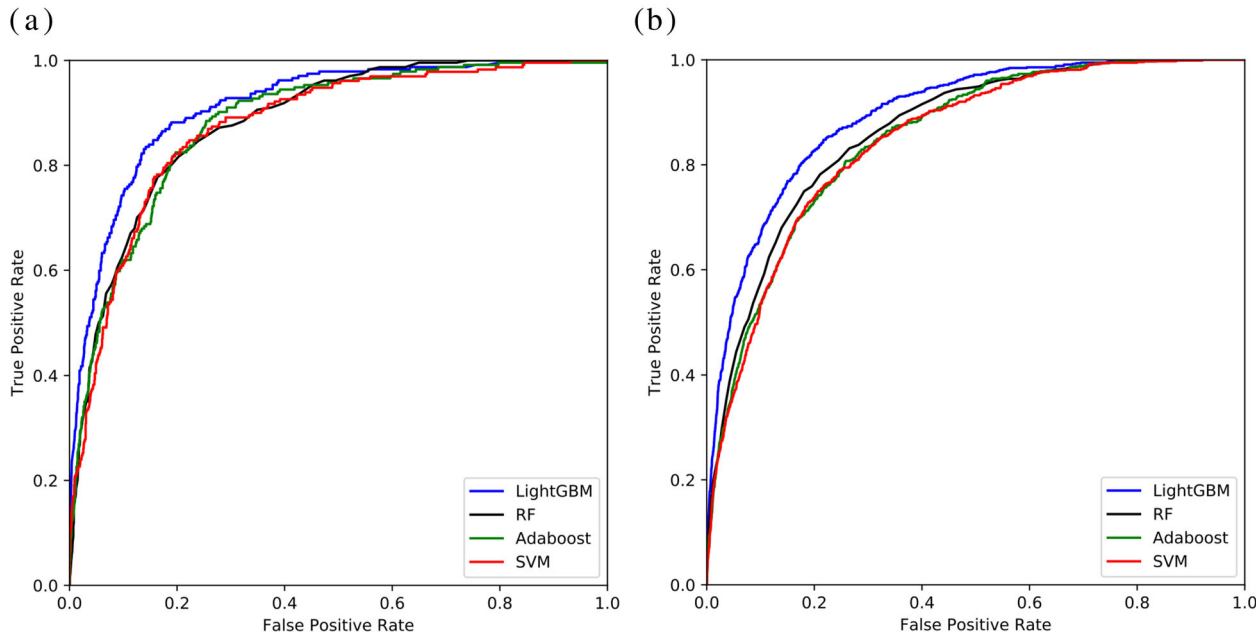

Fig. 4 Performanc of LightGBM, SVM, Random Forest and Adaboost on the benchmark datasets. **a** shows the ROC curves on the PDNA-62 dataset. **b** shows the ROC curves on the PDNA-224 dataset

Table 3 Performance comparison of various prediction methods on PDNA-62 with 5-fold cross-validation

Methods	ACC	SN	SP	ST	PRE	F1	MCC	AUC	*P*-value
Dps-pred	0.791	0.403	0.818	0.611	0.279	0.330	0.191	-	-
Dbs-pssm	0.664	0.682	0.660	0.671	0.210	0.376	0.249	-	-
BindN	0.703	0.694	0.705	0.700	0.291	0.410	0.297	0.752	-
Dp-bind	0.781	0.792	0.772	0.782	0.378	0.512	0.490	-	-
BindN-RF	0.782	0.781	0.782	0.782	0.385	0.516	0.436	0.861	-
BindN+	0.790	0.773	0.793	0.783	0.395	0.523	0.443	0.859	-
PreDNA	0.794	0.768	0.797	0.783	0.398	0.524	0.424	-	-
EL_PSSM-RT	0.808	0.854	0.801	0.826	0.428	0.569	0.507	0.901	**
PDRLGB	0.815	0.863	0.806	0.835	0.438	0.581	0.523	0.912	1.79×10^{-5}

The results are shown in Fig. 3. As the dimension of the features increases, the highest RC score of 0.85 is obtained when using the top 800 features. Finally, we select a subset of features (Top 800) that contribute the most to the classification as the optimal feature set.

Performance comparison with other machine learning techniques

In this section, we conduct a comparison experiment of LightGBM with existing machine learning techniques, including Support Vector Machine (SVM) [49], Random Forest (RF) [40] and AdaBoost [50]. The performance of these classifiers are listed in Table 2. It is worth emphasizing that these classifiers are trained on the same benchmark with the same feature set. The ROC curves are shown in Fig. 4. It is obvious that LightGBM achieves significant performance improvement on both PDNA-62 and PDNA-224 when it compares to these classifiers. Concretely, on the PDNA-62 dataset, LightGBM obtains at least 3.1% increase on ST, 2.2% increase on F1, 3.7% increase on MCC and 2.9% increase on AUC when comparing with SVM, RF and AdaBoost. As for the PDNA-224 dataset, LightGBM achieves at least 3.1% increase on ST, 0.6% increase on F1, 3.2% increase on MCC and 3.0% increase on AUC. Due to the imbalanced problem on both datasets, the ROC curve is regarded as the useful estimation for the overall performance. Higher ROC curve denotes better prediction performance. Figure 4a and b also show that LightGBM obtains the best ROC curves on the two datasets (PDNA-62 and PDNA-224). The results imply that the LightGBM algorithm we used is more superior than other widely used classifiers.

Performance comparison with other state-of-the-art predictors

There exists many DNA-binding site prediction methods which trained and tested either on PDNA-62 or PDNA-224, such as Dps-pred [9], Dbs-pssm [51], BindN [10], Dp-bind [52], BindN-RF [13], BindN+ [53], PreDNA [19] and EL_PSSM-RT [23]. Note that some of these methods are only trained and tested the PDNA-62 dataset, and others are trained and tested on the two datasets. We calculate *P*-values using the two-tailed, paired t-test [54]. The prediction performance of our PDRLGB approach and other methods on PDNA-62 and PDNA-224 are shown in Tables 3 and 4, respectively. The results on PDNA-62 are shown in Table 3, PDRLGB achieves the best performance, outperforming other approaches by 0.9%-21.4% on ST, 1.2%-25.1% on F1, 1.6%-33.2% on MCC and 1.1%-16% on AUC. The results on the PDNA-224 dataset are shown in Table 4, PDRLGB performs better than PreDNA and EL_PSSM-RT by 2.7%-6.9% on ST, 2.9%-4.8% on F1, 4.2%-9.4% on MCC and 3.1% on AUC. These enhancements on performance indicate that the LightGBM-based PDRLGB method based on the optimally selected features is beneficial for predicting DNA-binding residues.

Performance comparison on the independent test dataset

To further assess the performance, we compare PDRLGB with seven existing state-of-the-art protein-DNA binding site prediction methods, DNABR [22], BindN [10], BindN-RF [13], BindN+ [53], EL_PSSM-RT [23], DRNApred [55] and CNNsite [56] on the TS-72 dataset. DNABR [22] and BindN-RF [13] are built using random forest (RF). BindN [10] and BindN+ [53] are trained using support

Table 4 Performance of PDRLGB Compared with PreDNA and EL_PSSM-RT on PDNA-224 with 5-fold cross-validation

Methods	ACC	SN	SP	ST	PRE	F1	MCC	AUC	*P*-value
PreDNA	0.791	0.695	0.798	0.746	0.195	0.305	0.289	-	-
EL_PSSM-RT	0.781	0.796	0.780	0.788	0.203	0.324	0.341	0.865	**
PDRLGB	0.800	0.833	0.797	0.815	0.224	0.353	0.383	0.896	4.07×10^{-5}

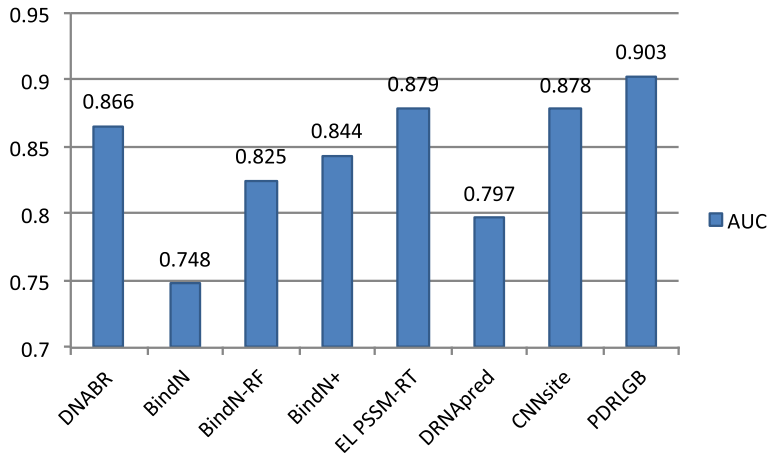

Fig. 5 Performance comparison on TS72

vector machines (SVMs). EL_PSSM-RT [23] is built using a ensemble learning classifier. DRNApred [55] is designed by using a two-layer predictor, which integrats hidden Markov model (HMM) and logistic regression models. CNNsite [56] is built using Convolutional Neural Network. The AUC scores of these approaches are shown in Fig. 5. DNABR, BindN, BindN-RF, BindN+, EL_PSSM-RT, DRNApred and CNNsite achieve AUC values of 0.866, 0.748, 0.825, 0.844, 0.879, 0.797 and 0.878, respectively. Comparing with these methods, our PDRLGB approach achives the highest AUC value of 0.903 and improves the AUC score by 2.4%-15.5% on the independent dataset TS-72.

We also compare our PDRLGB method with DP-Bind [57], EL_PSSM-RT [23] and DRNApred [55] on the independent dataset TS-61. DP-Bind is implemented using machine learning algorithms including SVM, kernel logistic regression and penalized logistic regression. DP-Bind also implements two ensemble classifiers by using majority voting (MAJ) and unanimity voting (STR) respectively. Here we only compare with DP-Bind (STR) since the unanimity voting approach achieves the best performance according to Hwang et al [57]. The results are depicted in Table 5. We observe that PDRLGB gains the highest AUC score of 0.850. Although DRNApred has the highest specificity, PDRLGB has a better balance between recall and specificity.

Computing time comparison

We present the training time cost comparisons in this subsection, which is shown in Fig. 6. Our experiments on the two datasets show that LightGBM speeds up the training process of classical methods by up to over 20 times faster than SVM and is also faster than Adaboost. Although random forest (RF) and LightGBM have similar calculation speed, in fact, the performance of the LightGBM-based method is far better than that of the RF classifier. Therefore, the PDRLGB is an accurate and fast model in the prediction of protein-DNA binding residues in the protein.

Case study

In order to further validate the usefulness of PDRLGB for DNA-binding residue prediction, we apply PDRLGB trained on PDNA-62 to distinguish the binding residues from non-binding residues for the ISDra2 transposase/IS end complex which is not in the training set, namely, 2XMA [58]. Here, we use PDRLGB to investigate the DNA-binding residues (2XMA:A). PDRLGB achieves 87.05% on ACC, 0.67 on MCC, 86.67% on SP, 87.16% on SN, 86.91% on ST, which is very precise when compared with the available experimental data in the PDB database. The experimentally determined DNA-binding sites and predicted sites by PDRLGB for complex 2XMA are shown in Fig. 7. Figure 7a denotes the experimentally determined

Table 5 Performance comparison on TS-61

Methods	ACC	SN	SP	ST	PRE	F1	MCC	AUC	*P*-value
DP-Bind(STR)	0.802	0.687	0.811	0.749	0.229	0.344	0.315	-	-
DRNApred	0.772	0.337	0.966	0.652	0.450	0.386	0.347	0.822	-
EL_PSSM-RT	0.773	0.726	0.777	0.752	0.211	0.327	0.305	0.839	**
PDRLGB	0.807	0.694	0.817	0.758	0.237	0.353	0.325	0.850	7.83×10^{-5}

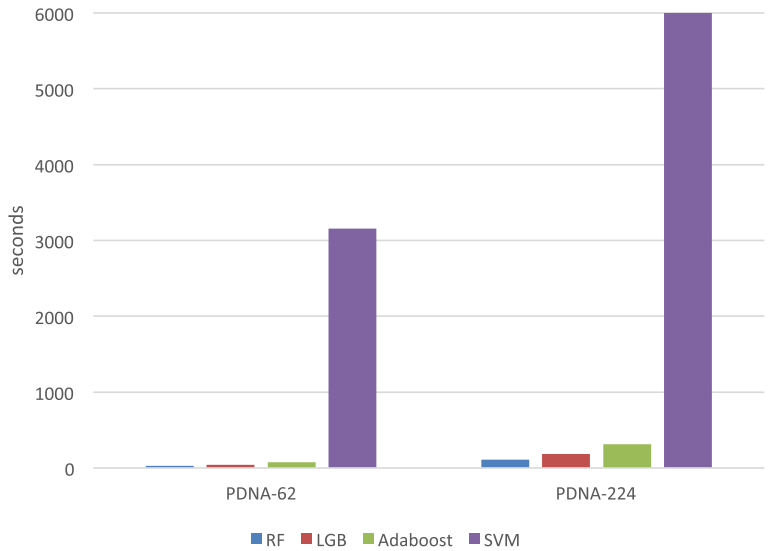

Fig. 6 Training time of LightGBM, SVM, Random Forest and Adaboost

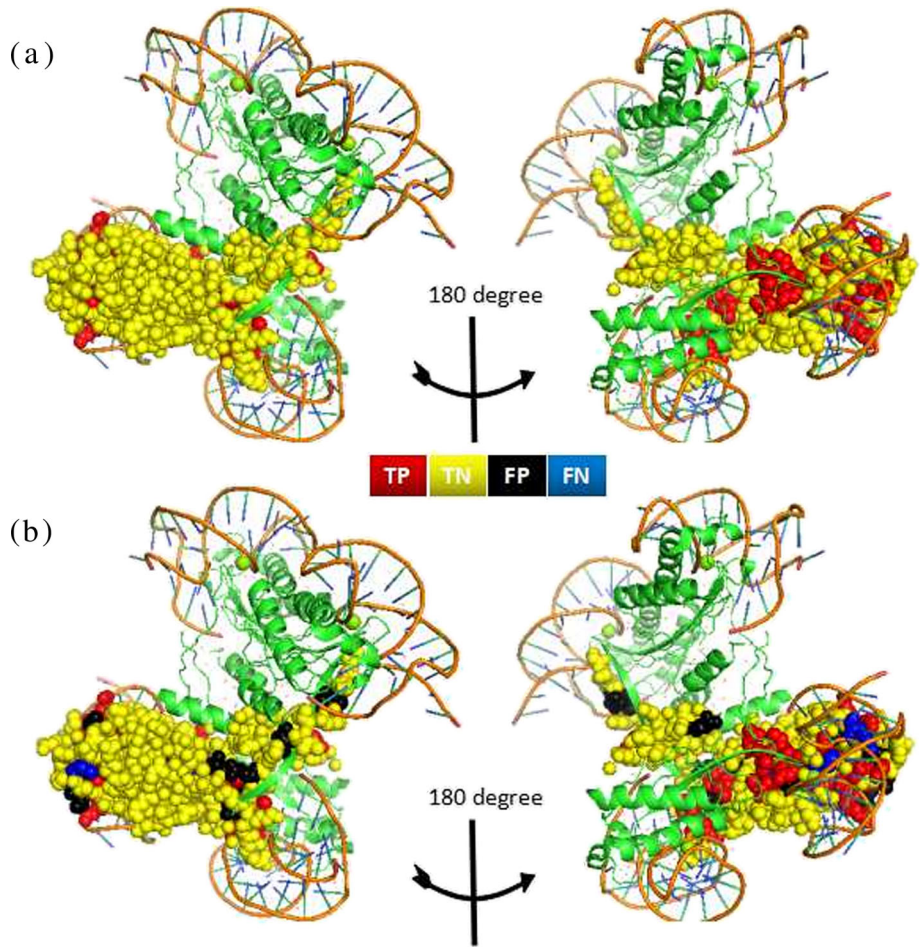

Fig. 7 Prediction results on the case study 2XMA. **a** shows the experimentally determined DNA-binding residues in protein 2XMA:A. **b** shows the predicted binding sites by PDRLGB, and the numbers of predicted TP, FP, TN and FN in 2XMA:A are 26, 14, 95, and 4, respectively. The true positive (TP), true negative (TN), false positive (FP) and false negative (FN) sites are displayed in red, yellow, black and blue, respectively

binding sites of protein 2XMA:A and the red spheres represent real DNA-binding sites. Figure 7b presents the predicting binding sites of protein 2XMA:A. The results show that the majority of the DNA-binding residues are correctly predicted by the PDRLGB model.

Discussion

Existing methods for predicting DNA-binding sites are mainly divided into sequence-based methods, structure-based methods and hybrid methods. In this study, we integrate both sequence and structural features to effectively predict DNA-binding residues. A limitation of our PDRLGB approach is that it requires the protein structural information, which may limit its application. However, with the increasing solved protein structures, protein homology modeling projects and predicted 3D structures, it is expected that PDRLGB can be used as a powerful tool to effectively identify DNA-binding residues. We believe that PDRLGB can be an effective tool for accurately predicting DNA-binding residues with the increasing availability of high-quality protein-DNA complex structures.

Conclusion

Targeting specific DNA-binding amino acids that contribute to the strength and specificity of protein-DNA interactions has broad applications ranging from rational drug design to the investigation of metabolic and signal transduction networks. In this paper, we have developed a novel LightGBM-based algorithm termed PDRLGB, for DNA-binding residue prediction. The sequence features and structural characteristics are combined to construct the feature space, and random forest combined with incremental feature selection is applied to make a feature selection. As a result, the prediction performance on the two datasets PDNA-62 and PDNA-224 with five-fold cross-validation demonstrate that PDRLGB can accelerate the training process and performs better when compared with other widely used machine learning classifiers. At the same time, performance comparisons between PDRLGB and other existing state-of-the-art DNA-binding site prediction methods demonstrate that our PDRLGB approach achieves the best performance. We have also employed our PDRLGB to identify binding sites on a protein-DNA complex 2XMA and obtained satisfactory results.

Acknowledgements
This work was supported by National Natural Science Foundation of China under grants No. 61672541 and No. 61672113, and Natural Science Foundation of Hunan Province under grant No. 2017JJ3287.

Funding
Publication costs are funded by National Natural Science Foundation of China under grant No. 61672541.

Availability of data and materials
The datasets used in this study is available at http://denglab.org/PDRLGB/.

About this supplement
This article has been published as part of *BMC Bioinformatics Volume 19 Supplement 19, 2018: Proceedings of the 29th International Conference on Genome Informatics (GIW 2018): bioinformatics*. The full contents of the supplement are available online at https://bmcbioinformatics.biomedcentral.com/articles/supplements/volume-19-supplement-19.

Authors' contributions
LD, JP, XX, WY, CL and HL designed the study and conducted experiments. LD, JP, XX, WY and CL performed statistical analyses. LD, JP and HL drafted the manuscript. JP prepared the experimental materials and benchmarks. All authors have read and approved the final manuscript.

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Publisher's Note

Author details
[1]School of Software, Central South University, 410075, Changsha, China. [2]Lab of Information Management, Changzhou University, 213164, Changzhou, China.

Published: 31 December 2018

References

1. Jones S, Heyningen PV, Berman HM, Thornton JM. Protein-dna interactions: a structural analysis. Nucleic Acids Res. 1999;29(4):943–54.
2. Jones S, Barker JA, Nobeli I, Thornton JM. Using structural motif templates to identify proteins with dna binding function. Nucleic Acids Res. 2003;31(11):2811.
3. Kono H, Sarai A. Structure-based prediction of dna target sites by regulatory proteins. Proteins Struct Funct Bioinforma. 2015;35(1):114–31.
4. Orengo CA, Michie AD, Jones S, Jones DT, Swindells MB, Thornton JM. Cath-a hierarchic classification of protein domain structures. Structure. 1997;5(8):1093–108.
5. Olson WK, Gorin AA, Lu XJ, Hock LM, Zhurkin VB. Dna sequence-dependent deformability deduced from protein-dna crystal complexes. Proc Natl Acad Sci U S A. 1998;95(19):11163–8.
6. Ponting CP, Schultz J, Milpetz F, Bork P. Smart: identification and annotation of domains from signalling and extracellular protein sequences. Nucleic Acids Res. 1999;27(1):229–32.
7. Wei L, Tang J, Zou Q. Local-dpp: An improved dna-binding protein prediction method by exploring local evolutionary information. Inf Sci. 2017;384:135–44.
8. Jones S, Shanahan HP, Berman HM, Thornton JM. Using electrostatic potentials to predict dna-binding sites on dna-binding proteins. Nucleic Acids Res. 2003;31(24):7189–98.
9. 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.
10. 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 issue):243–8.
11. Ferrercosta C, Shanahan HP, Jones S, Thornton JM. Hthquery: a method for detecting dna-binding proteins with a helix-turn-helix structural motif. Bioinformatics. 2005;21(18):3679–80.
12. Yan C, Terribilini M, Wu F, Jernigan RL, Dobbs D, Honavar V. Predicting dna-binding sites of proteins from amino acid sequence. BMC Bioinformatics. 2006;7(1):262.
13. Wang L, Yang MQ, Yang JY. Prediction of dna-binding residues from protein sequence information using random forests. BMC Genomics. 2009;10(S1):1.

14. Song L, Li D, Zeng X, Wu Y, Guo L, Zou Q. ndna-prot: identification of dna-binding proteins based on unbalanced classification. BMC Bioinformatics. 2014;15(1):298.
15. Carson MB, Langlois R, Lu H. Naps: a residue-level nucleic acid-binding prediction server. Nucleic Acids Res. 2010;38(Web Server issue):431–5.
16. Zou Q, Wan S, Ju Y, Tang J, Zeng X. Pretata: predicting tata binding proteins with novel features and dimensionality reduction strategy. BMC Syst Biol. 2016;10(4):114.
17. Ozbek P, Soner S, Erman B, Haliloglu T. Dnabindprot: fluctuation-based predictor of dna-binding residues within a network of interacting residues. Nucleic Acids Res. 2010;38(Web Server issue):417–23.
18. Chen YC, Wright JD, Lim C. Dr_bind: a web server for predicting dna-binding residues from the protein structure based on electrostatics, evolution and geometry. Nucleic Acids Res. 2012;40(Web Server issue): 249–56.
19. Li T, Li QZ, Liu S, Fan GL, Zuo YC, Peng Y. Predna: accurate prediction of dna-binding sites in proteins by integrating sequence and geometric structure information. Bioinformatics. 2013;29(6):678–85.
20. Friedman JH. Greedy function approximation: A gradient boosting machine. Ann Stat. 2001;29(5):1189–232.
21. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. The protein data bank, 1999–. Int Tables Crystallogr. 2000;67(Suppl):675–84.
22. Ma X, Guo J, Liu HD, Xie JM, Sun X. Sequence-based prediction of dna-binding residues in proteins with conservation and correlation information. IEEE/ACM Trans Biol Bioinforma. 2012;9(6):1766–75.
23. Zhou J, Lu Q, Xu R, He Y, Wang H. El_pssm-rt: Dna-binding residue prediction by integrating ensemble learning with pssm relation transformation. BMC Bioinformatics. 2017;18(1):379.
24. Fu L, Niu B, Zhu Z, Wu S, Li W. Cd-hit: accelerated for clustering the next-generation sequencing data. Bioinformatics. 2012;28(23):3150–2.
25. Swets JA. Measuring the accuracy of diagnostic systems. Science. 1988;240(4857):1285–93.
26. Bradley AP. The use of the area under the roc curve in the evaluation of machine learning algorithms. Pattern Recog. 1997;30(7):1145–59. https://doi.org/10.1016/S0031-3203(96)00142-2.
27. Altschul S, Madden T, Schäffer A, Zhang J, Zhang Z, Miller W, Lipman D. Gapped blast and psi-blast: a new generation of protein database search programs. Nucleic Acids Res. 1997;25(17):3389–402. https://doi.org/10.1093/nar/25.17.3389.
28. Miller S, Lesk AM, Janin J, Chothia C. The accessible surface area and stability of oligomeric proteins. Nature. 1987;328(6133):834–6.
29. Kawashima S, Ogata H, Kanehisa M. Aaindex: Amino acid index database. Nucleic Acids Res. 1999;27(1):368–9.
30. Linding R, Jensen LJ, Diella F, Bork P, Gibson TJ, Russell RB. Protein disorder prediction: implications for structural proteomics. Structure. 2003;11(11):1453.
31. Re A, Joshi T, Kulberkyte E, Morris Q, Workman CT. Rna-protein interactions: an overview. Methods Mol Biol. 2014;1097(1097):491.
32. Kabsch W, Sander C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers. 1983;22(12):2577–637.
33. Hubbard SJ, Naccess TM. Computer Program. London: Department of Biochemistry and Molecular Biology, University College of London; 1993.
34. Deng L, Zhang QC, Chen Z, Meng Y, Guan J, Zhou S. Predhs: a web server for predicting protein–protein interaction hot spots by using structural neighborhood properties. Nucleic Acids Res. 2014;42(W1): 290–5.
35. Tang Y, Liu D, Wang Z, Wen T, Deng L. A boosting approach for prediction of protein-rna binding residues. BMC Bioinformatics. 2017;18(13):465.
36. Pan Y, Wang Z, Zhan W, Deng L. Computational identification of binding energy hot spots in protein–rna complexes using an ensemble approach. Bioinformatics. 2017;34(9):1473–80.
37. Nie L, Deng L, Fan C, Zhan W, Tang Y. Prediction of protein s-sulfenylation sites using a deep belief network. Curr Bioinforma. 2018;13(5):461–7.
38. Mcdonald IK, Thornton JM. Satisfying hydrogen bonding potential in proteins. J Mol Biol. 1994;238(5):777–93.
39. Yuan Z, Bailey TL, Teasdale RD. Prediction of protein b-factor profiles. Proteins Struct Funct Bioinforma. 2005;58(4):905–12.
40. Breiman L. Random forests. Mach Learn. 2001;45(1):5–32.
41. Liaw A, Wiener M. Classification and regression by random forest. R News. 2002;2:18–22.
42. Pan Y, Liu D, Deng L. Accurate prediction of functional effects for variants by combining gradient tree boosting with optimal neighborhood properties. PloS ONE. 2017;12(6):0179314.
43. Kuang L, Yu L, Huang L, Wang Y, Ma P, Li C, Zhu Y. A personalized qos prediction approach for cps service recommendation based on reputation and location-aware collaborative filtering. Sensors. 2018;18(5): 1556.
44. Fan C, Liu D, Huang R, Chen Z, Deng L. Predrsa: a gradient boosted regression trees approach for predicting protein solvent accessibility. BMC Bioinf. 2016;17(Suppl 1):8.
45. Liao Z, Wan S, He Y, Zou Q. Classification of small gtpases with hybrid protein features and advanced machine learning techniques. Curr Bioinforma. 2018;13(5):492–500.
46. Li C, Zheng X, Yang Z, Kuang L. Predicting short-term electricity demand by combining the advantages of arma and xgboost in fog computing environment. Wirel Commun Mob Comput. 2018;2018:5018053.
47. Gan Y, Tao H, Zou G, Yan C, Guan J. Dynamic epigenetic mode analysis using spatial temporal clustering. BMC Bioinformatics. 2016;17(17):537.
48. Ke G, Meng Q, Finely T, Wang T, Chen W, Ma W, Ye Q, Liu T-Y. Lightgbm: A highly efficient gradient boosting decision tree. Adv Neural Inf Process Syst. 2017;30:3146–54.
49. Cai YD, Lin SL. Support vector machines for predicting rrna-, rna-, and dna-binding proteins from amino acid sequence. Biochim Biophys Acta. 2003;1648(1-2):127.
50. Lab R, Gunnar Rätsch PD. Soft margins for adaboost. Mach Learn. 2001;42(3):287–320.
51. Shandar A, Akinori S. Pssm-based prediction of dna binding sites in proteins. BMC Bioinformatics. 2005;6(1):1–6.
52. Kuznetsov IB, Gou Z, Li R, Hwang S. Using evolutionary and structural information to predict dna-binding sites on dna-binding proteins. Proteins Struct Funct Bioinforma. 2006;64(1):19.
53. Wang L, Huang C, Yang MQ, Yang JY. Bindn+ for accurate prediction of dna and rna-binding residues from protein sequence features. BMC Syst Biol. 2010;4(S1):3.
54. Dietterich TG. Approximate statistical tests for comparing supervised classification learning algorithms. Neural Comput. 1998;10(7):1895–923.
55. Yan J, Kurgan L. Drnapred, fast sequence-based method that accurately predicts and discriminates dna-and rna-binding residues. Nucleic Acids Res. 2017;45(10):84.
56. 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. In: Bioinformatics and Biomedicine (BIBM), 2016 IEEE International Conference On. Shenzhen: IEEE; 2016. p. 78–85.
57. 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(5):634–6.
58. Hickman AB, James JA, Barabas O, Pasternak C, Ton-Hoang B, Chandler M, Sommer S, Dyda F. Dna recognition and the precleavage state during single-stranded dna transposition in d. radiodurans. EMBO J. 2010;29(22): 3840–52.