CNNsite: Prediction of DNA-binding Residues in Proteins Using Convolutional Neural Network with Sequence Features

Jiyun Zhou*†, Qin Lu†, Ruifeng Xu*‡, Lin Gui*, Hongpeng Wang*
*School of Computer Science and Technology, Harbin Institute of Technology Shenzhen Graduate School, Shenzhen, China
†Department of Computing, Hong Kong Polytechnic University, Hung Hom, Hong Kong
Email addresses: zhoujiyun2010@gmail.com, csluqin@comp.polyu.edu.hk, xuruifeng@hitsz.edu.cn, guilin.nlp@gmail.com, wanghp@hit.edu.cn
‡Corresponding author

***Abstract*—Abstract‖Protein-DNA complexes play crucial roles in gene regulation. The prediction of the residues involved in protein-DNA interactions is critical for understanding gene regulation. Although many methods have been proposed, most of them overlooked motif features. Motif features are sub sequences and are important for the recognition between a protein and DNA. In order to efficiently use motif features for the prediction of DNA-binding residues, we first apply the Convolutional Neural Network (CNN) method to capture the motif features from the sequences around the target residues. CNN modeling consists of a set of learnable motif detectors that can capture the important motif features by scanning the sequences around the target residues. Then we use a neural network classifier, referred to as CNNsite, by combining the captured motif features, sequence features and evolutionary features to predict binding residues from sequences. The datasets PDNA-62 and PDNA-224 are used to evaluate the performance of CNNsite by five-fold cross-validation. Performance evaluation shows that the motif features performs better than sequence features and evolutionary features with at least 6.73% on ST, 0.097 on MCC and 0.069 on AUC. When comparing with previously published methods, CNNsite performs better with at least 0.019 on MCC, 4.37% on ST and 0.040 on AUC. CNNsite is also evaluated on an independent dataset TS-72 and CNNsite outperforms the previous methods by at least 0.012 on AUC. The discriminant powers of the motif features of size from 2 to 6 residues show that many motif features with large discriminant power are composed by the residues that play important roles in the DNA-protein interactions. The standalone version of the CNNsite is available at http://hlt.hitsz.edu.cn:8080/CNNsite/.**

I. Introduction

DNA-binding proteins are the proteins composed of DNA-binding domains. The interactions between these proteins and DNA play crucial roles in vital biological processes, including transcriptional regulation, DNA modification, DNA replication, DNA repair, DNA packing and DNA recombination. There are two mechanisms in the interactions between DNA-binding proteins and DNA: general interaction and specific interaction [1]. The interactions between histone and DNA are well-understood examples of general protein-DNA interactions [2]. Transcription factors are the most intensively studied DNA-binding proteins which form interactions with DNA by specific interactions. They activate or inhibit the transcription of genes by their interactions with the particular DNA sequences close to their promoters. Thus, the prediction of the residues involved in protein-DNA interactions is important for understanding the gene regulation process [3]. Many experimental methods were developed to identify DNA-binding residues from protein sequences, including electrophoretic mobility shift assays (EMSAs) [4], nuclear magnetic resonance (NMR) spectroscopy [5],and conventional chromatin immunoprecipitation (ChIP) [6]. However, they are costly and time-consuming. With the development of sequencing technology of protein, more and more DNA-binding proteins are sequenced. Thus, there is urgent need to propose computational methods that can automatically predict the DNA-binding residues on the genome scale.

So far, many computational methods have been proposed for the prediction of DNA-binding residue. Based on the features used for the prediction, these methods can be divided into three groups: evolutionary features based methods, sequence features based methods, and the ones based on structure features. Position specific score matrix (PSSM) is a common representation of evolutionary features and has been used in many bioinformatics problems. Ahmad et al. first used PSSM for the prediction of DNA-binding residue in proteins and got good performance on dataset PDNA62 [7]. Since then, PSSM was used by many works for the prediction of DNA-binding residue, for example, Wang et al. [8] proposed a prediction method, referred to as BindN+, by integrated PSSM and three physiochemical properties including hydrophobicity index, side chain pKa value and molecular mass. Other methods trained by PSSM include SVMCPSSM [9], Disis [10], BindN-RF [11], Ma et al.s SVM classifier [12], DBindR [13] and DNABR [14]. Sequence features, being very important features, include amino acid composition, predicted structure features and physiochemical properties. Sequence features are often used through combination with the evolutionary features. For example, Ofran et al. [10] proposed the predictor Disis by combining PSSM features and the local amino acid composition of its neighbors and Ma et al.[23] proposed DNABR by

978-1-5090-1610-5/16/$31.00 ©2016 IEEE

combining PSSM features and six physichemical properties including the pKa values of amino group, the pKa values of carboxyl group, the electron-ion interaction potential (EIIP), the number of lone electron pairs(LEPs), Wiener index and the molecular mass.

The functions of proteins are often affected by its structure features. Sequence features and evolutionary features are not sufficent for the prediction of DNA-binding residue. Thus more and more methods have incorporated structure features. The frequently used structure features include secondary structure, solvent accessible surface area, spatial neighbors, B-factor, protrusion index and depth index. Ahmad et al. [15] first proposed a SVM classifier, referred to as dbs-pred, by combining solvent accessibility surface area and sequence features. Kuznetsov et al. [16] developed a SVM classifier, referred to as dp-bind, by incorporating secondary structure and its spatial neighbors. Tsuchiya et al. [17] further proposed a prediction method based on the empirical preference of electrostatic potential and the shape of molecular surfaces. Tjong et al. [18] built a neural network predictor DISPLAR by utilizing solvent accessibility surface area and sequence features. There are also many other works in which structure features are used for the prediction [19]. Motif features has been indicated to be useful for the prediction of TF binding site in DNA, but it is rarely used for the prediction of DNA-binding residue. In this work, we first apply the Convolutional Neural Network (CNN) [20] to collect the important motif features from the training dataset and investigate their effects for the DNA-binding residue prediction. We also investigate its combined use with other types of features. In this work, we propose a neural network classifier, referred to as CNNsite, by combining the motif features captured by CNN, the sequence features and the evolutionary features. We did not consider structure features because they are not readily available for most of the proteins.

II. FEATURE EXTRACTION AND PROPOSED METHOD

A. Residue-wise data instances

n the problem of DNA-binding residue prediction, residue-wise data instances are the prediction targets. Since the biological function of a target residue is often influenced by its neighboring residues, a residue-wise data instance is defined as a window of length w with the target residue positioned in the middle and $(w-1)/2$ neighboring residues on either side. A residue-wise data instance is defined as a positive sample if the central residue is a DNA-binding residue or a negative sample if the central residue is a non-binding residue.

Given a protein sequence P with length of L formulated as

$$P = R_1 R_2 R_3 R_4 R_5 R_6 \dots R_{i-1} R_i R_{i+1} \dots R_L \qquad (1)$$

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

$$S_i = R_{i-\frac{w-1}{2}} R_{i-\frac{w-3}{2}} \dots R_{i-1} R_i R_{i+1} \dots R_{i-\frac{w-1}{2}} \qquad (2)$$

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 right side are termed as the left contextual residues and right contextual residues, respectively.

B. Feature Descriptors

Sequence features, evolutionary features and structure features are three kinds of commonly used features for the prediction of DNA-binding residue. Since the structure features for most of proteins are unavailable, methods using structure features cannot be applied on a genome-wide scale. Therefore, in this paper, we only use sequence features, evolutionary features and motif features for the prediction of DNA-binding residue.

Sequence features, denoted by SEQ, contain local amino acid composition, predicted second structure and predicted solvent accessible area. The local amino acid composition has components: amino acid composition over the left contextual residues and that over the left contextual residues. The predicted second structure for a residue is coded as a one-hot vector. The predicted second structure for a residue-wise data instance is represented by concatenating the one-hot vectors of all the residues in it. The predicted secondary structure and predicted solvent accessible area are obtained by PSIPRED [21] and SABLE [22], respectively.

PSSM, being a very important type of evolutionary features, denoted by EVO, is obtained by running the PSI-BLAST [23] program to search against the non-redundant (NR) database through three iterations with 0.001 as the E-value cutoff for multiple sequence alignment. In PSSM, there are 20 scores for each sequence position and each score means the conservation degree of a specific residue type on that position. Before feeding into a prediction engine, all the scores in PSSM need to be scaled between 0 and 1 using the following equation.

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

For every data instance, all the scaled scores in the PSSM are used as its evolutionary features.

Motif features, denoted by MOT, are sub sequences and extracted by the CNN algorithm introduced in the following subsection.

C. Convolutional Neural Network (CNN)

CNN is a type of feed-forward Artificial Neural Network in which the individual neurons are arranged in such a way that they respond to overlapping regions tiling the visual field, which has been applied in many fields, such as image and video recognition, recommender systems and natural language processing. In recent years, CNN has been introduced into bioinformatics problems to learn the protein sequence representation for the prediction of protein structure and function. For example, Alipanahi et al. [24] developed DeepBind for the prediction of the sequence specificities of DNA- and RNA-binding proteins. Moreover, Wang et al. [25] proposed a machine learning method DeepCNF for protein

second structure prediction by using CNN. In this work, we propose a novel method to identify important motif features from the sequences around the binding residues for DNA-binding residue prediction based on CNN and then develop a neural network classifier, referred to as CNNsite, by combining the important motif features, the sequence features and the evolutionary features.

The frame diagram of CNNsite is shown in Fig.1. CNNsite comprises four computational layers: the convolution layer, the rectification layer, the pooling layer and the neural network layer. In our prediction task, the first three layers can discover important motifs of the inputting residue-wise data instances and the last layer is used to get the prediction results. The convolution, rectification, and network layers have trainable motif detectors D, thresholds b, and weights W, respectively. For a residue-wise data instance S, CNNsite produces a real-valued score $f(S)$ for prediction by the following formula

$$f(S) = net_w(pool(rect_b(conv_D(S)))) \tag{4}$$

where $conv_D()$, $rect_b()$, $pool()$ and $net_W()$ denote the four layer in CNNsite, respectively

Convolution layer. In the convolution layer, 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 the motif scan operation in a PWM or a PSAM-based model. For a motif detector of size m, the residue-wise data instance S should be padded by concatenating $(m-1)$ unuseful residues on either sides. The padded sequence of S is represented as a matrix M in the following way:

$$M_{i,j} = \begin{cases} 0.5 & \text{if } i < m \text{ or } i > n-m \\ 1 & \text{if } S_{i-m+1} = j^{th} \text{ base} \\ 0 & \text{if otherwise} \end{cases} \tag{5}$$

where m is the size of the motif detector and n is the length of the residue-wise data instance. The output of the convolution layer is a matrix X where its element $X_{i,k}$ is essentially the score of the motif detector k aligned to position i of the padded sequence M. Given that the motif detectors are represented as an array D, where element $D_{k,j,l}$ is the coefficient of the motif detector k at motif position j and base l, the element $X_{i,k}$ of the output is calculated by the following formula

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

So, the column $X_{.,k}$ is the motif scan of motif detector k applied to the padded sequence M and row $X_{i,.}$ is the motif scan of all the motif detectors on position i of the padded sequence M.

Rectification layer. Rectification layer plays an important role in the deep learning. Its input is the matrix outputted by the convolution layer. The output $Y = rectb(X)$ is an matrix of the same size as X

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

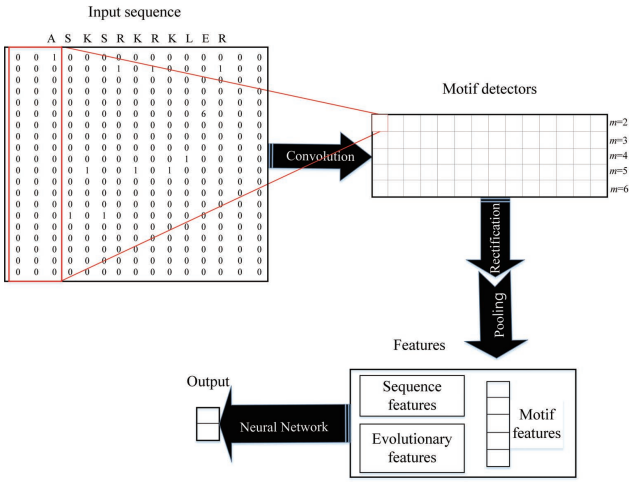

Fig. 1. The frame diagram of CNNsite.

where b_k is the activation threshold for the motif detector k, which is learned in the training process of CNNsite. The formula means that if score $X_{i,k}$ is greater than b_k, the relative score of the motif detector k at position i is passed to the next stage; Otherwise the motif detector k is deemed irrelevant at position i and so the relative score is zero. This layer is used to filter the unimportant motif features and keep only the motif features with scores larger than a specified threshold.

Pooling layer. The input of this layer is the matrix Y output by the rectification layer. Its output Z is a feature vector, of which the dimension depends on the number of motif detectors in the convolution layer. The pooling layer for a motif detector k $(1 \le k \le d)$ is formulated as

$$Z_k = max(Y_{1,k}, \cdots, Y_{n,k}) \tag{8}$$

For every residue-wise data instance, we can obtain a vector Z with dimension of d, where d is the number motifs used in the Convolution layer. The features contained in the vector Z are motif features captured by the d motif detectors in the convolution layer.

Neural network layer. The neural network layer is used for prediction. In this layer, three kinds of features are used as input: the motif features from the pooling layer, the sequence features, and the evolutionary features. In order to avoid CNNsite from overfitting, we use the recently proposed dropout technique before the hidden layer in the neural network layer. With the dropout technique, the entries of hidden representations are set to 0 with a dropout rate, which is tuned based on the development set.

III. EXPERIMENTS AND RESULTS

The purpose of the evaluation is to examine the effectiveness of the CNNsite for the prediction of DNA-binding residue. Since CNNsite uses a window based approach, the window size needs to be set properly. Due to the length of this paper, we skipped the parameter tuning and all the results shown in this section use the window size $w = 11$ that is the context size is 5 on both the left and right side of the window.

TABLE I
THE DETAILS OF THE THREE DATASETS

datasets	PDNA-62	PDNA-224	TS-72
binding residues	1,215	3,778	1,040
non-binding residues	6,948	53,570	13,226

Four sets of evaluations are conducted. The first set evaluates the performance of CNNsite with different combinations of the three kinds of features on PDNA-62. The second set evaluates the performance of CNNsite with different combinations of the three kinds of features on PDNA-224. The third one uses the datasets PDNA-62 and PDNA-224 to compare our CNNsite with previous published predictors. And the last one evaluates CNNsite on an independent test TS-72 compared with previous published methods.

A. Datasets

In order to objectively evaluate the performance of CNNsite for the prediction of DNA-binding residue, three datasets are used in this work.

PDNA-62 contains 67 protein chains from 62 protein-DNA complexes from the Protein Data Bank (PDB) [21]. The sequence identity between any two chains is less than 25%. PDNA-62 was first proposed by Ahmad et al. [15] and has been used by many predictors for the prediction of DNA-binding residue. PDNA-224 is a larger dataset recently proposed by Li et al. [26] from PDB. It contains 224 protein chains from 224 protein-DNA complexes and the sequence identity between any two chains from it is less than 25%. To compare with the predictors which are trained by different datasets from our training dataset, an independent dataset TS-72 is used in this work. TS-72 is proposed by Ma et al. [14] and used to compare DNABR with other three predictors including BindN [22], BindN-RF [11] and BindN+ [8]. This dataset contains 72 protein chains from 59 protein-DNA complexes and the sequence identity between any two chains is less than 25%. For these three datasets, a residue in the structure of the protein-DNA complexes is defined as a DNA-binding residue if its side chain or backbone atoms fall within a cutoff distance of 3.5$\mathring{A}$ from any atom of the nucleotides in DNA and the remaining residues are defined as the non-binding residues [8, 26]. Details of these three datasets are shown in TABLE I.

B. Evaluation metrics

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

$$SN = TP/(TP + FN) \tag{9}$$

$$SP = TN/(TN + FP) \tag{10}$$

$$ST = (SN + SP)/2 \tag{11}$$

$$ACC = (TP + TN)/(TP + FP + TN + FN) \tag{12}$$

$$MCC = \frac{TP * TN - FP * FN}{\sqrt{P * (TP + FP) * N * (TN + FN)}} \tag{13}$$

where TP is the number of true positives; TN is the number of true negatives; FP is the number of false positives; FN is the number of false negatives; P is the number of positives; N is the number of negatives.

Since the number of binding residues and that of non-binding residues in the datasets are unbalanced, ACC often provides very biased evaluating performance. Literatures [8, 26] have reported that the average of SN and SP can provide an appropriate evaluation for predictors when the dataset are unbalanced. Moreover, MCC can measure the matching degree between the prediction results and the real results. So ST and MCC are used as the main metrics while the remaining metrics are used only for references. Receiver Operating Characteristic ROC curve is a graphical plot that illustrates the performance of a binary classifier system as its discrimination threshold is varied. The curve is drawn by plotting the true positive rates (i.e. sensitivity) against the false positive rates (i.e. 1-specificity) by changing the classification threshold for predictors. AUC is the area under the ROC curve and is a fair metric for unbalanced problem.

C. The predicted results on PDNA-62

This set of experiments examines the contributions of the three different kinds of features in CNNSite for the DNA-binding residue prediction on PDNA-62. The performance is shown in TABLE II. As mentioned earlier, MCC, ST and AUC are the main metrics. Thus we shade the best performers of these three metrics for easy observation. It can be seen that the motif features achieve 0.459 for MCC, 80.12% for ST and 0.871 for AUC, outperforming the sequence features by 0.114 for MCC, 7.51% for ST and 0.101 for AUC and performs better than the evolutionary features with 0.097 for MCC, 6.73% for ST, 0.069 for AUC. It indicates that the motif features are more useful than the sequence features and the evolutionary features. When the motif features are combined with the sequence features, its performance is improved on all metrics with 0.014 for MCC, 0.97% for ST and 0.018 for AUC. When the motif features are combined with the evolutionary features, its performance is improved with 0.017 for MCC, 1.03% for ST and 0.026 for AUC. When the three kinds of features are combined, CNNsite achieves 0.509 for MCC, 82.67% for ST and 0.911 for AUC, outperforming other combinations with 0.033-0.164 for MCC, 1.52-10.06% for ST and 0.014-0.141 for AUC. Fig.2 also shows that the motif features gets better ROC curve than the sequence features and the evolutionary features and the combination of them gets the best ROC curve. It indicates that the motif features, the sequence features and the evolutionary features are complementary for each other.

TABLE II
THE PREDICTION PERFORMANCE ON PDNA-62 FOR VARIOUS FEATURES BY TEN-FOLD CROSS-VALIDATION

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
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
ALL	80.63	0.509	85.87	79.78	82.67	0.911

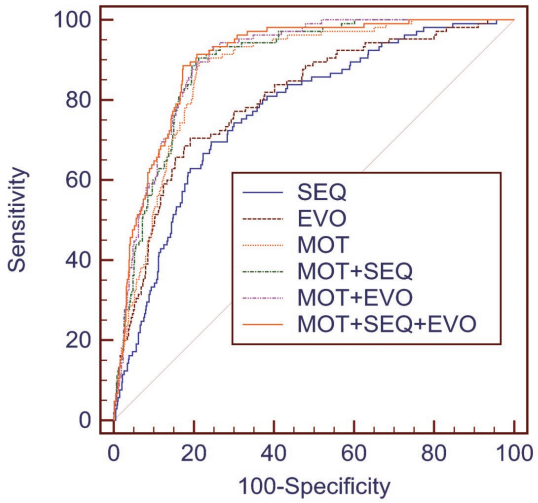

Fig. 2. ROC curves of CNNsite with different combination of features on PDNA-62.

D. The predicted results on PDNA-224

This set of experiments examines the contribution of the three kinds of features in CNNSite for the DNA- binding residue prediction on PDNA-224.

To further evaluate the performance of our proposed method CNNsite in predicting DNA-binding residues, we evaluate it on the recently proposed dataset PDNA-224. The results of CNNsite using various features are listed in TABLE III. Results show that the motif features achieve 0.367 for MCC, 78.38% for ST and 0.858 for AUC, performing better than the sequence features with 0.145 for MCC, 15.55% for ST and 0.102 for AUC and the evolutionary features with 0.116 for MCC, 14.99% for ST and 0.078 for AUC. When the motif features are combined with the sequence features, the performance is increased by 0.02 for MCC, 1.61% for ST and 0.011 for AUC. When the motif features are combined with the evolutionary features, the performance increases by 0.014 for MCC, 1.69% for ST and 0.014 for AUC. On this dataset, the best result (MCC of 0.397, ST of 80.66% and MCC of 0.397) is obtained when the three kinds of features are combined. It performs better than other combinations with 0.016-0.175 MCC, 0.59-17.83% ST and 0.02-0.136 AUC.

Although the combination of the motif features and the evolutionary features achieves higher value than the combination of the three kinds of features for SN, its SP value is lower than the latter. Fig.3 also shows that the motif features get better ROC curve than that of the sequence features and the evolutionary features. The combination of all the three features get the best ROC curve. The results on this dataset also indicate that the motif features are more useful than the sequence features and the evolutionary features for the prediction of DNA-binding residue and that these three kinds of features are complementary to each other in CNNsite.

TABLE III
THE PREDICTING PERFORMANCE ON PDNA-224 FOR VARIOUS FEATURES BY TEN-FOLD CROSS-VALIDATION

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
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
ALL	83.68	0.397	77.12	84.19	80.66	0.892

TABLE IV
THE PREDICTING PERFORMANCE COMPARED WITH OTHER COMPUTATIONAL METHODS ON PDNA-62

Method	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

E. Comparison with previous computational methods

This set of experiments evaluates the performance of our proposed CNNsite compared with previous published methods which have been trained and tested either on PDNA-62 or PDNA-224. Many predicting algorithms including Dps-pred [15], Dbs-pssm [7], BindN [27], Dp-bind [16], Dp-Bind [28], BindN-RF [11], BindN+ [8] and PreDNA [26] have been proposed for the prediction of DNA-binding residue, in which the former seven methods were trained and tested on PDNA-62 and the last one, PreDNA, was trained and tested on both data sets. PreDNA [26] was developed by integrating a SVM classifier and a template-based prediction protocol. The SVM classifier was trained by sequence information, evolutionary information and structure information. The template-based prediction protocol is completed by aligning the structure of the current protein-DNA complex and that in template library. Since CNNsite do not use any structure features for prediction, to compare the prediction performance of various methods fairly, we only consider PreDNA without using any structure features. The prediction performance of CNNsite and other methods on PDNA-62 and PDNA-224 are shown in TABLE IV and TABLE V, respectively. Since the performance of the existing methods is cited from their published papers, the values of some metrics are not known. TABLE IV shows that BindN+ achieves the best performance (MCC of 0.440, ST of 78.30% and AUC of 0.859) on PDNA-62 among the previous published methods. Among all the prediction

TABLE V
THE PREDICTING PERFORMANCE COMPARED WITH OTHER COMPUTATIONAL METHODS ON PDNA-224

Method	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

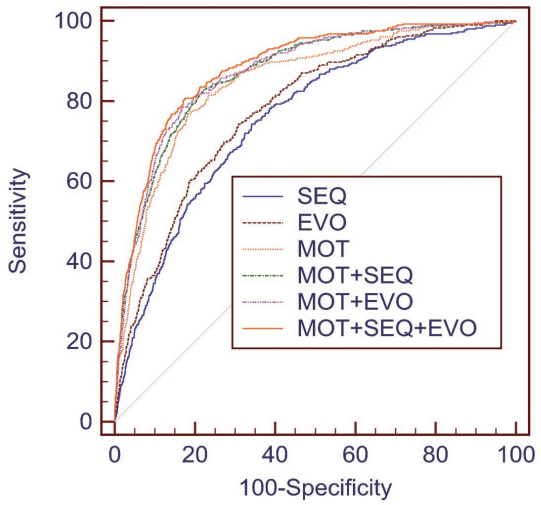

Fig. 3. ROC curves of CNNsite with different combinations of features on PDNA-224.

methods, CNNsite achieves the best performance (MCC of 0.509, ST of 82.67% and AUC of 0.911) outperforming the BindN+ on all the metrics with 0.069 on MCC, 4.37% on ST and 0.040 on AUC for PDNA-62.

TABLE V shows that when testing 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 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 on PDNA-224, we observe that the improvement on PDNA-224 is higher than the improvement on PDNA-62. This may be explained by the fact that PDNA-224 has more training data and CNN can make good use of the large number of training instances to improve its performance.

F. Independent test

This set of experiments evaluates the performance of CNNsite on the independent test TS-72. Since the performance on PDNA-62 and PDNA-224 are obtained by applying the ten-fold cross-validation and the test set and training set in the cross validation are drawn from the same population, the evaluating performances may need to be further justified. Moreover, there are also some other predicting methods which are no evaluated on PDNA-62 and PDNA-224. Therefore, to evaluate CNNsite more objectively, we conduct an experiment on the independent dataset TS-72. TS-72 is an independent dataset proposed by Ma et al. [14] to compare the performance of DNABR with that of three other predictors including BindN [27], BindN-RF [11] and BindN+ [8]. DNABR is a sequence based DNA-binding residue prediction method. BindN, BindN-RF and BindN+ are three methods proposed by using only sequence information. In this work, we use TS-72 to compare the performance of CNNsite with these four predictors. The AUC values of CNNsite, DNABR, BindN, BindN-RF, and BindN+ for TS-72 are 0.878, 0.866, 0.748, 825 and 0.844, respectively, where the AUC values of those four previous methods are reported in Ma et al.work [14]. In summary, our method increases the performance by 0.012-0.130 with at least 3.36E-5 of p-value (n = 20,1-tailed, one-sample t-test) on AUC for TS-72.

G. Further analysis of the important motif features

The evaluation on PDNA-62 and PDNA-224 shows that the motif features captured by CNNsite perform better than the sequence features and the evolutionary features, indicating that the motif features are more effective for DNA-binding residue prediction than the sequence features and the evolutionary features. In this section, we analyze discriminant powers of the motif features in CNNsite and give an explanation for the effectiveness of motif features for the prediction of DNA-binding residue. In the convolution layer of CNNsite, the raw input is convolved with many motif detectors. In CNNsite, 5 sets of motif detectors of length from 2 to 6 are used and every set contains 500 motif detectors. After CNNsite is trained by PDNA-62, the discriminant power of a motif t in CNNsite is calculated by the following formula

$$DP(t) = \sum_{i}^{p} \sum_{j}^{d} f_{i,j}(t) \tag{14}$$

$$f_{i,j}(t) = \begin{cases} Z_j & \text{if } argmax(Y_{1,j}, \cdots, Y_{n,j}) = pos \\ 0 & others \end{cases} \tag{15}$$

where p is the number of positive instances in PDNA-62; d is the number of motif detectors of the same length as motif t; and Z_j is the feature value of motif t for the motif detector j(for more entails about Z_j, refer to formula (8)); pos is the position of motif t in the instance i;.

The 15 top motif features with the largest discriminant power are shown in TALBE VI. For the motif features of 2 residues, TALBE VI shows that KR, GR, GN, GK, NR, EK, KT, RN, RT and KG are the top ten motif features. We find that the residues R, K, G are the important compositions of these motifs. This finding is consistent with the study of Szilgyi and Skolnick [29], 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 Sieber and Allemanns work [30] which states that R can indirectly interact with DNA by interacting with both the phosphate backbone and the carboxylate of E(345). Since these residues are important for the formation of protein-DNA interactions, we speculate that they often occur in the context of the DNA-binding residues and their occurrences are important features for prediction.

TABLE VI
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 VII
THE PROPOSITION OF R,A,G,K AND D IN THE TOP 15 MOTIF FEATURES OF VARIOUS RESIDUES WITH THE LARGEST DISCRIMINANT POWER

length	2(%)	3(%)	4(%)	5(%)	6(%)
A	0.00	2.22	1.67	0.00	0.00
G	16.67	4.44	11.67	16	13.33
K	20.00	13.33	15.00	12	16.67
D	3.33	2.22	0.00	0.00	0.00
Others	36.67	44.44	55.00	29.33	41.11

In the 15 top motif features of more than 2 residues with the largest discriminant power, most of them also contain these residues with high proportions. Motif features of 3 residues contain RNR, RMR, RGR, RKR and KTR, motif features of 4 residues contain CKGF, GHRF, FMKR, KRMR and RGHR, motif features of 5 residues contain NRRRK, KGNRS, GRRGR, TRKRK and SNRRRK, and motif features of 6 contain SNRRRK, KGRRGR, VSNRRR, VSRGRT and KKRRKT. It can be seen that the proportions of R, K and G in all these motif features are very high. The discriminant powers of all motif features of number residues from 2 to 6 is listed in the support information S1, which is an attached support information file of this paper and can be downloaded from our website.

The proportions of R, A, G, K and D in the top 15 motif features with the largest discriminant power are shown in TABLE VII. It can be seen that motif features of 5 residues have the highest proportion (78.67%) as the important residues, indicating that the motif features of 5 residues are more useful for DNA-binding residue prediction than other motif features. By observing the proportions of the five important residues separately, we found that the proportion of R is higher than that of other four important residues in all motifs features. It indicates that R is the most important feature for the formation of DNA-binding residues in protein chains, which is consistent with the findings in Sieber and Allemanns work [30].

IV. CONCLUSION

Protein-DNA complexes play crucial roles in gene regulation, the prediction of the residues involved in protein-DNA interaction is critical for understanding gene regulation. It is vitally important to develop some high-performance computational methods for the prediction of DNA-binding residue. Although many methods have been proposed, most of them overlooked the motif features. Motif features are sub sequences composed of various number of residues and are important for the recognition between protein and DNA. In order to efficiently use the motif features for DNA-binding residue prediction, we apply the Convolutional Neural Network (CNN) method to capture motif features from the sequence around the target residue. CNN is a model that consists of a set of learnable motif detectors that can capture the important motif features by scanning the sequences around target residues. We then develop a neural network classifier, referred to as CNNsite, by combining the learned motif features, the sequence features and the evolutionary features for the prediction of DNA-binding residue. Dataset PDNA-62 and PDNA-224 are used to evaluate the performance of CNNsite by the five-fold cross-validation. The performance shows that motif features perform better than sequence features and evolutionary features. When comparing with previous published methods, CNNsite performs better than them. CNNsite is also evaluated on an independent dataset TS-72 and the performance shows that CNNsite outperforms the previous methods. The discriminant powers of the motif features of size from 2 to 6 show that many motif features with the largest power are composed by the residues that are important for the formation of DNA-protein interactions. It indicates that CNNsite can capture the important motif features from the sequences around the target residues for its prediction.

ACKNOWLEDGMENT

This work was supported by the National Natural Science Foundation of China 61370165, 61632011, National 863 Program of China 2015AA015405, Shenzhen Peacock Plan Research Grant KQCX20140521144507925 and Shenzhen Foundational Research Funding JCYJ20150625142543470, Guangdong Provincial Engineering Technology Research Center for Data Science 2016KF09 and HK Polytechnic Universitys graduate student grant: PolyU-RUDD.

REFERENCES

[1] A. Travers and D. Suck, "Dna-protein interactions," Chapman & Hall London:, Tech. Rep., 1993.

[2] R. T. Dame, "The role of nucleoid-associated proteins in the organization and compaction of bacterial chromatin," *Molecular microbiology*, vol. 56, no. 4, pp. 858–870, 2005.

[3] S. E. Halford and J. F. Marko, "How do site-specific dna-binding proteins find their targets?" *Nucleic acids research*, vol. 32, no. 10, pp. 3040–3052, 2004.

[4] S. Jones, J. A. Barker, I. Nobeli, and J. M. Thornton, "Using structural motif templates to identify proteins with dna binding function," *Nucleic acids research*, vol. 31, no. 11, pp. 2811–2823, 2003.

[5] C. P. Ponting, J. Schultz, F. Milpetz, and P. Bork, "Smart: identification and annotation of domains from

signalling and extracellular protein sequences," *Nucleic acids research*, vol. 27, no. 1, pp. 229–232, 1999.

[6] H. Kono and A. Sarai, "Structure-based prediction of dna target sites by regulatory proteins," *Proteins: Structure, Function, and Bioinformatics*, vol. 35, no. 1, pp. 114–131, 1999.

[7] S. Ahmad and A. Sarai, "Pssm-based prediction of dna binding sites in proteins," *BMC bioinformatics*, vol. 6, no. 1, p. 1, 2005.

[8] L. Wang, C. Huang, M. Q. Yang, and J. Y. Yang, "Bindn+ for accurate prediction of dna and rna-binding residues from protein sequence features," *BMC Systems Biology*, vol. 4, no. 1, p. 1, 2010.

[9] S.-Y. Ho, F.-C. Yu, C.-Y. Chang, and H.-L. Huang, "Design of accurate predictors for dna-binding sites in proteins using hybrid svm–pssm method," *Biosystems*, vol. 90, no. 1, pp. 234–241, 2007.

[10] Y. Ofran, V. Mysore, and B. Rost, "Prediction of dna-binding residues from sequence," *Bioinformatics*, vol. 23, no. 13, pp. i347–i353, 2007.

[11] L. Wang, M. Q. Yang, and J. Y. Yang, "Prediction of dna-binding residues from protein sequence information using random forests," *Bmc Genomics*, vol. 10, no. 1, p. 1, 2009.

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

[13] J. Wu, H. Liu, X. Duan, Y. Ding, H. Wu, Y. Bai, and X. Sun, "Prediction of dna-binding residues in proteins from amino acid sequences using a random forest model with a hybrid feature," *Bioinformatics*, vol. 25, no. 1, pp. 30–35, 2009.

[14] X. Ma, J. Guo, H.-D. Liu, J.-M. Xie, and X. Sun, "Sequence-based prediction of dna-binding residues in proteins with conservation and correlation information," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 9, no. 6, pp. 1766–1775, 2012.

[15] S. Ahmad, M. M. Gromiha, and A. Sarai, "Analysis and prediction of dna-binding proteins and their binding residues based on composition, sequence and structural information," *Bioinformatics*, vol. 20, no. 4, pp. 477–486, 2004.

[16] I. B. Kuznetsov, Z. Gou, R. Li, and S. Hwang, "Using evolutionary and structural information to predict dna-binding sites on dna-binding proteins," *PROTEINS: Structure, Function, and Bioinformatics*, vol. 64, no. 1, pp. 19–27, 2006.

[17] Y. Tsuchiya, K. Kinoshita, and H. Nakamura, "Structure-based prediction of dna-binding sites on proteins using the empirical preference of electrostatic potential and the shape of molecular surfaces," *PROTEINS: structure, Function, and Bioinformatics*, vol. 55, no. 4, pp. 885–894, 2004.

[18] H. Tjong and H.-X. Zhou, "Displar: an accurate method for predicting dna-binding sites on protein surfaces," *Nucleic Acids Research*, vol. 35, no. 5, pp. 1465–1477, 2007.

[19] X. Zhu, S. S. Ericksen, and J. C. Mitchell, "Dbsi: Dna-binding site identifier," *Nucleic acids research*, p. gkt617, 2013.

[20] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[21] L. J. McGuffin, K. Bryson, and D. T. Jones, "The psipred protein structure prediction server," *Bioinformatics*, vol. 16, no. 4, pp. 404–405, 2000.

[22] R. Adamczak, A. Porollo, and J. Meller, "Accurate prediction of solvent accessibility using neural networks–based regression," *Proteins: Structure, Function, and Bioinformatics*, vol. 56, no. 4, pp. 753–767, 2004.

[23] A. A. Schäffer, L. Aravind, T. L. Madden, S. Shavirin, J. L. Spouge, Y. I. Wolf, E. V. Koonin, and S. F. Altschul, "Improving the accuracy of psi-blast protein database searches with composition-based statistics and other refinements," *Nucleic acids research*, vol. 29, no. 14, pp. 2994–3005, 2001.

[24] B. Alipanahi, A. Delong, M. T. Weirauch, and B. J. Frey, "Predicting the sequence specificities of dna-and rna-binding proteins by deep learning," *Nature biotechnology*, 2015.

[25] S. Wang, J. Peng, J. Ma, and J. Xu, "Protein secondary structure prediction using deep convolutional neural fields," *Scientific reports*, vol. 6, 2016.

[26] T. Li, Q.-Z. Li, S. Liu, G.-L. Fan, Y.-C. Zuo, and Y. Peng, "Predna: accurate prediction of dna-binding sites in proteins by integrating sequence and geometric structure information," *Bioinformatics*, vol. 29, no. 6, pp. 678–685, 2013.

[27] L. Wang and S. J. Brown, "Bindn: a web-based tool for efficient prediction of dna and rna binding sites in amino acid sequences," *Nucleic acids research*, vol. 34, no. suppl 2, pp. W243–W248, 2006.

[28] S. Hwang, Z. Gou, and I. B. Kuznetsov, "Dp-bind: a web server for sequence-based prediction of dna-binding residues in dna-binding proteins," *Bioinformatics*, vol. 23, no. 5, pp. 634–636, 2007.

[29] A. Szilágyi and J. Skolnick, "Efficient prediction of nucleic acid binding function from low-resolution protein structures," *Journal of molecular biology*, vol. 358, no. 3, pp. 922–933, 2006.

[30] M. Sieber and R. K. Allemann, "Arginine (348) is a major determinant of the dna binding specificity of transcription factor e12." *Biological chemistry*, vol. 379, no. 6, pp. 731–735, 1998.