Computational Biology and Chemistry 52 (2014) 51–59

Contents lists available at ScienceDirect

Computational Biology and Chemistry

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

Research Article

newDNA-Prot: Prediction of DNA-binding proteins by employing support vector machine and a comprehensive sequence representation

Yanping Zhang [a], Jun Xu [b], Wei Zheng [b], Chen Zhang [b], Xingye Qiu [b], Ke Chen [c,*], Jishou Ruan [b]

[a] Department of Mathematics, School of Science, Hebei University of Engineering, Handan 056038, PR China
[b] College of Mathematical Sciences and LPMC, Nankai University, No. 94 Weijin Road, Tianjin 300071, PR China
[c] School of Computer Science and Software Engineering, Tianjin Polytechnic University, No. 399 Binshui Road, Tianjin 300387, PR China

ARTICLE INFO

Article history:
Received 11 February 2014
Received in revised form 5 September 2014
Accepted 6 September 2014
Available online xxx

Keywords:
DNA-binding proteins
Features
Feature selection methods
SVM
ROC

ABSTRACT

Identification of DNA-binding proteins is essential in studying cellular activities as the DNA-binding proteins play a pivotal role in gene regulation. In this study, we propose newDNA-Prot, a DNA-binding protein predictor that employs support vector machine classifier and a comprehensive feature representation. The sequence representation are categorized into 6 groups: primary sequence based, evolutionary profile based, predicted secondary structure based, predicted relative solvent accessibility based, physicochemical property based and biological function based features. The mRMR, wrapper and two-stage feature selection methods are employed for removing irrelevant features and reducing redundant features. Experiments demonstrate that the two-stage method performs better than the mRMR and wrapper methods. We also perform a statistical analysis on the selected features and results show that more than 95% of the selected features are statistically significant and they cover all 6 feature groups. The newDNA-Prot method is compared with several state of the art algorithms, including iDNA-Prot, DNAbinder and DNA-Prot. The results demonstrate that newDNA-Prot method outperforms the iDNA-Prot, DNAbinder and DNA-Prot methods. More specific, newDNA-Prot improves the runner-up method, DNA-Prot for around 10% on several evaluation measures. The proposed newDNA-Prot method is available at http://sourceforge.net/projects/newdnaprot/

1. Introduction

DNA-binding proteins are functional molecules in a cell and play an important role in a variety of biological processes, i.e., transcription regulation, DNA replication, DNA packaging, DNA repair and DNA rearrangement (Gao and Skolnick, 2009). Moreover, proteins that interact with specific DNA sequences may be a potential therapeutics for genetic diseases. At present, both experimental and computational techniques have been developed for identification of the DNA-protein interactions. Although, the experimental techniques, including filter binding assays (Cajone et al., 1989), ChIP-chip (Buck and Lieb, 2004), genetic analysis (Freeman et al., 1995) and X-ray crystallography (Chou et al., 2003) provide atomic-level perspective on DNA-protein interaction, they are time-consuming and expensive. Therefore, there is an urgent need to develop computational methods that identify DNA-binding proteins with high success rates (Gromiha and Nagarajan, 2013). Presently, the in silico methods are mainly categorized into two class: methods based on protein structures (Ahmad and Sarai, 2004; Zhao et al., 2010; Tjiong and Zhao, 2007; Stawiski et al., 2003; Szilagyi and Skolnick, 2006) and methods based on primary sequence (Cai and Lin, 2003; Robert and Hui, 2010; Huang et al., 2011; Kumar et al., 2007; Shao et al., 2009; Lin et al., 2011). Although, the structure-based methods achieve higher accuracy when compared with the sequence-based techniques, they are not applicable for high-throughput annotation of DNA-binding proteins as structures are only determined for a small fraction of known proteins. Since it is widely recognized that primary sequence determines the tertiary structure which further determines the function of a protein (Cai and Doig, 2004), majority of the computational methods took primary sequence as the only input. The sequence-based methods mainly include two steps: firstly, transforms the primary sequence into a fixed number of numerical features, i.e., the composition of the amino acids; second, choose a machine learning algorithm and build a computation model for prediction. Previously, dozens of

* Corresponding author.
E-mail address: kchen1.tjpu@hotmail.com (K. Chen).

http://dx.doi.org/10.1016/j.compbiolchem.2014.09.002
1476-9271/

machine-learning algorithms such as support vector machine (SVM) (Zhao et al., 2010; Tjiong and Zhao, 2007; Brown and Akutsu, 2009; Wang et al., 2010; Xiong et al., 2011; Zou et al., 2013), artificial neural network (ANN) (Ahmad et al., 2004; Keil et al., 2004; Ahmad et al., 2009), random forest (Cai and Lin, 2003; Nimrod et al., 2010; Wang et al., 2009), naive Bayes (Yan et al., 2006) and nearest neighbor (Qian et al., 2006) have been proposed which performs prediction of DNA-binding proteins. Among these algorithms, the SVM is especially widely implemented.

The performance of SVM is largely dependent on the quality of the features (Liu et al., 2008). Although, plenty of feature representation and selection methods were proposed for protein sequence (Fang et al., 2008; Chou, 2011; Yuan et al., 2010; Song et al., 2008) and these methods were systematically surveyed (Nanni et al., 2010; Zhang et al., 2005), the underlying principle of protein-DNA interaction is still largely unknown. To this end, we propose a comprehensive feature representation, including the sequence information, evolutionary profiles, predicted secondary structural, predict relative solvent accessibility (RSA) information, physicochemical properties, and biological function information. Moreover, we employ three feature selection methods, mRMR, wrapper and two-stage (mRMR and wrapper), to remove the irrelevant features and reduce redundancy among the features. The proposed method outperforms existing sequence-based methods including the DNAbinder (Kumar et al., 2007), iDNA-Prot (Lin et al., 2011), DNA-Prot (Kumar et al., 2009) which either have a web server or provide standalone program. The framework of the proposed DNA-binding predictor is given in Fig. 1.

2. Materials and method

2.1. Datasets

The training set, namely DNAdset, contains 231 DNA-binding and 231 non-binding protein chains or domains which were obtained from a union of datasets used in previously related studies (Zou et al., 2013; Gao et al., 2012; Fang et al., 2008). By employing the CD-HIT program (Huang et al., 2010), any pair of the protein chains have a sequence similarity less than 40%.

Additionally, we build an independent test set, namely DNAiset, to perform an independent evaluation of the predictive model. The test set includes 97 DNA-binding proteins and 192 non-DNA binding proteins. The DNA-binding proteins are culled from PDB by keyword searching (released on 2014-06–01 and later) while the non-binding proteins were culled by Lin et al. (2011). We employ the CD-HIT (Wang and Dunbrack, 2003; Huang et al., 2010) program to remove the sequence similarity between DNAiset and DNAdset. Any sequence from DNAiset has a sequence similarity less than 30% to a sequence from DNAdset.

Many machine learning algorithms perform better when the number of positive samples is comparable to the number of negative samples. To this end, the numbers of DNA-binding proteins and non-binding proteins are equivalent in DNAdset and comparable in DNAiset. However, in real situation, the fraction of DNA-binding proteins is small when compared with the fraction of non-binding proteins. To this end, DNArset, which simulates the ratio between DNA-binding proteins and non-binding proteins in cellular environment, is created. DNArset contains 97 DNA-binding proteins taken from DNAiset and 1500 non-binding proteins culled by Kumar et al. (2007).

2.2. Features-based representation of the input sequence

The performance of a computational model is largely dependent on the quality of the feature representation (Chou, 2011; Brameier et al., 2006). In this study, we generate six groups of features that are derived from primary sequence, evolutionary information, predicted secondary structure, predicted RSA values (Faraggi et al., 2012), physicochemical properties and functional information.

2.2.1. AA sequence length and AA composition-based (101 features)

This group of features is directly calculated from primary sequence.

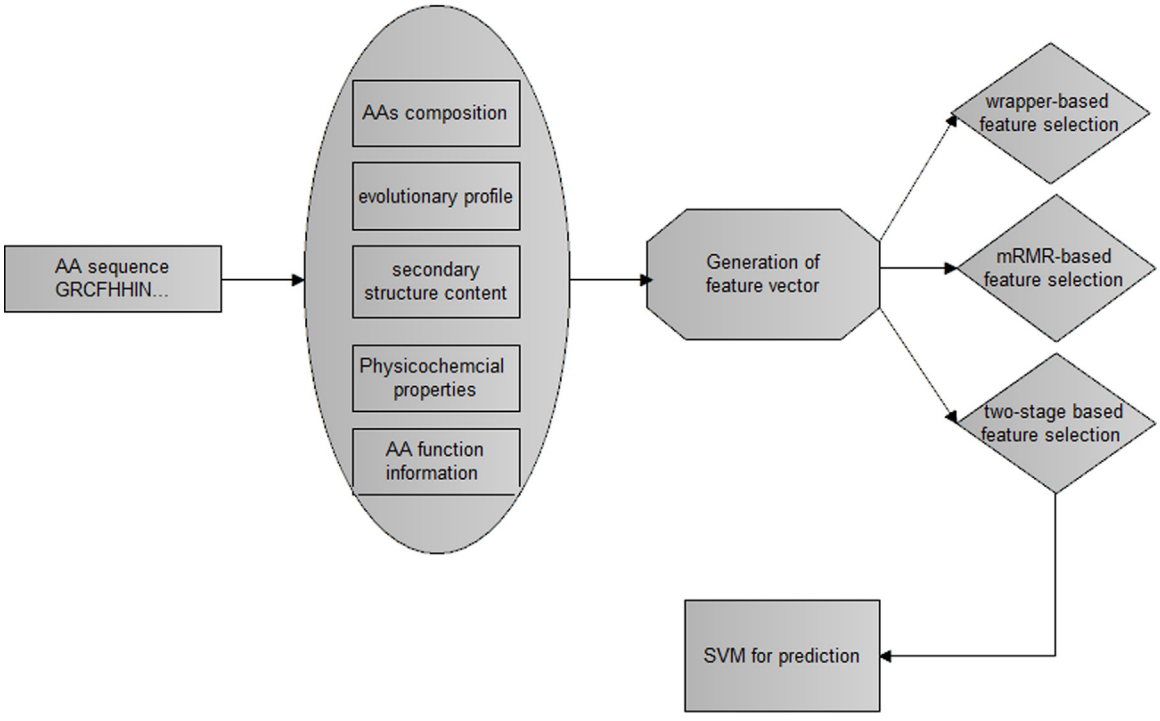

Fig. 1. The framework of the newDNA-Prot predictor.

- SeqLen is the length of a sequence (1 feature).
- Composition vector: $CV_i = AA_i/SeqLen$ represents the percentage of the ith type of AA in the sequence (20 features).
- The first and second order composition moment vector: $CMV_i^k = \sum_{i,j=1}^{x_i} n_{ij}^k / \Pi_{d=1}^k (N-k)$, where n_{ij} represents the jth position of the ith type of AA, n_i is the frequency of ith type of AA in the sequence, and $k = 1, 2$ is the order of the CMV.

2.2.2. PSSM matrix and RSA-based (218 features)

For an input protein sequence, its position-specific scoring matrix (PSSM) profile can be calculated from alignment profile generated by PSI-BLAST (Altschul et al., 1997). PSSM score indicates the conservation of a given AA position in the query sequence among its homologues (Wu and Zhang, 2008). Since functional sites, including DNA and RNA-binding residues, are more conserved than other residues, it was proved that PSSM profile are valuable in prediction of DNA-protein interactions (Kumar et al., 2007; Ahmad and Sarai, 2005).

The PSI-BLAST program generates two position specific scoring matrices, which are conservation scores and probability of occurrence of a given AA respectively. The two matrices are denoted as $PSSMcons_{lm}$ and $PSSMprob_{lm}$ respectively, where $l = 1,2,\ldots,SeqLen$ represents the position in the sequence and $m = 1,2,\ldots,20$ denotes the 20 substitution positions that correspond to the 20 AA types.

- Ach_CS_{AA_i}=$\sum_{l=1}^{SeqLen}$if $AA(l) = AA_i$, then $PSSMcons_{li}$ else 0/SeqLen, sum of all normalized $PSSMcons_{lm}$ values ("Ach_CS" stands for achieved conservation scores), where l includes only positions of AA_i and $m = AA_i$ (20 features). The other 2 features are the sum of Ach_CS_{AA_i} and the sum of Ach_CS_{AA_i} if AA locates on the solvent surface.
- Max_CS_{AA_i}=$\sum_{l=1}^{SeqLen}$if $AA(l) = AA_i$, then $\max_{i=1}^{20}(PSSMcons_{li})$ else 0/SeqLen sum of maximal, over m, $PSSMcons_{lm}$ values, where l includes only positions of AA_i, divided by the sequence length (20 features). The other 2 features are the sum of Max_CS_{AA_i} and the sum of Max_CS_{AA_i} if AA locates on the solvent surface.
- Max–Ach_CS_{AA_i}=$\sum_{l=1}^{SeqLen}$if $AA(l) = AA_i$, then $[\max_{i=1}^{20}(PSSMcons_{li}) - PSSMcons_{li}]$ else 0/SeqLen, sum of differences between maximal $PSSMcons_{lm}$ and $PSSMcons_{li}$ values where l includes only positions of AA_i and $i = AA_i$, divided by the sequence length (20 features). The other 2 features are the sum of Max−Ach_CS_{AA_i} and the sum of Max–Ach_CS_{AA_i} if AA locates on the solvent surface.
- Ach_Prob_{AA_i}=$\sum_{l=1}^{SeqLen}$if $AA(l) = AA_i$ then $PSSMBprob_{li}$ else 0/SeqLen, sum of all normalized $PSSMprob_{lm}$ values, where l includes only positions of AA_i and $i = AA_i$, divided by the sequence length (20 features). The other 2 features are the sum of Ach_Prob_{AA_i} and the sum of Ach_Prob_{AA_i} if AA locates on the solvent surface.
- Max_Prob_{AA_i}=$\sum_{l=1}^{SeqLen}$if $AA(l) = AA_i$ then $\max_{i=1}^{20}(PSSMprob_{li})$ else 0/SeqLen, sum of maximal, over m, $PSSMprob_{lm}$ values, where l includes only positions of AA_i, divided by the sequence length (20 features). The other 2 features are the sum of Max_Prob_{AA_i} and the sum of Max_Prob_{AA_i} if AA locates on the solvent surface.
- Max–Ach_prob_{AA_i}=$\sum_{l=1}^{SeqLen}$if $AA(l) =$ then $[\max_{i=1}^{20}(PSSMcons_{li}) - PSSMcons_{li}]$ else 0/SeqLen, sum of differences between maximal $PSSMprob_{lm}$ and $PSSMprob_{li}$ values where l includes only positions of AA_i and $i = AA_i$, divided by the sequence length (20 features). The other 2 features are the sum of Max–Ach_{AA_i} and the sum of Max−Ach_{AA_i} if AA locates on the solvent surface.
- CSSeq_{AA_i} = SeqLen, sum of normalized $PSSMcons_{lm}$ values where $l = 1,2,\ldots,SeqLen$ and $m = AA_i$, divided by the sequence length (20 features). The other 2 features are the sum of CSSeq_{AA_i} and the sum of CSSeq_{AA_i} if AA locates on the solvent surface.
- Seq_Prob_{AA_i}=$\sum_{l=1}^{SeqLen}PSSMprob_{li}$/SeqLen, sum of normalized $PSSMcons_{lm}$ values where $l = 1,2,\ldots,SeqLen$ and $m = AA_i$, divided by the sequence length (20 features). The other 2 features are the sum of Seq_Prob_{AA_i} and the sum of Seq_Prob_{AA_i} if AA locates on the solvent surface.
- Ent_{AA_i}=$-\sum_{l=1}^{SeqLen}[PSSMprob_{li} \times \log_2(PSSMprob_{li})]$, entropy of $PSSMprob_{lm}$ values, for $l = 1,2,\ldots,SeqLen$ and $m = AA_i$ (20 features). The other 22 features contain the average entropy per position in sequence (which is equivalent to conservation) (1 feature), the average entropy per position in sequence if all AA locates on the solvent surface and average entropy per position in sequence for each AA type (20 features).

2.2.3. Secondary structure-based (223 features)

Secondary structure is generated by the PSIPred program (McGuffin et al., 2000). PSIPred were successfully integrated into a number of computational methods that predict structural properties of proteins, i.e., structural class (Kurgan et al., 2008), beta-turns (Zheng and Kurgan, 2008), residue depth (Zhang et al., 2008), prediction fold (Reinhardt and Eisenberg, 2004) and contact orders (Song and Burrage, 2006).

- Ach_Prob_{SS_n}=$\sum_{l=1}^{SeqLen}$if $AA(l) = SS_n$ then PM_{li} else 0/SeqLen, sum of normalized PM_{lm} values (which represents the probability matrix (PM) by predict secondary structure), where l includes only positions of SS_n, which $SS_n = \{C,H,E\}$, divided by the sequence length and $m = SS_n$ (3 features).
- Content_{SS_n} = count (AA_1:AA_1 predicted as SS_n)/SeqLen, the number of residues predicted as SS_n where $l = 1,2,\ldots,SeqLen$, divided by the sequence length (3 features). The other 3 features are generated by AAs on the surface using this formula, which represent content of AAs related to binding.
- SegCount_{E,H}_L{L_i} = count(SEG:SEG(SS_n) AND length(SEG $\geq L_i$)/ $\sum_{SS\in\{E,H\}}$ count (SEG:SEG(SS)), the number of helix or strand segments which contain at least $L_i = 2,3,\ldots,20$ AAs divided by the total number of helix and strand segments in the input protein chain (38 features)
- SegCount_C_L{L_i} = count (SEG:SEG(SS_n) ANDlength(SEG $\geq L_i$)/ $\sum_{SS\in\{H,E,C\}}$ count (SEG:SEG(SS)), the number of coils which contain at least $L_i = 2,3,\ldots,20$ AAs divided by the number of all segments in a protein (19 features).
- SegCount_{E,H}_P{P_i} = count (SEG:SEG(SS_n) ANDlength(SEG$\geq P_i \times$ SegLen)/ $\sum_{SS\in\{E,H\}}$ count(SEG:SEG(SS)), the number of helix or strand segments which contain at least P_iAAs where $P_i = 2,4,\ldots,10\%$ of the sequence length, divided by the total number of helix and strand segments in the input protein chain (10 features).
- SegCount_C_P{P_i} = count (SEG:SEG(SS_n) ANDlength(SEG$\geq P_i \times$ SegLen)/ $\sum_{SS\in\{H,E,C\}}$ count(SEG:SEG(SS)), the number of coil segments which contain at least P_iAAs where $P_i = 2,4,\ldots,10\%$ of the sequence length, divided by the number of all segments (5 features).
- NormSegCount_{SS_n} = count (SEG:SEG(SS_n))/ $\sum_{SS\in\{H,E,C\}}$ count (SEG:SEG(SS)), the total number of SS_n segments divided by

the total number of all secondary structure segments in the input protein chains (3 features).

- MaxSegCount_{SS_n} = maxLen(SEG(SS_n)), the maximal SS_n segment length (3 features).
- NormMaxSegCount_{SS_n} = maxLen(SEG:SEG(SS_n))/SeqLen, the maximal SS_n, the maximal SS_n segment length divided by the sequence length (3 features).
- AvgSegCount_{SS_n} = avgLen(SEG(SS_n)), the average SS_n segment length (3 features).
- NormAvgSegLength_{SS_n} = avgLen(SEG:SEG(SS_n))/SeqLen, the average SS_n segment length divided by the sequence length (3 features).
- HH = count(HH), the number of helix–coil–helix motifs divided by the total number of the secondary structure segments in a protein (1 feature).
- EE = count(EE), the number of strand–coil–strand motifs divided by the total number of the secondary structure segments in a protein (1 feature).
- HE = count(HE) + count(EH), the number of strand–coil–helix or helix–coil–strand motifs by the total number of the secondary structure segments in a protein (1 feature).
- {HH,HE,EE} − L{L_i} = count({HH,HE,EE}:LEN(Coil) ≥ L_i)/ $\sum_{SS\in\{H,E,C\}}$count(SEG:SEG(SS)), the number of helix–coil–helix, helix–coil–strand/strand–coil–helix, or strand–coil–strand motifs which include at least L_i = 2,3, . . . ,20 AAs residues in the middle coil, divided by the total number of the secondary structure segments in a protein (57 features).
- {HH,HE,EE} − L{P_i} = count({HH,HE,EE}:LEN(Coil) ≥ P_i × SeqLen)/ $\sum_{SS\in\{H,E,C\}}$count(SEG:SEG(SS)), the number of helix–coil–helix, helix–coil–strand/strand–coil–helix, or strand–coil–strand motifs which include at least P_i = 2,3, . . . ,10% of a sequence length residues in the middle coil, divided by the total number of the secondary structure segments in a protein (15 features).
- MaxHCH = max(HC..*H*:count(*H*)), the maximal number of helices among all helix–coil–helix–coil . . . coil–helix motifs, i.e., the maximal number of helix segments separated only by coils (1 feature).
- MaxECE = max(EC..*E*:count(*E*)), the maximal number of helices among all strand–coil–strand–coil . . . coil–strand motifs, i.e., the maximal number of strand segments separated only by coils (1 feature).
- AvgHCH = avg(HC..(*H*))/ $\sum_{SS\in\{H,E,C\}}$ count(SEG:SEG(SS)), the average number of helices in all helix–coil–helix–coil . . . coil–helix motifs, divided by the total number of the secondary structure segments in a protein (1 feature).
- AvgECE = avg(EC..(*E*:count:(*E*))/ $\sum_{SS\in\{H,E,C\}}$ count(SEG:SEG(SS)), the average number of helices in all strand–coil–strand–coil . . . coil–strand motifs, divided by the total number of the secondary structure segments in a protein (1 feature).
- HCH_*L*{L_i} = count(HC..*H*:count(*H*) ≥ L_i)/ $\sum_{SS\in\{H,E,C\}}$ count(SEG:SEG (SS)), the number of helix–coil–helix–coil . . . coil–helix motifs with more than L_i = 1,2, . . . ,20 helices, divided by the total number of the secondary structure segments (19 features).
- HCH_*P*{P_i} = count(HC..*H*:count(*H*) ≥ P_i × SeqLen)/ $\sum_{SS\in\{H,E,C\}}$ count (SEG:SEG(SS)), the number of helix–coil–helix–coil . . . coil–helix motifs with more than P_i = 2,4, . . . ,10 of all helics in a protein, divided by the total number of the secondary structure segments (5 features).
- ECE_*L*{L_i} $\sum_{SS\in\{H,E,C\}}$ count(SEG:SEG(SS)), the number of strand–coil–strand–coil . . . coil-strand motifs with more than L_i = 1,2, . . . ,20 strands, divided by the total number of the secondary structure segments (19 features).
- ECE_*P*{P_i} = count(EC..E:count(E)≥P_i × SeqLen)/ $\sum_{SS\in\{H,E,C\}}$ count (SEG:SEG(SS)), the number of strand–coil–strand–coil . . . coil–strand motifs with more than P_i = 2,4, . . . ,10 of all helics in a protein, divided by the total number of the secondary structure segments (5 features).

2.2.4. Average RSA-based (23 features)

- Average RSA of the residues with AA type (20 features).
- Average RSA of the residues with secondary structure type {*C*,*H*,*E*} (3 features).

2.2.5. Physicochemical properties-based (203 features)

It is widely recognized that physicochemical properties of AAs play an important role in protein folding and protein-DNA interaction. Experimental and theoretical studies have proposed a wide variety of properties for AAs. Even for the same property, it might be assigned to different values by different studies. Based on the analysis of Huang et al. (2010), we include nine physicochemical indices, see in Table 1. Several physicochemical properties in Table 1 have been employed in prediction of protein-DNA interaction (Shao et al., 2009; Fang et al., 2008) while several other properties, i.e., the flexibility property, are utilized for prediction of DNA-binding proteins for the first time. The correlation between protein flexibility and biological function of proteins has demonstrated a linkage between protein-DNA interaction and flexibility of the corresponding domain (Gryk et al., 1996).

- $K^{\alpha} = \frac{1}{N}\sum_{i=1}^{N} \alpha_i$ where α represents the corresponding physicochemical AA index; charge, p*K*-C, p*K*-(—COOH), polarity (2), p*K*-N, pK_a (RCOOH) and normalized index were used, and these indexes values listed in the Table 1 (8 features).
- $A_n^{\alpha} = \frac{1}{N-n}\sum_{i=1}^{N-n} \alpha_i\alpha_{i+n}$, where α defines the corresponding physicochemical AA index; hydrophobicity index (4) n = 1, 2,...,6, flexibility (2) n = 1, 2,...,6, Secondary structure (4) n = 1, 2,...,6, and solvent accessibility (5) n = 1, 2,...,6 were used (90 features).
- $\mathrm{Acum}_n^{\alpha} = \sum_{i=1}^{N-n}\left(\sum_{j=1}^{i}\alpha_j\right) \times \left(\sum_{j=1}^{i+n}\alpha_j\right)/\mathrm{N} - \mathrm{n}$, where α is the hydrophobicity index (4) with n = 1, 2,...,6, flexibility (2) with

Table 1
Some typical properties for analyzing DNA-binding domains.

AA index ID	PCP	AA index ID	PCP
BHAR880101	Flexibility	FASG760105	p*K*-C
BURA740101	Secondary structure	JOND750102	p*K*-(—COOH)
CHOC760103	Solvent accessibility	RADA880108	Polarity
HOPT810101	Hydrophobicity	PRAM900101	Hydrophobicity
FAUJ880111	Charge	FUKS010104	Solvent accessibility
KARP850101	Flexibility	KUMS000103	Secondary structure
PALI810115	Secondary structure	PONP800107	Solvent accessibility
ROSM880101	Hydrophobicity	GRAR740102	Polarity
KUHL950101	Solvent accessibility	FASG760104	p*K*-N
ZIMJ680101	Hydrophobicity	FAUJ880113	pK_a (RCOOH)
EISD860101	Solvent accessibility	FAUJ880103	Normalized van der Walls volume
GEIM800101	Secondary structure		

n = 1, 2,...,6, secondary structure (4) with n = 1, 2,...,6, and solvent accessibility (5) with n = 1, 2,...,6 (90 features).

- $H^{a}_{\text{sum}} = \sum_{i=1}^{N} \alpha_i$, where α is the hydrophobicity index (4), flexibility (2), secondary structure (4), and solvent accessibility (5) (15 features).

2.2.6. Disorder scores-based (13 features)

The last decade had witnessed a growing recognition that a large fraction of proteins are intrinsically or natively disorder proteins (Vuzman and Levy, 2012). Plenty of intrinsically (or natively) disordered proteins (IDRs) are associated with DNA-protein interactions and play a crucial role in the interactions by increasing the affinity and specificity of DNA binding (Vuzman and Levy, 2012; Zheng and Kurgan, 2008). The ability of IDRs to interact with DNA is due to the high content of charged residues in IDRs. In general, IDRs that interact with DNA are rich in positively charged residues and these residues are spatially clustered. The IUPred (Dosztanyi et al., 2005) method is employed to calculate the pairwise energy profile along the input AA sequence. The IUPred method generates a series of scores, which stands for energy values, ranging from 0 (completely ordered) to 1 (completely disordered). In order to produce the same number of features for protein sequences with varied length, we employ the CTD method, which was firstly proposed by Dubchak et al. (1995). The CTD method is widely used in prediction of protein folding class and protein function (Cai et al., 2003; Zou et al., 2013). The letter C, which is the abbreviation of composition, stands for the percentage of AAs with a certain property (such as disorder). The letter T, which is the abbreviation of transition, stands for the percentage of AAs followed by an AA with different property, i.e., an ordered residue followed by a disordered residue. The letter D, which is the abbreviation of distribution, measures the range of the first, 25, 50, 75 and 100% of the AAs with a certain property respectively. The CTD method generates 13 features for each protein sequence.

3. Performance evaluation

Several quality measures, including accuracy (ACC), sensitivity, precision, specificity, F-measure, and Matthews correlation coefficient (MCC) (Baldi et al., 2000), are employed for evaluation of the proposed method. These measures are defined as follows:

$$\text{ACC} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \times 100\% \tag{1}$$

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}} \times 100\% \tag{2}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \times 100\% \tag{3}$$

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}} \times 100\% \tag{4}$$

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}} \times 100\% \tag{5}$$

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

where TP and TN are the number of correctly predicted DNA-binding and non-binding proteins, respectively, FP denotes the non-binding proteins predicted as the DNA-binding proteins, and the FN denotes DNA-binding proteins predicted as non-binding proteins. The F-measure is used to balance the sensitivity and precision (Gromiha and Yabuki, 2008). Higher values of these measures indicate better quality of predictions. The MCC value ranges between −1 and 1 and other indices range between 0 and 1. Higher value of these measures imply better predictions.

We also employ the area under the ROC curve (AUC) to evaluate the predicted probabilities. The above mentioned measures are dependent on a threshold that discretizes the predicted probabilities into positives and negatives, while the AUC takes all thresholds into consideration. Therefore, the AUC index provides more comprehensive perspective on the predictions than other measures. The AUC value ranges between 0 and 1 and higher values imply better predictions.

4. Feature selection methods

The generated features may be irrelevant to the prediction of DNA-binding domains and redundant with each other. To this end, we employ mRMR, wrapper and the two-stage (mRMR and wrapper) feature selection algorithms to remove the irrelevant features and reduce redundancy among the features.

4.1. Maximum relevance, minimum redundancy (mRMR) feature selection

The maximum relevance, minimum redundancy (mRMR) method was proposed by Peng for processing microarray data (Peng et al., 2005). The approach selects the features having maximum dependency, minimum redundancy, and maximum relevance. The redundancy and relevance are measured by mutual information defined as following:

$$I(x,y) = \iint p(x,y)\frac{p(x,y)}{p(x)p(y)} \tag{7}$$

where $p(x)$, $p(y)$ and $p(x,y)$ are the probabilistic density functions of two variables x and y. Detailed description of the algorithm is given in (Peng et al., 2005).

The mRMR algorithm generates a feature set S in which features are arranged in certain order:

$$S = \{f'_1, f'_2, L, f'_h, K, f'_N\} \tag{8}$$

4.2. Wrapper-based feature selection

The wrapper-based method convolves with a classifier (Kohavi and John, 1997). The features are first sorted by the absolute value of biserial correlation coefficient (BCC), which is defined as:

$$\text{BCC} = (M_b - M_{nb}) \times \text{sqrt}(\text{nb} \times n_{nb}/n)/(\text{stdev}) \tag{9}$$

where M_b and M_{nb} are the mean values of the feature values of DNA-binding and non DNA-binding chains, respectively, stdev is the standard deviation of the feature, nb and n_{nb} are the numbers of DNA-binding and non DNA-binding chains, respectively, and n is the total number of chains.

Next, we start with a single feature that provides the highest MCC based on the 5-fold cross validation on training set. The remaining features are added in the order determined by BCC. A feature is retained if the addition of the feature improves MCC and discarded otherwise. After the completion of feature selection, we optimize the parameters of the classification model. The wrapper-based feature selection method was employed in a number of

Table 2
The performance of prediction models by employing mRMR-based feature selection method.

Method	AUC	ACC	MCC	Sensitivity	Specificity	Precision	F-measure
SVM model 200[a] features	0.9500	0.9026	0.8055	0.8874	0.9177	0.9152	0.9011
SVM model 300[b] features	0.9536	0.9091	0.8185	0.8961	0.9221	0.9200	0.9079
SVM model 400[c] features	0.9538	0.9004	0.8019	0.8745	0.9264	0.9224	0.8978
SVM model 500[d] features	0.9553	0.8983	0.7982	0.8658	0.9307	0.9259	0.8949

[a] Results based on 5-fold cross validation for the SVM model ($C=2^3$,$\gamma=2^{-5}$) that uses the selected top 200 features.
[b] Results based on 5-fold cross validation for the SVM model ($C=2^3$, $\gamma=2^{-5}$) that uses the selected top 300 features.
[c] Results based on 5-fold cross validation for the SVM model ($C=2^3$, $\gamma=2^{-5}$) that uses the selected top 400 features.
[d] Results based on 5-fold cross validation for the SVM model ($C=2^3$, $\gamma=2^{-5}$) that uses the selected top 500 features.

related studies (Chen et al., 2011; Mizianty et al., 2010; Mizianty and Kurgan, 2011; Gao et al., 2012).

4.3. Two-stage (mRMR and wrapper) feature selection

mRMR method is also called filter method (Langley, 1994) as it identifies a subset of the features and ignore the remaining features. In the two-stage method, we first use mRMR-based feature selection method, which identifies a small set of candidate features. In this study, the mRMR-based method select 300 candidate features out of 781 features, which constitute the complete feature representation.

In the second stage, the wrapper-based method is employed to identify a compact feature subset from the candidate feature set. By adding the first stage, the computational complexity of wrapper-based method is reduced and predictive performance is improved. After completion of feature selection, we optimize the parameters of the classification model.

5. Results and discussion

5.1. mRMR-based results

The mRMR program can be downloaded from http://penglab.janelia.org/proj/mRMR/#c ++. The mRMR program ranks all features and generates the top 200, 300, 400 and 500 candidate features from the initial 781 features for DNAdset (462 samples). Features with higher mRMR score implies better predictive performance. The performance by employing different number of features is compared in Table 2. We note that employing top 300 features results in better predictions. This model provides high values for all evaluation indices, i.e., 0.9536 for AUC, 0.9091 for ACC, 0.8185 for MCC, 0.8961 for sensitivity, 0.9221 for specificity, 0.92 for precision and 0.9011 for F-measure.

5.2. The results by employing different feature selection methods

We also assess the performance of the prediction models by employing different feature selection methods, including the mRMR-based, wrapper-based and the two-stage method. Table 3 demonstrates that the model that integrates two-stage method generates better predictions. The model with the two-stage method achieves higher values for all indices, i.e., 0.9654 for AUC, 0.9221 for ACC, 0.8441 for MCC, 0.9091 for sensitivity, 0.9351 for specificity, 0.9333 for precision and 0.9211 for F-measure. The results demonstrate that the combination of the mRMR-based and wrapper-based methods enhance the performance of individual feature selection method.

5.3. Performance of individual feature groups

We first investigate statistical significance of the differences between the feature values of DNA-binding and feature values of non-binding proteins on DNAdset. The two-sided t test is employed and the P value is the probability that the difference between feature values of DNA-binding and non-binding groups occurs by a random chance. We note that at a significance level of 0.05, 118 out of 128 (118/128 = 92.2%) features have P values less than 0.05. Therefore, the difference between feature values of DNA-binding and non-binding groups are statistically significant for 118 features. The results prove that majority of the selected features have statistically significant differences between the DNA-binding and non-binding proteins.

We also perform an analysis on the selected features, see Table 4. It should be emphasized that the selected features cover all 6 feature groups, which indicates that all considered feature groups contribute to prediction of DNA-protein interactions. The largest group of the selected 128 features is PM + RA, of which 48 features are selected by the two-stage feature selection method. It demonstrates that the evolutionary information generated by PSI-BLAST plays a pivotal role in prediction of DNA-binding proteins. The second largest group, AA sequence length and AA composition group, consists of 28 selected features. The remaining selected features includes 24 features calculated from predicted secondary structure, 18 features calculated from the physiochemical properties, 7 features are calculated from predicted RSA values and 3 features are calculated from predicted disordered information. The results demonstrate that the selected 128 features are valuable in prediction of DNA-binding proteins.

We also assess the performance of predictive models by employing individual feature groups. The ROC curves by utilizing individual feature groups are given in Fig. 2. The corresponding AUC values are given in Table 5. The AUC values for each of the 6 groups are above 0.5. We note that the PM + RA group provides

Table 3
Comparison of the predictive performance by employing different feature selection methods.

Method	AUC	ACC	MCC	Sensitivity	Specificity	Precision	F-measure
SVM model 781[a] features	0.9524	0.9048	0.8110	0.8745	0.9351	0.9309	0.9018
SVM model 300[b] features (mRMR-based)	0.9536	0.9091	0.8185	0.8961	0.9221	0.9200	0.9079
SVM model 271[c] features (wrapper-based)	0.9567	0.9158	0.8352	0.9134	0.9217	0.9214	0.9174
SVM model 128[d] features (two-stage)	0.9654	0.9221	0.8441	0.9091	0.9351	0.9333	0.9211

[a] Results based on 5-fold cross validation for the SVM model ($C=2^2$, $\gamma=2^{-5}$) that uses all 781 features.
[b] Results based on 5-fold cross validation for the SVM model ($C=2^3$, $\gamma=2^{-5}$) that uses the selected 300 features.
[c] Results based on 5-fold cross validation for the SVM model ($C=2^4$, $\gamma=2^{-4}$) that uses the selected 271 features.
[d] Results based on 5-fold cross validation for the SVM model ($C=2^4$, $\gamma=2^{-4}$) that uses the selected 128 features.

Table 4
Number of selected features by the two-stage method for each of the 6 feature groups.

Feature group	Abbreviated name	Number of features	Number of selected features (128) with two-stage
AA sequence length and AAs composition	AAC	101	28
PSSM matrix and RSA	PM + RA	218	48
Secondary structure	SS	223	24
Average RSA	ARA	23	7
Physicochemical properties	PP	203	18
Disorder scores	DS	13	3
Total number of features		781	128

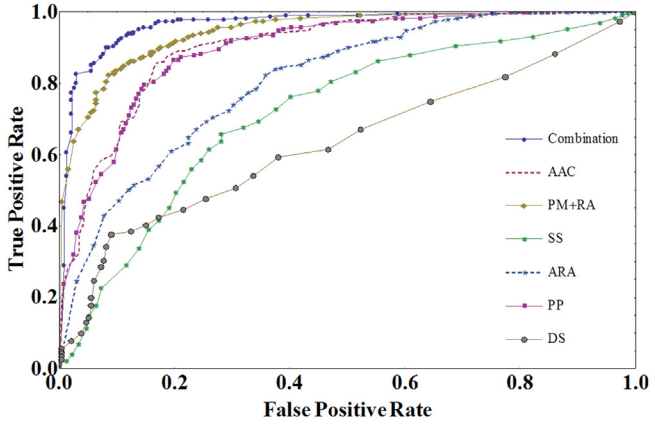

Fig. 2. ROC curves with the six groups and their combination.

Table 5
Comparison between predictive models by employing individual feature groups on the DNAdset. The results are calculated by using 5-fold cross validation. The predictive model employs SVM classifier and features are selected by the two-stage method. The abbreviations of the feature groups are given in Table 3.

Scale	ACC	AUC	MCC
AAC	0.8602	0.8843	0.7041
PM + RA	0.9032	0.9381	0.7876
SS	0.7527	0.771	0.5049
ARA	0.7634	0.84	0.5618
PP	0.7527	0.8354	0.5411
DS	0.6739	0.7052	0.4048

higher AUC values than other groups which is likely because the two-stage method selects more features for the PM + RA group than other groups.

Table 7
Comparison of the predictive performance between newDNA-Prot and the competing methods on the DNArset.

Methods	ACC	MCC	F-measure	Specificity
Our method	0.840	0.166	0.222	0.871
iDNA-Prot	0.614	0.132	0.172	0.611
DNAbinder	0.384	0.101	0.143	0.355
DNA-Prot	0.735	0.152	0.197	0.747

5.4. Comparison with existing methods

The proposed method is also compared with several existing methods (DNAbinder (Kumar et al., 2007), iDNA-Prot (Lin et al., 2011), and DNA-Prot (Kumar et al., 2009)). All methods are evaluated on an independent test set, namely DNAiset, which contains 97 DNA-binding proteins and 192 non-binding proteins. Table 6 demonstrates the predictive performance of our method and the competing methods.

As shown in Table 6, our method provides higher accuracy, AUC value, MCC and F-measure than the DNAbinder and DNA-Prot methods. However, the iDNA-Prot (Lin et al., 2011) generates higher accuracy of 0.889, MCC of 0.752, specificity of 1.000, and F-measure of 0.795 than our method. The better performance of iDNA-Prot is due to the fact that a number of sequences in the test DNAiset have more than 35% sequence similarity to the sequences of the training set of S^{Bench}, which was used to build the classification model of iDNA-Prot. Therefore, the performance of the iDNA-Prot method should be overestimated. The DNAbinder method is developed by Kumar et al. (2007) and can be accessed at www.imtech.res.in/raghava/dnabinder/. The method was introduced by Kumar et al. (2007). DNAbinder employs SVM classifier and evolutionary information in the form of PSSM profiles. DNAbinder identifies DNA-binding proteins at a success rate of 70.9% in the DNAiset. The DNA-Prot method was also introduced by Kumar et al. (2009). This method integrates random forest classifier and features calculated from primary sequence. The DNA-Prot method achieves a success rate of 82.4% in prediction of DNA-binding proteins in DNAiset.

We also compare our newDNA-Prot method with iDNA-Prot, DNAbinder and DNA-Prot in DNArset, see Table 7. As discussed in Section 2, DNArset simulates the ration between DNA-binding and non-binding proteins in cellular environment and contains more non-binding proteins than DNAiset. Table 7 demonstrates that our method achieves higher ACC, MCC, F-measure and specificity than the iDNA-Prot, DNAbinder and DNA-Prot methods. Specifically, newDNA-Prot improves on the runner-up method DNA-Prot for (0.84−0.735)/0.735 = 14.3% for ACC, (0.166−0.152)/0.152 = 9.2% for MCC, (0.222−0.197)/0.197 = 12.7% for F-measure and (0.871−0.747)/0.747 = 16.6% for specificity.

These results demonstrate, that using our comprehensive protein sequence information and two-stage feature selection method can better predict the DNA-binding proteins with SVM model than some previous studies.

Table 6
Comparison of the predictive performance between newDNA-Prot and the competing methods on the DNAiset.

Method	ACC	AUC	MCC	Sensitivity	Specificity	Precision	F-measure
DNAbinder	0.709	0.809	0.459	0.845	0.643	0.536	0.656
iDNA-Prot	0.889	N/A	0.752	0.659	1.000	1.000	0.795
DNA-Prot	0.824	0.732	0.589	0.526	0.969	0.894	0.662
Our method	0.848	0.881	0.652	0.753	0.894	0.777	0.764

6. Conclusions

DNA-binding proteins play an important role in a variety of biological processes. Therefore, development of computational method that identifies DNA-binding proteins at high success rate is valuable. This study proposes a novel DNA-binding protein predictor that integrates a comprehensive set of feature descriptors, including primary sequence based, evolutionary profile based, predicted secondary structure based, predicted relative solvent accessibility based, physicochemical property based and biological function based features. To the best of our knowledge, the predicted secondary structure based, predicted relative solvent accessibility based and biological function based features are first employed in prediction of DNA-binding proteins. We have also assessed the performance of three feature selection methods,i.e., the mRMR, wrapper and two-stage methods in prediction of DNA-binding proteins. Results demonstrate that the two-stage method performs better than the other two methods. Statistical analysis shows that more than 95% of the selected features are statistically significant and the selected features cover all 6 feature groups. This implies all 6 feature groups contribute to the prediction of DNA-binding proteins. The proposed newDNA-Prot method is compared with three state of the art algorithms, DNAbinder, DNA-Prot and iDNA-Prot on two independent datasets. The results show that the newDNA-Prot method outperforms existing methods.

Acknowledgements

This work is supported by the National Natural Science Foundation of China (Grant no. 11201334), Science and Technology Commission of Tianjin Municipality (Grant no. 12JCYBJC31900) to KC. JRuan group is supported by Natural Science Fund of China (NSFC) (10671100, 68075049 and 31050110432) and the aid of a grant (No. 104519-010) from the International Development Research Center, Ottawa, Canada.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.compbiolchem.2014.09.002.

References

Ahmad, S., Sarai, A., 2004. Moment-based prediction of DNA-binding proteins. J. Mol. Biol. 341, 65–71.
Ahmad, S., Sarai, A., 2005. PSSM-based prediction of DNA binding sites in proteins. BMC Bioinform. 6, 33–38.
Ahmad, S., Gromiha, M.M., Sarai, A., 2004. Analysis and prediction of DNA-binding proteins and their binding residues based on composition, sequence and structural information. Bioinformatics 20, 477–486.
Ahmad, S., Andrabi, M., Mizuguchi, K., Sarai, A., 2009. Prediction of mono- and dinucieotide-specific DNA-binding sites in proteins using neural networks. BMC Struct. Biol. 9, 30–47.
Altschul, S.F., Madden, T.L., Schaffer, A.A., et al., 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25, 3389–3402.
Baldi, P., Brunak, S., Chauvin, Y., Andersen, C.A., Nielsen, H., 2000. Assessing the accuracy of prediction algorithms for classification: an overview. Bioinformatics 16, 412–424.
Brameier, M., Haan, J., Krings, A., MacCallum, R.M., 2006. Automatic discovery of cross-family sequence features associated with protein function. BMC Bioinform. 7, 16–49.
Brown, J.B., Akutsu, T., 2009. Identification of novel DNA repair proteins via primary sequence, secondary structure, and homology. BMC Bioinform. 10, 25–46.
Buck, M.J., Lieb, J.D., 2004. ChIP-chip: considerations for the design, analysis, and application of genome-wide chromatin immunoprecipitation experiments. Genomics 83, 349–360.
Cai, Y.D., Doig, A.J., 2004. Prediction of *Saccharomyces cerevisiae* protein functional class from functional domain composition. Bioinformatics 20, 1292–1300.
Cai, Y.D., Lin, S.L., 2003. Support vector machines for predicting rRNA-, RNA-, and DNA-binding proteins from amino acid sequence. Biochim. Biophys. Acta 1648, 127–133.
Cai, C.Z., Han, L.Y., Ji, Z.L., Chen, X., Chen, Y.Z., 2003. SVM-Prot: web-based support vector machine software for functional classification of a protein from its primary sequence. Nucleic Acids Res. 31, 3692–3697.
Cajone, F., Salina, M., Benelli-Zazzera, A., 1989. 4-hydroxynonenal induces a DNA-binding protein similar to the heat-shock factor. Biochem. J. 262, 977–979.
Chen, K., Mizianty, M.J., Kurgan, L., 2011. ATPsite: sequence-based prediction of ATP-binding residues. Proteome Sci. 9, S4.
Chou, C.C., Lin, T.W., Chen, C.Y., Wang, A.H., 2003. Crystal structure of the hyperthermophilic archaeal DNA-binding protein Sso 10b2 at a resolution of 1.85 angstroms. J. Bacteriol. 185, 4066–4073.
Chou, K.C., 2011. Some remarks on protein attribute prediction and pseudo amino acid composition. J. Theor. Biol. 273, 236–247.
Dosztanyi, Z., Crizmok, V., Tompa, P., Simon, I., 2005. IUPred: web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content. Bioinformatics 21, 3433–3434.
Dubchak, I., Muchnik, I., Holbrook, S.R., Kim, S.H., 1995. Prediction of protein folding class using global description of amino acid sequence. Proc. Natl. Acad. Sci. U. S. A. 92, 8700–8704.
Fang, Y., Guo, Y., Feng, Y., Li, M., 2008. Predicting DNA-binding proteins: approached from Chou's pseudo amino acid composition and other specific sequence features. Amino Acids 34, 103–109.
Faraggi, E., Zhang, T., Yang, Y.D., Kurgan, L., Zhou, Y.Q., 2012. SPINE X: improving protein secondary structure prediction by multistep learning coupled with prediction of solvent accessible surface area and backbone torsion angles. J. Comput. Chem. 33, 259–267.
Freeman, K., Gwadz, M., Shore, D., 1995. Molecular and genetic analysis of the toxic effect of RAP1 overexpression in yeast. Genetics 141, 1253–1262.
Gao, M., Skolnick, J., 2009. A threading-based method for the prediction of DNA-binding proteins with application to the human genome. PLoS Comput. Biol. 5, e1000567.
Gao, J., Faraggi, E., Zhou, Y., Ruan, J., Kurgan, L., 2012. BEST: improved prediction of B-cell epitopes from antigen sequences. PloS One 7, e40104.
Gromiha, M.M., Nagarajan, R., 2013. Computational approaches for predicting the binding sites and understanding the recognition mechanism of protein-DNA complexes. Adv. Protein Chem. Struct. Biol. 91, 65–99.
Gromiha, M.M., Yabuki, Y., 2008. Functional discrimination of membrane proteins using machine learning techniques. BMC Bioinform. 9, 135–142.
Gryk, M.R., Jardetzky, O., Kilg, L.S., Yanofsky, C., 1996. Flexibility of DNA binding domain of trp repressor required for recognition of different operator sequences. Protein Sci. 5, 1195–1197.
Huang, Y., Niu, B., Gao, Y., Fu, L., Li, W., 2010. CD-HIT Suite: a web server for clustering and comparing biological sequences. Bioinformatics 26, 680–682.
Huang, H.L., Lin, I.C., Liou, Y.F., Isai, C.T., Hsu, K.T., Huang, W.L., Ho, S.J., Ho, S.Y., 2011. Predicting and analyzing DNA-binding domains using a systematic approach to identifying a set of informative physicochemical and biochemical properties. BMC Bioinform. 12, S47.
Keil, M., Exner, T.E., Brickmann, J., 2004. Pattern recognition strategies for molecular surfaces: III: binding site prediction with a neural network. J. Comput. Chem. 25, 779–789.
Kohavi, R., John, G., 1997. Wrapper for feature subset selection. Artif. Intell. 97, 273–324.
Kumar, M., Gromiha, M.M., Raghava, G.P., 2007. Identification of DNA-binding proteins using support vector machines and evolutionary profiles. BMC Bioinform. 8, 463–472.
Kumar, K.K., Pugalenthi, G., Suganthan, P.N., 2009. DNA-Prot: identification of DNA binding proteins from protein sequence information using random forest. J. Biomol. Struct. Dyn. 26, 679–686.
Kurgan, L., Cios, K., Chen, K., 2008. SCPRED: accurate prediction of protein structural class for sequences of twilight-zone similarity with predicting sequences. BMC Bioinform. 9, 226–240.
Langley, P., 1994. Selection of relevant features in machine learning. Proc. AAAI Fall Symp. Relevance .
Lin, W.Z., Fang, J.A., Xiao, X., Chou, K.C., 2011. iDNA-prot: identification of DNA-binding proteins using random forest with grey model. PLoS One 6, e24756.
Liu, Z.P., Wu, L.Y., Wang, Y., Zhang, X.S., Chen, L.N., 2008. Bridging protein local structures and protein functions. Amino Acids 35, 627–650.
McGuffin, L.J., Bryson, K., Jones, D.T., 2000. The PSIPRED protein structure prediction server. Bioinformatics 16, 404–405.
Mizianty, M.J., Kurgan, L., 2011. Sequence-based prediction of protein crystallization, purification and production propensity. Bioinformatics 27, i24–33.
Mizianty, M.J., Stach, W., Chen, K., Kedarisetti, K.D., Disfani, F.M., Kurgan, L., 2010. Improved sequence-based prediction of disordered regions with multilayer fusion of multiple information sources. Bioinformatics 26, i489–496.
Nanni, L., Brahnam, S., Lumini, A., 2010. High performance set of PseAAC and sequence based descriptors for protein classification. J. Theor. Biol. 266, 1–10.
Nimrod, G., Schushan, M., Szilagyi, A., Leslie, C., Ben-Tal, N., 2010. iDBPs: a web server for the identification of DNA binding proteins. Bioinformatics 26, 692–693.

Peng, H., Long, F., Ding, C., 2005. Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy. IEEE Trans. Pattern Anal. Mach. Intell. 27, 1226–1238.
Qian, Z.L., Cai, Y.D., Li, Y.X., 2006. A novel computational method to predict transcription factor DNA binding preference. Biochem. Biophys. Res. Commun. 348, 1034–1037.
Reinhardt, A., Eisenberg, D., 2004. DPANN: improved sequence to structure alignments following fold recognition. Proteins 56, 528–538.
Robert, E.L., Hui, L., 2010. Boosting the prediction and understanding of DNA binding domains from sequence. Nucleic Acids Res. 38, 3149–3185.
Shao, X., Tian, Y., Wu, L., Wang, Y., Jing, L., Deng, N., 2009. Prediction DNA- and RNA-binding proteins from sequences with kernel methods. J. Theor. Biol. 258, 289–293.
Song, J., Burrage, K., 2006. Predicting residue-wise contact orders in proteins by support vector regression. BMC Bioinform. 7, 425–439.
Song, J., Tan, H., Takemoto, K., Akutsu, T., 2008. HSEpred: predict half-sphere exposure from protein sequences. Bioinformatics 24, 1489–1497.
Stawiski, E.W., Gregoret, L.M., Mandel-Gutfreund, Y., 2003. Annotating nucleic acid-binding function based on protein structure. J. Mol. Biol. 326, 1065–1079.
Szilagyi, A., Skolnick, J., 2006. Efficient prediction of nucleic acid binding function from low-resolution protein structures. J. Mol. Biol. 358, 922–933.
Tjiong, H., Zhao, H.X., 2007. DISPLAR: an accurate method for prediction DNA-binding sites on protein surfaces. Nucleic Acids Res. 35, 1465–1477.
Vuzman, D., Levy, Y., 2012. Intrinsically disordered regions as affinity tuners in protein-DNA interactions. Mol. Biosyst. 8, 47–57.
Wang, G., Dunbrack Jr., R.L., 2003. PISCES: a protein sequence culling server. Bioinformatics 19, 1589–1591.
Wang, L.J., Yang, M.Q., Yang, J.Y., 2009. Prediction of DNA-binding residues from protein sequence information using random forests. BMC Genomics 10, S1.
Wang, L., Huang, C., Yang, M.Q., Yang, J.Y., 2010. BindN+ for accurate prediction of DNA and RNA-binding residues from protein sequence features. BMC Syst. Biol. 1, S3.
Wu, S., Zhang, Y., 2008. MUSTER: improving protein sequence profile–profile alignments by using multiple sources of structure information. Proteins 72, 547–556.
Xiong, Y., Liu, J., Wei, D.Q., 2011. An accurate feature-based method for identifying DNA-binding residues on protein surfaces. Proteins 79, 509–517.
Yan, C., Terribilini, M., Wu, F., Jernigan, R.L., Dobbs, D., Honavar, V., 2006. Predicting DNA-binding sites of proteins from amino acid sequence. BMC Bioinform. 7, 262–271.
Yuan, Y., Shi, X., Li, X., Lu, W., Cai, Y., Gu, L., Liu, L., Li, M., Kong, X., Xing, M., 2010. Prediction of interactiveness of proteins and nucleic acids based on feature selections. Mol. Divers. 14, 627–633.
Zhang, Z., Kochhar, S., Grigorov, M.G., 2005. Descriptor-based protein remote homology identification. Protein Sci. 14, 431–444.
Zhang, H., Zhang, T., Chen, K., Shen, S., Ruan, J., Kurgan, L., 2008. Sequence based residue depth prediction using evolutionary information and predicted secondary structure. BMC Bioinform. 9, 388–409.
Zhao, H., Yang, Y., Zhou, Y., 2010. Structure-based prediction of DNA-binding proteins by structural alignment and a volume-fraction corrected DFIRE-based energy function. Bioinformatics 26, 1857–1863.
Zheng, C., Kurgan, L., 2008. Prediction of beta-turns at over 80% accuracy based on an ensemble of predicted secondary structures and multiple alignments. BMC Bioinform. 9, 430–443.
Zou, C.X., Gong, J.Y., Li, H.L., 2013. An improved sequence based prediction protocol for DNA-binding proteins using SVM and comprehensive feature analysis. BMC Bioinform. 14, 90–103.