Information Sciences 384 (2017) 135–144

Contents lists available at ScienceDirect

Information Sciences

journal homepage: www.elsevier.com/locate/ins

Local-DPP: An improved DNA-binding protein prediction method by exploring local evolutionary information

Leyi Wei[a], Jijun Tang[a,b], Quan Zou[a,*]

[a] School of Computer Science and Technology, Tianjin University, Tianjin, China
[b] Department of Computer Science and Engineering, University of South Carolina, USA

ARTICLE INFO

Article history:
Received 5 November 2015
Revised 6 June 2016
Accepted 19 June 2016
Available online 23 June 2016

Keywords:
DNA-binding protein prediction
Random forest
Local evolutionary information
Machine learning-based method
Feature representation algorithm

ABSTRACT

Increased knowledge of DNA-binding proteins would enhance our understanding of protein functions in cellular biological processes. To handle the explosive growth of protein sequence data, researchers have developed machine learning-based methods that quickly and accurately predict DNA-binding proteins. In recent years, the predictive accuracy of machine learning-based predictors has significantly advanced, but the predictive performance remains unsatisfactory. In this paper, we establish a novel predictor named Local-DPP, which combines the local Pse-PSSM (Pseudo Position-Specific Scoring Matrix) features with the random forest classifier. The proposed features can efficiently capture the local conservation information, together with the sequence-order information, from the evolutionary profiles (PSSMs). We evaluate and compare the Local-DPP predictor with state-of-the-art predictors on two stringent benchmark datasets (one for the jackknife test, the other for an independent test). The proposed Local-DPP significantly improved the accuracy of the existing predictors, from 77.3% to 79.2% and 76.9% to 79.0% in the jackknife and independent tests, respectively. This demonstrates the efficacy and effectiveness of Local-DPP in predicting DNA-binding proteins. The proposed Local-DPP is now freely accessible to the public through the user-friendly webserver http://server.malab.cn/Local-DPP/Index.html.

1. Introduction

DNA-binding proteins play critical roles in various molecular functions, such as the detection of DNA damage, DNA replication, the combination and separation of single-stranded DNA, and transcriptional regulation [21,36]. Given the importance of DNA-binding proteins, methods for identifying members of this protein class are highly desired. In early research, DNA-binding proteins were determined by experimental approaches; typically by filter binding assays, genetic analysis, chromatin immune precipitation on microarrays, and X-ray crystallography [30]. However, experimental methods are costly in terms of time and resources [21]. With the development and application of next-generation high-throughput DNA sequencing techniques [38], the number of new protein sequences has exploded. Since 1986, the Swiss-Prot [3] database has expanded its protein sequence repository by more than 100 times. To handle such large-scale protein sequence data, we must replace the experimental methods with fast and effective computational methods. In recent years, computational methods based on machine learning (ML) algorithms have received much attention for their encouraging performance. Given a protein sequence as input, ML-based methods automatically predict whether that protein sequence binds to DNA.

* Corresponding author.
E-mail address: zouquan@nclab.net (Q. Zou).

http://dx.doi.org/10.1016/j.ins.2016.06.026
0020-0255/

The predictive performance of ML-based methods depends mainly on their feature representation and classification algorithms. Feature representation numerically formulates the best representation of a query protein sequence [45]. Feature representation methods employed in ML-based predictors are broadly classified into two groups; (1) structure-based predictors (i.e., [1,4,5,13,14,34,40-42,50,51]), and (2) sequence-based predictors (i.e., [7,11,12,19,25,30-33,35,37,39,43,46-49,52]).

Structure-based predictors rely heavily on the structural information (i.e. high-resolution 3-dimensional (3D) structure) of protein sequences. The method of Ahmad and Sarai [1] represents proteins with 62 structural features from the following three structural perspectives: the protein's net charge, electric dipole moment and quadrupole moment tensors. Similarly, Nimrod et al. [34] computed various structural characteristics of proteins from their average surface electrostatic potentials, dipole moments and cluster-based amino acid conservation patterns. Other predictors are based on both structural and sequential features. An example is the logistic regression (LR)-based predictor of Szilágyi and Skolnick [42], which uses the relative proportions of certain amino acids, the spatial distribution asymmetry of certain other amino acids, and the dipole moment of the whole molecule. However, structure-based predictors are inapplicable to protein sequences without known structural information. This limits the use of structure-based predictors in the post-genomic age, where next-generation sequencing techniques have yielded a huge number of uncharacterized genomic and proteomic sequences.

To successfully predict these sequences, we require sequence-based predictors that are free of structural information. Direct feature representation from primary sequences (amino acid sequences) has been recently developed. For instance, Cai and Lin [7] formulated a 40-dimensional (40D) feature vector that represents DNA-binding proteins from the pseudo amino acid composition (PseAAC) of proteins. Liu et al. [28] accelerated the computational time of Cai et al.'s algorithm by reducing the dimension of the PseAAC vector using a reduced-alphabet approach. To further improve DNA-binding protein prediction from the PseAAC vector, they also combined the PseAAC with a physicochemical distance transformation [27]. Besides PseAAC, DNA-binding proteins are represented by other commonly used sequence-based features, such as physicochemical properties [7,27,39,47], amino acid composition [7,42,49], autocross-covariance transformation [11,12], dipeptide composition [12,32], and other hybrid features [25]. Kumar et al. [20] newly incorporated evolutionary information into sequence-based methods. Evolutionary information is embedded in the sequence profiles that are automatically generated by PSI-BLAST [2]. The features containing the evolutionary information of PSI-BLAST profiles are called evolutionary features. Kumar et al. [20] combined the evolutionary and sequential features into a SVM predictor called DNAbinder. The evolutionary features significantly improved the predictive accuracy of their algorithm [20], suggesting that the evolutionary information is important for distinguishing DNA-binding proteins from non-DNA-binding proteins. Similar results were reported by Ho et al. [16]. Liu et al. [25] proposed a new method for DNA-binding protein prediction called iDNAPro-PseAAC, which integrates the profile-based representation of the evolutionary information retrieved by PSI-BLAST [29] into the classical PseAAC. Interestingly, they found that negative samples in the training model improved the predictive performance. Xu et al. [46] also proposed a SVM-based predictor that incorporates the evolutionary information into a general PseAAC vector via the top-n-gram approach. Recently, Song et al. [39] reported that the number of non-DNA-binding proteins in datasets far outweighs the number of DNA-binding proteins. They addressed the data imbalance problem by a novel ensemble classifier (imDC; see [39]). Furthermore, they programmed their imDC classifier in an improved DNA-binding protein predictor based on the 188D sequence physicochemical features.

As mentioned above, developing a feature representation algorithm that effectively encodes each query protein sequence as a feature vector is a challenging task. Most of the current multi-perspective efforts (sequence- and structure-based) consider only the global features, which may not be sufficiently informative to distinguish between DNA-binding proteins and non-DNA-binding ones. The major difference between DNA-binding and non-DNA-binding proteins is the presence of functional binding sites in the former, which are lacking in the corresponding local regions of protein space in the latter. Moreover, protein functions in these local regions are likely to be evolutionarily conserved. Consequently, a perfect classification must capture this local functional conservation information and quantize it with a feature vector.

To address this problem, we propose a novel feature representation algorithm that efficiently extracts the local features from the profiles (PSSM). Within the framework of the proposed algorithm, we first capture the locally conserved protein information by fragmenting the PSSMs into several equally sized sub-PSSMs. For each sub-PSSM, we compute the local features by the Pse-PSSM feature extraction algorithm. Finally, we combine the local Pse-PSSM features from all sub-PSSMs to form the features. Based on the proposed features, we develop our Local-DPP machine learning-based method, which predicts DNA-binding proteins by a RF classifier. Evaluated on two stringent benchmark datasets (one for the jackknife test, and the other for an independent test), the Local-DPP demonstrated superior performance to state-of-the-art predictors. Local-DPP is freely downloadable from the user-friendly website http://server.malab.cn/Local-DPP/Index.html. Local-DPP is expected to become a useful tool for predicting and analyzing large-scale DNA-binding proteins.

2. Materials and methods

2.1. Framework of the proposed method

Fig. 1 illustrates the overall framework of the Local-DPP method for DNA-binding protein prediction. The two stages of Local-DPP are model training and protein prediction. In the training phase, the training samples are first encoded by the proposed local Pse-PSSM feature representation algorithm, obtaining the meaningful feature vectors for the training set. These feature vectors are then fed into the RF classifier to generate the training model. The feature representation algorithm

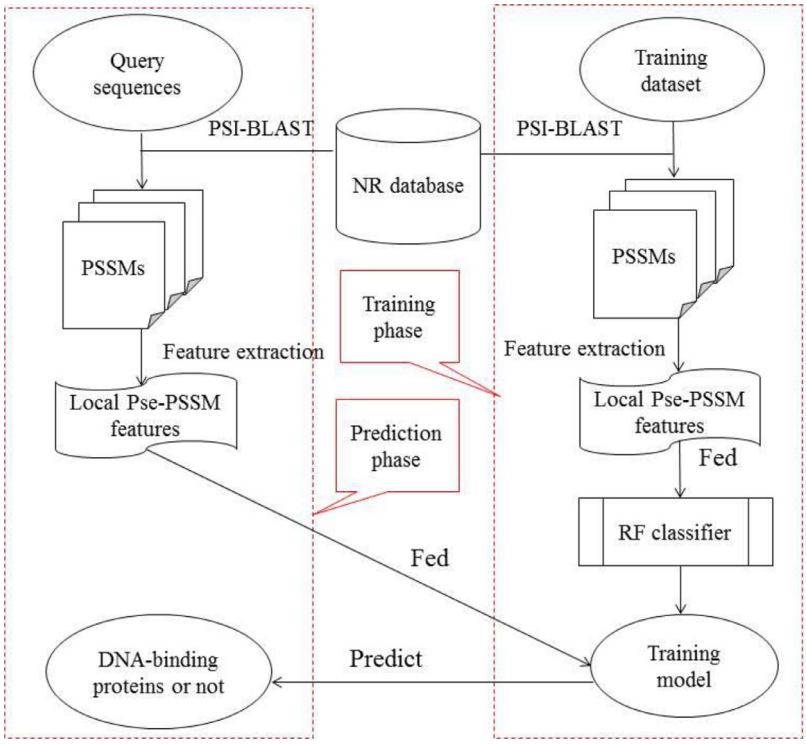

Fig. 1. Overall framework of the proposed Local-DPP predictor. NR denotes the non-redundant protein database, PSSM represents Position-Specific Scoring Matrix, and RF denotes Random Forest. The overall framework contains two phases: (1) model training phase and (2) prediction phase. In the model training phase, training samples are fed into the proposed feature representation algorithm to generate the local Pse-PSSM features. The resulting features are then fed into the RF classifier to generate the training model. The prediction phase bases its predictions on the feature representations of the query sequences.

also encodes a query protein sequence as a 120D feature vector in the protein prediction stage. This feature vector is fed into the training model, which predicts whether the query sequence binds to DNA.

2.2. Datasets

The reliability and accuracy of a predictor must be evaluated on a well-established dataset. After reviewing the recent literature on DNA-binding protein prediction, we noticed that dataset construction generally follows certain common procedures. In the first step, the DNA-binding protein sequences are acquired from the Protein Data Bank (PDB: http://www.rcsb.org/pdb/home/home.do) dataset by searching the relevant keywords (such as "DNA-binding protein", "Protein-DNA complex" or "DNA-binding") in the Advanced Search interface. The second step removes short sequences (less than 50 amino acids) and sequences containing the consecutive character "X". The third step eliminates the redundancy and homology bias that likely leads to overestimated performance. To this end, it removes sequences with $\geq$25% pairwise sequence identity to any other sequences in the dataset using the program CD-HIT [18]. Sequences that pass these three steps are assembled into a stringent benchmark dataset.

In this study, the predictor accuracy was tested on two stringent benchmark datasets. The first benchmark dataset called PDB1075, originally compiled by Liu et al. [28], contains 525 DNA-binding proteins (positive samples) and 550 non-DNA-binding proteins (negative samples) selected from PDB (version released in December 2013). As reported in [28], the PDB1075 dataset includes the highest number of protein sequences with low similarity, which is desirous for model evaluation. The other benchmark dataset, called PDB186, was recently constructed by Lou et al. [30], and contains 93 actual DNA-binding and 93 non-DNA-binding proteins also collected from PDB. The PDB186 dataset provides an independent test for validating the predictors. All sequences in these two benchmark datasets are currently downloadable from our webserver (http://server.malab.cn/Local-DPP/Datasets.html).

2.3. Classification algorithm

The RF algorithm is a popular machine learning algorithm proposed by Breiman [6]. The RF algorithm is an ensemble of tree predictors. Each tree is grown by two factors: (1) a random feature vector sampled from the original feature space,

and (2) random bootstrap data sampled from the original data. It should be noted that all trees are independent. The number of features for each tree is determined by computing the generalization error, classifier strength and dependence. The prediction result of the RF algorithm is the ensemble of the results of all trained trees combined with the majority voting strategy. The RF algorithm is detailed in [6]. Our proposed method employs the RF algorithm as the underlying classification algorithm. The RF algorithm is implemented in a data mining tool called WEKA (Waikato Environment for Knowledge Analysis) [15], an ensemble package of several machine learning algorithms. All experiments in this paper were carried out in version 3.7 of WEKA.

2.4. Feature representation algorithm

Evolutionary information embedded in the profiles (PSSMs) has been widely applied in protein fold prediction [45], protein structural class prediction [44], protein remote homology detection [26,29], and other similar fields. Our novel feature representation algorithm efficiently maps the query protein sequences onto a discriminative feature space by incorporating both evolutionary and local conservation information. In the following subsections, we briefly introduce the profile, and describe its application in the proposed feature representation algorithm.

Position-specific scoring matrix (PSSM). A given protein sequence S is represented as $S_1S_2 \dots S_L$, where $S_i (1 \le i \le L)$ represents the amino acid (residue) appearing in the ith position of S, and L is the length of S. The so-called evolutionary profile of S is the position-specific scoring matrix (PSSM), generated by three iterations of a PSI-BLAST [2] search of the protein database *nrdb90* [17]. The E-value (expectation value) cutoff for the multiple sequence alignment was 0.001. The PSSM contains the probability that each type of amino acid is found at each residue position of the protein sequence during the evolutionary process. Hence, the PSSM measures the residue conservation at a given location. The evolutionary information in the PSSM is stored in a matrix of dimensions $L \times 20$ (L rows and 20 columns), formulated as follows:

$$\boldsymbol{P}_{\boldsymbol{original}} = \begin{bmatrix} p_{1,1} & p_{1,2} & \cdots & p_{1,20} \\ p_{2,1} & p_{2,2} & \cdots & p_{2,20} \\ \vdots & \vdots & \vdots & \vdots \\ p_{i,1} & p_{i,2} & \cdots & p_{i,20} \\ \vdots & \vdots & \vdots & \vdots \\ p_{L,1} & p_{L,2} & \cdots & p_{L,20} \end{bmatrix}_{L\times 20} \tag{1}$$

where a row denotes the corresponding position of the sequence S. For example, the first, second and Lth rows refer to the first, second and Lth positions of S, respectively. The columns represent the corresponding residue type of the 20 amino acids {A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W, Y}; for example, the first, second and 20th columns refer to "A", "C" and "Y", respectively. An entry $p_{i,j}$ represents the score of the residue at the ith position of S being mutated to residue type j during the evolutionary process ($1 \le i \le L,\ 1 \le j \le 20$). The higher the $p_{i,j}$ score, the more frequent the mutation (in general). Residues at highly mutable sites are likely to be functional.

Local Pse-PSSM features. The Pse-PSSM features proposed by Chou and Shen [8] were targeted at membrane protein prediction. These features can sufficiently explore the evolutionary information and the sequence-order information embedded in PSSMs [8]. However, if the DNA-binding protein sequences are directly represented by Chou's Pse-PSSM features, all of the local conservation information during the evolutionary process would be lost. To preserve the local conservation information, we modified Chou's Pse-PSSM features by the following steps.

Step 1. Normalize PSSM. The original PSSM (see Eq. (1)) is normalized as follows:

$$f_{i,j} = \frac{p_{i,j} - \frac{1}{20}\sum_{k=1}^{20} p_{i,k}}{\sqrt{\frac{1}{20}\sum_{l=1}^{20}\left(p_{i,l} - \frac{1}{20}\sum_{k=1}^{20} p_{i,k}\right)^2}},\ (i = 1, 2, \dots, L;\ j = 1, 2, \dots, 20) \tag{2}$$

where $p_{i,j}$ represent the original scores of PSSM. The normalized scores ($f_{i,j}$) have a zero mean over the 20 amino acids. A positive (negative) score indicates that the corresponding mutation occurs more (less) frequently in the multiple alignment than expected by chance. The normalized PSSM is represented by

$$\boldsymbol{P}_{\boldsymbol{normalized}} = \begin{bmatrix} f_{1,1} & f_{1,2} & \cdots & f_{1,20} \\ f_{2,1} & f_{2,2} & \cdots & f_{2,20} \\ \vdots & \vdots & \vdots & \vdots \\ f_{i,1} & f_{i,2} & \cdots & f_{i,20} \\ \vdots & \vdots & \vdots & \vdots \\ f_{L,1} & f_{L,2} & \cdots & f_{L,20} \end{bmatrix}_{L\times 20} \tag{3}$$

Step 2. Fragment the normalized matrix $\boldsymbol{P}_{\boldsymbol{normalized}}$. This step row-fragments the normalized matrix $\boldsymbol{P}_{\boldsymbol{normalized}}$ into n (≥ 1) sub-matrices. For convenience, we denote the kth sub-matrix as $\boldsymbol{P}^{\boldsymbol{k}}_{\boldsymbol{normalized}} (1 \le k \le n)$. Each of the first $(n-1)$ sub-matrices has L/n rows and 20 columns; the final sub-matrix ($\boldsymbol{P}^{\boldsymbol{n}}_{\boldsymbol{normalized}}$) has $(L - (n-1) * L/n)$ rows and 20 columns. Note

that the sizes of the first $(n-1)$ sub-matrices and the last sub-matrix are equal if and only if L/n is an integer. Moreover, every sub-matrix retains the evolutionary information contained in the original PSSM. Importantly, the fragmentation operation captures the local conservation information, because this information is always embedded in the local regions.

Step 3. Compute the local Pse-PSSM features for all sub-matrices. To quantize the local conservation information of the protein S, we computed the local Pse-PSSM features for all sub-matrices. However, $\boldsymbol{P^k_{normalized}}(1 \le k \le n-1)$ and $\boldsymbol{P^n_{normalized}}$ are not necessarily equal in size. Therefore, the features were separately computed for $\boldsymbol{P^k_{normalized}}(1 \le k \le n-1)$ and $\boldsymbol{P^n_{normalized}}$.

For each of the first $(n-1)$ sub-matrices ($\boldsymbol{P^k_{normalized}}(1 \le k \le n-1)$), we computed 20 local features by incorporating the evolutionary information as follows:

$$Part_1 = \{F_j(k) = \frac{1}{L/n} \sum_{i=(k-1)L/n}^{k*L/n} f_{i,j} | 1 \le k \le n-1; j = 1, 2, \ldots, 20\} \tag{4}$$

where $F_j(k)$ denotes the average probability that each residue position in the k^{th} fragmented sequence mutates to residue type j during the evolutionary process. Thus, we obtained $(n-1) \times 20$ local features containing evolutionary information for the first $(n-1)$ sub-matrices.

To incorporate the sequence-order information, we represent the protein S by

$$Part_2 = \left\{ \Phi_j^{\xi}(k) = \frac{1}{\frac{L}{n} - \xi} \sum_{i=(k-1)\frac{L}{n}}^{k*\frac{L}{n}-\xi} \left(f_{i,j} - f_{(i+\xi),j}\right)^2 | \ \xi = 1, \ldots, \lambda; 1 < \lambda < \frac{L}{n} \right\} \tag{5}$$

where $\Phi_j^{\xi}(k)$ is the average correlation between two coupled residues separated by ξ for amino acid type j in the kth sub-matrix. For example, $\Phi_j^1(k)$ and $\Phi_j^2(k)$ are the correlation factors obtained by coupling contiguous residues and every two residues along the protein chain, respectively, for amino acid type j in the kth sub-matrix. The maximum L should be the minimum length of the sequences in the dataset.

After combining the local features containing evolutionary information ($Part_1$) and sequence-order information ($Part_2$), we obtained $20(n-1)(1+\lambda)$ local Pse-PSSM features for the first $(n-1)$ sub-matrices. The space representation of the features is given by

$$FV(n-1) = (Part_1, Part_2) \tag{6}$$

The local Pse-PSSM for the last sub-matrix ($\boldsymbol{P^n_{normalized}}$) is given by

$$FV(n) = \left(F_1(n), \ldots, F_{20}(n), \Phi_1^1(n), \ldots, \Phi_{20}^1(n), \ldots, \Phi_1^{\lambda}(n), \ldots, \Phi_{20}^{\lambda}(n)\right) \tag{7}$$

where $F_j(n)$ and $\Phi_j^{\xi}(n)$ are computed as described for the first $(n-1)$ sub-matrices.

The final feature vector combines the feature vectors $FV(n-1)$ and $FV(n)$ to give

$$FV = (FV(n-1), FV(n)) \tag{8}$$

Here, we selected the best-performing parameters ($\lambda = 1$ and $n = 3$) as the default parameters. The parameter optimization is detailed in subsection 3.5 (Parameter o ptimization). The protein sequence is finally represented as a 120D feature vector.

2.5. Measurements

The effectiveness of a predictor can be rigorously analyzed by leave-one-out cross-validation (LOOCV). In the LOOCV test, every protein is removed one-by-one from the training set, and the predictor is trained by the remaining proteins in the learning dataset. The isolated protein is then tested by the trained predictor. The LOOCV test (also known as the jackknife test) is widely used for evaluating DNA-binding protein predictors (e.g., [19,20,24,27-29,33,39,47,52]). In the present study, the LOOCV test is carried out for a fair comparison with existing methods.

To comprehensively examine the prediction quality of our predictor, we employ four commonly used evaluation metrics; Sensitivity (SE), Specificity (SP), Accuracy (ACC), and Mathew's correlation coefficient (MCC). These metrics are respectively formulated as follows:

$$\text{SE} = \frac{TP}{TP+FN} * 100\%$$

$$\text{SP} = \frac{TN}{TN+FP} * 100\%$$

$$\text{ACC} = \frac{TP+TN}{TP+TN+FN+FP} * 100\%$$

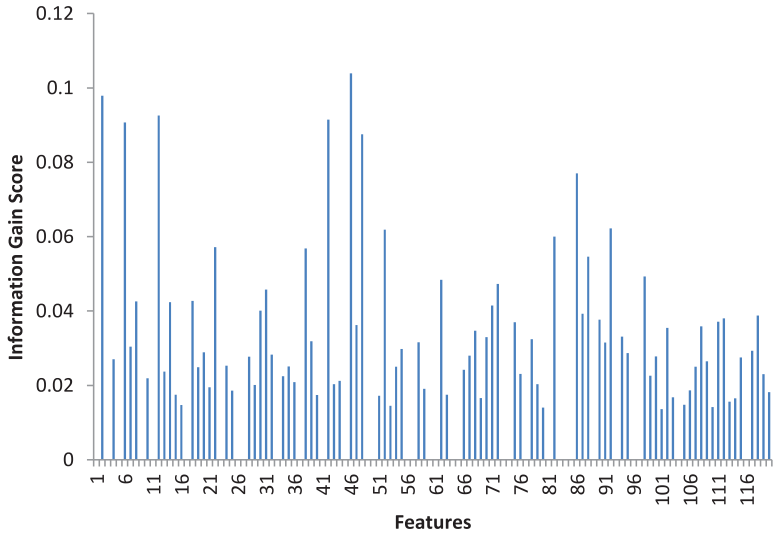

Fig. 2. Information gain scores of the proposed Pse-PSSM features. The parameters of the proposed features were $n = 3$ and $\lambda = 1$. The x-axis represents the numbers (1–120) of the features in the 120D feature vector.

Table 1
Twenty top-scoring features among the proposed features evaluated on the benchmark dataset PDB1075.

Rank	Features	IG[a]	Rank	Features	IG[a]
1	Feature_46	0.1039	11	Feature_22	0.0572
2	Feature_2	0.0979	12	Feature_38	0.0568
3	Feature_12	0.0926	13	Feature_88	0.0546
4	Feature_42	0.0915	14	Feature_98	0.0493
5	Feature_6	0.0907	15	Feature_62	0.0484
6	Feature_48	0.0875	16	Feature_72	0.0473
7	Feature_86	0.077	17	Feature_31	0.0458
8	Feature_92	0.0622	18	Feature_18	0.0427
9	Feature_52	0.0619	19	Feature_8	0.0426
10	Feature_82	0.06	20	Feature_14	0.0424

[a] Denotes the information gain score.

$$\mathrm{MCC} = \frac{TP * TN - FP * FN}{\sqrt{(TP+FN)(TP+FP)(TN+FP)(TN+FN)}} * 100\%$$

where *TP*, *TN*, *FP*, and *FN* represent the numbers of true positive, true negative, false positive, and false negative, respectively. The SE (SP) measures the ratio of predicted DNA-binding proteins (non-DNA-binding proteins) that are true DNA-binding proteins (true non-DNA-binding proteins), ACC measures the ratio of the correct predictions among all true DNA-binding and non-DNA-binding proteins, and MCC measures the degree of overlap between all predictions and true predictions. The MCC ranges from -1 (all predictions are incorrect) to $+1$ (all predictions are correct). In particular, the MCC score of 0 corresponds to random predictions.

3. Results and discussion

3.1. Feature importance analysis

In this section, we analyze the importance of the proposed 120 local Pse-PSSM features for DNA-binding protein prediction. Feature importance is measured by the information gain score IG(c, x), which denotes the information gain of feature x relative to class attribute c [9]. The IG scores of the proposed 120 features were evaluated on the benchmark dataset PDB1075, and the obtained scores are presented in Fig. 2. The top 20 important features and their IG scores are summarized in Table 1. Among the top 20 scoring features, Feature_46 was most important, with an IG score of 0.1039. To analyze the distribution of the local features, we divided the proposed 120 features into three intervals: [1, 40), [41, 80), and [81, 120). The distribution of the top 20 features in each interval is presented in Fig. 3. This figure reveals nine important features distributed in the interval [1, 40), six in [41, 80), and five in [81, 120). This indicates that features from the interval [1, 40)

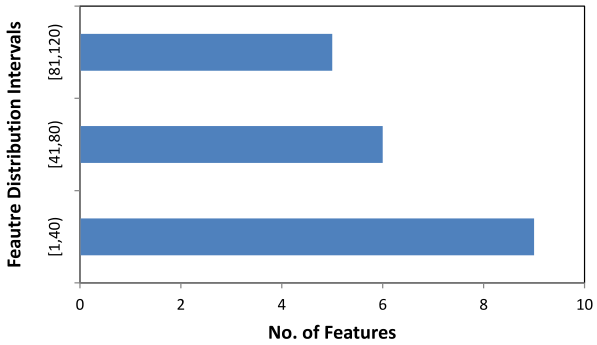

Fig. 3. Distribution of the 20 top-scoring features among the proposed features evaluated on the benchmark dataset PDB1075.

Table 2
Evaluation results of local Pse-PSSM features and Chou's Pse-PSSM features on the PDB1075 dataset (Jackknife validation test).

	ACC (%)	MCC	SE (%)	SP (%)
Chou's Pse-PSSM features	77.3	0.549	81.3	73.5
Local Pse-PSSM features ($n=2$)	78.5	0.574	83.2	**74.0**
Local Pse-PSSM features ($n=3$)	**79.1**	**0.587**	**84.8**	73.6
Local Pse-PSSM features ($n=4$)	78.2	0.569	83.4	73.3
Local Pse-PSSM features ($n=5$)	78.0	0.564	82.5	73.8
Local Pse-PSSM features ($n=6$)	76.8	0.540	81.3	72.5

n represents the number of fragments of the PSSM.

are more informative than features from the other two intervals. In fact, the three intervals correspond to the three locally fragmented regions of the profile from which we extracted the proposed features. Thus, we infer that the first fragment of the profile contains more discriminative information (i.e., more conservation information) than the other two fragments. The information gain scores of all 120 features are detailed in Supplement A.

3.2. Comparisons of local and global features

The local Pse-PSSM features were obtained from Chou's Pse-PSSM by varying the parameter n (see Section 2.4; Feature representation algorithm). When $n > 1$, our features contained the local conservation information; when $n = 1$, they contained the global information. To investigate the importance of the local conservation information, we compared the proposed local Pse-PSSM features (local features) and Chou's Pse-PSSM features (global features). The comparison results are presented in Table 2. Before discussing these results, we should note that the parameter n in the table represents the number of fragmented sub-matrices; for example, $n = 2$ and $n = 3$ denote that the PSSM is row-divided into two and three sub-matrices of equal size, respectively. According to Table 2, most of the local features performed better than the global features. The local features for $n = 3$ obtained the highest ACC (79.1%), MCC (0.587), SE (84.8%), and the third highest SP (73.6%). These scores were 1.8%, 0.038, 3.5%, and 0.1% higher than the maximum scores of the global features, respectively. This indicates that features derived from local regions are more discriminative than features derived from the whole region.

Why are the features extracted from local regions more informative than their global counterparts? We can reasonably expect that the fragmented local regions contain the functional conservation information that discriminates between DNA-binding proteins and non-DNA-binding proteins. This supposition is indirectly confirmed by other information in Table 2; in particular, the predictive performance decreased as the number of fragmented regions increased from four to six. In fact, the local features for $n = 6$ performed worse than the global features (see Table 2). Dividing the original region into many small fragments might degrade the performance by breaking the local conservation information.

3.3. Comparisons with state-of-the-art predictors on the benchmark dataset PDB1075

In this subsection, the performance of Local-DPP is evaluated on the benchmark dataset PDB1075, and compared with the performances of several state-of-the-art predictors; namely, iDNA-Prot|dis [28], iDNA-Prot [24], DNA-Prot [19], PseDNA-Pro [27], DNAbinder [20], iDNAPro-PseAAC [25], and Kmer1+ACC [11]. The predictive results of the jackknife validation test are presented in Table 3. Among the evaluated methods, the proposed Local-DPP (with $n = 3$ and $\lambda = 1$) achieved the best predictive performance on three metrics: ACC (79.20%), MCC (0.59) and SE (84.00%). The ACC and MCC of the proposed method (with $n = 3$ and $\lambda = 1$) was 1.8% and 0.05 higher, respectively, than the best-performing predictor iDNA-Prot|dis (ACC = 77.30% and MCC = 0.54). In summary, the proposed method outperforms existing state-of-the-art methods in the prediction of DNA-binding proteins, demonstrating the superiority and effectiveness of the proposed method.

Table 3
Results of the proposed method and state-of-the-art predictors on the benchmark dataset PDB1075 (Jackknife test evaluation).

Methods	ACC (%)	MCC	SE (%)	SP (%)
iDNA-Prot|dis	77.30	0.54	79.40	75.27
PseDNA-Pro	76.55	0.53	79.61	73.63
iDNA-Prot	75.40	0.50	83.81	64.73
DNA-Prot	72.55	0.44	82.67	59.76
DNAbinder (dimension = 400)	73.58	0.47	66.47	**80.36**
DNAbinder (dimension = 21)	73.95	0.48	68.57	79.09
iDNAPro-PseAAC	76.56	0.53	75.62	77.45
Kmer1 + ACC	75.23	0.50	76.76	73.76
The proposed method ($n = 3$, $\lambda = 1$)	79.10	**0.59**	**84.80**	73.60
The proposed method ($n = 2$, $\lambda = 2$)	**79.20**	**0.59**	84.00	74.50

Table 4
Results of the proposed method and state-of-the-art predictors on the independent dataset PDB186.

Methods	ACC (%)	MCC	SE (%)	SP (%)
iDNA-Prot|dis	72.0	0.445	79.5	64.5
iDNA-Prot	67.2	0.344	67.7	66.7
DNA-Prot	61.8	0.240	69.9	53.8
DNAbinder	60.8	0.216	57.0	64.5
DNABIND	67.7	0.355	66.7	68.8
DNA-Threader	59.7	0.279	23.7	**95.7**
DBPPred	76.9	0.538	79.6	74.2
iDNAPro-PseAAC-EL[a]	71.5	0.442	82.8	60.2
Kmer1+ACC	71.0	0.431	82.8	59.1
The proposed method ($n = 3$, $\lambda = 1$)	**79.0**	**0.625**	**92.5**	65.6
The proposed method ($n = 2$, $\lambda = 2$)	**77.4**	**0.568**	90.3	64.5

[a] iDNAPro-PseAAC-EL denotes the iDNAPro-PseAAC method using the ensemble learning algorithm.

3.4. Comparisons with state-of-the-art predictors on an independent dataset PDB186

To examine the robustness of the proposed method, we evaluated Local-DPP on an independent dataset (PDB186), and again compared its performance with those of existing methods. PDB186 contains 93 DNA-binding and 93 non-DNA-binding proteins. To avoid homology bias between the training set (PDB1075) and the independent set (PDB186), we followed the procedure of Liu et al. [28], removing those proteins in the PDB1075 dataset with more than 25% sequence identity to any protein in the PDB186 dataset using BLASTCLUST [10], and rebuilding the proposed method on the removed PDB1075 dataset. The independent test results are presented in Table 4. The proposed method (with $n = 3$ and $\lambda = 1$) achieved the highest ACC, MCC, and SE among the evaluated methods, and outperformed the existing best-performing predictor DBPPred (ACC = 76.9%, MCC = 0.538, and SE = 79.6%) by 2.1% in ACC, 0.087 in MCC, and 12.9% in SE. The independent test corroborates the previous test results, confirming that our proposed predictor effectively identifies DNA-binding proteins. Because the proposed method robustly performed in the independent test, it should effectively predict novel DNA-binding proteins.

3.5. Parameter optimization

In this subsection, we optimize the parameters of the proposed feature representation algorithm. The parameters of the proposed local Pse-PSSM feature extraction are n and λ (see Section 2.4, Feature representation algorithm, for details). To optimize these parameters, we implemented the proposed method on the benchmark dataset PDB1075, varying λ from 1 to 7 and n from 1 to 2, and evaluated the predictive performance by the jackknife test. Table 5 presents the predictive results of the proposed method for different values of n and λ. The performance is maximized at two parameter combinations; $n = 3$ and $\lambda = 1$ (ACC = 79.1%; MCC = 0.587), and $n = 2$ and $\lambda = 2$ (ACC = 79.2%; MCC = 0.587). Therefore, both combinations were set as the default parameter values for generating the proposed features. We note that either combination ($n = 3$, $\lambda = 1$ or $n = 2$, $\lambda = 2$) generates a 120D feature vector for a query protein.

4. Conclusions

This paper presented a novel machine learning-based method called Local-DPP for DNA-binding protein prediction. Within the framework of Local-DPP, we proposed a novel feature representation algorithm that addresses the challenging problem of discretizing protein sequences such that DNA-binding proteins and non-DNA-binding proteins are effectively

Table 5
Results of different n and λ in local Pse-PSSM feature selection (evaluated on the benchmark dataset PDB1075).

Metrics		$n=1$	$n=2$	$n=3$	$n=4$	$n=5$	$n=6$	$n=7$
ACC (%)	$\lambda=1$	77.3	78.5	**79.1**	78.2	78.0	76.8	76.4
	$\lambda=2$	78.4	**79.2**	78.4	77.9	77.1	76.9	76.4
MCC	$\lambda=1$	0.549	0.574	**0.587**	0.569	0.564	0.540	0.530
	$\lambda=2$	0.570	**0.587**	0.572	0.561	0.545	0.541	0.529
SE (%)	$\lambda=1$	81.3	83.2	**84.8**	83.4	82.5	81.3	80.4
	$\lambda=2$	81.1	**84.0**	83.2	82.7	81.1	80.2	78.7
SP (%)	$\lambda=1$	73.5	74.0	**73.6**	73.3	73.8	72.5	72.5
	$\lambda=2$	75.8	**74.5**	73.8	73.3	73.3	73.8	74.2

discriminated. The proposed feature representation algorithm extracts local features by segmenting the original large PSSM matrix into several sub-matrices of equal size. In experimental evaluations, the evolutionary features containing local information outperformed the evolutionary features containing global information. This indicates that the local information embedded in the profiles (PSSMs) improves the prediction of DNA-binding proteins. To investigate the prediction quality of Local-DPP, we compared its performance with those of state-of-the-art predictors on two stringent benchmark datasets (PDB1075 and PDB186, evaluated by the jackknife test and an independent test, respectively). In the jackknife test, Local-DPP yielded the best ACC (79.1%) and MCC (0.587), leading the existing methods by 1.9 – 6.7% and 0.05 – 0.15, respectively. Similarly, in the independent test, Local-DPP achieved the highest ACC (79.0%) and MCC (0.625), leading the existing predictors by 2.1 – 19.3% and 0.087 – 0.409 respectively. The superior performance on both datasets, especially in the independent test, confirms the potential effectiveness of Local-DPP in predicting DNA-binding proteins. Our proposed webserver can also predict DNA-binding proteins in large-scale datasets for practical applications.

In future work, we will improve the predictive performance of Local-DPP by refining the feature representation and classification algorithm. For feature representation, we will consider the incorporation of other biologically relevant features (such as predicted secondary structures and amino acid compositions) into the proposed local PSSM-based features. This additional information might generate new discriminators for the classification. To improve the predictive power of classification algorithm, we will consider a well-established ensemble classifier, such as that developed by Lin et al. [22,23]. Their powerful ensemble classifier, namely LibD3C, adopts a clustering and dynamic selection strategy. LibD3C is more effective than single basic classifiers in the fields of protein fold prediction [23] and cytokine prediction [53].

Acknowledgement

The work is supported by the National Natural Science Foundation of China (61370010).

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.ins.2016.06.026.

References

[1] S. Ahmad, A. Sarai, Moment-based prediction of DNA-binding proteins, J. Mol. Biol. 341 (2004) 65–71.
[2] S.F. Altschul, T.L. Madden, A.A. Schaffer, J. Zhang, Z. Zhang, W. Miller, D.J. Lipman, Gapped BLAST and PSI-BLAST: a new generation of protein database search programs, Nucleic Acids Res. 25 (1997) 3389–3402.
[3] A. Bairoch, R. Apweiler, The SWISS-PROT protein sequence data bank and its supplement TrEMBL, Nucleic Acids Res. 25 (1997) 31–36.
[4] N. Bhardwaj, R.E. Langlois, G. Zhao, H. Lu, Kernel-based machine learning protocol for predicting DNA-binding proteins, Nucleic Acids Res. 33 (2005) 6486–6493.
[5] N. Bhardwaj, H. Lu, Residue-level prediction of DNA-binding sites and its application on DNA-binding protein predictions, FEBS Lett. 581 (2007) 1058–1066.
[6] L. Breiman, Random forests, Mach. Learn. 45 (2001) 5–32.
[7] Y.-d. Cai, S.L. Lin, Support vector machines for predicting rRNA-, RNA-, and DNA-binding proteins from amino acid sequence, Biochim. Biophys. Acta 1648 (2003) 127–133.
[8] K.-C. Chou, H.-B. Shen, MemType-2L: a web server for predicting membrane proteins and their types by incorporating evolution information through Pse-PSSM, Biochem. Biophys. Res. Commun. 360 (2007) 339–345.
[9] H. Deng, G. Runger, E. Tuv, Bias of importance measures for multi-valued attributes and solutions, in: Artificial Neural Networks and Machine Learning, Springer, 2011, pp. 293–300.
[10] I. Dondoshansky, Y. Wolf, Blastclust (NCBI Software Development Toolkit), NCBI, Bethesda, Md, 2002.
[11] Q. Dong, S. Wang, K. Wang, X. Liu, B. Liu, Identification of DNA-binding proteins by auto-cross covariance transformation, in: Bioinformatics and Biomedicine (BIBM), 2015 IEEE International Conference on, IEEE, 2015, pp. 470–475.
[12] Y. Fang, Y. Guo, Y. Feng, M. Li, Predicting DNA-binding proteins: approached from Chou's pseudo amino acid composition and other specific sequence features, Amino. Acids. 34 (2008) 103–109.
[13] M. Gao, J. Skolnick, DBD-Hunter: a knowledge-based method for the prediction of DNA–protein interactions, Nucleic Acids Res. 36 (2008) 3978–3992.
[14] M. Gao, J. Skolnick, A threading-based method for the prediction of DNA-binding proteins with application to the human genome, PLoS Comput. Biol. 5 (2009) e1000567.
[15] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, I.H. Witten, The WEKA data mining software: an update, ACM SIGKDD Explor. Newslett. 11 (2009) 10–18.

[16] S.-Y. Ho, F.-C. Yu, C.-Y. Chang, H.-L. Huang, Design of accurate predictors for DNA-binding sites in proteins using hybrid SVM–PSSM method, Biosystems 90 (2007) 234–241.
[17] L. Holm, C. Sander, Removing near-neighbour redundancy from large protein sequence collections, Bioinformatics. 14 (1998) 423–429.
[18] Y. Huang, B. Niu, Y. Gao, L. Fu, W. Li, CD-HIT Suite: a web server for clustering and comparing biological sequences, Bioinformatics 26 (2010) 680–682.
[19] K.K. Kumar, G. Pugalenthi, P. Suganthan, DNA-Prot: identification of DNA binding proteins from protein sequence information using random forest, J. Biomol. Struct. Dyn. 26 (2009) 679–686.
[20] M. Kumar, M.M. Gromiha, G.P. Raghava, Identification of DNA-binding proteins using support vector machines and evolutionary profiles, BMC Bioinform. 8 (2007) 463.
[21] R.E. Langlois, H. Lu, Boosting the prediction and understanding of DNA-binding domains from sequence, Nucleic Acids Res. (2010) gkq061.
[22] C. Lin, W. Chen, C. Qiu, Y. Wu, S. Krishnan, Q. Zou, LibD3C: ensemble classifiers with a clustering and dynamic selection strategy, Neurocomputing 123 (2014) 424–435.
[23] C. Lin, Y. Zou, J. Qin, X. Liu, Y. Jiang, C. Ke, Q. Zou, Hierarchical classification of protein folds using a novel ensemble classifier, PLoS One 8 (2013) e56499.
[24] W.-Z. Lin, J.-A. Fang, X. Xiao, K.-C. Chou, iDNA-Prot: identification of DNA binding proteins using random forest with grey model, PLoS One 6 (2011) e24756.
[25] B. Liu, S. Wang, X. Wang, DNA binding protein identification by combining pseudo amino acid composition and profile-based protein representation, Scient. Rep. 5 (2015) 15479.
[26] B. Liu, X. Wang, Q. Zou, Q. Dong, Q. Chen, Protein Remote Homology Detection by Combining Chou's Pseudo Amino Acid Composition and Profile-Based Protein Representation, Mol. Inf. 32 (2013) 775–782.
[27] B. Liu, J. Xu, S. Fan, R. Xu, J. Zhou, X. Wang, PseDNA-Pro: DNA-Binding Protein Identification by Combining Chou's PseAAC and Physicochemical Distance Transformation, Mol. Inf. 34 (2015) 8–17.
[28] B. Liu, J. Xu, X. Lan, R. Xu, J. Zhou, X. Wang, K.-C. Chou, iDNA-Prot| dis: Identifying DNA-binding proteins by incorporating amino acid distance-pairs and reduced alphabet profile into the general pseudo amino acid composition, PLoS One 9 (2014) e106691.
[29] B. Liu, D. Zhang, R. Xu, J. Xu, X. Wang, Q. Chen, Q. Dong, K.-C. Chou, Combining evolutionary information extracted from frequency profiles with sequence-based kernels for protein remote homology detection, Bioinformatics 30 (2014) 472–479.
[30] W. Lou, X. Wang, F. Chen, Y. Chen, B. Jiang, H. Zhang, Sequence based prediction of DNA-binding proteins based on hybrid feature selection using random forest and Gaussian naive Bayes, PLoS One 9 (2014).
[31] G.B. Motion, A.J. Howden, E. Huitema, S. Jones, DNA-binding protein prediction using plant specific support vector machines: validation and application of a new genome annotation tool, Nucleic Acids Res. 43 (2015) e158-e158.
[32] L. Nanni, A. Lumini, Combing ontologies and dipeptide composition for predicting DNA-binding proteins, Amino. Acids. 34 (2008) 635–641.
[33] L. Nanni, A. Lumini, An ensemble of reduced alphabets with protein encoding based on grouped weight for predicting DNA-binding proteins, Amino. Acids. 36 (2009) 167–175.
[34] G. Nimrod, M. Schushan, A. Szilágyi, C. Leslie, N. Ben-Tal, iDBPs: a web server for the identification of DNA binding proteins, Bioinformatics. 26 (2010) 692–693.
[35] A.K. Patel, S. Patel, P.K. Naik, Binary Classification of Uncharacterized Proteins into DNA Binding/Non-DNA Binding Proteins from Sequence Derived Features Using Ann, Dig. J. Nanomat. Biostruct. (DJNB) 4 (2009).
[36] A. Sarai, H. Kono, Protein-DNA recognition patterns and predictions, Annu. Rev. Biophys. Biomol. Struct. 34 (2005) 379–398.
[37] X. Shao, Y. Tian, L. Wu, Y. Wang, L. Jing, N. Deng, Predicting DNA-and RNA-binding proteins from sequences with kernel methods, J. Theor. Biol. 258 (2009) 289–293.
[38] J. Shendure, H. Ji, Next-generation DNA sequencing, Nat. Biotechnol. 26 (2008) 1135–1145.
[39] L. Song, D. Li, X. Zeng, Y. Wu, L. Guo, Q. Zou, nDNA-prot: identification of DNA-binding proteins based on unbalanced classification, BMC Bioinform. 15 (2014) 298.
[40] E.W. Stawiski, L.M. Gregoret, Y. Mandel-Gutfreund, Annotating nucleic acid-binding function based on protein structure, J. Mol. Biol. 326 (2003) 1065–1079.
[41] A. Szabóová, O. Kuželka, F. Železný, J. Tolar, Prediction of DNA-binding propensity of proteins by the ball-histogram method using automatic template search, BMC Bioinform. 13 (2012) S3.
[42] A. Szilágyi, J. Skolnick, Efficient prediction of nucleic acid binding function from low-resolution protein structures, J. Mol. Biol. 358 (2006) 922–933.
[43] M. Waris, K. Ahmad, M. Kabir, M. Hayat, Identification of DNA binding proteins using evolutionary profiles position specific scoring matrix, Neurocomputing 199 (2016) 154–162.
[44] L. Wei, M. Liao, X. Gao, Q. Zou, An improved protein structural classes prediction method by incorporating both sequence and structure information, IEEE Trans. Nanobiosci. 14 (2015) 339–349.
[45] L. Wei, M. Liao, X. Gao, Q. Zou, Enhanced Protein Fold Prediction Method Through a Novel Feature Extraction Technique, Nanobiosci. IEEE Trans. 14 (2015) 649–659.
[46] R. Xu, J. Zhou, B. Liu, Y. He, Q. Zou, X. Wang, K.-C. Chou, Identification of DNA-binding proteins by incorporating evolutionary information into pseudo amino acid composition via the top-n-gram approach, J. Biomol. Struct. Dyn. 33 (2015) 1720–1730.
[47] R. Xu, J. Zhou, B. Liu, L. Yao, Y. He, Q. Zou, X. Wang, enDNA-Prot: identification of DNA-Binding Proteins by applying ensemble learning, BioMed Res. Int. 2014 (2014).
[48] R. Xu, J. Zhou, H. Wang, Y. He, X. Wang, B. Liu, Identifying DNA-binding proteins by combining support vector machine and PSSM distance transformation, BMC Syst. Biol. 9 (2015) S10.
[49] X. Yu, J. Cao, Y. Cai, T. Shi, Y. Li, Predicting rRNA-, RNA-, and DNA-binding proteins from primary structure with support vector machines, J. Theor. Biol. 240 (2006) 175–184.
[50] H. Zhao, Y. Yang, Y. Zhou, Structure-based prediction of DNA-binding proteins by structural alignment and a volume-fraction corrected DFIRE-based energy function, Bioinformatics 26 (2010) 1857–1863.
[51] W. Zhou, H. Yan, Prediction of DNA-binding protein based on statistical and geometric features and support vector machines, Proteome Sci. 9 (2011) 1–6.
[52] C. Zou, J. Gong, H. Li, An improved sequence based prediction protocol for DNA-binding proteins using SVM and comprehensive feature analysis, BMC Bioinform. 14 (2013) 90.
[53] Q. Zou, Z. Wang, X. Guan, B. Liu, Y. Wu, Z. Lin, An approach for identifying cytokines based on a novel ensemble classifier, BioMed Res. Int. 2013 (2013).