Prediction of DNA-binding residues from sequence information using convolutional neural network

Jiyun Zhou

School of Computer Science and Technology,
Harbin Institute of Technology Shenzhen Graduate School,
Shenzhen, Guangdong, China
and
Department of Computing,
The Hong Kong Polytechnic University,
Hung Hom, Hong Kong
Email: zhoujiyun2010@gmail.com

Qin Lu

Department of Computing,
The Hong Kong Polytechnic University,
Hung Hom, Hong Kong
Email: csluqin@comp.polyu.edu.hk

Ruifeng Xu*

School of Computer Science and Technology,
Harbin Institute of Technology Shenzhen Graduate School,
Shenzhen, Guangdong, China
and
Shenzhen Engineering Laboratory of Performance Robots at Digital Stage,
Shenzhen Graduate School,
Harbin Institute of Technology,
Shenzhen, China
Email: xuruifeng@hit.edu.cn
*Corresponding author

Lin Gui and Hongpeng Wang

School of Computer Science and Technology,
Harbin Institute of Technology Shenzhen Graduate School,
Shenzhen, Guangdong, China
Email: guilin.nlp@gmail.com
Email: wanghp@hit.edu.cn

Abstract: Most DNA-binding residue prediction methods overlooked the motif features which are important for the recognition between protein and DNA. In order to efficiently use the motif features for prediction, we first propose to use Convolutional Neural Network (CNN) in deep learning to extract discriminant motif features. We then propose a neural network classifier, referred to as

CNNsite, by combining the extracted motif features, sequence features and evolutionary features. The evaluation on PDNA-62, PDNA-224 and TR-265 shows that motif features perform better than sequence features and evolutionary features. The evaluation on PDNA-62, PDNA-224 and an independent data set shows that CNNsite also outperforms the previous methods. We also show that many motif features composed by the residues which play important roles in DNA–protein interactions have large discriminant powers. It indicates that CNNsite has very good ability to extract important motif features for DNA-binding residue prediction.

Keywords: DNA; protein; interaction; residue; CNN; motif; sequence; PSSM; evolutionary; binding; neural network.

Reference to this paper should be made as follows: Zhou, J., Lu, Q., Xu, R., Gui, L. and Wang, H. (2017) 'Prediction of DNA-binding residues from sequence information using convolutional neural network', *Int. J. Data Mining and Bioinformatics*, Vol. 17, No. 2, pp.132–152.

Biographical notes: Jiyun Zhou obtained his BEng degree from Northeast Forestry University, China, and MEng degree from Harbin Institute of Technology, China, respectively. He is now a PhD candidate in School of Computer Science and Technology, Harbin Institute of Technology (Shenzhen) and Department of Computing, Hong Kong Polytechnic University. His main research interests are bioinformatics, natural language processing and machine learning.

Qin Lu obtained her BEng degree from Beijing Normal University, and MSc and PhD degrees from University of Illinois at Urbana-Champaign, respectively. She is now a Full Professor in Department of Computing, The Hong Kong Polytechnic University, Hong Kong. Her research interests are computational linguistics, ontology, text mining, knowledge discovery and bioinformatics.

Ruifeng Xu obtained his BEng degree from Harbin Institute of Technology, China, and MPhil and PhD degrees from The Kong Polytechnic University, respectively. He is now a Full Professor and PhD Supervisor in School of Computer Science and Technology, Harbin Institute of Technology (Shenzhen), China. His research interests are bioinformatics, natural language processing, emotion computing and text mining.

Lin Gui obtained his BS degree from Nankai University, China, and MEng degree in from Harbin Institute of Technology, China, respectively. He is now a PhD candidate in School of Computer Science and Technology, Harbin Institute of Technology (Shenzhen). His research interests are natural language processing, machine learning and emotion computing.

Hongpeng Wang obtained his BEng, MEng and PhD degrees from Harbin Institute of Technology, China. He is now a Full Professor and PhD Supervisor in School of Computer Science and Technology, Harbin Institute of Technology (Shenzhen). His research interests are intelligent robot, computer vision, artificial intelligence and bioinformatics.

This paper is a revised and expanded version of a paper entitled 'CNNsite: prediction of DNA-binding residues in proteins using convolutional neural network with sequence features' presented at the 'IEEE BIBM 2016', Shenzhen, China, 15–18 December 2016.

1 Introduction

DNA-binding proteins are the proteins containing DNA-binding domains (Pabo and Sauer, 1984) which can form interactions with DNA. There are two mechanisms by which protein and DNA form the interactions: general interaction and specific interaction (Travers and Suck, 1993). The interactions between histone and DNA are well-understood examples of general protein–DNA interactions (Dame, 2005; Luger et al., 1997). A good example of specific interaction is the interaction between DNA and transcription factor which can activate or inhibit the transcription of genes (Ptashne, 2005; Kang et al., 2010). DNA-binding residues are important for the functions of the target protein and mutations of some DNA-binding residues may predispose individuals to disease. For example, the mutations of some binding residues on the tumour repressor protein P53 may predispose individuals to cancer (Bullock and Fersht, 2001). Thus, the prediction of the residues involved in protein–DNA interactions is not only important for understanding the gene regulation process but also helpful for annotating the function of proteins (Halford and Marko, 2004). Many experimental methods were developed to identify DNA-binding residues from protein sequences, such as Electrophoretic Mobility Shift Assays (EMSAs) (Jones et al., 2003; Jones et al., 1999) and Nuclear Magnetic Resonance (NMR) spectroscopy (Ponting et al., 1999). However, they are highly costly and laborious procedures. With the development of sequencing technology of proteins, more and more DNA-binding proteins with unknown function are sequenced. Therefore, it is urgent to propose computational methods for the prediction of DNA-binding residue automatically on a genome scale.

So far, many computational methods have been proposed for the prediction of DNA-binding residue. In these prediction methods, the features used in the prediction include three groups: sequence features, evolutionary features and structure features. The sequence features include amino acid composition, predicted structure features and physiochemical properties. Evolutionary features are used to represent the conservation characteristic for every residue in the protein sequences and are important features for the prediction of the protein functions and structures. Position Specific Score Matrix (PSSM) is a common representation of evolutionary features and has been used in many bioinformatics problems (Dehzangi et al., 2014; Fang et al., 2015). Both sequence features and evolutionary features can be extracted from the residue sequence of the target protein. However, in the early stage, only sequence features are used in the prediction due to the limitation of computing power. For example, Ofran et al. (2007) proposed predictor DISIS by using only sequence features, such as the local amino acid composition of its neighbours. With the advancement in computing power, evolutionary features are used by more and more methods for prediction. For example, Ahmad and Sarai (2005) used PSSM for the prediction of DNA-binding residues in proteins and got good performance on the data set PDNA-62. Ma et al. (2012) proposed DNABR by using six physiochemical properties and PSSM. Wang et al. (2010) proposed a prediction method, referred to as BindN+, by integrating PSSM and three physiochemical properties (Ahmad and Sarai, 2005). For predictions of DNA-binding residues, the classifiers are mostly trained by machine learning models, such as SVM, Random Forest and Neural Network. For example, the predictors SVM-PSSM (Ho et al., 2007) and Ma et al.'s SVM classifier (Ma et al., 2009) used SVM classifiers. The predictor DISIS (Ofran et al., 2007) used a Neural Network to predict DNA-binding residues using features from PSSM, level of conservation, and the predicted secondary structure. DBindR (Wu et al., 2009),

BindN-RF (Wang et al., 2009) and DNABR (Ma et al., 2012) are trained from PSSM and the sequence features by using Random Forest. There are also works using an ensemble classifier which combines the results from multiple classifiers to determine the classification results for residues (Si et al., 2011). The base classifiers used include BindN (Wang and Brown, 2006), BindN-RF (Wang et al., 2009), DBS-PRED (Ahmad et al., 2004), DISIS (Ofran et al., 2007), DNABindR (Wu et al., 2009) and DP-Bind (Hwang et al., 2007).

Since the functions of proteins are closely related to their 3D structure, structure features are more important than both sequence features and evolutionary features for function determination of residues and sequences. The secondary structure, the solvent accessible surface area, the spatial neighbours, the B-factor, the protrusion index and the depth index are several commonly used structural features. In recent years, structure features extracted from 3D structures of proteins are used by many methods. Zhu et al. (2013) developed the predictor DBSI by some structural features including polar atom availability, electrostatic potential, surface curvature, local atomic density, residue microenvironments, and non-local polar. Ozbek et al. (2010) proposed the method DNABINDPROT by first selecting candidate residues based on the fluctuations of residues in high-frequency modes and then filtering the selected residues with their evolutionary conservation profiles. Chen et al. (2012) proposed the predictor DR_bind by first calculating geometry features, electrostatics features as well as conservation features and then selecting the three patches with the largest features as biding residues. Li et al. (2013) proposed PreDNA by integrating an SVM classifier and a structure template-based prediction protocol and Liu et al. (2015) proposed the predictor DNABind by integrating an SVM classifier and a different template-based prediction protocol. For structure features-based methods, there are also some predictors that are developed by machine learning algorithms including SVM and neural network. For example, SVM was used by DBS-PRED, proposed by Ahmad et al. (2004), DP-Bind, proposed by Kuznetsov et al. (2006), PreDNA, proposed by Li et al. (2013), DNABind developed by Liu et al. (2015) and DBSI, proposed by Zhu et al. (2013). DISPLAR, proposed by Tjong and Zhou (2007), was trained by a neural network.

Motif features are defined as short residue sequences with uncertain length. Motif features often occur in the regions around the binding residues by which the corresponding DNA fragments recognise the binding residues. Thus, they may be very important discriminant for the prediction of DNA-binding residues. However, due to the lack of effective methods to extract motif features, there is no related works using motif features as far as we know.

In this work, we investigate the use motif features for the prediction of DNA-binding residues. Since motif features are usually located in the regions around the binding residues and can be extracted from protein sequences (Ahmad et al., 2004), we apply Convolutional Neural Network (CNN) (Krizhevsky et al., 2012; Simard et al., 2003; Lawrence et al., 1997) to extract important motif features from the training data set. CNN uses learnable detectors, which can be learned from the training data set, to capture important motif features for DNA–protein residue recognition. Since sequence features, evolutionary features and structure features are also important for proteins, they should also be included in the prediction. However, structure features for most of the protein data are unavailable (Sakar et al., 2014). Methods based on structure features cannot be used on a genomic scale. Thus, in this work, only sequence features and evolutionary

features are combined with our extracted motif features for prediction of DNA-binding residues. By combining the three groups of features, we develop a neural network classifier, referred to as CNNsite.

2 Feature extraction and proposed method

In this section, we first introduce the data sets used in this study. We then introduce the definition of residue-wise data instance for each residue, followed by feature extraction methods. Finally, we introduce the CNNsite algorithm for motif features extraction.

2.1 Data sets

In order to evaluate the prediction performance of CNNsite and compare it with other existing state-of-the-art prediction classifiers, we first introduce the three benchmarking data sets and one independent data set used in this study.

PDNA-62, the first benchmarking data set, was constructed by Ahmad et al. (2004). The similarity between any two proteins in PDNA-62 is less than 25%. PDNA-224, the second benchmarking data set, is a recently developed data set for DNA-binding residue prediction (Li et al., 2013) and contains 224 protein sequences with sequence similarity less than 25%. These two benchmarking data sets were used in our previous work (Zhou et al., 2016) for evaluating DNA-binding residue prediction methods. However, these two data sets are small and cannot provide very persuasive conclusion. In order to give a more comprehensive study in this subject, a new benchmarking data set is used in this work. TR-265, the third benchmarking data set, was proposed by Ma et al. (2012) and it contains 265 proteins from PDB. The sequence identity between any two chains in TR-265 is less than 25%. To compare with other methods that were not evaluated on the above three data sets, our previous work used an independent data set, TS-72, extracted from Ma et al. (2012). But TS-72 was relatively old which did not include some of the new proteins. So, in this work, we build a new independent data set, referred to as TS-61, to evaluate the performance of our method. TS-61 is a novel data set constructed in this study by first retrieving protein–DNA complexes from PDB and then screening the sequences with the cut-off pairwise sequence similarity of 25%. The result contains 61 sequences and the similarity between a sequence in TS-61 and any sequence in PDNA-62, PDNA-224 and TR-265 is less than 25%. The PDB id and the chain id of the 61 protein sequences in TS-61 are listed in part A of addition file 1, which can be obtained from our website http://hlt.hitsz.edu.cn/CNNsite/.

In the above four data sets, positive and negative samples are defined by the following criterion (Wang and Brown, 2006; Chen et al., 2012; Yan et al., 2006): 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 cut-off distance of 3.5 Å from any atom of the partner DNA molecule in the complex; otherwise, the residue is considered as a non-binding residue. The number of positive samples and negative samples of the four data sets are listed in Table 1. Note that the data are quite imbalanced. There are much more non-binding residues than binding residues. In order to hander the imbalance between the binding and non-binding residues in these data sets, we use the same method as most of the methods (Ahmad et al., 2004; Wang et al., 2010) randomly select approximately the same number of non-binding residues as the binding residues to construct the training data set.

Table 1 Number of the positive samples and negative samples of the four data sets

Data set	*PDNA-62*	*PDNA-224*	*TR-265*	*TS-61*
Binding residue	1215	3778	4054	1078
Non-binding residue	6948	53,570	52,658	13,175
Total	8163	57,348	56,712	14,253

2.2 *Residue-wise data instances*

In DNA-binding residue prediction, residues are the prediction targets. Since the biology function of a residue is often influenced by its neighbouring residues, the residue-wise data instance for each residue is defined as a window composed of w residues, where the target residue is positioned in the middle and $(w-1)/2$ neighbouring residues are located on either side. A residue-wise data instance is defined as positive if the central residue is a DNA-binding residue or negative if the central residue is a non-binding residue.

Given a protein sequence P with length of L formulated as:

$$P = R_1 R_2 R_3 R_4 R_5 R_6 \cdots R_{i-1} R_i R_{i+1} \cdots R_L \tag{1}$$

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

$$S_i = R_{i-\frac{w-1}{2}} R_{i-\frac{w-3}{2}} \cdots R_{i-1} R_i R_{i+1} \cdots R_{i+\frac{w-3}{2}} R_{i+\frac{w-1}{2}} \tag{2}$$

where all the residues in sequence fragment S_i except the target residue R_i are the contextual residues. The $(w-1)/2$ contextual residues on the left side and the right side are termed as the left contextual residues and right contextual residues, respectively.

2.3 *Feature descriptors*

In this paper, sequence features and evolutionary features are combined with motif features for the prediction of DNA-binding residue.

(1) Sequence features, denoted by SEQ, contain local amino acid composition, predicted second structure and predicted solvent accessible area. The predicted secondary structure and predicted solvent accessible area are obtained by PSIPRED (McGuffin et al., 2000) and SABLE (Adamczak et al., 2004; Adamczak et al., 2005; Wagner et al., 2005), respectively. (2) PSSM is a common representation of evolutionary features, denoted by EVO, which is obtained by running the PSI-BLAST (Schaffer et al., 2001) program to search against the non-redundant (NR) database through three iterations with 0.001 as the E-value cut-off for multiple sequence alignment. Before being fed into a prediction engine, all the scores in PSSM need to be scaled between 0 and 1 using the following equation.

$$NPSSM(i,j) = \frac{1}{1+e^{-PSSM(i,j)}} \tag{3}$$

Motif features, denoted as MOT, are short residue sequences with uncertain length. We consider five groups of motif features in our work and their length are 2, 3, 4, 5 and 6, respectively.

PSSM

Feature = SEQ + EVO + MOT

AAC + second structure + accessible area

Sequence features and evolutionary features can be obtained directly from the protein sequence, while motif features need to be extracted by CNN which will introduced in the following text.

2.4 Convolutional neural network (CNN)

In this work, we propose to use CNN to identify important motif features from the sequences around the binding residues. The extracted features can then be used by a neural network classifier, by combining with both sequence features and evolutionary features for the DNA-binding residue prediction. The system is referred to as CNNsite.

The frame diagram of CNNsite is shown in Figure 1. CNNsite comprises of four computational layers typical of a CNN: the convolution, the rectification, the pooling, and the neural network classifier. In our prediction task, the first three layers are all used to extract useful motif features from the inputting residue-wise data instances and the last layer is used for prediction. For the four layers, training is for three sets of parameters, the motif detectors M in the convolution layer, the thresholds b for the rectification layer, and the weights W for the network layer, respectively. For a residue-wise data instance S, CNNsite produces a real-valued score $f(S)$ by the following formula:

$$f(S) = net_W\left(pool\left(rect_b\left(conv_M(S)\right)\right)\right) \tag{4}$$

where $conv_M()$, $rect_b()$, $pool()$ and $net_W()$ denote the four layers in CNNsite, respectively. This real-valued score is used for prediction.

Convolution. In the convolution process, several filters, called motif detectors, are used to convolve the raw input. For a residue-wise data instance, the convolution of a motif detector over it can play the same role as 'motif scan' operation in a PWM or a PSAM-based model. For a motif detector of length m, the residue-wise data instance S should be padded by concatenating $(m - 1)$ unuseful residues on either side. So, a residue-wise data instance S is represented as a $(n+2m-2)\times 20$ matrix R in the following way:

$$R_{i,j} = \begin{cases} 0.05 & \text{if } i<m \text{ or } i>n-m \\ 1 & \text{if } S_{i-m+1} = j \\ 0 & \text{otherwise} \end{cases} \tag{5}$$

where m is the length of the motif detector, n is the length of the residue-wise data instance, i is the position in S and j denotes the j-th residue type. The convolution output $X = conv_M(R)$ is an $(n+m-1)\times d$ matrix where element $X_{i,k}$ is essentially the score of the motif detector k aligned to position i of R. Given that the motif detectors are represented as a $d\times m\times 20$ array M where d is the number of motif detectors and element $M_{k,j,l}$ is the coefficient of motif detector k at motif position j and residue type l, the element $X_{i,k}$ of the output is calculated by the following formula:

$$X_{i,k} = \sum_{j=1}^{m}\sum_{l=1}^{20} R_{i+j,l} M_{k,j,l} \tag{6}$$

The column X_k is the motif scan of motif detector k applied to R and row X_i is the motif scan of all the motif detectors at position i of R.

Rectification. Rectification plays an important role in deep learning. Its input is the $(n+m-1)\times d$ matrix from the convolution. The output is a matrix of the same size $Y = rect_b(X)$.

$$Y_{i,k} = \max\left(0, X_{i,k} - b_k\right) \quad (7)$$

where b_k is the activation threshold for motif detector k, learned in the training process. The formula means that if score $X_{i,k} \geq b_k$ then the relative score of motif detector k at position i is passed to the next stage; otherwise, motif detector k is deemed irrelevant at position i and so the relative score is zero. This layer is used to identify the important motif features by keeping only the motif features with scores larger than a specified threshold.

Pooling. Pooling takes the $(n+m-1)\times d$ matrix Y outputted by the rectification layer. For every residue-wise data instance, we can obtain a vector Z with dimension of d. The features contained in vector Z are motif features captured by the d motif detectors in the convolution layer. Z, as an output vector, is a feature vector composed by the maximum values of the d motif detectors in the residue-wise data instance S. Thus, the dimension of Z depends on the number of motif detectors in the convolution. Each feature in Z is the maximum value of the corresponding motif detector in S. The feature Z_k ($1 \leq k \leq d$) in feature vector Z is formulated as:

$$Z_k = \max\left(Y_{1,k}, \ldots, Y_{n,k}\right) \quad (8)$$

where $Y_{1,k},\ldots,Y_{n,k}$ are the values of the motif detector k at the n positions in the residue-wise data instance S.

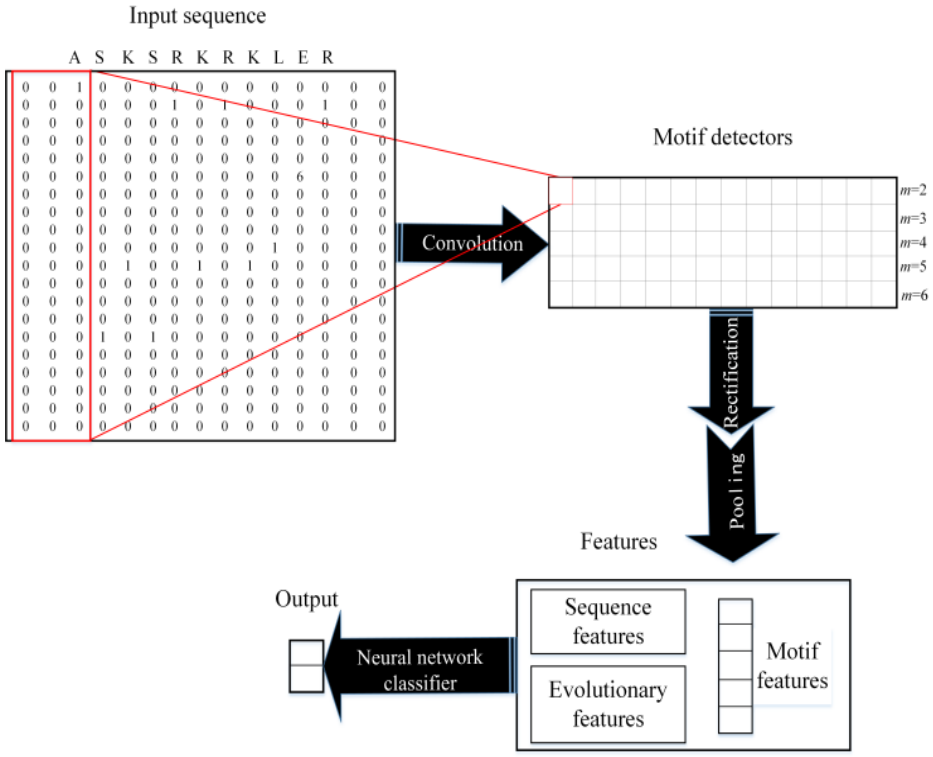
Neural network classifier. Most of the prediction methods which use CNN to learn the representation for protein sequence or residues often use neural network as the classifier (Alipanahi et al., 2015; Wang et al., 2016) and neural network classifier usually has similar performance compared to an SVM classifier and a random forest (RF) classifier for the features learned from CNN (Ciregan et al., 2012; Collobert and Weston, 2008). Since the motif features are learned from CNN, we need to use a neural network classifier in this layer. So in this work, the neural network classifier is used for prediction. In CNNsite, three kinds of features are used as input in this layer including motif features, sequence features, and evolutionary features, where sequence features and evolutionary features are directly obtained from the protein sequence, while motif features are learned from the above three layers of CNN. The neural network classifier contains a hidden layer with p rectified-linear units. The prediction for the residue-wise data instance S is completed by the following formula:

$$h_j = \max\left(0, W_{j,d+d_1+d_2+1} + \sum_{k=1}^{d} W_{j,k} Z_k + \sum_{k}^{d_1} W_{j,(d+k)} Q_k + \sum_{k}^{d_2} W_{j,(d+d_1+k)} E_k\right) \quad \text{for } j = 1 \ldots p \quad (9)$$

$$f = w_{l+1} + \sum_{j=1}^{32} w_j h_j \quad (10)$$

where W is the weight matrix of the hidden layer, w is the weight vector used for prediction, Z represents the motif features outputted by pooling, Q represents the sequence features and E represents the evolutionary features. In order to avoid overfitting for CNNsite, we use the recently proposed dropout technique after pooling before the data are fed into the neural network classifier. With the dropout technique, the entries of the hidden representations are set to 0 with a dropout rate, which is tuned based on the development set.

Figure 1 The frame diagram of CNNsite (see online version for colours)



3 Experiments and results

The purpose of the evaluation is to examine the performance of CNNsite for the prediction of DNA-binding residue. Since CNNsite uses a window-based approach, the window size needs to be set properly. Owing to the length of this paper, we skipped the parameter tuning experiment and all the results shown in this section use window size $w = 11$ which was experimentally set. Four sets of evaluations are conducted here. The first set compares the performance of motif features with that of sequence features and evolutionary features on the most commonly used to data sets: PDNA-62 and PDNA-224. The second set conducts the same evaluation on the more comprehensive data set:

TR-265. The third experiment compares CNNsite with previous published predictors on both PDNA-62 and PDNA-224. The last set of experiments evaluates CNNsite on a novel independent test TS-61 compared with previous published methods.

3.1 Evaluation metrics

In this work, five common metrics: Sensitivity (SN), Specificity (SP), Strength (ST), Accuracy (ACC), and Mathews Correlation Coefficient (MCC) are used to evaluate the performance of CNNsite for DNA-binding residue prediction. They are formulated as follows:

$$\mathrm{SN}=TP/(TP+FN) \tag{11}$$

$$\mathrm{SP}=TN/(TN+FP) \tag{12}$$

$$\mathrm{ST}=(\mathrm{SN}+\mathrm{SP})/2 \tag{13}$$

$$\mathrm{ACC}=(TP+TN)/(TP+FP+TN+FN) \tag{14}$$

$$\mathrm{MCC}=\frac{(TP*TN-FP*FN)}{\sqrt{(TP+FN)*(TP+FP)*(TN+FP)*(TN+FN)}} \tag{15}$$

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

Since the number of binding residues and that of non-binding residues in the data sets are unbalanced, ACC often provides very biased evaluating performance. Literatures (Wang and Brown, 2006; Wang et al., 2010; Li et al., 2013) have reported that ST, the average of SN and SP, can provide a more appropriate evaluation for predictors when the data set are unbalanced. MCC can measure the matching degree between prediction results and real results. So ST and MCC are often used as the main metrics, while the remaining metrics are used only for references especially when the data set are unbalanced. Receiver Operating Characteristic (ROC) curve is a standard representation of the trade-off between false positive rate and sensitivity. The curve is drawn by plotting the true positive rates (i.e. sensitivity) against the false positive rates (i.e. 1-specificity) based on the classification threshold for predictors. AUC is the area under the ROC curve and is a fair metric for unbalanced problem.

3.2 The effectiveness of motif features on PDNA-62 and PDNA-224

This set of experiments compares the performance of motif features with that of sequence features and evolutionary features for the DNA-binding residue prediction by PDNA-62 and PDNA-224. For a fair comparison, the neural network classifier (Breiman, 2001; Holmes et al., 1994) is used for all features and feature combinations in this experiment. The performance on PDNA-62 and PDNA-224 are shown in Tables 2 and 3, respectively. As mentioned earlier, since MCC, ST and AUC are the main performance metrics, we shade the best performers of these three metrics for easy observation. From Tables 2 and 3 we can see that motif features, when used as single set of features, outperforms both sequence features and evolutionary features (for example, by at least

0.114 on MCC, 7.51% on ST and 0.101 on AUC to sequence features in the PDNA-62 data set and by at least 0.097 on MCC, 6.73% on ST, and 0.069 on AUC to evolutionary features in the PDNA-62 data set). This indicates that motif features are more discriminant than sequence features and evolutionary features for the prediction as single features. When motif features are combined either with sequence features or evolutionary features, the performance is improved on all metrics. More specifically, the increase for sequence features is at least 0.014 on MCC, 0.97% on ST and 0.011 on AUC in the PDNA-62 data set and the increase for evolutionary features is 0.014 on MCC, 1.03% on ST and 0.014 on AUC in the PDNA-224 data set.

We also show the performance of using the combinations of only sequence and evolutionary features. But, it does not perform better than using motif features with either of them. This also indirectly proves that motif features are more dominant than either the sequence features or the revolutionary features. When all three types of features are used, CNNsite achieves the best result and outperforms other types of combinations quite significantly on both data sets. Figure 2 (A) and Figure 2 (B) show the ROC curves of different feature combinations on PDNA-62 and PDNA-224, respectively. It shows that motif features have better ROC curve than sequence features and evolutionary features, and the combination of them gets the best ROC curve. In other words, the three types of features are complementary for prediction.

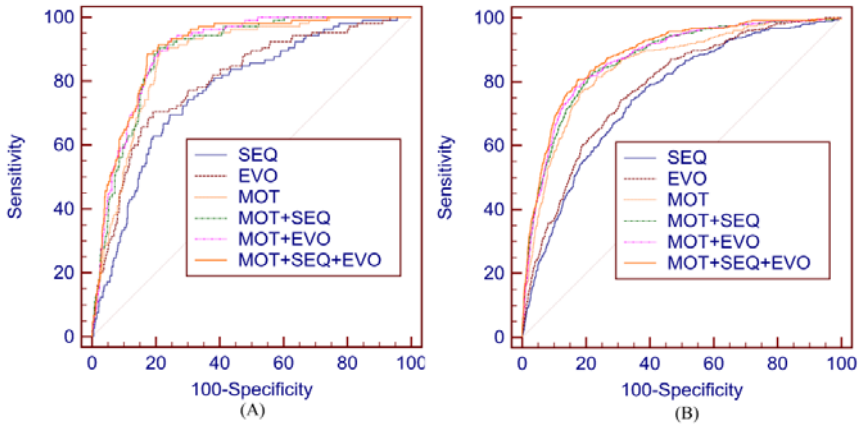
Table 2 The predicting performance on PDNA-62 for various features by tenfold cross-validation using the neural network classifier

Method	*ACC (%)*	*MCC*	*SN (%)*	*SP (%)*	*ST (%)*	*AUC*
SEQ	73.78	0.345	70.94	74.29	72.61	0.770
EVO	75.27	0.362	70.74	76.04	73.39	0.802
MOT	77.48	0.459	83.89	76.36	80.12	0.871
SEQ + EVO	75.63	0.396	76.41	75.45	75.93	0.823
MOT + SEQ	78.15	0.473	85.25	76.92	81.09	0.889
MOT + EVO	78.57	0.476	84.81	77.48	81.15	0.897
MOT + EVO + SEQ	**80.63**	**0.509**	**85.87**	**79.78**	**82.67**	**0.911**

Table 3 The predicting performance on PDNA-224 for various features by tenfold cross-validation using the neural network classifier

Method	*ACC (%)*	*MCC*	*SN (%)*	*SP (%)*	*ST (%)*	*AUC*
SEQ	87.58	0.222	33.85	91.80	62.83	0.756
EVO	89.16	0.251	33.23	93.35	63.39	0.780
MOT	83.09	0.367	72.85	83.91	78.38	0.858
SEQ + EVO	85.99	0.290	49.73	88.84	69.29	0.832
MOT + SEQ	82.85	0.382	76.63	83.34	79.99	0.869
MOT + EVO	82.40	0.381	**77.35**	82.79	80.07	0.872
MOT + SEQ + EVO	**83.68**	**0.397**	77.12	**84.19**	**80.66**	**0.892**

Figure 2 The ROC curve of different groups of features using the neural network classifier



3.4 The performance on TR-265

This set of experiments compares the performance of the extracted motif features with that of sequence features and evolutionary features for the DNA-binding residue prediction by TR-265. TR-265 contains more samples than both the PDNA-62 data set and PDNA-224 data set, thus it is a more comprehensive data set. To get the comprehensive performance of our proposed method CNNsite for predicting DNA-binding residues, we evaluate it on TR-265. The results of CNNsite using various features by using the neural network classifier are listed in Table 4. The results show two similar conclusions as that on the PDNA-62 data set and the PDNA-224 data set: (1) the extracted motif features achieve better performance than both sequence features and evolutionary features, and (2) motif features combined with either sequence features or evolutionary features perform better than sequence features and evolutionary features individually or the combination of them on all metrics. The ROC curves of the three groups of features and their different combinations on TR-265 are shown in Figure 3. It shows that the extracted motif features get better ROC curve than the other two groups of features and the combination of all the three groups of features gets the best ROC curve.

Table 4 The predicting performance on TR-265 for various features by tenfold cross-validation using the neural network classifier

Method	*ACC (%)*	*MCC*	*SN (%)*	*SP (%)*	*ST (%)*	*AUC*
SEQ	88.83	0.217	29.94	93.36	61.65	0.778
EVO	**90.04**	0.281	34.89	**94.29**	64.59	0.810
MOT	84.79	0.385	72.02	85.78	78.90	0.870
SEQ + EVO	87.80	0.323	50.50	90.67	70.58	0.831
MOT + SEQ	84.52	0.405	76.98	85.11	81.05	0.887
MOT + EVO	84.19	0.401	76.94	84.74	80.84	0.888
MOT + SEQ + EVO	84.90	**0.421**	**79.07**	85.35	**82.21**	**0.901**

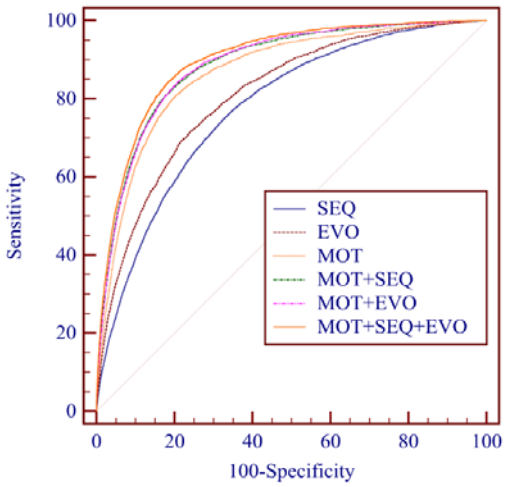
Furthermore, when comparing the performance of motif features for PDNA-62, PDNA-224 and TR-265, we see that the AUC of motif features are 0.871, 0.858 and 0.870 for

PDNA-62, PDNA-224 and TR-265, respectively. In these three data sets, the ratios of the binding residues are 1:3.11:3.96. However, PDNA-62, the one with about one-fourth of data samples against TR-265, has similar performance to TR-265, and it also outperforms PDNA-224 by 0.013 on AUC. This result is against our general intuition that the more the samples are, the better the performance should be. Generally speaking, the interactions of binding residues can be categorised into two classes, the ones for specific interactions and the others are for general interactions. We note that the portions of binding residues involved in specific interactions against the total number of binding residues for the three data sets PDNA-62, PDNA-224 and TR-265 are 55.96%, 22.49% and 30.24% as shown in Table 5. Specific interaction is one mechanism by which proteins and DNAs interact with each other (Travers and Suck, 1993), which are formed by the binding between proteins and specific DNA sequences. The interaction between DNA and transcription factor is a good example of specific interaction (Ptashne, 2005). General interactions are the binding between structure proteins and non-specific DNAs. The interactions between histone and DNA are well-understood general protein–DNA interactions (Dame, 2005; Luger et al., 1997). We hypothesise that the binding residues involved in specific interactions contain specific motif features, while the binding residues involved in general interactions do not include specific motif features. Therefore, the performance of the three data sets is likely to be linked to the portions of the binding residues involved in specific interactions. Note that the portion of binding residues involved in specific interactions for PDNA-62 is higher than that of TR-265. But, TR-265 has much larger sample size. This may explain why they end up with similar performance. In our future work, investigations will conduct to validate this hypothesis.

Table 5 Portions of binding residues involved in specific interactions and general interactions for the three benchmarking data sets

Data set	*PDNA-62*	*PDNA-224*	*TR-265*
Total no. of binding residue	1215	3778	4054
Portion in specific interactions	55.96%	22.49%	30.24%
Portion in general interactions	44.04%	77.51%	69.76%

Figure 3 The ROC curve of different groups of features on TR-265 using the neural network classifier



3.5 Comparison with previous computational methods

This set of experiments compares the performance of CNNsite with that of previous published methods which either used PDNA-62 or PDNA-224. In this experiment, we include seven algorithms in the evaluation: (1) DPS-PRED (Ahmad et al., 2004), DBS-PSSM (Ahmad and Sarai, 2005), (2) BindN (Wang and Brown, 2006), (3) DP-Bind (Kuznetsov et al., 2006), (4) DP-Bind (Hwang et al., 2007), (5) BindN-RF (Wang et al., 2009), (6) BindN+ (Wang et al., 2010), and (7) PreDNA. The first six algorithms were trained and tested on PDNA-62. The last one, PreDNA, was trained and tested on both data sets. PreDNA (Li et al., 2013) was developed by integrating an SVM classifier and a template-based prediction protocol. The SVM classifier was trained by sequence features, evolutionary features and structure features. The template-based prediction protocol is completed by aligning the structure of the target protein–DNA complex with that of the protein–DNA complexes in a template library. Since CNNsite does not use structure features in the prediction, comparison to PreDNA takes its version without using structure features. The prediction performance of CNNsite and other methods on PDNA-62 and PDNA-224 are shown in Tables 6 and 7, respectively.

Table 6 shows that among all the previously published works, BindN+ achieves the best performance (MCC of 0.440, ST of 78.30% and AUC of 0.859) on PDNA-62. CNNsite outperforms BindN+ on all the metrics with increase of 0.069 on MCC, 4.37% on ST and 0.040 on AUC for PDNA-62. Table 7 shows that, when tested on PDNA-224, CNNsite also achieves the best performance (MCC of 0.397, ST of 80.66% and AUC of 0.892) and performs better than PreDNA with increase of 0.107 on MCC, 6.06% on ST for PDNA-224. By comparing the improvement of our proposed CNNsite over previous methods on PDNA-62 and PDNA-224, we observe that the improvement on PDNA-224 is higher than the improvement on PDNA-62. This result may be explained by the fact that the instances in PDNA-224 are much more than that in PDNA-62 and CNNsite can make good use of the large number of training instances to improve its performance.

Table 6 The predicting performance of different computational methods on PDNA-62

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	0.490	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	0.440	77.30	79.30	78.30	0.859
PreDNA	79.40	0.420	76.80	79.70	78.30	–
CNNsite	**80.63**	**0.509**	**85.87**	**79.78**	**82.67**	**0.911**

PDNAsite 846 0.56

Table 7 The predicting performance of different computational methods on PDNA-224

Methods	*ACC (%)*	*MCC*	*SN (%)*	*SP (%)*	*ST (%)*	*AUC*
PreDNA	79.10	0.290	69.50	79.80	74.60	–
CNNsite	**83.68**	**0.397**	**77.12**	**84.19**	**80.66**	**0.892**

3.6 Evaluation on the independent data sets

For the independent data set TS-61, we compare our proposed method with DB-Bind (Hwang et al., 2007). DB-Bind (Hwang et al., 2007) is a web server for predicting DNA-binding sites in a DNA-binding protein from its amino acid sequence. The web server implements three 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 penalised logistic regression. DB-Bind (Hwang et al., 2007) also implements two types of consensus methods. 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.

Since the strict consensus in DB-Bind is very strict, for many residues in proteins, the DP-Bind with the strict consensus cannot provide prediction results for them. Therefore, in the performance comparison on TS-61, we do not conduct the comparison with the strict consensus. The performance of CNNsite trained by PDNA-224 and the different DB-Bind methods is shown in Table 8. From Table 8, we can see that the our method gets the best performance and outperforms DP-Bind with different machine learning methods or consensus methods with 0.02–0.05 on MCC, 2.26–6.48% on ST and 0.038–0.056 on AUC.

Table 8 Performance of CNNsite compared with DP-Bind by independent testing 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	–
CNNsite	77.37	**0.32**	**74.49**	77.61	**76.05**	**0.840**

3.7 Further analysis of the important motif features

The evaluation on PDNA-62, PDNA-224 and TR-265 shows that the motif features captured by CNNsite perform better than sequence features and evolutionary features, indicating that the motif features are more useful for DNA-binding residue prediction than sequence features and evolutionary features. In this section, we analyse the discriminant powers of motif features in the prediction and give an explanation for their usefulness in the prediction. In CNNsite, five sets of motif features with length from 2 to 6 are used. After CNNsite is trained by PDNA-62, the discriminant power of a motif *m* in CNNsite is calculated by the following formula:

$$DP(m) = \sum_{i}^{n} \sum_{j}^{d} Z_j * \mathbf{1}\left(argmax\left(Y_{1,j},\ldots,Y_{n,j}\right) = p_i\right) \tag{16}$$

where **1**(.) is an indicator function, n is the number of positive instances in PDNA-62, d is the number of motif detectors with the same length as motif m, p_i is the position of motif m in positive instance i, and Z_j is the feature value of motif m in instance i (for more entails on Z_j, please refer to formula (8)).

The discriminant powers of all the motif features of two residues are shown as a heat map in Figure 4, where the row represents the first residue of the motif features and the column represents the second residue of the motif features. From Figure 4, we can see that the rows corresponding to the residues *R*, *G*, *K*, *D* and the columns corresponding to the residues *R*, *G*, *K* have large weights, which indicate that residues *R*, *K*, *G* and *D* are the important compositions of these motifs. This finding is consistent with the study by Szilágyi and Skolnick (2006), in which they found that *R*, *A*, *G*, *K* and *D* are important for the formation of protein–DNA interactions. The importance of *R* for the formation of protein–DNA interactions is further confirmed by the work of Sieber and Allemann (1998) which states that *R* can indirectly interact with DNA by interacting with both the phosphate backbone and the carboxylate of E(345). The 15 top motif features with the largest discriminant power for the motif features with length of 2 are shown in Table 9.

Table 9 The top 15 motif features of various length with the largest discriminant power

Length	*2*	*3*	*4*	*5*	*6*
1	KR	RNR	KNWV	NRRRK	SNRRRK
2	GR	RMR	WVSN	KGNRS	KGRRGR
3	GN	RGR	CKGF	TRGRV	VSNRRR
4	GK	RLP	KGFF	GRRGR	VSRGRT
5	NR	RKR	GHRF	TRKRK	TTRKRK
6	EK	KTR	HSPA	RGHRF	KKRRKT
7	KT	HSP	VSNR	KRVRG	GIGNIT
8	RN	LKG	YRPG	VSNRR	YKGNRS
9	RT	TRK	KTRK	SNRRR	KSIGRI
10	KG	ALR	IKNW	RGRVK	MKRVRG
11	GT	IQI	FGKM	KGRRG	RKSIGR
12	IS	DSL	SIGR	KTRGR	GSGNTT
13	DK	RKT	FMKR	RVRGS	NKRMRS
14	TR	MRN	KRMR	KRMRS	SKTRKT
15	SR	RKE	RGHR	SRGRT	KTRGRV

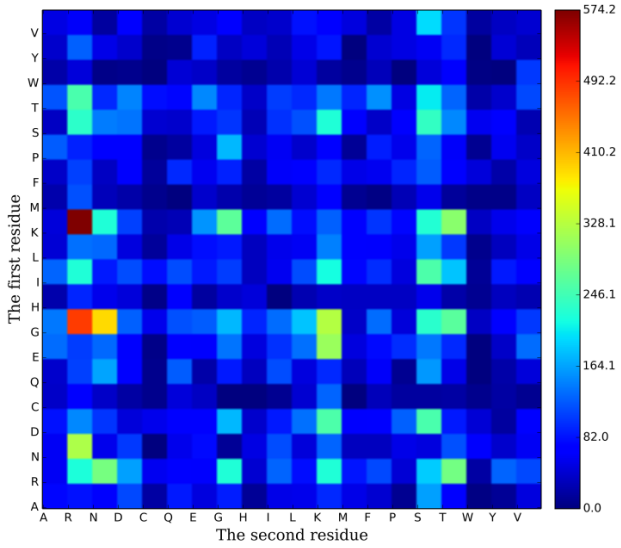
Table 9 shows that *KR*, *GK* and *KG* are three of the motif features with the largest discriminant power. It means that the combinations between *K* and *R* and between *K* and *G* are very important for the interaction between protein and DNA. This conclusion is consistent with the conclusion conducted by Ahmad et al.'s study, in which they concluded that *K* and *R* can enhance *R*'s ability to bind DNA and that the *K* residues within binding regions seem to favour *G* as their immediate neighbour on both sides. Table 9 also shows the 15 top features with the largest discriminant power for the motif features of length from 3 to 6 residues. Note that the proportions of *R*, *K* and *G* in all these motif features are very high. It also indicates that *R* residues are frequently surrounded by residues such as *K* and *R* and *K* residues within binding regions also seem to favour *G* as their immediate neighbour on both sides. The discriminant powers of all motif features with length from 2 to 6 is listed in the part B of addition file 1, which can be obtained from our website http://hlt.hitsz.edu.cn/CNNsite/.

Since *R*, *A*, *G*, *K* and *D* are important for the recognition between protein and DNA, we further examine the proportions of *R*, *A*, *G*, *K* and *D* for the top 15 motif features with the largest discriminant power for length from 2 to 6. The data are shown in Table 10 with the proportions of the five important residues listed individually against the remaining 15 residues in the top 15 features of all the motifs features. Note that the proportion of *R* (144.89%) not only doubles the next residue *K* (77.00%), but also the sum of the four residues *K*, *G*, *D*, *A* (148.55%). It is a clear indication that *R* is the most important feature for the prediction of DNA-binding residues. This is consistent with the findings in the work of Sieber and Allemann (1998). Table 10 also shows that by comparing with residue *R*, *G*, and *K*, the proportions of *A* and *D* are very low. By analysing the electrical properties of the five important residues, we found that residue *R* and *K* are positively charged, residue *G* is neutral and residue *A* and *D* are negatively charged. Since DNA is negatively charged, we think that the negative electricity of residues *A* and *D* may influence the recognition between protein and DNA. Therefore, the electrical properties of the five residues can be used to explain their different proportions in the motif features and different contributions for the prediction of DNA-binding residues. Our conclusion is that the electrical properties of residues are very important factors for the recognition between DNA and protein.

Table 10 The proportions of *R*, *A*, *G*, *K* and *D* in the top 15 motif features with the largest discriminant power for every length

Length	*2 (%)*	*3 (%)*	*4 (%)*	*5 (%)*	*6 (%)*	*Sum (%)*
R	23.33	33.33	16.67	42.67	28.89	144.89
K	20.00	13.33	15.00	12	16.67	77.00
G	16.67	4.44	11.67	16	13.33	62.11
D	3.33	2.22	0.00	0.00	0.00	5.55
A	0.00	2.22	1.67	0.00	0.00	3.89
Others	36.67	44.46	54.99	29.33	41.11	206.56

Figure 4 Heat map to show the discriminant powers of all the motif features of length 2



4 Conclusion and future works

Protein–DNA complexes play crucial roles in gene regulation. The prediction of residues involved in protein–DNA interaction is critical for the understanding of gene regulation. There is an urgent to develop high-performance computational methods for the prediction of DNA-binding residues. Although many methods have been proposed, most of them overlooked the motif features. In this paper, we proposed a deep learning method using the CNN to capture motif features from the sequence around the target residue. A CNN model consists of a set of learnable motif detectors that can capture important motif features by scanning the sequences around target residues using training data. Then we proposed a neural network classifier, referred to as CNNsite, by combining the learned motif features, sequence features and evolutionary features for the prediction of DNA-binding residue. Evaluation on PDNA-62, PDNA-224 and TR-265 shows that motif features perform better than sequence features and evolutionary features by at least 6.73% on ST, 0.097 on MCC and 0.069 on AUC. CNNsite also outperforms state-of-the-art methods with at least 0.019 on MCC, 4.37% on ST and 0.040 on AUC. Evaluation on an independent data set shows that CNNsite also outperforms the previous methods by 0.02–0.05 on MCC, 2.26–6.48% on ST and 0.038–0.056 on AUC for AUC. The discriminant power of the learned motif features with length from 2 to 6 shows that many motif features composed by the residues that are important for the formation of DNA–protein interactions indeed have larger discriminant power. It indicates computationally that *R* is the most important feature for the prediction of DNA-binding residues, which can be explained by its positive charged electrical properties. Evaluation indicates that our proposed CNNsite has very good ability to extract important motif features for DNA-binding residue prediction. The standalone version of the CNNsite is freely available at http://hlt.hitsz.edu.cn/CNNsite/. This work provides computational evidence for the effectiveness of different residues. One direction for future work is to further examine the effect of the difference in the proportions of the binding residues involved in specific interactions and general interaction related to the motif features. Collaboration with biomedical staff may provide more insight to further explain the computational findings of the importance of different residues.

Acknowledgement

This work was supported by the National Natural Science Foundation of China 61370165, U1636103, 61632011, National 863 Program of China 2015AA015405, Shenzhen Foundational Research Funding JCYJ20150625142543470 and Guangdong Provincial Engineering Technology Research Centre for Data Science 2016KF09 and HK Polytechnic University's graduate student grant: PolyU-RUDD.

Reference

Adamczak, R., Porollo, A. and Meller, J. (2004) 'Accurate prediction of solvent accessibility using neural networks based regression', *Proteins*, Vol. 56, pp.753–767.

Adamczak, R., Porollo, A. and Meller, J. (2005) 'Combining prediction of secondary structure and solvent accessibility in proteins', *Proteins*, Vol. 59, pp.467–475.

Ahmad, S., Gromiha, M.M. and Sarai, A. (2004) 'Analysis and prediction of DNA-binding proteins and their binding residues based on composition, sequence and structural information', *Bioinformatics*, Vol. 20, pp.477–486.

Ahmad, S. and Sarai, A. (2005) 'PSSM-based prediction of DNA binding sites in proteins', *BMC Bioinformatics*, Vol. 6, p.33.

Alipanahi, B., Delong, A., Weirauch, M.T. and Frey, B.J. (2015) 'Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning', *Nature Biotechnology*, Vol. 33, pp.831–838.

Breiman, L. (2001) 'Random forests', *Machine Learning*, Vol. 45, pp.5–32.

Bullock, A.N. and Fersht, A.R. (2001) 'Rescuing the function of mutant p53', *Nature reviews: Cancer*, Vol. 1, pp.68–76.

Chen, Y.C., Wright, J.D. and LIM, C. (2012) 'DR_bind: a web server for predicting DNA-binding residues from the protein structure based on electrostatics, evolution and geometry', *Nucleic Acids Research*, Vol. 40, pp.W249–W256.

Ciregan, D., Meier, U. and Schmidhuber, J. (2012) 'Multi-column deep neural networks for image classification', *2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, New York, pp.3642–3649.

Collobert, R. and Weston, J. (2008) 'A unified architecture for natural language processing: deep neural networks with multitask learning', *Proceedings of the 25th International Conference on Machine Learning*, ACM, New York, pp.160–167.

Dame, R.T. (2005) 'The role of nucleoid-associated proteins in the organization and compaction of bacterial chromatin', *Molecular Microbiology*, Vol. 56, pp.858–870.

Dehzangi, A., Sharma, A., Lyons, J., Paliwal, K.K. and Sattar, A. (2014) 'A mixture of physicochemical and evolutionary-based feature extraction approaches for protein fold recognition', *International Journal of Data Mining and Bioinformatics*, Vol. 11, pp.115–138.

Fang, C., Noguchi, T. and Yamana, H. (2015) 'Condensing position-specific scoring matrixs by the Kidera factors for ligand-binding site prediction', *International Journal of Data Mining and Bioinformatics*, Vol. 12, pp.70–84.

Halford, S.E. and Marko, J.F. (2004) 'How do site-specific DNA-binding proteins find their targets?', *Nucleic Acids Research*, Vol. 32, pp.3040–3052.

Ho, S.Y., Yu, F.C., Chang, C.Y. and Huang, H.L. (2007) 'Design of accurate predictors for DNA-binding sites in proteins using hybrid SVM-PSSM method', *Biosystems*, Vol. 90, pp.234–241.

Holmes, G., Donkin, A. and Witten, I.H. (1994) 'Weka: a machine learning workbench', *Proceedings of the Second Australia and New Zealand Conference on Intelligent Information Systems*, Brisbane, Australia.

Hwang, S., Gou, Z. and Kuznetsov, I.B. (2007) 'DP-Bind: a web server for sequence-based prediction of DNA-binding residues in DNA-binding proteins', *Bioinformatics*, Vol. 23, pp.634–636.

Jones, S., Barker, J.A., Nobeli, I. and Thornton, J.M. (2003) 'Using structural motif templates to identify proteins with DNA binding function', *Nucleic Acids Research*, Vol. 31, pp.2811–2823.

Jones, S., Heyninge, P., Berman, H.M. and Thornton, J.M. (1999) 'Protein-DNA interactions: a structural analysis', *Journal of Molecular Biology*, Vol. 287, pp.877–896.

Kang, H., Oh, E., Choi, G. and Lee, D. (2010) 'Genome-wide DNA-binding specificity of PIL5, an Arabidopsis basic Helix-Loop-Helix (bHLH) transcription factor', *International Journal of Data Mining and Bioinformatics*, Vol. 4, pp.588–599.

Krizhevsky, A., Sutskever, I. and Hinton, G.E. (2012) 'Imagenet classification with deep convolutional neural networks', *Advances in Neural Information Processing Systems*, Lake Tahoe, Nevada, pp.1097–1105.

Kuznetsov, I.B., Gou, Z., Li, R. and Hwang, S. (2006) 'Using evolutionary and structural information to predict DNA-binding sites on DNA-binding proteins', *Proteins*, Vol. 64, pp.19–27.

Lawrence, S., Giles, C.L., Tsoi, A.C. and Back, A.D. (1997) 'Face recognition: a convolutional neural-network approach', *IEEE Transactions on Neural Networks*, Vol. 8, pp.98–113.

Li, T., Li, Q.Z., Liu, S., Fan, G.L., Zuo, Y.C. and Peng, Y. (2013) 'PreDNA: accurate prediction of DNA-binding sites in proteins by integrating sequence and geometric structure information', *Bioinformatics*, Vol. 29, pp.678–685.

Liu, B., Liu, F., Fang, L., Wang, X. and Chou, K.C. (2015) 'repDNA: a Python package to generate various modes of feature vectors for DNA sequences by incorporating user-defined physicochemical properties and sequenceorder effects', *Bioinformatics*, Vol. 31, pp.1307–1309.

Luger, K., Mäder, A.W., Richmond, R.K., Sargent, D.F. and Richmond, T.J. (1997) 'Crystal structure of the nucleosome core particle at 2.8 Å resolution', *Nature*, Vol. 389, pp.251–260.

Ma, X., Guo, J., Liu, H.D., Xie, J.M. and Sun, X. (2012) 'Sequence-based prediction of DNA-binding residues in proteins with conservation and correlation information', *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, Vol. 9, pp.1766–1775.

Ma, X., Wu, J.S., Liu, H.D., Yang, X.N., Xie, J.M. and Sun, X. (2009) 'SVM-based approach for predicting DNA-binding residues in proteins from amino acid sequences', *International Joint Conference on Bioinformatics, Systems Biology and Intelligent Computing, 2009. IJCBS'09*, IEEE, Washington, DC, pp.225–229.

Mcguffin, L.J., Bryson, K. and Jones, D.T. (2000) 'The PSIPRED protein structure prediction server', *Bioinformatics*, Vol. 16, pp.404–405.

Ofran, Y., Mysore, V. and Rost, B. (2007) 'Prediction of DNA-binding residues from sequence', *Bioinformatics*, Vol. 23, pp.i347–i353.

Ozbek, P., Soner, S., Erman, B. and Haliloglu, T. (2010) 'DNABINDPROT: fluctuation-based predictor of DNA-binding residues within a network of interacting residues', *Nucleic Acids Research*, doi:10.1093/nar/gkq396.

Pabo, C.O. and Sauer, R.T. (1984) 'Protein-DNA recognition', *Annual Review of Biochemistry*, Vol. 53, pp.293–321.

Ponting, C.P., Schultz, J., Milpetz, F. and Bork, P. (1999) 'SMART: identification and annotation of domains from signalling and extracellular protein sequences', *Nucleic Acids Res*, Vol. 27, pp.229–232.

Ptashne, M. (2005) 'Regulation of transcription: from lambda to eukaryotes', *Trends in Biochemical Sciences*, Vol. 30, pp.275–279.

Sakar, C.O., Kursun, O., Seker, H. and Gurgen, F. (2014) 'Combining multiple clusterings for protein structure prediction', *International Journal of Data Mining and Bioinformatics*, Vol. 10, pp.162–174.

Schaffer, A.A., Aravind, L., Madden, T.L., Shavirin, S., Spouge, J.L., Wolf, Y.I., Koonin, E.V. and Altschul, S.F. (2001) 'Improving the accuracy of PSI-BLAST protein database searches with composition-based statistics and other refinements', *Nucleic Acids Research*, Vol. 29, pp.2994–3005.

Si, J., Zhang, Z., Lin, B., Schroeder, M. and Huang, B. (2011) 'MetaDBSite: a meta approach to improve protein DNA-binding sites prediction', *BMC Systems Biology*, Vol. 5, doi:10.1186/1752-0509-5-S1-S7.

Sieber, M. and Allemann, R.K. (1998) 'Arginine (348) is a major determinant of the DNA binding specificity of transcription factor E12', *Biological Chemistry*, Vol. 379, pp.731–735.

Simard, P.Y., Steinkraus, D. and Platt, J.C. (2003) 'Best practices for convolutional neural networks applied to visual document analysis', *ICDAR*, Vol. 3, pp.958–962.

Szilágyi, A. and Skolnick, J. (2006) 'Efficient prediction of nucleic acid binding function from low-resolution protein structures', *Journal of Molecular Biology*, Vol. 358, pp.922–923.

Tjong, H. and Zhou, H.X. (2007) 'DISPLAR: an accurate method for predicting DNA-binding sites on protein surfaces', *Nucleic Acids Research*, Vol. 35, pp.1465–1477.

Travers, A. and Suck, D. (1993) *DNA–Protein Interactions*, Chapman & Hall, London.

Wagner, M., Adamczak, R., Porollo, A. and Meller, J. (2005) 'Linear regression models for solvent accessibility prediction in proteins', *Journal of Computational Biology*, Vol. 12, pp.355–369.

Wang, L. and Brown, S.J. (2006) 'BindN: a web-based tool for efficient prediction of DNA and RNA binding sites in amino acid sequences', *Nucleic Acids Research*, Vol. 34, pp.W243–W248.

Wang, L., Huang, C., Yang, M.Q. and Yang, J.Y. (2010) 'BindN+ for accurate prediction of DNA and RNA-binding residues from protein sequence features', *BMC Systems Biology*, Vol. 4, p.S3.

Wang, L., Yang, M.Q. and Yang, J.Y. (2009) 'Prediction of DNA-binding residues from protein sequence information using random forests', *BMC Genomics*, Vol. 10, p.S1.

Wang, S., Peng, J., Ma, J. and Xu, J. (2016) 'Protein secondary structure prediction using deep convolutional neural fields', *Scientific Reports*, Vol. 6, doi:10.1038/srep18962.

Wu, J.S., Liu, H.D., Duan, X.Y., Ding, Y., Wu, H.T., Bai, Y.F. and Sun, X. (2009) 'Prediction of DNA-binding residues in proteins from amino acid sequences using a random forest model with a hybrid feature', *Bioinformatics*, Vol. 25, pp.30–35.

Yan, C., Terribilini, M., Wu, F., Jernigan, R.L., Dobbs, D. and Honavar, V. (2006) 'Predicting DNA-binding sites of proteins from amino acid sequence', *BMC Bioinformatics*, Vol. 7, p.262.

Zhou, J., Lu, Q., Xu, R., Gui, L. and Wang, H. (2016) 'CNNsite: prediction of DNA-binding residues in proteins using convolutional neural network with sequence features', *2016 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, IEEE, Shenzhen, China, pp.78–85.

Zhu, X., Ericksen, S.S. and Mitchell, J.C. (2013) 'DBSI: DNA-binding site identifier', *Nucleic Acids Research*, doi:10.1093/nar/gkt617.