RESEARCH ARTICLE

Open Access

EL_PSSM-RT: DNA-binding residue prediction by integrating ensemble learning with PSSM Relation Transformation

Jiyun Zhou[1,2], Qin Lu[2], Ruifeng Xu[1,3*] (iD), Yulan He[4] and Hongpeng Wang[1]

Abstract

Background: Prediction of DNA-binding residue is important for understanding the protein-DNA recognition mechanism. Many computational methods have been proposed for the prediction, but most of them do not consider the relationships of evolutionary information between residues.

Results: In this paper, we first propose a novel residue encoding method, referred to as the Position Specific Score Matrix (PSSM) Relation Transformation (PSSM-RT), to encode residues by utilizing the relationships of evolutionary information between residues. PDNA-62 and PDNA-224 are used to evaluate PSSM-RT and two existing PSSM encoding methods by five-fold cross-validation. Performance evaluations indicate that PSSM-RT is more effective than previous methods. This validates the point that the relationship of evolutionary information between residues is indeed useful in DNA-binding residue prediction. An ensemble learning classifier (EL_PSSM-RT) is also proposed by combining ensemble learning model and PSSM-RT to better handle the imbalance between binding and non-binding residues in datasets. EL_PSSM-RT is evaluated by five-fold cross-validation using PDNA-62 and PDNA-224 as well as two independent datasets TS-72 and TS-61. Performance comparisons with existing predictors on the four datasets demonstrate that EL_PSSM-RT is the best-performing method among all the predicting methods with improvement between 0.02–0.07 for MCC, 4.18–21.47% for ST and 0.013–0.131 for AUC. Furthermore, we analyze the importance of the pair-relationships extracted by PSSM-RT and the results validates the usefulness of PSSM-RT for encoding DNA-binding residues.

Conclusions: We propose a novel prediction method for the prediction of DNA-binding residue with the inclusion of relationship of evolutionary information and ensemble learning. Performance evaluation shows that the relationship of evolutionary information between residues is indeed useful in DNA-binding residue prediction and ensemble learning can be used to address the data imbalance issue between binding and non-binding residues. A web service of EL_PSSM-RT (http://hlt.hitsz.edu.cn:8080/PSSM-RT_SVM/) is provided for free access to the biological research community.

Keywords: DNA-protein interaction, DNA-binding residue, PSSM, Ensemble learning, SVM, Random forest, Relation transformation

* Correspondence: xuruifeng@hit.edu.cn
[1]School of Computer Science and Technology, Harbin Institute of Technology Shenzhen Graduate School, HIT Campus Shenzhen University Town, Xili, Shenzhen, Guangdong 518055, China
[3]Shenzhen Engineering Laboratory of Performance Robots at Digital Stage, Shenzhen Graduate School, Harbin Institute of Technology, Shenzhen, China
Full list of author information is available at the end of the article

Zhou *et al. BMC Bioinformatics* (2017) 18:379

Page 2 of 16

Background

DNA-binding proteins play an important role in many essential biological processes such as DNA replication, recombination, repair, gene transcription and expression [1]. It has been reported that 2–3% of a prokaryotic genome and 6–7% of a eukaryotic genome encode DNA-binding proteins [2, 3]. As the interactions between proteins and DNAs are mainly formed by the immediate contacts [4], the identification of residues involved in the contacts is important for understanding the mechanism between them. Moreover, the identification of DNA-binding residues can also contribute to the understanding of the pathogenesis of diseases. Bullock and Fersht [5] have reported that mutations of some DNA-binding residues on proteins may be associated with some diseases. For example, the mutations of DNA-binding residues on the tumor repressor protein P53 may predispose individuals to cancer [5]. Many experimental techniques have been proposed to distinguish DNA-binding residues including electrophoretic mobility shift assays (EMSAs) [6, 7], nuclear magnetic resonance (NMR) spectroscopy [8], X-ray crystallography [9], peptide nucleic acid (PNA)-assisted identification of RNA binding proteins (RBPs) (PAIR) [10], MicroChIP [11], Fast ChIP [12], and conventional chromatin immunoprecipitation (ChIP) [13]. However, the experimental methods are very expensive and time-consuming. With the rapid accumulation of protein sequences, there is an urgent need to develop computational methods for the identification of DNA-binding residues.

For DNA-binding residue prediction, many computational methods have been proposed in recent years. The features used in these prediction methods include three types: sequence features, structure features and evolutionary features. In the early stage, the evolutionary features are not easy to get due to the limitation of computing power, so the predictors were developed mainly based on either structure information or sequence features, or a combination of them. For instance, the Support Vector Machine (SVM) classifier developed by Ahmad et al. [14] utilized only sequence features, such as the local amino acid composition and solvent accessible surface area. The classifier built by Tsuchiya et al. [15] used only structure features, such as electrostatic potential on the surface and the shape of the molecular surface. The DNA-binding residue classifier proposed by Bhardwaj [16] et al. used both sequence and structure information, such as solvent accessibility, local composition, net charge, and electrostatic potentials. The later SVM classifier proposed by Bhardwaj et al. [17] used structure features such as the net charge of a residue, occurrence in a cationic patch, and the average potential on a residue in addition to

the features used in their previous work [16]. The major limitation for the methods described above is that they did not use any evolutionary information which has been reported to be helpful for protein function prediction [18–20]. Thus, incorporating evolutionary information into the identification of DNA-binding residues may potentially improve its identification accuracy.

With the improvement of computing power, the use of evolutionary features becomes easier. Thus, more methods are now using evolutionary features for the prediction. Position Specific Score Matrix (PSSM) is a common representation of the evolutionary features and is used in the prediction methods in two ways: (1) combination methods that encode residues by combining evolutionary information and physiochemical properties and (2) concatenation methods that encode residues by concatenating the PSSM (Position Specific Score Matrix) scores in the sliding window. In combination methods, PSSM is combined with physiochemical properties to calculate the feature values for every residue. For instance, the encoding method proposed by Wang et al. [21] combined the BLAST-based conservation scores generated by sequence alignment and several biochemical properties to calculate the feature values for residues. The later encoding method proposed by his group [19] combined three physicochemical features including hydrophobicity, side chain pKa value and molecular mass and frequency profile to calculate the physicochemical feature values for the target residue and its context residues. The mean and the standard deviation of the three physicochemical features are used to construct the feature space. The encoding method proposed by Ma et al. [22] combined PSSM and four physicochemical properties including the lone electron pairs, hydrophobicity, side chain pKa value and molecular mass.

Concatenation methods usually concatenate the PSSM scores of all the residues in the sliding window to encode residues. For instance, Ahmad and Sarai's work [20] concatenated all the PSSM scores of residues within the sliding window of the target residue to construct the feature vector. Then the concatenation method proposed by Ahmad and Sarai [20] were used by many classifiers. For example, the SVM classifier proposed by Kuznetsov et al. [23] was developed by combining the concatenation method, sequence features and structure features. The predictor, called SVM-PSSM, proposed by Ho et al. [24] was developed by the concatenation method. The SVM classifier proposed by Ofran et al. [1] was developed by integrating the concatenation method and sequence features including predicted solvent accessibility, and predicted secondary structure.

Zhou *et al. BMC Bioinformatics* (2017) 18:379

Page 3 of 16

It should be noted that both current combination methods and concatenation methods did not include the relationships of evolutionary information between residues. However, many works on protein function and structure prediction have already shown that the relationships of evolutionary information between residues are important [25, 26], we propose a method to include the relationship of evolutionary information as features for the prediction of DNA-binding residue. The novel encoding method, referred to as the PSSM Relationship Transformation (PSSM-RT), encodes residues by incorporating the relationships of evolutionary information between residues. In addition to evolutionary information, sequence features, physicochemical features and structure features are also important for the prediction. However, as the structure features for most of the proteins are unavailable, we do not include structure feature in this work. In this paper, we include PSSM-RT, sequence features and physicochemical features to encode residues. Additionally, for DNA-binding residue prediction, there are much more non-binding residues than binding residues in protein sequences. However, most of the previous methods cannot take advantages of the abundant number of non-binding residues for the prediction. In this work, we propose an ensemble learning model by combining SVM and Random Forest to make good use of the abundant number of non-binding residues. By combining PSSM-RT, sequence features and physicochemical features with the ensemble learning model, we develop a new classifier for DNA-binding residue prediction, referred to as EL_PSSM-RT. A web service of EL_PSSM-RT (http://hlt.hitsz.edu.cn:8080/PSSM-RT_SVM/) is made available for free access by the biological research community.

Methods

As shown by many recently published works [27–30], a complete prediction model in bioinformatics should contain the following five components: validation benchmark dataset(s), an effective feature extraction procedure, an efficient predicting algorithm, a set of fair evaluation criteria and a web service to make the developed predictor publicly accessible. In the following text, we will describe the five components of our proposed EL_PSSM-RT in details.

Datasets

In order to evaluate the prediction performance of EL_PSSM-RT for DNA-binding residue prediction and to compare it with other existing state-of-the-art prediction classifiers, we use two benchmarking datasets and two independent datasets.

The first benchmarking dataset, PDNA-62, was constructed by Ahmad et al. [14] and contains 67 proteins from the Protein Data Bank (PDB) [31]. The similarity between any two proteins in PDNA-62 is less than 25%. The second benchmarking dataset, PDNA-224, is a recently developed dataset for DNA-binding residue prediction [32], which contains 224 protein sequences. The 224 protein sequences are extracted from 224 protein-DNA complexes retrieved from PDB [31] by using the cut-off pair-wise sequence similarity of 25%. The evaluations on these two benchmarking datasets are conducted by five-fold cross-validation. To compare with other methods that were not evaluated on the above two datasets, two independent test datasets are used to evaluate the prediction accuracy of EL_PSSM-RT. The first independent dataset, TS-72, contains 72 protein chains from 60 protein-DNA complexes which were selected from the DBP-337 dataset. DBP-337 was recently proposed by Ma et al. [33] and contains 337 proteins from PDB [31]. The sequence identity between any two chains in DBP-337 is less than 25%. The remaining 265 protein chains in DBP-337, referred to as TR265, are used as the training dataset for the testing on TS-72. The second independent dataset, TS-61, is a novel independent dataset with 61 sequences constructed in this paper by applying a two-step procedure: (1) retrieving protein-DNA complexes from PDB [31]; (2) screening the sequences with cut-off pair-wise sequence similarity of 25% and removing the sequences having > 25% sequence similarity with the sequences in PDNA-62, PDNA-224 and TS-72 using CD-HIT [34]. CD-HIT is a local alignment method and short word filter [35, 36] is used to cluster sequences. In CD-HIT, the clustering sequence identity threshold and word length are set as 0.25 and 2, respectively. By using the short word requirement, CD-HIT skips most pair-wise alignments because it knows that the similarity of two sequences is below certain threshold by simple word counting. For the testing on TS-61, PDNA-62 is used as the training dataset. The PDB id and the chain id of the protein sequences in these four datasets are listed in the part A, B, C, D of the Additional file 1, respectively.

In the above 4 datasets, positive and negative samples are defined by the following criterion [18, 37, 38]: a residue in a protein is regarded as a binding residue if the side chain or the backbone atoms of the residue falls within a cutoff distance of 3.5 Å from any atom of the partner DNA molecule in the complex; Otherwise, the residue is considered a non-binding residue. The number of positive samples and negative samples of the four datasets are shown in Table 1.

Evaluation metrics

In order to evaluate the performance of EL_PSSM-RT for DNA-binding residue prediction, Sensitivity (SN), Specificity (SP), Strength (ST), Accuracy (ACC), and

Zhou *et al. BMC Bioinformatics* (2017) 18:379

Page 4 of 16

Table 1 Number of the positive samples and negative samples of the four datasets

Dataset	PDNA-62	PDNA-224	TS-72	TS-61
Binding residue	1215	3778	1040	1078
Non-binding residue	6948	53,570	13,226	13,175

Mathews Correlation Coefficient (MCC) are used as performance metrics. They are typical evaluation metrics in bioinformatics and have been widely used by many works. The five metrics can be calculated according to the following formula

$$SN = TP/(TP + FN) \qquad (1)$$

$$SP = TN/(TN + FP) \qquad (2)$$

$$ST = (SN + SP)/2 \qquad (3)$$

$$ACC = (TP + TN)/(TP + FP + TN + FN) \qquad (4)$$

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

where TP is the number of true positives, TN is the number of true negatives, FP is the number of false positives, and FN is the number of false negatives.

Since all the four datasets have much more negative training examples than positive training examples, using ACC alone may produce biased results, for example simply classifying all the test samples as non-binding residues will give a very high ACC value. Many literatures have indicated that ST, the average of SN and SP, can give a more appropriate evaluation for a classifier when the numbers of positive and negative samples are unbalanced [14, 38, 39]. Additionally, since MCC can measure the matching degree between the predicted results and the real results, it is also an appropriate evaluation metric. Moreover, the Receiver Operating Characteristic (ROC) curve [40] and the area under ROC curve (AUC) [41] are two more commonly used metrics for performance evaluation on imbalanced data sets. The ROC curve is drawn by plotting the true positive rates (i.e. sensitivity) against the false positive rates (i.e. 1-specificity) calculated by changing the classification threshold for predictors. AUC is the area under the ROC curve with values limited to the closed interval between −1.0 and 1.0. An AUC of 1.0 and 0.5 indicate the best performance and a random performance, respectively. Therefore, ST, MCC, AUC and ROC are used as the main performance measures and the other three metrics are used for references only.

Sequence context

In DNA-binding residue prediction, residues are the samples for training and testing [16, 42]. Apart from a target residue, its adjacent residues also have a significant impact on its function. So, the sequence context of the target residue needs to be considered in the prediction. In order to use the sequence context for prediction, we define residue-wise data instances by a sliding window with size w. The sliding window is a sequence fragment with the target residue positioned in the middle and $(w\text{-}1)/2$ neighboring residues on either side. All the residues in the sliding window except the target residue are considered as the sequence context. The $(w\text{-}1)/2$ neighboring residues on the left side and the right side are referred to as the left sequence context and the right sequence context, respectively. The length of the sliding window, w, should be an odd number to be set experimentally.

Given a protein sequence P of length L denoted as

$$P = R_1R_2R_3R_4R_5R_6\cdots R_{i-1}R_iR_{i+1}\cdots R_L, \qquad (6)$$

where R_1 represents the first residue of the protein sequence P, R_2 represents the second residue and so forth. The residue-wise instance F_i for target residue R_i can be represented as

$$F_i = R_{i-\frac{w-1}{2}}R_{i-\frac{w-3}{2}}\cdots R_{i-1}R_iR_{i+1}\cdots R_{i+\frac{w-3}{2}}R_{i+\frac{w-1}{2}}, \qquad (7)$$

where all the residues in the residue-wise instance F_i except the target residue R_i define its sequence context.

Features of data instance

Evolutionary information is produced by the evolutionary processes and it is important for protein structure and function prediction. PSSM is a common representation for evolutionary information and has been used in many bioinformatics studies including protein functionality annotation and protein structure prediction [43–47]. For every protein sequence in this study, its PSSM is calculated from multiple sequence alignments produced by running the PSI-BLAST program [48] to search the non-redundant (NR) database through three iterations with the E-value cutoff at 0.001. For a protein with length L, PSSM is usually represented as a matrix with $L \times 20$ dimensions. 20 denote the 20 standard types of residues. For the sequence fragment F_i using representation defined in Formula (7), its PSSM can be represented as a matrix with dimensions $w \times 20$. Thus, the PSSM of the residue-wise instance F_i for the target residue R_i can be formulated as

Zhou *et al. BMC Bioinformatics* (2017) 18:379

Page 5 of 16

$$\text{PSSM}_{\text{F}_i} = \begin{bmatrix} S_{i-\frac{w-1}{2},1} & \cdots & S_{i-\frac{w-1}{2},r} & \cdots & S_{i-\frac{w-1}{2},20} \\ \vdots & \cdots & \vdots & \cdots & \vdots \\ S_{i,1} & \cdots & S_{i,r} & \cdots & S_{i,20} \\ \vdots & \cdots & \vdots & \cdots & \vdots \\ S_{i+\frac{w-1}{2},1} & \cdots & S_{i+\frac{w-1}{2},r} & \cdots & S_{i+\frac{w-1}{2},20} \end{bmatrix},$$

(8)

where $S_{i,r}$ is the conservative score of residue type r at position i in the sequence fragment.

Before PSSM-RT is calculated, the conservative scores in PSSM should be normalized between 0 and 1. Thus, for a given $S_{i,r}$, its normalized value $S_{i,r}^{(\text{N})}$ can be expressed by a logistic function given below

$$S_{i,r}^{(\text{N})} = \frac{1}{1 + e^{-S_{i,r}}},$$

(9)

PSSM-RT contains three categories of features: residue conservations, pair-relationships and multi-relationships. The residue conservations contain the PSSM scores of the target residue and its context residues. The pair-relationship is defined as the relationship of evolutionary information between two positions, for example, the pair-relationship between the residue r_1 of position i and the residue r_2 of position j is calculated as

$$\text{PSSM-RT}(i, j, r_1, r_2) = S_{i,r_2}^{(\text{N})} * S_{j,r_1}^{(\text{N})},$$

(10)

As every position in a residue-wise data instance has conservative scores for the 20 standard residue type, 400 types of relationships can be calculated for any two positions.

As the target position in a residue-wise data instance is influenced by all its context positions, the all pair-relationships between the target position and its context positions needs to be included in the prediction. Thus the pair-relationship for a residue-wise data instance is defined as the sum of pair-relationship between the target position and all its context positions. For example, the pair-relationship between residue r_1 and residue r_2 for a residue-wise data instance with i as its target position is formulated as

$$\text{PSSM-RT}(i, r_1, r_2) = \sum_j \text{PSSM-RT}(i, j, r_1, r_2), \quad (11)$$

where j is the context position of the target position.

Multi-relationships are the evolutionary information relationships between multiple residues. We consider two kinds of multi-relationships: the left multi-relationships that include the relationships between the target residue and its left context residues and the right multi-relationships that include the relationships between the target residue and its right context residues. For residue r, the left multi-relationship of residue-wise data instance at target position i is formulated as

$$\text{PSSM-RT}_{left}(i, r) = \sum_{k=i-\frac{w-1}{2}}^{i} S_{k,r}^{(\text{N})}. \quad (12)$$

For residue r, the right multi-relationship of residue-wise data instance at target position i is formulated as

$$\text{PSSM-RT}_{right}(i, r) = \sum_{k=i}^{i+\frac{w-1}{2}} S_{k,r}^{(\text{N})}. \quad (13)$$

Thus, the dimension of the feature space constructed by PSSM-RT is (20*w + 20*20 + 2*20).

In addition to PSSM-RT, there are two other types of features that are used in this work: sequence features and physiochemical features. Sequence features given in the datasets include amino acid composition, predicted secondary structure, predicted solvent accessible area, and identity of the target residue. Physiochemical features include pKa values of amino group, pKa values of carboxyl group, electron-ion interaction potential (EIIP) [49], number of lone electron pairs(LEPs), Wiener index [50], molecular mass [50], side chain pKa value, and hydrophobicity index. The predicted secondary structure and predicted solvent accessible area are obtained by applying PSIPRED [51] and SABLE [52–54], respectively.

Ensemble learning

Ensemble learning is now an active area of research in machine learning and pattern recognition. Ensemble learning first learns several base predictors from the training dataset and then combines them into an ensemble predictor. Ensemble learning aims to take advantage of the different learning ability of the different base predictors. There are three widely used ensemble strategies to train base predictors: training by different data subsets, training from different feature subsets and training by different classification algorithms.

In DNA-binding residue prediction, non-binding residues outnumber binding residues by a large margin. In order to get a balanced dataset for training, many predictors chose to discard a large part of non-binding residues [33]. However, discarded non-binding residues may potentially be useful information to improve prediction performance. In order to better use all the data available, we propose to use ensemble learning by combining all the three ensemble strategies. And then use our proposed method, referred to as EL_PSSM-RT, to combine

the ensemble learning model with PSSM-RT. The system architecture of EL_PSSM-RT is shown in Fig. 1. Note that EL_PSSM-RT contains 4 steps: Dataset Partition, Feature Extraction, Base Classifier Training and Base Classifier Selection. In Step 1 of Dataset Partition, the non-binding residues in the training dataset are first partitioned into n non-overlapping subsets with the number of samples approximately equal to that of all the binding residues. Then, n new balanced training datasets are formed by adding the binding residues into the n subsets non-binding residues. In Step 2 of Feature Extraction, three categories of features are extracted for residues including sequence features, physiochemical features, and evolutionary information extracted by PSSM-RT. In Step 3 of Base Classifier Training, both the SVM classifier and the Random Forest classifier are used by each category of features on every newly formed training dataset. SVM and Random Forest are used because they are proven to have good predicting performances for DNA-binding residue prediction [18, 19, 55]. Thus, $6*n$ ($2*3*n$) base predictors are trained in this step. In Step 4 of Base Classifier Selection, a diversity based dynamic ranking and selecting method is designed based on diversity to build the ensemble predictor using an iterative approach. In our dynamic ranking and selecting method, a base predictor is initially selected at random. Then in each iteration, all the unselected base predictors are first ranked based on their diversity with the selected base predictor(s), followed by the selection step in which the one with the largest diversity will be added into the set of selected predictors. Diversity between two base classifiers is measured by the proportion of the number of samples with different labels from the two classifiers to the total number of samples in validation dataset. The iteration is terminated when the addition of diversity for the set of selected predictors is less than a specified criterion. The exact stopping criterion for a dataset is determined by a validation dataset which is separated from the dataset of interest. Finally, the selected base predictors are combined to construct an ensemble predictor using a simple majority vote strategy.

Results and discussion

The purpose of the evaluation is to examine the effectiveness of our proposed PSSM-RT over other methods. Four
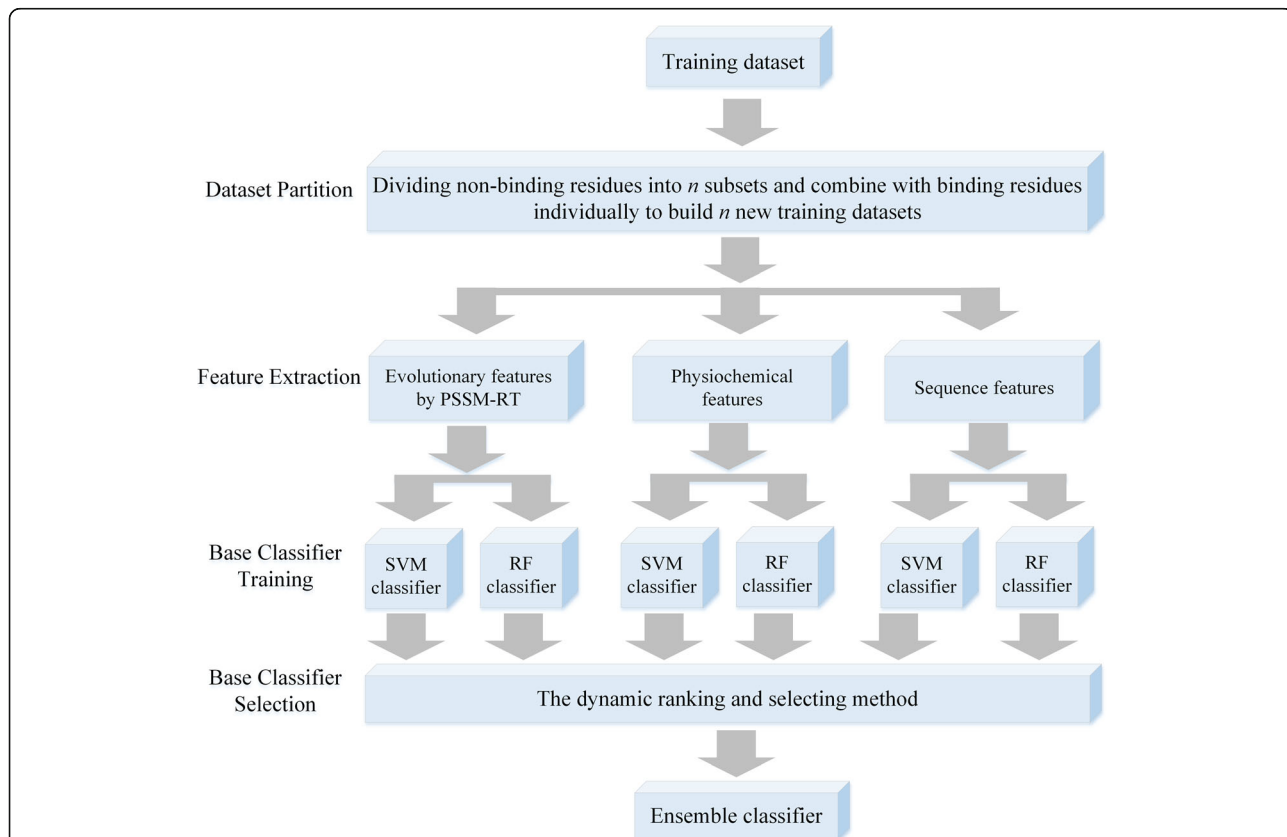

Fig. 1 The framework diagram of EL_PSSM-RT. EL_PSSM-RT contains 4 steps. The first step is to divide the non-binding residues in the training dataset into n subsets and to construct n new training datasets by combining the n subsets of non-binding residues and binding residues individually. The secondary step is to extract the three categories of features for all the residues. The third step is to train both SVM classifier and Random Forest classifier by each category of features on every training subset. The fourth step is to use a dynamic ranking and selecting method to select the based predictors with the largest diversity between each other to build the ensemble predictor

Zhou *et al. BMC Bioinformatics* (2017) 18:379

Page 7 of 16

sets of evaluations are conducted here. Experiment 1 compares PSSM-RT with previous encoding methods. Experiment 2 compares the ensemble learning model with the base classifiers. Experiment 3 compares our proposed predictor EL_PSSM-RT with previous predictors, and Experiment 4 evaluates EL_PSSM-RT on two independent datasets. Based on the obtained data, we further analyse the relation-pairs of amino acids followed a case study of two proteins in the binding-residues identified by our method. In order to assess the significance of statistic comparison between pairs of methods, we calculate the *p*-values of statistic comparisons by the Wilcoxon signed-rank test. As AUC is the most appropriate metric for performance evaluation, for Experiment 1 and Experiment 2, we use AUC to assessed the significance. Because in Experiment 3 and Experiment 4, the AUC values for some methods are not available, so we use ST to calculate the significance. Note that as the evaluation on TS-72 in Experiment 4 have AUC values for models, we use AUC to assessed the significance.

Window size of PSSM-RT
Since PSSM-RT uses a window based approach, the window size needs to be set properly. For the SVM classifier which uses PSSM-RT as features, the performance of the SVM classifier with different window size w is shown in Fig. 2. It can be seen that the ST value continues to increase and peaks when w reaches 13. So, the window size $w = 13$ is used for all the SVM classifiers.

Experiment 1: Comparison of PSSM-RT with previous encoding methods
This set of experiments first compares PSSM-RT to PSSM by using SVM and Random Forest (RF). PSSM-RT is then compared with two types of existing encoding methods: the combination methods and the concatenation methods. The comparison between PSSM-RT and PSSM by using SVM and RF is shown in Table 2. The top performers of the three major performance indicators ST, MCC and AUC are shaded for easy observation. This table shows that PSSM-RT outperforms PSSM significantly with *p*-value of at least less than 2.33e-6 for both SVM and RF on PDNA-62. It also shows that PSSM-RT outperform PSSM significantly on PDNA-224 with *p*-value less than 7.69e-5 for both SVM and RF. As there are a number of combination methods and concatenation methods, we only consider the state-of-the-art works for the respective groups. Consequently, Ma et al.'s work using combination method [56] and Li et al.'s work [32] using the concatenation methods are used for comparison. In Ma et al.'s work, it used PSSM with four physicochemical properties including the lone electron pairs, hydrophobicity, side chain pKa value and molecular mass are combined to calculate the feature representation for residues. In Li et al.'s work, the PSSM scores of residues within the sliding window of the target residue are concatenated to construct the feature vector. So PSSM-RT, Ma et al.'s work and Li et al.'s work all use the same set of features except that Ma et al.'s work uses additional physicochemical features. Both Ma et al.'s work [56] and Li et al.'s work [32] used SVM as the classifier, so we also use SVM as the classifier in this experiment. Note that all the SVM classifiers in this paper used the radial kernel and the parameters of all the SVM classifiers were tuned by the grid method. Since both Ma et al.'s work and Li et al.'s work did not provide the performance for evolutionary features and combination with sequence features on PDNA-62
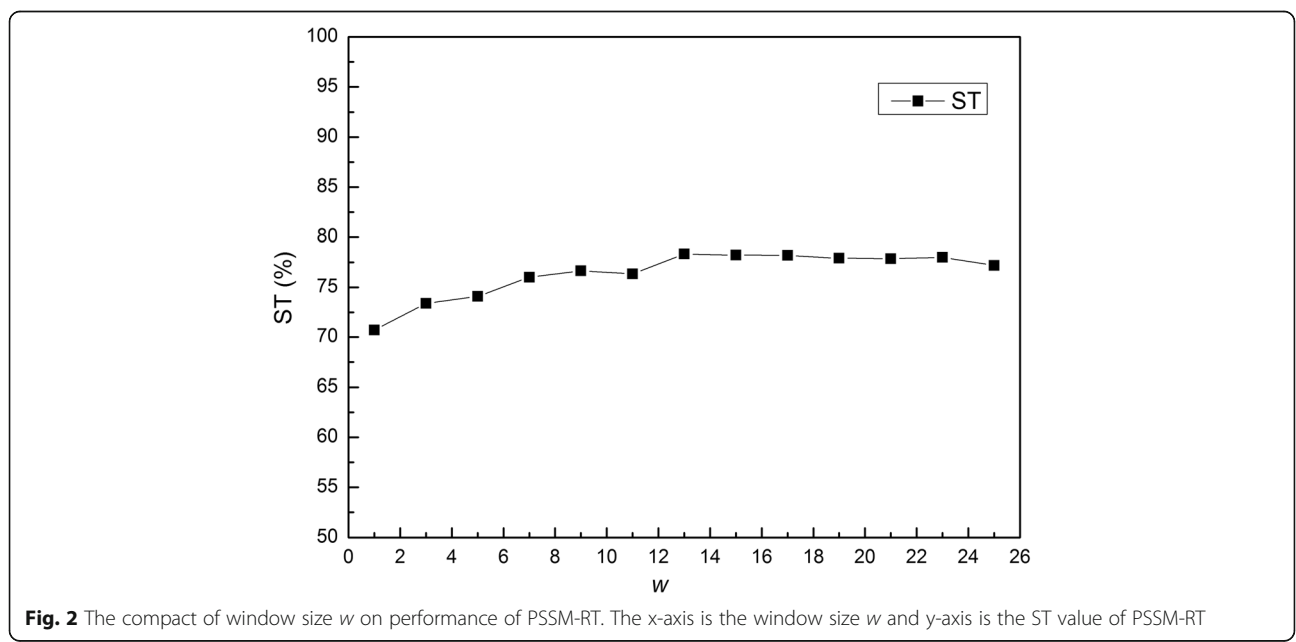

Fig. 2 The compact of window size *w* on performance of PSSM-RT. The x-axis is the window size *w* and y-axis is the ST value of PSSM-RT

Zhou *et al. BMC Bioinformatics* (2017) 18:379

Page 8 of 16

Table 2 Performance comparison between PSSM-RT and PSSM by multiple classification algorithms on benchmark datasets

Dataset	Methods	ACC (%)	MCC	SN (%)	SP (%)	ST (%)	AUC
PDNA-62	SVM(PSSM)	77.51	0.39	77.66	78.55	75.11	0.821
	RF(PSSM)	77.32	0.40	72.00	78.27	75.14	0.821
	SVM(PSSM-RT)	76.45	0.43	80.23	75.80	78.02	0.845
	RF(PSSM-RT)	78.46	0.44	77.53	78.66	78.10	0.855
PDNA-224	SVM(PSSM)	79.18	0.29	67.21	80.09	73.65	0.813
	RF(PSSM)	80.45	0.29	65.24	81.59	73.42	0.813
	SVM(PSSM-RT)	80.39	0.31	68.11	81.32	74.72	0.826
	RF(PSSM-RT)	80.07	0.30	67.63	81.01	74.32	0.825

and PDNA-224, their methods are implemented in this study to obtain evaluation data. The performances on both datasets PDNA-62 and PDNA-224 are shown in Table 3. The corresponding ROC curves are shown in Fig. 3(a) and (b), respectively.

From Table 3, we can see that PSSM-RT performs better than Ma et al.'s work on both datasets with *p*-values less than 3.05e-5, which means the improvement is quite significant. More specifically, the increase in the PDNA-62 dataset is 0.17 on MCC, 11.06% on ST and 0.111 on AUC and 0.13 on MCC, 9.99% on ST and 0.103 on AUC for the PDNA-224 dataset. PSSM-RT outperforms Li et al.'s work quite significantly on both datasets with *p*-value less than 4.71e-5. More specifically, the increase in the PDNA-62 dataset is 0.03 on MCC, 2.88% on ST and 0.024 on AUC and 0.02 on MCC, 1.07% on ST and 0.013 on AUC on the PDNA-224 dataset. Figure 3(a) and (b) show that PSSM-RT has the best ROC curve on both PDNA-62 and PDNA-224.

When both sequence features and physiochemical features are added, the performances of the three methods on PDNA-62 and PDNA-224 are shown in Table 4. The corresponding ROC curves are shown in Fig. 4(a) and (b).

Table 4 shows the same performance trends as that in Table 3. Figure 4(a) and (b) also show that PSSM-RT has the best ROC curve on both PDNA-62 and PDNA-224 when the three types of features are combined. This clearly indicates that PSSM-RT outperforms both Ma et al.'s work and Li et al.'s work when all three types of features are used. When comparing Tables 3 and 4, we observe that the performance of PSSM-RT is improved by 0.05 on MCC, 1.52% on ST and 0.028 on AUC for

Table 3 Performance for evolutionary features on benchmark datasets by SVM

Dataset	Methods	ACC (%)	MCC	SN (%)	SP (%)	ST (%)	AUC
PDNA-62	Ma et al.	72.23	0.26	59.45	74.48	66.96	0.734
	Li et al.	77.32	0.40	72.00	78.27	75.14	0.821
	PSSM-RT	76.45	0.43	80.23	75.80	78.02	0.845
PDNA-224	Ma et al.	76.88	0.18	50.59	78.87	64.73	0.723
	Li et al.	79.18	0.29	67.21	80.09	73.65	0.813
	PSSM-RT	80.39	0.31	68.11	81.32	74.72	0.826

PDNA-62 and 1.92% on ST and 0.017 on AUC for PDNA-224. This shows that PSSM-RT is complementary to the other two features. This set of experiments indicates that the relationships of evolutionary information between residues perform better than the two previous state-of-the-art encoding methods.

Experiment 2: Comparison of EL_PSSM-RT with base classifiers

This set of experiments compares EL_PSSM-RT with the base classifiers. The performances of EL_PSSM-RT, the SVM classifier and the Random Forest (RF) classifier are shown in Table 5, where the performances for the SVM classifier and the RF classifier are their best performances, respectively. Note that EL_PSSM-RT, the SVM classifier and the RF classifier shown in Table 5 are trained from the same set of features. The corresponding ROC curves are shown in Fig. 5(a) and (b).

Table 5 shows that compares to both the SVM classifier and the RF classifier, EL_PSSM-RT achieves significant performance improvement on both PDNA-62 with *p*-value less than 6.52e-5 and PDNA-224 with *p*-value less than 7.25e-5. More specifically, on the PDNA-62 dataset, the increase to the SVM classifier is 0.03 on MCC, 3.03% on ST and 0.028 on AUC and 0.04 on ACC, 3.14% on ST and 0.021 on AUC to the RF classifier. For the PDNA-224 dataset, the increase to the SVM classifier is 0.03 on MCC, 2.13% on ST and 0.022 on AUC and to the RF classifier is 0.02 on MCC, 2.38% on ST and 0.021 on AUC. Figure 5(a) and (b) also show that EL_PSSM-RT obtains the best ROC curve on both PDNA-62 and PDNA-224. This indicates that ensemble learning makes EL_PSSM-RT more superior than both the SVM classifier and the RF classifier. Furthermore, Table 5 shows that the performance of the RF classifier is good. As RF can examine the learning model and quantify the importance of features used to train the classifier, it helps us to analyze the importance of different features. In this paper, among the 200 top features with the largest importance, we counted the number of features with respect to different categories. The analysis shows that 12 features come from PSSM scores, 134 features

Zhou *et al. BMC Bioinformatics* (2017) 18:379
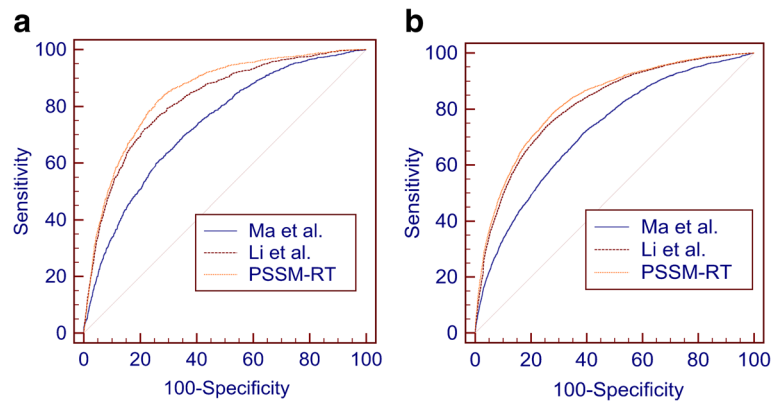
Page 9 of 16

Fig. 3 Comparison between different encoding methods. **a** The ROC curves of PSSM-RT, Ma et al.'s method and Li et al.'s method on PDNA-62. **b** The ROC curves of PSSM-RT, Ma et al.'s method and Li et al.'s method on PDNA-224

come from pair-relationships in PSSM-RT and 26 feature come from multi-relationships in PSSM-RT and 30 come from other features including sequence features and physiochemical features. It validates the importance of pair-relationships and multi-relationships in PSSM-RT for prediction of DNA-binding residues.

For proteins which share less than 25% identical residues, they can may still belong to the same homologous family and have very similar 3D structures, referred to as homologous proteins. Thus, predictors trained on datasets compiled using only the 25% identity threshold can be over-fitted towards large over-represented homologous families. In order to evaluate the influence of homologous proteins on predictors, we construct two novel benchmark datasets based on PDNA-62 and PDNA-224 by keeping only one protein for every homologous families. The novel PDNA-62 contains 35 sequences and the novel PDNA-224 contains 163 sequences. This means that the original PDNA-62 contains approximate 50% homologous proteins and PDNA-224 contains 25% homologous proteins. The results of EL_PSSM-RT, the SVM classifier and the RF classifier on the novel benchmark datasets are shown in Table 6.

Compared to the results on the original benchmark datasets shown in Table 5, the performances of all

classifiers are decreased by at least 1.46% ST and 0.012 AUC on novel PDNA-62. On novel PDNA-224, ST and AUC of EL_PSSM-RT and the SVM classifier do not have obvious decrease. But, MCC decrease by at least 0.02. For the RF classifier, ST, AUC and MCC are decreased by 0.05, 1.30% and 0.013, respectively. It indicates that over-representation of some homologous families indeed bias the performance towards those families which leads to biased evaluation. By comparing the performance decreases on PDNA-62 with 50% structure redundant proteins and that on PDNA-224 with 25% redundant proteins, we found that a higher rate of structure redundant proteins leads to more biased evaluation. However, in order to make a fair comparison with state of art methods, we still use the original PDNA-62 and PDNA-224 as datasets for evaluation in the following text.

Experiment 3: Comparison with previous predictors

This set of experiments evaluates the performance of our proposed ensemble learning based EL-PSSM-RT compared to other state-of-the art methods trained and tested either on PDNA-62 or PDNA-224 including eight algorithms: (1) Dps-pred [14], (2) Dbs-pssm [20], (3) BindN [18], (4) Dp-bind [23], (5) DP-Bind [57], (6) BindN-RF [55], (7) BindN+ [19], and (8) PreDNA [32]. The first seven methods were trained and tested on PDNA-62. The last method, PreDNA [32], was trained and tested on both datasets. PreDNA was proposed recently and achieved the best performance for DNA-binding residue prediction so far. In addition to sequence features and evolutionary information, PreDNA [32] also used structure features. As we have pointed out earlier, structure features of most proteins are unavailable and the experimental 3D structure is very expensive to obtain. Thus, PreDNA [32] cannot be used as a general method at the current time for DNA-binding residue prediction on a genomic scale. For this reason,

Table 4 Performance for all features on benchmark datasets by SVM

Dataset	Methods	ACC (%)	MCC	SN (%)	SP (%)	ST (%)	AUC
PDNA-62	Ma et al.	75.11	0.40	78.22	74.58	76.40	0.837
	Li et al.	77.81	0.42	75.50	78.24	76.87	0.851
	PSSM-RT	81.50	0.48	76.74	82.34	79.54	0.873
PDNA-224	Ma et al.	76.66	0.27	68.95	77.25	73.10	0.808
	Li et al.	78.65	0.29	69.48	79.34	74.41	0.825
	PSSM-RT	78.14	0.31	74.92	78.38	76.65	0.843

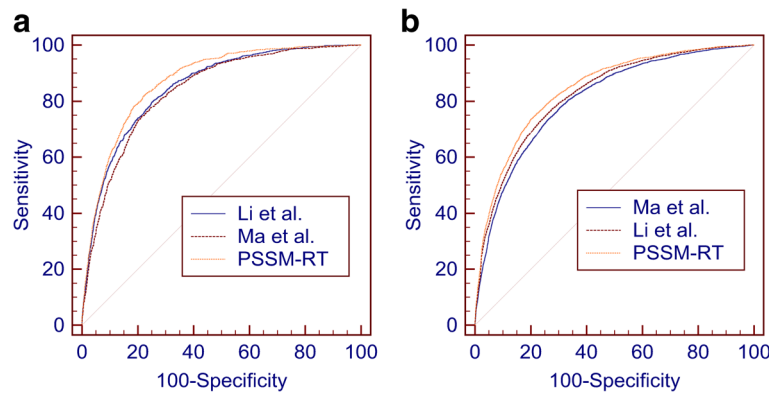Zhou *et al. BMC Bioinformatics* (2017) 18:379

Page 10 of 16

Fig. 4 Comparison between different encoding methods when combining with sequence features and physicochemical features. **a** The ROC curves of PSSM-RT, Ma et al.'s method and Li et al.'s method on PDNA-62. **b** The ROC curves of PSSM-RT, Ma et al.'s method and Li et al.'s method on PDNA-224

EL_PSSM-RT does not use any structure feature, similar to many other methods. In order to fairly compare the prediction performance of various methods, the version of PreDNA without using structure features is used in this evaluation. The prediction accuracy of EL_PSSM-RT and other methods on PDNA-62 and PDNA-224 are shown in Tables 7 and 8, respectively.

Table 7 shows that EL_PSSM-RT achieves the best performance with significant improvement with p-value less than 3.06e-5 for PDNA-62, outperforming other methods by 0.02–0.07 on MCC, 4.27%–21.47% on ST and 0.040–0.149 on AUC. Table 8 shows that, for the PDNA-224 dataset, EL_PSSM-RT performs better than PreDNA by 0.05 on MCC, 4.18% on ST with p-value less than 3.64e-5. The results on both datasets indicate that the effect use of relation information and ensemble learning is superior to other existing methods.

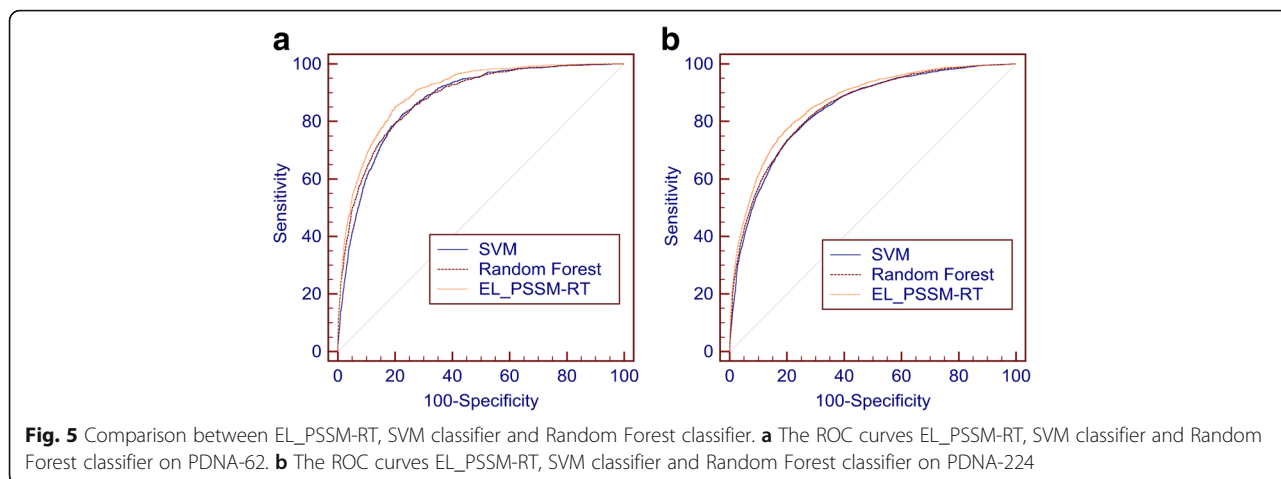### Experiment 4: Independent tests use TS-72 and TS-61

We evaluate the performance of our EL-PSSM-RT on the TS-72 dataset so we can compare it with the previous published DNABR method [33] and the BindN series [18, 19, 55]. DNABR is a sequence based DNA-binding residue prediction method proposed by Ma et al. [33].

BindN, BindN-RF and BindN+ are three methods proposed by Wang et al. using only sequence information [18, 19, 55]. the AUC values of the four published methods are 0.866, 0.748, 0.825 and 0.844, respectively according to Ma et al.' work [33] which are trained on TR265. The AUC value for EL_PSSM-RT, is 0.879. Our method increases the performance by 0.013–0.131 on AUC with p-value less than 8.37e-4 for the independent dataset TS-72.

For the second independent dataset TS-61, we compare our proposed method with DP-Bind[57]. DP-Bind [57] is a web server for predicting DNA-binding sites in a DNA-binding protein from its amino acid sequence. The web server implements three individual machine learning classifiers: DP-Bind(SVM) that uses support vector machine, DP-Bind(KLR) that use kernel logistic regression and DP-Bind(PLR) that uses penalized logistic regression. DP-Bind [57] also implements two types of consensus classifiers. One is majority consensus on the results of three machine learning methods by majority vote, referred to as DP-Bind(MAJ). The other is strict consensus obtained by unanimous agreement, referred to as DP-Bind(STR). The performance of EL_PSSM-RT trained by PDNA-62 and the different DP-Bind methods is shown in Table 9. Table 9 shows that our method outperforms DP-Bind(SVM), DP-Bind(KLR), DP-Bind(PLR) and DP-Bind(MAJ) with 0.05–0.06 on MCC, 3.78–4.43% on ST and 0.027–0.049 on AUC. By comparing to DP-Bind(STR), we found that our method only outperform it marginally. Note that DP-Bind(STR) is based on the assumption that if DP-Bind(SVM), DP-Bind(KLR) and DP-Bind(PLR) assign the same label to a given residue. So it can provide more correct prediction results for residues. The results shown in Table 8 also demonstrates that DP-Bind(STR) achieves better performance than other four models.

Table 5 Comparison of EL_PSSM-RT with base classifiers on benchmark datasets

Dataset	Methods	ACC (%)	MCC	SN (%)	SP (%)	ST (%)	AUC
PDNA-62	SVM	81.50	0.48	76.74	82.34	79.54	0.873
	RF	80.90	0.47	77.43	81.42	79.43	0.880
	EL_PSSM-RT	80.82	0.51	85.04	80.10	82.57	0.901
PDNA-224	SVM	78.14	0.31	74.92	78.38	76.65	0.843
	RF	80.95	0.32	71.11	81.69	76.40	0.844
	EL_PSSM-RT	78.09	0.34	79.58	77.98	78.78	0.865

Zhou *et al. BMC Bioinformatics* (2017) 18:379

Page 11 of 16

Fig. 5 Comparison between EL_PSSM-RT, SVM classifier and Random Forest classifier. **a** The ROC curves EL_PSSM-RT, SVM classifier and Random Forest classifier on PDNA-62. **b** The ROC curves EL_PSSM-RT, SVM classifier and Random Forest classifier on PDNA-224

However, as DP-Bind(STR) only identify a subset of residues with similar label from the three individual classifiers, it cannot provide prediction results for the other residues. In TS-61, among the 14,253 residues (including 1078 binding residues and 13,175 non-binding residues), DP-Bind(STR) cannot provide prediction results for 2206 residues (including 213 binding residues and 1993 non-binding residues). Therefore, our method is a more general prediction classifier when comparing to DP-Bind(STR).

Analysis of important pair-relationships

To further understand the importance of PSSM-RT for DNA-binding residue prediction, we analyze the important pair-relationships found by the learning algorithm. Since the importance of the relations can be reflected by the discriminant weight vector of the pair-relationships extracted by PSSM-RT, the values in the discriminant weight vector indicates the discriminant powers of the features in the feature space. Following the published works in [58–60], the discriminant weight vector $\mathbf{W}$ is calculated as follows: first, we obtain the classification weight vector $\mathbf{A}$ from the ensemble learning classifier during the training process. $\mathbf{W}$ is calculated by applying the following formulae:

$$\mathbf{W} = \mathbf{A}^{\mathrm{T}} \cdot \mathbf{M} = \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{bmatrix}^{\mathrm{T}} \begin{bmatrix} m_{11} & m_{12} & \cdots & m_{1d} \\ m_{21} & m_{22} & \cdots & m_{2d} \\ \vdots & \vdots & \cdots & \vdots \\ m_{n1} & m_{n2} & \cdots & m_{nd} \end{bmatrix} \quad (14)$$

where $\mathbf{A}$ is the classification weight vector of the training dataset by the ensemble learning classifier trained on PDNA-62 and $\mathbf{M}$ is the feature vectors of all training data instances; d is the dimension of the feature space and n is the number of data instances in the training dataset. The analysis results are shown in Fig. 6 based on the data given in the part E of the Additional file 1 which lists all the discriminant weights of the 400 pair-relationships between the target residue and its neighboring residue. Figure 6 includes a heatmap showing the discriminant weight of every pair-relationship and a diagram of binding residues showing the pair-relationships between important residues. Figure 6a shows that the

Table 6 Comparison of EL_PSSM-RT with base classifiers on novel benchmark datasets

Dataset	Methods	ACC (%)	MCC	SN (%)	SP (%)	ST (%)	AUC
PDNA-62	SVM	76.81	0.37	79.65	76.51	78.08	0.861
	RF	79.44	0.36	71.83	80.29	76.06	0.847
	EL_PSSM-RT	80.11	0.43	81.91	79.92	80.92	0.881
PDNA-224	SVM	73.66	0.28	79.68	73.25	76.47	0.839
	RF	76.13	0.27	73.91	76.28	75.10	0.831
	EL_PSSM-RT	76.74	0.32	81.45	76.42	78.93	0.863

Table 7 Performance comparison of various prediction methods on PDNA-62 by five-fold cross-validation

Methods	ACC (%)	MCC	SN (%)	SP (%)	ST (%)	AUC
Dps-pred	79.10	–	40.30	81.80	61.10	–
Dbs-pssm	66.40	–	68.20	66.00	67.10	–
BindN	70.30	–	69.40	70.50	69.95	0.752
Dp-bind	78.10	0.49	79.20	77.20	78.20	–
DP-Bind	77.20	–	76.40	76.60	76.50	–
BindN-RF	78.20	–	78.10	78.20	78.15	0.861
BindN+	79.00	0.44	77.30	79.30	78.30	0.859
PreDNA[a]	79.40	0.42	76.80	79.70	78.30	–
EL_PSSM-RT	80.82	0.51	85.04	80.10	82.57	0.901

[a]denotes PreDNA without using structure features

Zhou *et al. BMC Bioinformatics* (2017) 18:379

Page 12 of 16

Table 8 Performance of EL_PSSM-RT Compared with PreDNA on PDNA-224 by five-fold cross-validation

Methods	ACC (%)	MCC	SN (%)	SP (%)	ST (%)	AUC
PreDNA[a]	79.10	0.29	69.50	79.80	74.60	–
EL_PSSM-RT	78.09	0.34	79.58	77.98	78.78	0.865

[a]denotes PreDNA without using structure features

relationships between amino acid pairs (K, K), (K, R), (R, R), (Q, K), (Q, R), (S, K), (S, R), (R, Q), (S, S), (S, Q), (T, R), (E, K), (E, R), (E, R),(E, Q) are the fifteen relationships with larger discriminant weights. This means that the amino acids K, R, Q, S, T and E are important in the interaction between proteins and its corresponding DNA molecular. This feature analysis result is consistent with many other works for DNA-binding proteins research which stated that R, K, E and S are important for the interaction between DNA-binding proteins and its responsible DNA molecules [61, 62]. Figure 6b shows eight DNA-binding residues and its context residues extracted from the structure of a protein-DNA complex (PDB id: 1u1q). As we can see from Fig. 6b, the relationship between R and K has the highest occurrence frequency among the eight DNA-binding residues and is the most important feature for DNA-binding residue prediction for this protein. The second most important feature is the relationship between R and K. The relationships between E and Q and between E and R are the third most important features. The analysis result validates the usefulness of PSSM-RT for the representation of DNA-binding residues.

Case study

In order to further validate the usefulness of EL_PSSM-RT for DNA-binding residue prediction, we apply EL_PSSM-RT trained on PDNA-62 to distinguish the binding residues from non-binding residues for two protein-DNA complexes which are not in the training set, namely, 1s40 and 1b3t. The proteins in these two complexes are two typical DNA-binding proteins and the sequences have sequence similarity

Table 9 Performance of EL_PSSM-RT Compared with DP-Bind on TS-61

Methods	ACC (%)	MCC	SN (%)	SP (%)	ST (%)	AUC
DP-Bind(SVM)	75.90	0.26	65.99	76.70	71.34	0.794
DP-Bind(KLR)	76.45	0.25	64.22	77.45	70.83	0.790
DP-Bind(PLR)	75.46	0.25	65.24	76.29	70.76	0.812
DP-Bind(MAJ)	76.64	0.26	65.24	77.57	71.41	–
DP-Bind(STR)	80.21	0.31	68.74	81.11	74.92	–
EL_PSSM-RT	77.33	0.31	72.64	77.73	75.19	0.839

of less than 25% for all the sequences in the training set. On 1s40, EL_PSSM-RT achieves 96.71% on ACC, 0.74 on MCC, 92.06% on SN, 96.96% on SP and 94.51% on ST, respectively. This means that 34 residues out of a total of 39 actual binding residues are correctly predicted by EL_PSSM-RT and only 24 residues in the 264 non-binding residues are incorrectly predicted as binding residues. The actual residues and predicted residues in 1s40 are shown in Fig. 7a and b, respectively. The two figures show that most of the real binding residues overlap with the predicted binding residues. This provides a visual indication that most of binding residues are correctly predicted.. In the case of 1b3t, EL_PSSM-RT achieves 90.02% on ACC, 0.60 on MCC, 79.17% on SN, 91.35% and 85.25% on ST, respectively. In other words, 40 residues out of 48 binding residues are correctly predicted and only 32 residues out of 244 non-binding residues are incorrectly predicted as binding residues. Figure 7c and d depict the actual binding regions and predicted binding regions on 1b3t, respectively. We can see that most of the actual binding residues overlap with the predicted binding residues and only very few non-binding residues are wrongly identified as the binding residues.

Web service description

A user-friendly web service of EL_PSSM-RT is made available in order to make our proposed predictor freely accessible to the public. For the convenience of users, we provide a step-by-step guideline to use EL_PSSM-RT below.

Step 1. Using the URL (http://hlt.hitsz.edu.cn:8080/PSSM-RT_SVM/) to access the web service (as shown in Fig. 8). The **Read Me** button on the homepage supplies more details of EL_PSSM-RT.
Step 2. Type or copy and paste a query sequence in the input box at the center of the homepage. The query sequence must be in the fasta format [63]. By clicking the **Example button**, some examples for sequences in the fasta format [63] will be returned. As EL_PSSM-RT needs to apply PSI_BLAST, PSIPRED and SABLE to retrieve PSSMs, predicted secondary structures as well as predicted solvent accessible area for the query sequence, it will take quite some time. Therefore, only one sequence is allowed for submission at a time.
Step 3. Click the **Submit** button to get the prediction from the server. The predicted results of all residues in the sequence will be displayed on the result page, where '+' denotes a binding residue and '-' denotes a non-binding residue.
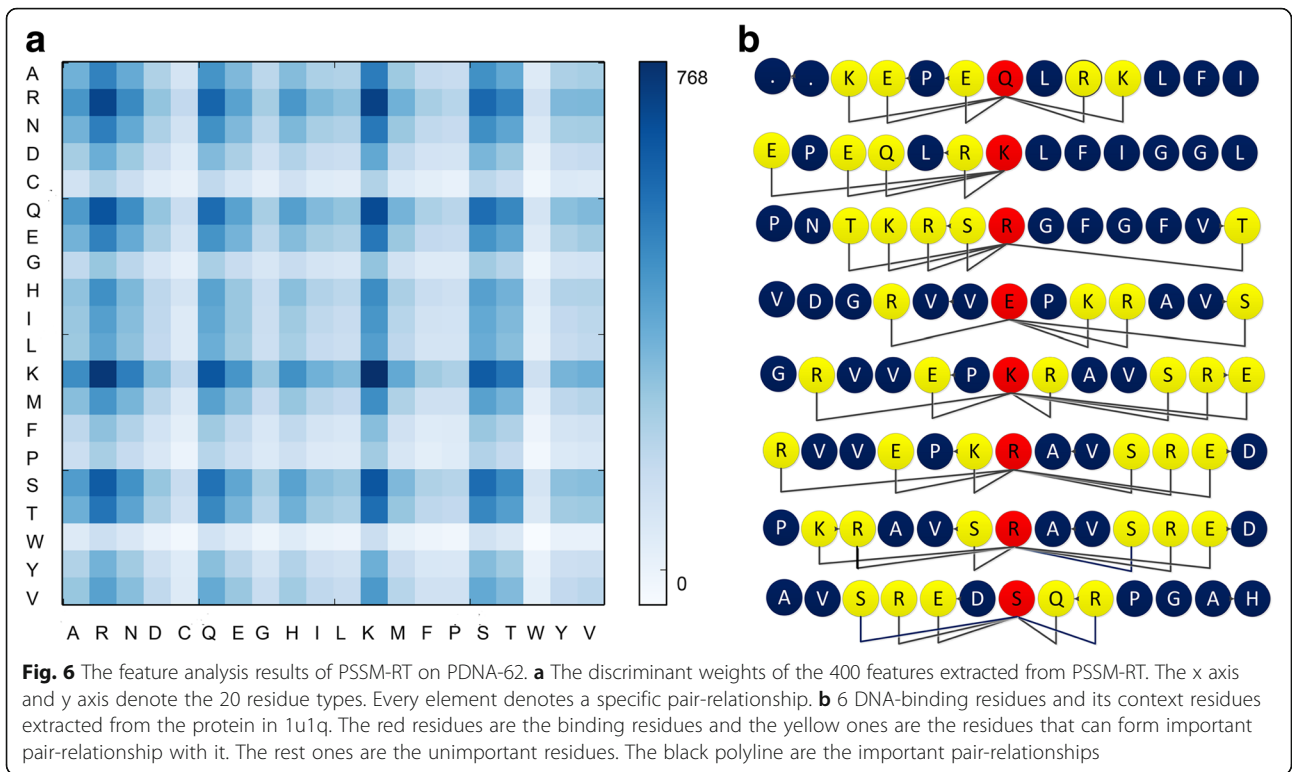
Zhou *et al. BMC Bioinformatics* (2017) 18:379

Page 13 of 16

Fig. 6 The feature analysis results of PSSM-RT on PDNA-62. **a** The discriminant weights of the 400 features extracted from PSSM-RT. The x axis and y axis denote the 20 residue types. Every element denotes a specific pair-relationship. **b** 6 DNA-binding residues and its context residues extracted from the protein in 1u1q. The red residues are the binding residues and the yellow ones are the residues that can form important pair-relationship with it. The rest ones are the unimportant residues. The black polyline are the important pair-relationships

Conclusion

As the interactions between proteins and DNAs are mainly formed by the immediate contacts between them, the identification of residues involved in the contacts is important for understanding the mechanism between amino acids and nucleotides. Many methods have been proposed to use evolutionary information for DNA-binding residues prediction. The combination methods and the concatenation methods are two commonly used methods. Both of them used only the evolutionary information of residues, yet the relationships of evolutionary information between residues are overlooked. In this paper, we propose a novel PSSM encoding method, referred to as PSSM-RT, which
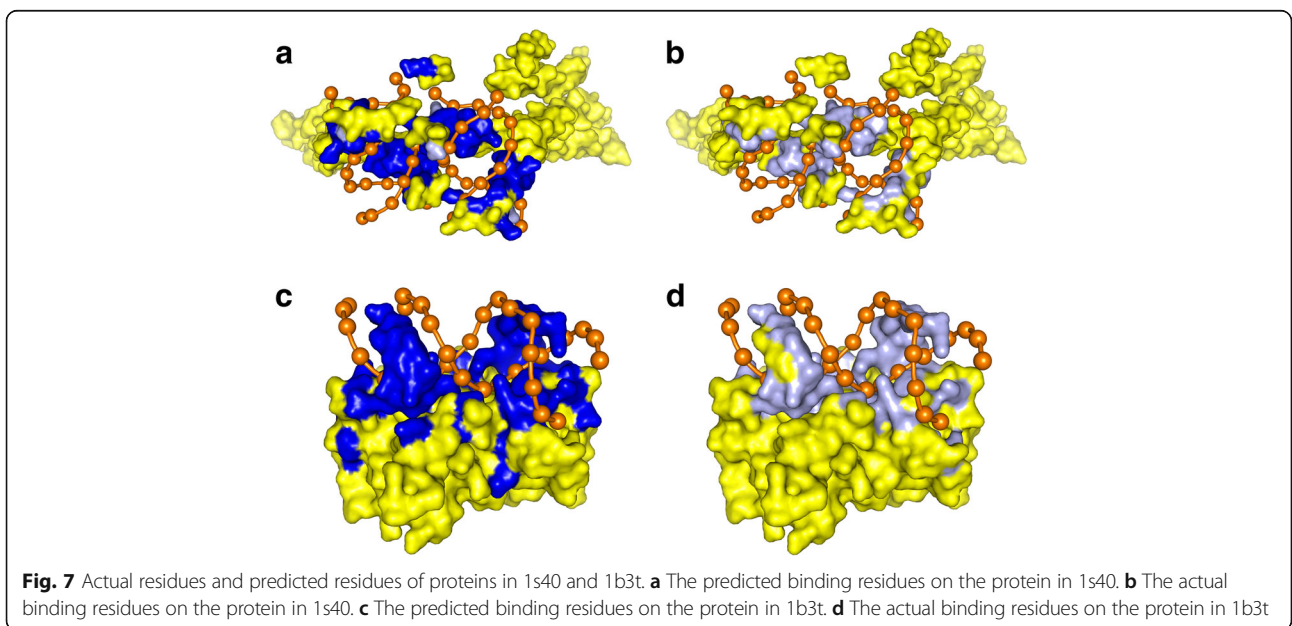

Fig. 7 Actual residues and predicted residues of proteins in 1s40 and 1b3t. **a** The predicted binding residues on the protein in 1s40. **b** The actual binding residues on the protein in 1s40. **c** The predicted binding residues on the protein in 1b3t. **d** The actual binding residues on the protein in 1b3t
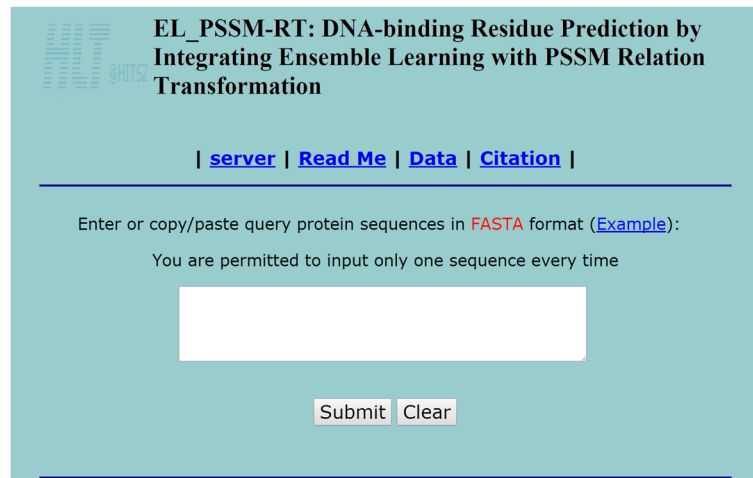
Zhou *et al. BMC Bioinformatics* (2017) 18:379

Page 14 of 16

Fig. 8 The homepage of the web service of EL_PSSM-RT. The web address of this webserver is http://hlt.hitsz.edu.cn:8080/PSSM-RT_SVM/. See the description in the server description for further explanation

includes the relationships of evolutionary information between residues. On both the PDNA-62 dataset and the PDNA-224 dataset, PSSM-RT performs better than the combination methods and the concatenation methods. When sequence features and physicochemical features are added, the prediction performance is further improved. This indicates that the evolution information, sequence features and physiochemical features are complementary for predictions. By combining ensemble learning and PSSM-RT, we propose a novel classifier EL_PSSM-RT to better handle the imbalance between binding and non-binding residues in datasets. The comparison of EL_PSSM-RT with the SVM classifier and the RF classifier on PDNA-62 and PDNA-224 indicates that ensemble learning is indeed useful for DNA-binding residue prediction. Performance comparisons between EL_PSSM-RT and existing predictors on two commonly used datasets and two independent datasets demonstrate that EL_PSSM-RT is more effective than the published works. Feature analysis of PSSM-RT on the PDNA-62 dataset demonstrates that PSSM-RT can extract many useful pair-relationships for DNA-binding residue prediction. The case study on 1s40 and 1b3t indicates that EL_PSSM-RT can correctly predict most of the binding residues with very low false positive rate. The performance evaluation and the case study on 1s40 and 1b3t show that the relationship of evolutionary information between residues is indeed useful in DNA-binding residue prediction and ensemble learning can be used to address the data imbalance issue between binding and non-binding residues in training datasets. Furthermore, we construct two novel benchmark datasets based on PDNA-62 and PDNA-224 by keeping only one protein for every homologous families and evaluate EL_PSSM-RT as well as the SVM classifier and the RF classifier on these two novel benchmark datasets. By comparing the performance on the novel benchmark datasets with that on the original benchmark datasets, we observe that over-representation of some homologous families indeed bias the performance towards those families. So our future works will study the details of DNA-binding residues for homologous proteins and the influence of homologous proteins on the predicting performance of EL_PSSM-RT and state-of-the-art methods.

Additional file

Additional file 1: The PDB id and the chain id of the protein sequences in the four datasets and the discriminant weights of the 400 pair-relationships between the target residue and its neighboring residue extracted from PSSM-RT. (DOC 141 kb)

Abbreviations
ACC: Accuracy; AUC: Area under ROC curve; ChIP: Chromatin immunoprecipitation; EIIP: Electron-ion interaction potential; EMSAs: Electrophoretic mobility shift assays; KLR: Kernel logistic regression; LEPs: Lone electron pairs; MAJ: Majority vote; MCC: Mathews Correlation Coefficient; NMR: Nuclear magnetic resonance; NR: Non-redundant; PAIR: Peptide nucleic acid-assisted identification of RNA binding proteins; PDB: Protein Data Bank; PLR: Penalized logistic regression; PNA: Peptide nucleic acid; PSSM: Position Specific Score Matrix; PSSM-RT: Position Specific Score Matrix Relation Transformation; RBPs: RNA binding proteins; RF: Random Forest; ROC: Receiver Operating Characteristic; SN: Sensitivity; SP: Specificity; ST: Strength; STR: Strict consensus; SVM: Support Vector Machine

Zhou *et al. BMC Bioinformatics* (2017) 18:379

Page 15 of 16

Availability of data and materials

All data generated or analysed during this study are included in this published article and its supplementary information files. The webserver developed by this article is available at http://hlt.hitsz.edu.cn:8080/PSSM-RT_SVM/.

Authors' contributions

JZ initiated and designed the study. RX made substantial contributions to acquisition of data, analysis and interpretation of data. JZ drafted the manuscript. RX and QL involved in drafting the manuscript or revising it. YH and HW provided valuable insights on biomolecular interactions and systems biology modeling, participated in result interpretation and manuscript preparation. All authors read and approved the final manuscript.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Author details

[1]School of Computer Science and Technology, Harbin Institute of Technology Shenzhen Graduate School, HIT Campus Shenzhen University Town, Xili, Shenzhen, Guangdong 518055, China. [2]Department of Computing, the Hong Kong Polytechnic University, Kowloon, Hong Kong. [3]Shenzhen Engineering Laboratory of Performance Robots at Digital Stage, Shenzhen Graduate School, Harbin Institute of Technology, Shenzhen, China. [4]School of Engineering and Applied Science, Aston University, Birmingham, UK.

References

1. Ofran Y, Mysore V, Rost B. Prediction of DNA-binding residues from sequence. Bioinformatics. 2007;23(13):i347–53.
2. Luscombe NM, Austin SE, Berman HM, Thornton JM. An overview of the structures of protein–DNA complexes. Genome Biol. 2000;1(1):1–37.
3. Walter MC, Rattei T, Arnold R, Guldener U, Munsterkotter M, Nenova K, Kastenmuller G, Tischler P, Wolling A, Volz A, et al. PEDANT covers all complete RefSeq genomes. Nucleic Acids Res. 2009;37:D408–11.
4. Luscombe NM, Thornton JM. Protein-DNA interactions: amino acid conservation and the effects of mutations on binding specificity. J Mol Biol. 2002;320(5):991–1009.
5. Bullock AN, Fersht AR. Rescuing the function of mutant p53. Nat Rev Cancer. 2001;1:68–76.
6. Jones S, Barker JA, Nobeli I, Thornton JM. Using structural motif templates to identify proteins with DNA binding function. Nucleic Acids Res. 2003;31: 2811–23.
7. Jones S, Heyninge P, Berman HM. Thornton JM: protein-DNA interactions: a structural analysis. J Mol Biol. 1999;287:877–96.
8. Ponting CP, Schultz J, Milpetz F, Bork P. SMART: identification and annotation of domains from signalling and extracellular protein sequences. Nucleic Acids Res. 1999;27:229–32.
9. Orengo CA, Michie AD, Jones S, Jones DT, Swindells MB, Thornton JM. CATH—Ahierarchic classification of protein domain structures. Structure. 1997;5:1093–108.
10. Olson WK, Gorin AA, Lu XJ, Hock LM, Zhurkin VB. DNA sequence-dependent deformability deduced from protein-DNA crystal complexes. Proc Natl Acad Sci U S A. 1998;95:11163–8.
11. Luscombe NM, Laskowski RA, Thornton JM. Amino acid-base interactions: a three-dimensional analysis of protein-DNA interactions at an atomic level. Nucleic Acids Res. 2001;29:2860–74.
12. Mandel-Gutfreund Y, Margalit H. Quantitative parameters for amino acid-base interaction: implications for prediction of protein-DNA binding sites. Nucleic Acids Res. 1998;26:2306–12.
13. Kono H, Sarai A. Structure-based prediction of DNA target sites by regulatory proteins. Proteins. 1999;35:114–31.
14. Ahmad S, Gromiha MM, Sarai A. Analysis and prediction of DNA-binding proteins and their binding residues based on composition, sequence and structural information. Bioinformatics. 2004;20(4):477–86.
15. Tsuchiya Y, Kinoshita K, Nakamura H. Structure-based prediction of DNA-binding sites on proteins using the empirical preference of electrostatic potential and the shape of molecular surfaces. Proteins. 2004;55(4):885–94.
16. Bhardwaj N, Langlois RE, Zhao G, Lu H. Structure based prediction of binding residues on DNA-binding proteins. Conf Proc IEEE Eng Med Biol Soc. 2005;3:2611–4.
17. Bhardwaj N, Lu H. Residue-level prediction of DNA-binding sites and its application on DNA-binding protein predictions. FEBS Lett. 2007;581(5): 1058–66.
18. Wang L, Brown SJ. BindN: a web-based tool for efficient prediction of DNA and RNA binding sites in amino acid sequences. Nucleic Acids Res. 2006; 34(suppl 2):W243–8.
19. Wang L, Huang C, Yang MQ, Yang JY. BindN+ for accurate prediction of DNA and RNA-binding residues from protein sequence features. BMC Syst Biol. 2010;4(Suppl 1):S3.
20. Ahmad S, Sarai A. PSSM-based prediction of DNA binding sites in proteins. BMC Bioinformatics. 2005;6(1):33.
21. Wang L, Brown SJ. Prediction of DNA-binding residues from sequence features. J Bioinforma Comput Biol. 2006;4:1141–58.
22. Ma X, Wu JS, Liu HD, Yang XN, Xie JM, Sun X. A SVM-based approach for predicting DNA-binding residues in proteins from amino acid sequences, 2009 International Joint Conference on Bioinformatics, Systems Biology and Intelligent Computing, Proceedings; 2009. p. 225–9.
23. Kuznetsov IB, Gou Z, Li R, Hwang S. Using evolutionary and structural information to predict DNA-binding sites on DNA-binding proteins. Proteins. 2006;64(1):19–27.
24. Ho SY, Yu FC, Chang CY, Huang HL. Design of accurate predictors for DNA-binding sites in proteins using hybrid SVM–PSSM method. Biosystems. 2007; 90(1):234–41.
25. Xu R, Zhou J, Wang H, He Y, Wang X, Liu B. Identifying DNA-binding proteins by combining support vector machine and PSSM distance transformation. BMC Syst Biol. 2015;9(Suppl 1):S10.
26. Liu B, Xu J, Lan X, Xu R, Zhou J, Wang X, Chou KC. 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. 2014;9(9):e106691.
27. Liu B, Xu J, Zou Q, Xu R, Wang X, Chen Q. Using distances between Top-n-gram and residue pairs for protein remote homology detection. BMC Bioinformatics. 2014;15(Supple 2):S3.
28. Liu B, Zhang D, Xu R, Xu J, Wang X, Chen Q, Chou KC. Combining evolutionary information extracted from frequency profiles with sequence-based kernels for protein remote homology detection. Bioinformatics. 2014; 30(4):472–9.
29. Xu R, Zhou J, Liu B, Yao L, He Y, Zou Q, Wang X. enDNA-Prot: identification of DNA-binding proteins by applying ensemble learning. Biomed Res Int. 2014;2014:294279.
30. Xu R, Zhou J, Liu B, He Y, Zou Q, Wang X, Chou KC. Identification of DNA-binding proteins by incorporating evolutionary information into pseudo amino acid composition via the top-n-gram approach. J Biomol Struct Dyn. 2014;33(8):1720–30.
31. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. The protein data bank. Nucleic Acids Res. 2000;28:235–42.
32. Li T, Li QZ, Liu S, Fan GL, Zuo YC, Peng Y. PreDNA: accurate prediction of DNA-binding sites in proteins by integrating sequence and geometric structure information. Bioinformatics. 2013;29(6):678–85.
33. Veljkovic V, Veljkovic N, Este JA, Huther A, Dietrich U. Application of the EIIP/ ISM bioinformatics concept in development of new drugs. Curr Med Chem. 2007;14(4):441–53.
34. Li W, Godzik A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics. 2006;22:1658–9.
35. Li W, Jaroszewski L, Godzik A. Clustering of highly homologous sequences to reduce the size of large protein databases. Bioinformatics. 2001;17(3):282.
36. Li W, Jaroszewski L, Godzik A. Sequence clustering strategies improve remote homology recognitions while reducing search times. Protein Eng. 2002;15(8):643.

Zhou *et al. BMC Bioinformatics* (2017) 18:379

Page 16 of 16

37. Chen YC, Wright JD, Lim C. DR_bind: a web server for predicting DNA-binding residues from the protein structure based on electrostatics, evolution and geometry. Nucleic Acids Res. 2012;40(W1):W249–56.

38. Yan C, Terribilini M, Wu F, Jernigan RL, Dobbs D, Honavar V. Predicting DNA-binding sites of proteins from amino acid sequence. BMC Bioinformatics. 2006;7(1):262.

39. Terribilini M, Lee JH, Yan C, Jernigan RL, Honavar V, Dobbs D. Prediction of RNA binding sites in proteins from amino acid sequence. RNA. 2006;12:450–1462.

40. Swets JA. Measuring the accuracy of diagnostic systems. Science. 1998;240:1285–93.

41. Bradley AP. The use of the area under the ROC curve in the evaluation of machine learning algorithms. Pattern Recogn. 1997;30(7):1145–59.

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

43. Kumar M, Gromiha M, Raghava G. Identification of DNA-binding proteins using support vector machines and evolutionary profiles. BMC Bioinformatics. 2007;8(1):563.

44. Jones DT. Improving the accuracy of transmembrane protein topology prediction using evolutionary information. Bioinformatics. 2007;23:538–44.

45. Biswas AK, Noman N, Sikder AR. Machine learning approach to predict protein phosphorylation sites by incorporating evolutionary information. BMC Bioinformatics. 2010;11(1):273.

46. Ruchi V, Grish CV, Raghava GPS. Prediction of mitochondrial proteins of malaria parasite using split amino acid composition and PSSM profile. Amino Acids. 2010;39:101–10.

47. Zhao XW, Li XT, Ma ZQ, Yin MH. Prediction of lysine ubiquitylation with ensemble classifier and feature selection. Int J Mol Sci. 2011;12:8347–61.

48. Schaffer AA, Aravind L, Madden TL, Shavirin S, Spouge JL, Wolf YI, Koonin EV, Altschul SF. Improving the accuracy of PSI-BLAST protein database searches with composition-based statistics and other refinements. Nucleic Acids Res. 2001;29(14):2994–3005.

49. Bonchev D. The overall Wiener index - a new tool for characterization of molecular topology. J Chem Inf Comput Sci. 2001;41(3):582–92.

50. Vapnik VN. Statistical learning theory. New York: John Wiley; 1998.

51. McGuffin LJ, Bryson K, Jones DT. The PSIPRED protein structure prediction server. Bioinformatics. 2000;16(4):404–5.

52. Adamczak R, Porollo A, Meller J. Accurate prediction of solvent accessibility using neural networks based regression. Proteins. 2004;56(4):753–67.

53. Adamczak R, Porollo A, Meller J. Combining prediction of secondary structure and solvent accessibility in proteins. Proteins. 2005;59(3):467–75.

54. Wagner M, Adamczak R, Porollo A, Meller J. Linear regression models for solvent accessibility prediction in proteins. J Comput Biol. 2005;12(3):355–69.

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

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

57. Hwang S, Gou Z, Kuznetsov IB. DP-Bind: a web server for sequence-based prediction of DNA-binding residues in DNA-binding proteins. Bioinformatics. 2007;23(5):634–6.

58. Liu B, Wang X, Chen Q, Dong Q, Lan X. Using amino acid physicochemical distance transformation for fast protein remote homology detection. PLoS One. 2012;7(9):e46633.

59. Park KJ, Kanehisa M. Prediction of protein subcellular locations by support vector machines using compositions of amino acids and amino acid pairs. Bioinformatics. 2003;19(13):1656–63.

60. Yu CS, Chen YC, Lu CH, Hwang JK. Prediction of protein subcellular localization. Proteins. 2006;64(3):643–51.

61. Sieber M, Allemann RK. Arginine (348) is a major determinant of the DNA binding specificity of transcription factor E12. Biol Chem. 1998;379(6):731–5.

62. Szilágyi A, Skolnick J. Efficient prediction of nucleic acid binding function from low-resolution protein structures. J Mol Biol. 2006;358(3):922–3.

63. Lipman DJ, Pearson WR. Rapid and sensitive protein similarity searches. Science. 1985;227(4693):1435–1.