SCIENTIFIC REPORTS

OPEN

Received: 29 May 2015
Accepted: 28 September 2015
Published: 20 October 2015

DNA binding protein identification by combining pseudo amino acid composition and profile-based protein representation

Bin Liu[1,2], Shanyi Wang[1] & Xiaolong Wang[1,2]

DNA-binding proteins play an important role in most cellular processes. Therefore, it is necessary to develop an efficient predictor for identifying DNA-binding proteins only based on the sequence information of proteins. The bottleneck for constructing a useful predictor is to find suitable features capturing the characteristics of DNA binding proteins. We applied PseAAC to DNA binding protein identification, and PseAAC was further improved by incorporating the evolutionary information by using profile-based protein representation. Finally, Combined with Support Vector Machines (SVMs), a predictor called iDNAPro-PseAAC was proposed. Experimental results on an updated benchmark dataset showed that iDNAPro-PseAAC outperformed some state-of-the-art approaches, and it can achieve stable performance on an independent dataset. By using an ensemble learning approach to incorporate more negative samples (non-DNA binding proteins) in the training process, the performance of iDNAPro-PseAAC was further improved. The web server of iDNAPro-PseAAC is available at http://bioinformatics.hitsz.edu.cn/iDNAPro-PseAAC/.

DNA-binding proteins have diverse functions in the cell, and play vital roles in various cellular processes, such as gene regulation, DNA replication, and repair[1]. Identification of DNA-binding proteins is one of the most important tasks in the annotation of protein functions. In recent years, DNA-binding proteins can be identified by several experimental techniques, including filter binding assays[2], X-ray crystallography[3] and NMR[4]. However, it is time-consuming and expensive to identify DNA-binding proteins by experimental approaches. Facing the avalanche of new protein sequences generated in the post-genomic and big data age[5,6], it is highly desired to develop automated methods for rapidly and effectively identifying DNA-binding proteins basing on the protein sequence information alone.

The computational methods for DNA binding protein identification can be grouped into two categories: (i) methods based on structures (ii) methods based on sequences. The first type makes use of both the structural and sequential information of target proteins (see, e.g.,[7–10]). Although these methods show promising predictive performance, the structural information of proteins is not always available, particularly for the huge amount of proteins, which prevents the application of these methods. In contrast, the second type methods overcome this shortcoming by only requiring the sequence information as input for the prediction[11–20].

A key to improve the performance of the sequence-based methods is to find suitable feature extraction algorithms that can capture the characteristics of DNA binding proteins and non DNA binding proteins. Motivated by the successful application of Chou's pseudo amino acid composition (PseAAC) to many important tasks in the field of computational proteomics, here we are to propose a new approach

[1]School of Computer Science and Technology, Harbin Institute of Technology Shenzhen Graduate School, Shenzhen, Guangdong, China. [2]Key Laboratory of Network Oriented Intelligent Computation, Harbin Institute of Technology Shenzhen Graduate School, Shenzhen, Guangdong, China. Correspondence and requests for materials should be addressed to B.L. (email: bliu@insun.hit.edu.cn)

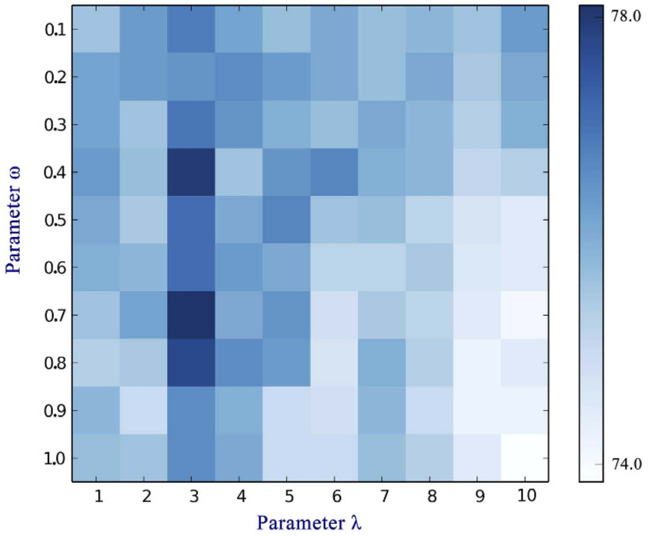

Figure 1. The performance of iDNAPro-PseAAC with different λ and ω combinations. iDNAPro-PseAAC achieves the best performance with $\lambda = 3$ and $\omega = 0.7$.

for DNA binding protein identification called **iDNAPro-PseAAC**, which extends the classic PseAAC approach by incorporating the evolutionary information in the form of profile-based protein representation[21]. The **iDNAPro-PseAAC** has the following advantages compared with other currently available approaches: (i) It is able to incorporate the global or long range sequence-order effects by means of PseAAC. (ii) The evolutionary information imbedded in the profile-based protein representation is employed by **iDNAPro-PseAAC**. (iii) It considers the various physical-chemical properties of amino acids.

To establish a really useful statistical predictor for a protein system, we need to consider the following procedures: (i) Construct or select a valid benchmark dataset to train and test the predictor. (ii) Formulate the protein samples with an effective mathematical expression that can truly reflect their intrinsic correlation with the attribute to be predicted. (iii) Introduce or develop a powerful algorithm (or engine) to operate the prediction. (iv) Perform properly cross-validation tests to objectively evaluate the anticipated accuracy of the predictor. Below, we are going to describe how to build the new predictor according to the four procedures.

Results

The influence of λ and ω on the performance of iDNAPro-PseAAC.

There are two parameters λ and ω in **iDNAPro-PseAAC**, which would influence its performance (see method section). λ can be any integer between 1 and L-1, where L is the shortest length of sequences in the benchmark dataset. The range of ω is 0-1. The performance of **iDNAPro-PseAAC** with different λ and ω combinations is shown in Fig. 1 and Supplementary S1, from which we can see that **iDNAPro-PseAAC** achieves the best performance with $\lambda = 3$ and $\omega = 0.7$. These parameter values are used in the following experiments.

Discriminant Visualization.

To further study the discriminant power of features, we calculate the discriminant weight vectors in the feature space. In the SVM training process, we can get the sequence-specific weights, which can be used to calculate the discriminant weight of each feature. The feature discriminant weight vector **W** can be calculated as following:

$$\mathbf{W} = \mathbf{A} \cdot \mathbf{M} = \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_N \end{bmatrix}^{\mathbf{T}} \begin{bmatrix} m_{11} & m_{12} & \cdots & m_{1j} \\ m_{21} & m_{22} & \cdots & m_{2j} \\ \vdots & \vdots & \ddots & \vdots \\ m_{N1} & m_{N2} & \cdots & m_{Nj} \end{bmatrix} \quad (1)$$

where the weight vector **A** of the training set with N samples obtained from the kernel-based training; **M** is the matrix of sequence representatives; j is the dimension of the feature vector. The element in **W** represents the discriminative power of the corresponding feature.

The discriminative weights of all the 23 features are shown in Fig. 2. We can see that 9 amino acids show positive values, while the other 11 amino acids show negative values. Interestingly, most of the amino acids with positive values, such as R, K, have been reported as important residues in DNA binding proteins, and they are redundant in DNA protein binding regions[22]. **iDNAPro-PseAAC** is able to capture

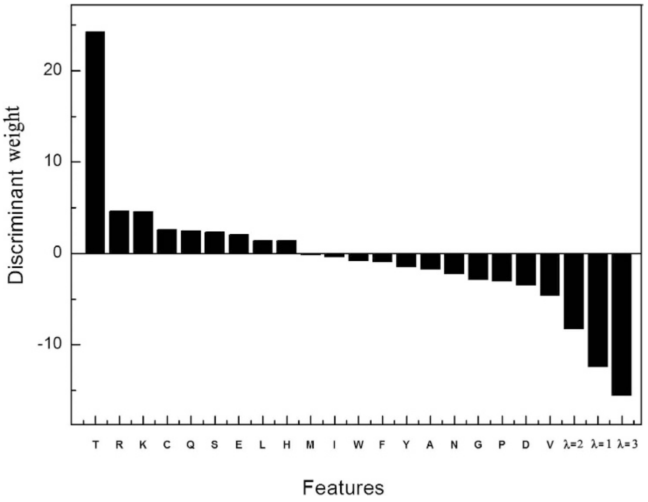

Figure 2. The discriminative weights of all the 23 features.

Method	Acc(%)	MCC	Sn(%)	Sp(%)	AUC(%)
iDNAPro-PseAAC	76.56	0.53	75.62	77.45	83.92
DNAbinder (dimension 21)[a]	73.95	0.48	68.57	79.09	81.40
DNAbinder (dimension 400)[b]	73.58	0.47	66.47	80.36	81.50
DNA-Prot[c]	72.55	0.44	82.67	59.76	78.90
iDNA-Prot[d]	75.40	0.50	83.81	64.73	76.10

Table 1. A comparison of the jackknife test results by iDNAPro-PseAAC with the other methods on the benchmark dataset of Eq. 2 (cf. Supporting Information S3).

this kind of features of DNA binding proteins, which could explain the reason for its better performance. Another interesting pattern is that all the three features capturing the sequence-order effects ($\lambda = 1, 2, 3$) show negative values, indicating that this kind of features is useful for representing the features of non DNA binding proteins.

Results on the benchmark dataset. Table 1 shows the predictive results of **iDNAPro-PseAAC** on the benchmark dataset by using Jackknife test. For comparison, the results of four state-of-the-art methods are also listed, including DNAbinder (dimension 21)[23], DNAbinder (dimension 400)[23], DNA-Prot[24] and iDNA-Prot[25]. The reason why we select these four methods is that they have public available software tools with reported optimized parameters. Their optimized results on the benchmark dataset can be easily obtained by using these tools and parameter settings.

From Table 1 we can see that **iDNAPro-PseAAC** achieves the best performance. In order to further study the performance of the proposed method, the ROC curve is employed to evaluate the performance of different methods. ROC curve is a graphical plot that illustrates the performance of a binary classifier system with its discrimination threshold varying. The horizontal coordinate is false-positive rate and the vertical is true-positive rate. The true-positive rate is also known as sensitivity in biomedical informatics, or recall in machine learning[5,26]. The false-positive rate is also known as the fall-out and can be calculated as 1 -specificity. The area under the curve (AUC) is the evaluation criteria for the classifier. Figure 3 shows the ROC curve of the five methods, from which we can see that **iDNAPro-PseAAC** outperforms other four approaches in terms of AUC.

Performance comparison with other related computational predictors. To further evaluate the performance of **iDNAPro-PseAAC** and facilitate the comparison against previous predictors, an independent test dataset PDB186 constructed by Lou *et al.* is used[27], where 93 proteins are DNA-binding proteins and 93 proteins are non-DNA-binding proteins. To avoid the homology bias, we use the NCBI's BLASTCLUST[28] to remove those proteins from the benchmark dataset that have more than 25% sequence identity to any protein within a same subset of the PDB186 dataset. The **iDNAPro-PseAAC** is re-trained on the resulting benchmark dataset, and then this model is used to predict the samples in the independent dataset. The results are shown in Table 2, and the ROC curves of various methods are plotted in Fig. 4. Compared with the results listed in Table 2, we can see that **iDNAPro-PseAAC** can achieve stable performance on the independent dataset, indicating that the proposed method is a

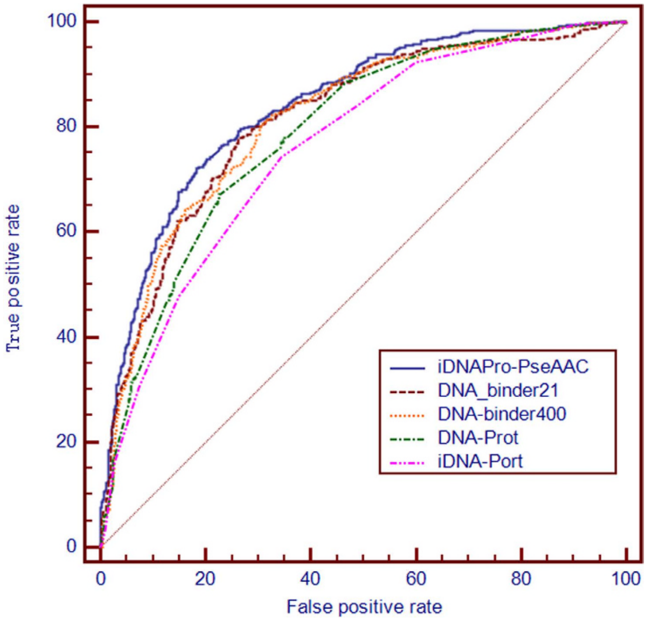

Figure 3. The ROC curves obtained by different methods on the benchmark dataset using the jackknife tests. The areas under the ROC curves or AUC are 0.839, 0.826, 0.814, 0.815, 0.789 and 0.761 for iDNAPro-PseAAC, DNAbinder (dimension 21), DNAbinder(dimension 400), DNA-Prot and iDNA-Prot, respectively. See the main text for further explanation.

Methods	Acc(%)	MCC	Sn(%)	Sp(%)	AUC(%)
iDNAPro-PseAAC	69.89	0.402	77.41	62.37	77.54
iDNA-Prot	67.20	0.344	67.70	66.70	N/A[b]
DNA-Prot	61.80	0.240	69.90	53.80	N/A
DNAbinder	60.80	0.216	57.00	64.50	60.70
DNABIND	67.70	0.355	66.70	68.80	69.40
DNA-Threader	59.70	0.279	23.70	95.70	N/A
DBPPred	76.90	0.538	79.60	74.20	79.10

Table 2. A comparison of the results[a] obtained by iDNAPro-PseAAC and the other methods on the independent dataset PDB186. [a]The results of iDNA-Prot[25], DNA-Prot[24], DNAbinder[23], DNABIND[47], DNA-Threader[48], and DBPPred[27] were obtained from[27]. [b]N/A represents the unreported values.

useful tool for DNA binding protein identification. **iDNAPro-PseAAC** outperforms other approaches except for DBPPred. However, our method is more efficient than DBPPred. DBPPred uses 1486 features derived from predicted secondary structure, predicted relative solvent accessibility, and position specific scoring matrix. These features are calculated with the help of two software tools, including SPINE-X and Psi-Blast. All these tools require a time consuming multiple sequence alignment process[29]. Furthermore, these features contain several parameters, which should be optimized on a validate dataset. This requires additional running time for DBPPred and raises the risk of over-fitting problem caused by this parameter optimization process. In contrast, the 23 features used in **iDNAPro-PseAAC** can be easily generated only based on the protein sequences. Therefore, **iDNAPro-PseAAC** is more efficient than DBPPred, and avoids the risk of over-fitting.

Influence of Negative Samples on the Predictive Performance. In real world application, there are more non DNA binding proteins (negative samples) than the DNA binding proteins (positive samples)[30]. However, in order to avoid the classifier biased problem, a balanced benchmark dataset $\mathbb{S}$ is used to construct **iDNAPro-PseAAC**. Therefore, it is interesting to explore the influence of different negative sets on the predictive performance of **iDNAPro-PseAAC**. In this regard, we conduct the following experiments. First we extend the size of the negative set $\mathbb{S}^{-}$ in the the benchmark dataset $\mathbb{S}$ by selecting more non DNA binding proteins from PDB[31]. After removing the redundant proteins sharing more than 25%

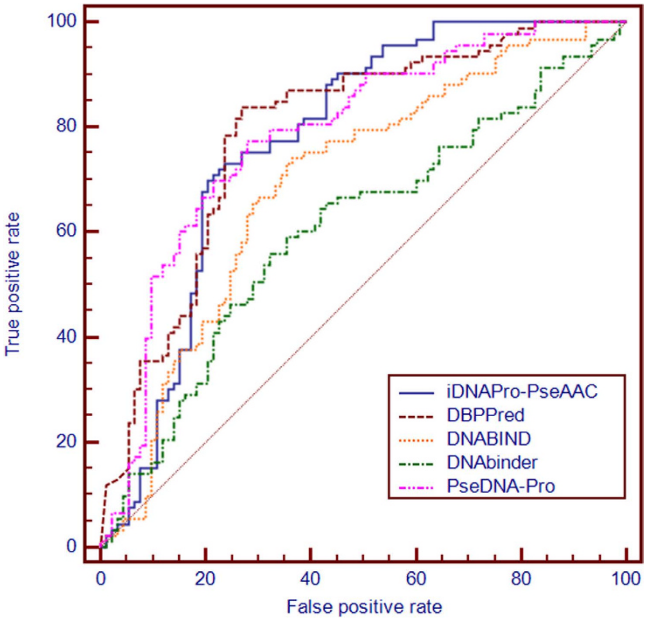

Figure 4. The ROC curves obtained by different methods on the independent dataset PDB186. The areas under the ROC curves or AUC are 0.775, 0.607, 0.694, and 0.791 for iDNAPro-PseAAC, DNAbinder, DNABIND and DBPPred, respectively. See the main text for further explanation.

Methods	Acc(%)	MCC	Sn(%)	Sp(%)	AUC(%)
iDNAPro-PseAAC-1	69.89	0.406	79.57	60.22	75.00
iDNAPro-PseAAC-2	71.50	0.449	86.02	56.99	75.00
iDNAPro-PseAAC-3	69.89	0.409	81.72	58.06	74.10
iDNAPro-PseAAC-4	69.35	0.407	84.95	53.76	70.20
iDNAPro-PseAAC-EL	71.50	0.442	82.76	60.22	77.80

Table 3. Results on independent dataset PDB186 achieved by iDNAPro-PseAAC trained with different datasets.

similarity with the independent dataset, we obtain 2059 negative samples, which were listed in Supplementary S2. The extended negative set is then randomly divided into 4 subsets. For each subset, its size is approximately equal to that of the positive set $\mathbb{S}^+$ in the the benchmark dataset $\mathbb{S}$. These four subsets are respectively combined with the positive set $\mathbb{S}^+$, and four new datasets are generated. Four predictors of **iDNAPro-PseAAC** trained with these four datasets can be represented as **iDNAPro-PseAAC-1, iDNAPro-PseAAC-2, iDNAPro-PseAAC-3, and iDNAPro-PseAAC-4**, respectively. Their performance is then evaluated on the independent dataset. Table 3 shows the results of the four methods, and the corresponding ROC curves are plotted in Fig. 5, from which, we can see that the four predictors show similar performance, indicating that different subsets of negative samples don't have significant impact on the performance of **iDNAPro-PseAAC**. Next, we investigate if these four predictors can be combined to further improve the performance. In this regard, we employ a simple ensemble learning approach to combine them[32,33]. For each test sample, it is predicted by the four predictors respectively, and the final class label of the test sample is assigned based on the average values of the four probability values calculated by the four predictors. The results of **iDNAPro-PseAAC-EL** (**iDNAPro-PseAAC** with the ensemble learning approach) are shown in Table 3 and Fig. 5. Performance improvement can be observed. This is because by using ensemble learning method, more negative samples are used to train **iDNAPro-PseAAC**, leading to a more accurate predictor.

Discussion

Because of the importance of DNA binding protein identification, computational predictors only using the sequence information for DNA binding protein identification is highly desired. In this study, we proposed a method called **iDNAPro-PseAAC** for DNA binding protein identification, which combines the pseudo amino acid composition with profile-based protein representation. Experimental results show

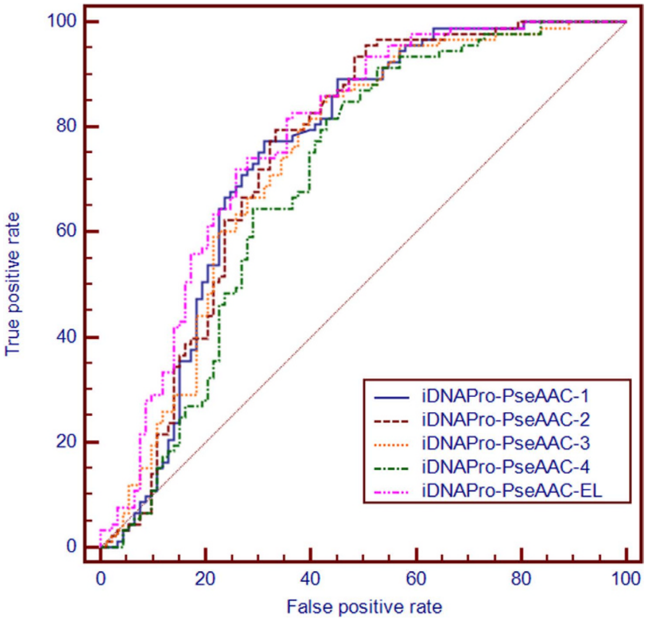

Figure 5. The ROC curves obtained by different iDNAPro-PseAAC predictors on the independent dataset PDB186. The areas under the ROC curves or AUC are 0.750, 0.750, 0.741, 0.702, 0.778 for **iDNAPro-PseAAC-1, iDNAPro-PseAAC-2, iDNAPro-PseAAC-3, iDNAPro-PseAAC-4, and iDNAPro-PseAAC-EL**, respectively.

that it outperform other approaches in both benchmark dataset and independent dataset. Furthermore, the discriminative model can be analyzed to reveal the in-depth features of DNA binding proteins, which would benefit the researchers who want to investigate the characteristics of DNA binding proteins. Some recent studies have shown that DNA-binding proteins also regulate the microRNA targets, and involve in the noncoding RNA-protein-disease network [34–39]. We believe that this predictor would be a high throughput tool for DNA binding protein investigation.

Methods

Benchmark Dataset. A reliable and stringent benchmark dataset is necessary to build and evaluate a statistical predictor. In this regard, an updated benchmark dataset for this study is constructed based on the latest version of Protein Data Bank (PDB)[31], which can be formulated as:

$$\mathbb{S} = \mathbb{S}^{+} \cup \mathbb{S}^{-} \tag{2}$$

where $\mathbb{S}^{+}$ represents the subset containing DNA binding proteins (positive samples), $\mathbb{S}^{-}$ represents the subset containing non DNA binding proteins (negative samples), and the symbol ∪ is the "union" in the set theory. DNA-binding protein sequences are collected from the PDB by searching the mmCIF keyword of 'DNA binding protein', 'protein-DNA complex', and other key words with similar meaning. To construct a high quality and non-redundant benchmark dataset, the protein sequences obtained should be filtered by the following 2 criteria. (1) Proteins with length less than 50 AA were removed, which might be fragment. (2) To reduce redundancy and homology bias, the sequence similarity lower than 25% between any two proteins were cut off by using PISCES[40]. Finally, we obtained 525 DNA binding proteins for the subset of $\mathbb{S}^{+}$. 550 non DNA binding proteins were randomly selected from the PDB according to the above criteria. The accession codes and sequences of the 525 positive and 550 negative samples are given in the Supplementary S3.

Profile-based protein representation. Profile-based protein representation[21] is an efficient approach to extract the evolutionary information from frequency profiles. Its main steps are as follows.

Given the protein sequence **P** consisting *L* amino acids as formulated as:

$$\mathrm{P} = \mathrm{R}_1\mathrm{R}_2\mathrm{R}_3\mathrm{R}_4\mathrm{R}_5\mathrm{R}_6 \cdots \mathrm{LR}_L \tag{3}$$

where R_1 represents the 1st residue, R_2 represents the 2nd residue and so forth. The frequency profile of sequence **P** generated by PSI-BLAST[28] with default parameters can be represented as a matrix **M**:

$$\mathbf{M} = \begin{bmatrix} m_{1,1} & m_{1,2} & \cdots & m_{1,L} \\ m_{2,1} & m_{2,2} & \cdots & m_{2,L} \\ \vdots & \vdots & \vdots & \vdots \\ m_{20,1} & m_{20,2} & \cdots & m_{20,L} \end{bmatrix} \tag{4}$$

where 20 is the number of standard amino acids; m_{ij} is the target frequency representing the probability of amino acid i ($i = 1, 2, \ldots, 20$) appearing in sequence position j ($j = 1, 2, 3\ldots, L$) of protein **P** during evolutionary process. The m_{ij} is calculated as:

$$m_{i,j} = \frac{(\alpha f_{ij} + \beta g_{ij})}{(\alpha + \beta)} \tag{5}$$

where f_{ij} represents the observed frequency of amino acid i in column j, α is the number of different amino acids in column j-1. β is a free parameter set to a constant value of 10, which is initially used by PSI-BLAST. g_{ij} is the pseudo-count for standard amino acid i in position j. It is calculated as follows:

$$g_{ij} = \sum_{k=1}^{20} \frac{f_{kj} q_{ik}}{p_k} \tag{6}$$

where p_k is the background frequency of amino acid k, q_{ik} is the score of amino acid i being aligned to amino acid j in BLOSUM62 substitution matrix, which is the default score matrix of PSI-BLAST.

For each column in **M**, the amino acids are sorted in descending order according to their frequency values. Thus the sorted matrix $\mathbf{M}^{\downarrow}$ can be represented as:

$$\mathbf{M}^{\downarrow} = \begin{bmatrix} m_{1,1}^{\downarrow} & m_{1,2}^{\downarrow} & \cdots & m_{1,L}^{\downarrow} \\ m_{2,1}^{\downarrow} & m_{2,2}^{\downarrow} & \cdots & m_{2,L}^{\downarrow} \\ \vdots & \vdots & \vdots & \vdots \\ m_{20,1}^{\downarrow} & m_{20,2}^{\downarrow} & \cdots & m_{20,L}^{\downarrow} \end{bmatrix} \tag{7}$$

where

$$m^{\downarrow}{}_{1,j} \geq m^{\downarrow}{}_{2,j} \geq \cdots \geq m^{\downarrow}{}_{20,j} (j = 1, 2, \ldots L) \tag{8}$$

The profile-based protein representation $\mathbf{P'}$ of protein **P** can be generated by combining the most frequent amino acids in all the columns of **M**, and can be represented as:

$$\mathrm{P}' = \mathrm{R}'_1 \mathrm{R}'_2 \mathrm{R}'_3 \mathrm{R}'_4 \mathrm{R}'_5 \mathrm{R}'_6 \cdots \mathrm{R}'_\mathrm{L} \tag{9}$$

where R'_i represents the most frequent amino acid in the i-th column of $\mathbf{M}^{\downarrow}$, whose frequency value is $m^{\downarrow}{}_{1,i}$.

Pseudo amino acid composition (PseAAC). One of the most important but also most difficult problems in computational biology and biomedicine is how to formulate a biological sequence with a discrete model or a vector, yet still keep considerable sequence order information. This is because all the existing operation engines, such as SVM (Support Vector Machine) and NN (Neural Network), can only handle vector but not sequence samples, as elaborated in[41,42]. However, a vector defined in a discrete model may completely lose all the sequence-order information. To avoid completely losing the sequence-order information for proteins, the pseudo amino acid composition or PseAAC was proposed[43].

The PseAAC approach then performs on the profile-based protein representation $\mathbf{P'}$ (c.f. Eq. 9) to convert it into a fixed length feature vector by using PseAAC:

$$\mathbf{P'} = [x_1, x_2, \ldots, x_u, \ldots, x_{20+\lambda}]^{\mathrm{T}} \tag{10}$$

where T is transpose operator, λ is the distance parameter considering the sequence-order effects of residues in proteins. x_u can be calculated by

$$x_\mathrm{u} = \begin{cases} \dfrac{f(u)}{\sum_{i=1}^{20} f(u) + \omega \sum_{k=1}^{\lambda} \theta_k} & (1 \leq u \leq 20) \\ \dfrac{\omega \theta_u}{\sum_{i=1}^{20} f(u) + \omega \sum_{k=1}^{\lambda} \theta_k} & (20 + 1 \leq u \leq 20 + \lambda) \end{cases} \tag{11}$$

where f_u is the occurrence frequency of 20 standard amino acid in profile-based protein representation $\mathbf{P'}$. ω is the weight factor for the sequence-order effect. θ_k is the sequence-order correlation factor, which can be calculated as:

$$\theta_k = \frac{\sum_{i=1}^{L-k}\Theta(\mathrm{R'}_i, \mathrm{R'}_{i+k})}{L - k} \qquad (k \le \lambda) \tag{12}$$

where $\mathrm{R'}_i$ is the *i*-th amino acid in $\mathbf{P'}$. L is the length of $\mathbf{P'}$. k is the distance between two amino acids along $\mathbf{P'}$. $\Theta(\mathrm{R'}_i, \mathrm{R'}_{i+k})$ represents the scores calculated according to seven kinds of physical-chemical properties of amino acids (their values are listed in Supplementary S4, which can be calculated by:

$$\Theta(\mathrm{R'}_i, \mathrm{R'}_{i+k}) = \frac{1}{7}\sum_{j=1}^{7}\left(\mathrm{I}_j(\mathrm{R'}_i) - \mathrm{I}_j(\mathrm{R'}_{i+k})\right)^2 \tag{13}$$

where $\mathrm{I}_j(\mathrm{R'}_i)$ and $\mathrm{I}_j(\mathrm{R'}_{i+k})$ are the normalized physicochemical property values of amino acid $\mathrm{R'}_i$ and $\mathrm{R'}_{i+k}$ in property *j*, which can be calculated by the following equation:

$$\mathrm{I}_j(a_i) = \frac{\tilde{\mathrm{I}}_j(a_i) - \frac{1}{20}\sum_{k=1}^{20}\tilde{\mathrm{I}}_j(a_k)}{\sqrt{\frac{1}{20}\sum_{k=1}^{20}\left(\tilde{\mathrm{I}}_j(a_k) - \frac{1}{20}\sum_{k=1}^{20}\tilde{\mathrm{I}}_j(a_k)\right)^2}} \tag{14}$$

where $\tilde{\mathrm{I}}_j(a_i)$ represents the raw physicochemical property value of amino acid a_j in property *j*. a_k ($k = 1, 2, 3, 4, \ldots, 20$) represents the 20 standard amino acids.

Support Vector Machine. In machine learning, support vectors are supervised learning models with associated learning algorithms[44]. For a given training samples, the basic mission of SVM is constructing a separating hyper-plane to maximize the margin of different samples in training set. An SVM model is a representation of examples as points in space, mapped so that the examples of the separate categories are divided by a clear gap.

In this study, we adopt the Lib-SVM package. The kernel function was set as Radial Basis Function(RBF) which can be defined as:

$$\mathrm{K}(\mathrm{X}_{i,}\mathrm{X}_{j}) = \exp(-\gamma\|\mathrm{X}_\mathrm{i} - \mathrm{X}_\mathrm{j}\|^2) \tag{15}$$

The two parameters C and γ were were optimized on the benchmark dataset by the grid tools in the LIBSVM. After optimizing, C is set as 8192 and γ is set as 8.0.

The flowchart of generating the feature vectors and constructing the SVM classifier for **iDNAPro-PseAAC** is shown in Fig. 6.

Evaluation methodology. How to evaluate the performance of a new predictor is a key component. There are three cross-validation methods, which are often used: independent dataset, subsampling or K-fold(such as 5-fold, 7-fold, or 10-fold) test, and Jackknife test. However, there are considerable arbitrariness exists in the independent dataset test and the K-fold cross validation. Jackknife can make the least arbitrary and has been widely used in computational genomics and proteomics. In the jackknife test, each of the proteins sequence in the benchmark is singled out as an independent test sample in turn.

Also, four metrics called the sensitivity(Sn), specificity(Sp), accuracy(Acc), and Mathew's correlation coefficient(MCC), are often used to measure the test quality of a predictor from different angles[45].

$$\begin{cases} \mathrm{Sn} = \dfrac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \\ \mathrm{Sp} = \dfrac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FP}} \\ \mathrm{Acc} = \dfrac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}} \\ \mathrm{Mcc} = \dfrac{(\mathrm{TP} \times \mathrm{TN}) - (\mathrm{FP} \times \mathrm{FN})}{\sqrt{(\mathrm{TP} + \mathrm{FP})(\mathrm{TP} + \mathrm{FN})(\mathrm{TN} + \mathrm{FP})(\mathrm{TN} + \mathrm{FN})}} \end{cases} \tag{16}$$

where TP represents the number of the true positive; TN, the number of the true negative; FP, the number of the false positive; FN, the number of the false negative; SN, the sensitivity; Sp, the specificity; Acc, the accuracy; MCC, the Mathew's correlation coefficient.

In the study, we also use the metrics receiver operation characteristics (ROC) score. ROC curve is a graphical plot that illustrates the performance of a binary classifier system as its discrimination threshold

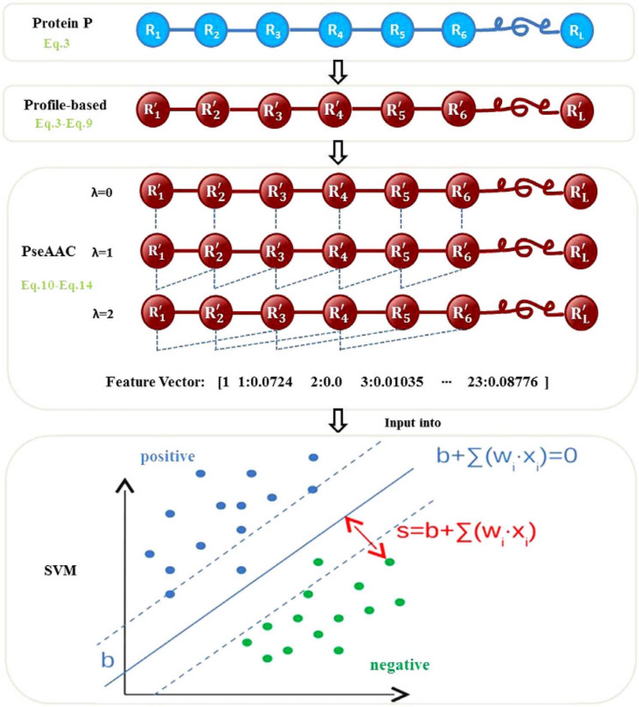

Figure 6. The flowchart of iDNAPro-PseAAC. First, a profile-based protein is generated based on the frequency profile. Pseudo amino acid composition is then performed on this profile-based protein to convert it into a fixed feature vector, which will be fed into SVM to train the classification model.

is varied[46]. A score 1 denotes perfect separation of positive samples from negative ones, whereas a score of 0 indicates that none of the sequences selected by the algorithm is positive.

References

1. Jones, K. A., Kadonaga, J. T., Rosenfeld, P. J., Kelly, T. J. & Tjian, R. A cellular DNA-binding protein that activates eukaryotic transcription and DNA replication. *Cell* **48,** 79–89, doi: http://dx.doi.org/10.1016/0092-8674(87)90358-8 (1987).
2. Helwa, R. & Hoheisel, J. Analysis of DNA–protein interactions: from nitrocellulose filter binding assays to microarray studies. *Anal Bioanal Chem* **398,** 2551–2561, doi: 10.1007/s00216-010-4096-7 (2010).
3. Jaiswal, R., Singh, S. K., Bastia, D. & Escalante, C. R. Crystallization and preliminary X-ray characterization of the eukaryotic replication terminator Reb1-Ter DNA complex. *Acta Crystallographica Section F* **71,** 414–418, doi: doi: 10.1107/S2053230X15004112 (2015).
4. Omichinski, J. *et al.* NMR structure of a specific DNA complex of Zn-containing DNA binding domain of GATA-1. *Science* **261,** 438–446, doi: 10.1126/science.8332909 (1993).
5. Lin, C. *et al.* LibD3C: Ensemble Classifiers with a Clustering and Dynamic Selection Strategy. *Neurocomputing* **123,** 424–435 (2014).
6. Li, P., Guo, M., Wang, C., Liu, X. & Zou, Q. An overview of SNP interactions in genome-wide association studies. *Briefings in Functional Genomics* **14,** 143–155 (2015).
7. Bowie, J., Luthy, R. & Eisenberg, D. A method to identify protein sequences that fold into a known three-dimensional structure. *Science* **253,** 164–170, doi: 10.1126/science.1853201 (1991).
8. Gao, M. & Skolnick, J. DBD-Hunter: a knowledge-based method for the prediction of DNA–protein interactions. *Nucleic Acids Research* **36,** 3978–3992, doi: 10.1093/nar/gkn332 (2008).
9. Ohlendorf, D. H., Anderson, W. F., Fisher, R. G., Takeda, Y. & Matthews, B. W. The molecular basis of DNA-protein recognition inferred from the structure of cro repressor. *Nature* **298,** 718–723 (1982).
10. Stawiski, E. W., Gregoret, L. M. & Mandel-Gutfreund, Y. Annotating Nucleic Acid-Binding Function Based on Protein Structure. *Journal of Molecular Biology* **326,** 1065–1079, doi: http://dx.doi.org/10.1016/S0022-2836(03)00031-7 (2003).
11. Liu, B. *et al.* PseDNA-Pro: DNA-Binding Protein Identification by Combining Chou's PseAAC and Physicochemical Distance Transformation. *Molecular Informatics* **34,** 8–17, (2015).
12. Wang, L. & Brown, S. J. BindN: a web-based tool for efficient prediction of DNA and RNA binding sites in amino acid sequences. *Nucleic Acids Research* **34,** W243–W248, doi: 10.1093/nar/gkl298 (2006).
13. Hwang, S., Gou, Z. & Kuznetsov, I. B. DP-Bind: a web server for sequence-based prediction of DNA-binding residues in DNA-binding proteins. *Bioinformatics* **23,** 634–636, doi: 10.1093/bioinformatics/btl672 (2007).
14. Ofran, Y., Mysore, V. & Rost, B. Prediction of DNA-binding residues from sequence. *Bioinformatics* **23,** i347–i353, doi: 10.1093/bioinformatics/btm174 (2007).
15. Wu, J. *et al.* Prediction of DNA-binding residues in proteins from amino acid sequences using a random forest model with a hybrid feature. *Bioinformatics* **25,** 30–35, doi: 10.1093/bioinformatics/btn583 (2009).
16. Kern, S. *et al.* Identification of p53 as a sequence-specific DNA-binding protein. *Science* **252,** 1708–1711, doi: 10.1126/science.2047879 (1991).
17. Cai, Y.-d. & Lin, S. L. Support vector machines for predicting rRNA-, RNA-, and DNA-binding proteins from amino acid sequence. *Biochimica et Biophysica Acta (BBA) - Proteins and Proteomics* **1648,** 127–133, doi: http://dx.doi.org/10.1016/S1570-9639(03)00112-2 (2003).

18. Lin, C. *et al.* Hierarchical Classification of Protein Folds Using a Novel Ensemble Classifier. *PLoS ONE* **8,** e56499 (2013).
19. Wei, L., Liao, M., Gao, X. & Zou, Q. An Improved Protein Structural Prediction Method by Incorporating Both Sequence and Structure Information. *IEEE Transactions on Nanobioscience* **14,** 339–349 (2015).
20. Liu, B. *et al.* 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,** e106691 (2014).
21. Liu, B. *et al.* Combining evolutionary information extracted from frequency profiles with sequence-based kernels for protein remote homology detection. *Bioinformatics* **30,** 472–479 (2014).
22. Andrea, S., Ondřej, K., Filip, Ž. & Jakub, T. Prediction of DNA-binding propensity of proteins by the ball-histogram method using automatic template search. *BMC Bioinformatics* **13,** S3 (2012).
23. Kumar, M., Gromiha, M. & Raghava, G. Identification of DNA-binding proteins using support vector machines and evolutionary profiles. *BMC Bioinformatics* **8,** 463 (2007).
24. Kumar, K. K., Pugalenthi, G. & Suganthan, P. N. DNA-Prot: Identification of DNA Binding Proteins from Protein Sequence Information using Random Forest. *Journal of Biomolecular Structure and Dynamics* **26,** 679–686, doi: 10.1080/07391102.2009.10507281 (2009).
25. Lin, W.-Z., Fang, J.-A., Xiao, X. & Chou, K.-C. iDNA-Prot: Identification of DNA Binding Proteins Using Random Forest with Grey Model. *PLoS ONE* **6,** e24756, doi: 10.1371/journal.pone.0024756 (2011).
26. Wei, L. *et al.* Improved and Promising Identification of Human MicroRNAs by Incorporating a High-quality Negative Set. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* **11,** 192–201 (2014).
27. Lou, W. *et al.* Sequence Based Prediction of DNA-Binding Proteins Based on Hybrid Feature Selection Using Random Forest and Gaussian Naïve Bayes. *PLoS ONE* **9,** e86703, doi: 10.1371/journal.pone.0086703 (2014).
28. Altschul, S. F. *et al.* Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research* **25,** 3389–3402, doi: 10.1093/nar/25.17.3389 (1997).
29. Zou, Q., Hu, Q., Guo, M. & Wang, G. HAlign: Fast Multiple Similar DNA/RNA Sequence Alignment Based on the Centre Star Strategy. *Bioinformatics*, doi: 10.1093/bioinformatics/btv177 (2015).
30. Song, L., Li, D., Zeng, X., Yunfeng Wu, L. G. & Zou, Q. nDNA-prot: Identification of DNA-binding Proteins Based on Unbalanced Classification. *BMC Bioinformatics* **15,** 298 (2014).
31. Berman, H. M. *et al.* The Protein Data Bank. *Nucleic Acids Research* **28,** 235–242 (2000).
32. Wang, C., Hu, L., Guo, M., Liu, X. & Zou, Q. imDC: an ensemble learning method for imbalanced classification with miRNA data. *Genetics and Molecular Research* **14,** 123–133 (2015).
33. Zhao, X., Zou, Q., Liu, B. & Liu, X. Exploratory predicting protein folding model with random forest and hybrid features. *Current Proteomics* **11,** 289–299 (2014).
34. Zou, Q., Li, J., Song, L., Zeng, X. & Wang, G. Similarity computation strategies in the microRNA-disease network: A Survey. *Briefings in Functional Genomics*, doi: 10.1093/bfgp/elv024 (2015).
35. Zeng, X., Zhang, X. & Zou, Q. Integrative approaches for predicting microRNA function and prioritizing disease-related microRNA using biological interaction networks. *Briefings in Bioinformatics*, doi: 10.1093/bib/bbv033 (2015).
36. Zou, Q. *et al.* Prediction of microRNA-disease associations based on social network analysis methods. *BioMed Research International* **2015,** 810514 (2015).
37. Shi, H., Wu, Y., Zeng, Z. & Zou, Q. A Discussion of MicroRNAs in Cancers. *Current Bioinformatics* **9,** 453–462 (2014).
38. Zou, Q., Li, J., Wang, C. & Zeng, X. Approaches for recognition disease genes based on Network. *BioMed Research International* **2014,** 416323 (2014).
39. Wang, Q. *et al.* Briefing in family characteristics of microRNAs and their applications in cancer research. *BBA–Proteins and Proteomics* **1844,** 191–197 (2014).
40. Wang, G. & Dunbrack, R. L. PISCES: recent improvements to a PDB sequence culling server. *Nucleic Acids Research* **33,** W94–W98, doi: 10.1093/nar/gki402 (2005).
41. Liu, B. *et al.* Pse-in-One: a web server for generating various modes of pseudo components of DNA, RNA, and protein sequences. *Nucleic Acids Research* **W1,** W65–W71 (2015).
42. Liu, B., Liu, F., Fang, L., Wang, X. & Chou, K.-C. repDNA: a Python package to generate various modes of feature vectors for DNA sequences by incorporating user-defined physicochemical properties and sequence-order effects. *Bioinformatics* **31,** 1307–1309 (2015).
43. Chou, K.-C. Prediction of protein cellular attributes using pseudo-amino acid composition. *Proteins: Structure, Function, and Bioinformatics* **43,** 246–255, doi: 10.1002/prot.1035 (2001).
44. Suykens, J. A. K. & Vandewalle, J. Least Squares Support Vector Machine Classifiers. *Neural Processing Letters* **9,** 293–300, doi: 10.1023/A:1018628609742 (1999).
45. Liu, B. *et al.* Identification of real microRNA precursors with a pseudo structure status composition approach. *PLoS ONE* **10,** e0121501 (2015).
46. Liu, B., Chen, J. & Wang, X. Application of Learning to Rank to protein remote homology detection. *Bioinformatics*, doi: 10.1093/bioinformatics/btv413 (2015).
47. Szilágyi, A. & Skolnick, J. Efficient Prediction of Nucleic Acid Binding Function from Low-resolution Protein Structures. *Journal of Molecular Biology* **358,** 922–933, doi: http://dx.doi.org/10.1016/j.jmb.2006.02.053 (2006).
48. Gao, M. & Skolnick, J. A Threading-Based Method for the Prediction of DNA-Binding Proteins with Application to the Human Genome. *PLoS Computational Biology* **5,** e1000567, doi: 10.1371/journal.pcbi.1000567 (2009).

Acknowledgements

This work was supported by the National Natural Science Foundation of China (No. 61300112 and 61272383), the Scientific Research Foundation for the Returned Overseas Chinese Scholars, State Education Ministry, the Natural Science Foundation of Guangdong Province (2014A030313695), Strategic Emerging Industry Development Special Funds of Shenzhen (JCYJ20140508161040764), and National High Technology Research and Development Program of China (863 Program) [2015AA015405].

Author Contributions

B.L. conceived of the study and designed the experiments, participated in designing the study, drafting the manuscript and performing the statistical analysis. S.Y.W. participated in coding the experiments and drafting the manuscript. X.L.W. participated in performing the statistical analysis. All authors read and approved the final manuscript.

Additional Information

Supplementary information accompanies this paper at http://www.nature.com/srep

Competing financial interests: The authors declare no competing financial interests.

How to cite this article: Liu, B. *et al.* DNA binding protein identification by combining pseudo amino acid composition and profile-based protein representation. *Sci. Rep.* **5**, 15479; doi: 10.1038/srep15479 (2015).