Mathematical Biosciences and Engineering

MBE, 21(1): 170–185.
DOI: 10.3934/mbe.2024008
Received: 12 October 2023
Revised: 18 November 2023
Accepted: 28 November 2023
Published: 11 December 2023

http://www.aimspress.com/journal/MBE

Research article

Identification of DNA-protein binding residues through integration of Transformer encoder and Bi-directional Long Short-Term Memory

Haipeng Zhao[1], Baozhong Zhu[1], Tengsheng Jiang[2], Zhiming Cui[1] and Hongjie Wu[1,*]

[1] School of Electronic and Information Engineering, Suzhou University of Science and Technology, Suzhou, China
[2] Gusu School, Nanjing Medical University, Suzhou, China

* **Correspondence:** Email: hongjiewu@usts.edu.cn.

Abstract: DNA-protein binding is crucial for the normal development and function of organisms. The significance of accurately identifying DNA-protein binding sites lies in its role in disease prevention and the development of innovative approaches to disease treatment. In the present study, we introduce a precise and robust identifier for DNA-protein binding residues. In the context of protein representation, we combine the evolutionary information of the protein, represented by its position-specific scoring matrix, with the spatial information of the protein's secondary structure, enriching the overall informational content. This approach initially employs a combination of Bi-directional Long Short-Term Memory and Transformer encoder to jointly extract the interdependencies among residues within the protein sequence. Subsequently, convolutional operations are applied to the resulting feature matrix to capture local features of the residues. Experimental results on the benchmark dataset demonstrate that our method exhibits a higher level of competitiveness when compared to contemporary classifiers. Specifically, our method achieved an MCC of 0.349, SP of 96.50%, SN of 44.03% and ACC of 94.59% on the PDNA-41 dataset.

Keywords: DNA-protein binding residues identification; Transformer encoder; BiLSTM; deep learning

1. Introduction

The interaction between DNA and proteins plays a crucial role in various biological processes, including DNA transcription, repair and protein synthesis [1–3]. The accurate identification of DNA -

protein binding sites holds significant importance in gaining a thorough understanding of multiple critical biological processes. This encompasses unraveling gene regulation, the interaction between proteins and nucleic acids and cellular signal transduction mechanisms. This is essential for a profound understanding of cellular function and the mechanisms of disease onset [4]. Furthermore, research on identifying DNA-protein binding sites also holds practical value in the field of drug design, aiding researchers in the discovery of new drug targets and the design of drug molecules for these sites [5]. DNA-protein binding sites represent pivotal locations for the interaction between DNA and proteins, and through these interactions, we can infer the function of proteins and predict their roles [6,7].

Given the significance of accurately identifying DNA-protein binding sites, numerous researchers in the past have developed various experimental methods to precisely identify sites where proteins bind to DNA. These methods include nuclear magnetic resonance (NMR) spectroscopy [8], conventional chromatin immunoprecipitation (ChIP) [9] and MicroChIP [10]. While the experimental methods exhibit good accuracy, they come with significant financial costs and time burdens. Moreover, the intricate nature of the experimental analysis process renders them unsuitable for large-scale data analysis.

The rapid advancement of protein sequencing technology has led to a significant increase in the number of identified protein sequences. However, their structures and functions continue to pose unresolved mysteries. In view of the above factors, there is an urgent need for a more precise and efficient method to accurately identify the binding sites of proteins with DNA. Recently, there have been several computational methods for discerning DNA-protein binding sites based on sequence, structure, or a combination of both.

Sequence-based methods typically employ protein sequence features and features extracted from the sequence. The formidable computational power of modern computers now allows us to extract evolutionary features from protein sequences. Combining high-quality evolutionary features with other sequence attributes often yields favorable experimental results. Although these methods lack structural information, the increasing prevalence of proteins with only sequence data and no other information has led to a growing interest in sequence-based approaches. A multitude of sequence-based methodologies have emerged, including but not limited to DNApred [11], DNABR [12] and DRNApred [13]. DNApred introduces an ensemble support vector machine (SVM) with all improvements, where the method initially employs hyperplane distance-based under-sampling technique to generate training subsets for training meta-classifiers. Subsequently, the base classifiers are integrated into the system. DNABR is a random forest prediction algorithm that utilizes evolutionary information, physicochemical features and the relationships between neighboring residues as joint inputs. DRNApred was designed using a penalty cross-prediction regression [14] and a novel two-layer structure, which effectively reduces cross-predictions and accurately identifies high-quality false positives. Additionally, it can accurately predict DNA- or RNA-binding proteins.

The spatial structure of proteins to some extent dictates their function [15]. This is one of the reasons why structure-based methods or hybrid approaches that combine spatial information with sequence information often exhibit superior performance compared to sequence-based methods. For example, the structural method GraphBind [16] constructs a graph based on intrinsic properties of residues and the environment. The edge features of the graph are computed using geometric knowledge, while node information encompasses physicochemical properties, geometric knowledge and evolutionary conservation. It introduces a graph neural network (GNN) [17] where each module can learn more advanced feature representations. The hybrid method DNABind [18] introduces a

complementary strategy that combines machine learning with a template-based approach [19,20]. In this approach, the machine learning method involves a pure sequence predictor, whose results are linearly combined with the results of a support vector machine structure predictor, using some structural properties and evolutionary conservation as inputs. Additionally, the template method detects the best reference structure in a template library.

However, it is worth noting that most studies have only considered certain structural or sequence features of individual residues and have not taken into account the relationships between residues. In other areas of interaction, some studies [21–23] have considered the interdependencies between elements and have achieved significant progress in their respective fields. For example, Wang et al. proposed DMFGAM [21], which initially fuses multiple molecular fingerprint features and utilizes attention mechanism to extract molecular graph features. Subsequently, a fully connected neural network is employed to determine whether a molecule is a human ether-a-go-go-related gene (hERG) blocker. Their research findings have become a powerful tool for predicting hERG blockers.

Inspired by some prior work, in our study, we integrated Bi-directional Long Short-Term Memory (BiLSTM) with the Transformer encoder to capture the interactions between residues in proteins. We did not overlook the local features of residues; instead, we subsequently employed a convolutional neural network to process the protein feature matrix. It is known that secondary structure information can offer insights into the local environment and geometric configuration of residues, which is crucial for determining the types of protein residues. In characterizing proteins, we combined position-specific scoring matrix (PSSM) information with predicted protein secondary structure information to enhance the richness of protein features. On the PDNA-543 and PDNA-41 datasets, we executed a series of experiments, comparing our method with existing methods to gauge its efficacy. The outcomes suggest that our methodology demonstrates competitiveness, and in certain instances, surpasses the performance of prevailing state-of-the-art methods.

Significant accomplishments in our study encompass: 1) The combination of Transformer encoder with BiLSTM allows for the collaborative extraction of long-range dependencies between residues at deeper and more complex levels. 2) Our methodology considers not only the global interplays between residues but also the localized feature attributes of target residues.

2. Materials and methods

2.1. Benchmark datasets

We utilized the PDNA-543 and PDNA-41 datasets for construction and evaluation of our approach. These two datasets were initially introduced in the TargetDNA [24]. At the outset, the dataset consisted of 7186 protein sequences. To mitigate sequence similarity, CD-Hit software [25] was employed, ensuring that the similarity between sequences in the dataset remained below 30%. The resulting 584 sequences were divided into two parts: One named PDNA-543, comprising 543 protein sequences and the other named PDNA-41, consisting of 41 protein sequences. In this work, PDNA-543 will be utilized as the training dataset to iteratively optimize the model parameters, whereas PDNA-41 will be employed to evaluate the generalization performance of the model. Table 1 provides an overview of the information related to both datasets.

Table 1. Data distribution of benchmark datasets.

Dataset	Proteins	Residues			% of binding residues
		Binding	Non-Binding	Total	
PDNA-543	543	9549	134,995	144,544	7.07
PDNA-41	41	734	14,021	14,755	5.24

2.2. Feature representation

The performance can vary significantly when different feature representations are used, even when the data is the same. Therefore, the representativeness and richness of data are crucial. The PSSM is widely employed in protein functional annotation and sequence analysis, leading to excellent performance in these tasks. Describing proteins solely based on sequence features is limited. Therefore, predicted protein secondary structure information has been incorporated into the sequence features to enhance predictive capability.

2.2.1. PSSM

In previous studies on predicting protein function [26] and protein-protein interactions [27], the effectiveness of PSSM in the field of bioinformatics has been demonstrated. In our research, we utilized PSI-BLAST [28], a multiple sequence alignment tool, to generate PSSM features. We adjusted the value of E to 10^{-3} and then performed three iterations on the Uniprot [29] database. This tool can represent a protein of length L as a feature matrix of size L*20. The number 20 represents the likelihood of an amino acid at that position mutating to one of the 20 different amino acids. To standardize units across different features, the PSSM values were normalized using a scaling formula that maps them to the interval (0, 1). The formula used to normalize the PSSM values is as follows:

$$y = \frac{1}{1+e^{-x}} \tag{1}$$

where x represents the original values of the matrix elements, and y denotes the normalized results.

2.2.2. Predicted secondary structure

The spatial arrangement and relative positions of protein regions are described by the secondary structure, which typically includes coiled, α-helix and β-fold. When binding with DNA, proteins adopt specific secondary structure elements to interact with specific DNA sites. Therefore, in the study of identifying binding sites between proteins and DNA, the secondary structure information of protein is of significant assistance. Measuring the protein's secondary structure directly requires advanced techniques and may be influenced by factors such as temperature, making it relatively challenging. Computational methods offer high precision at a low cost. In this study, PSIPRED [30] was employed to obtain the protein's secondary structure. Here, we represent a protein with a sequence length of L as a feature matrix of dimensions L*3. Each row contains three elements representing the probability of the target amino acid being of the corresponding secondary structure type.

2.3. Model architecture

In this work, an approach has been proposed that can capture long-term dependency relationships among DNA-binding protein residues while also extracting local features of the residues. Figure 1 illustrates the model's overarching structure.

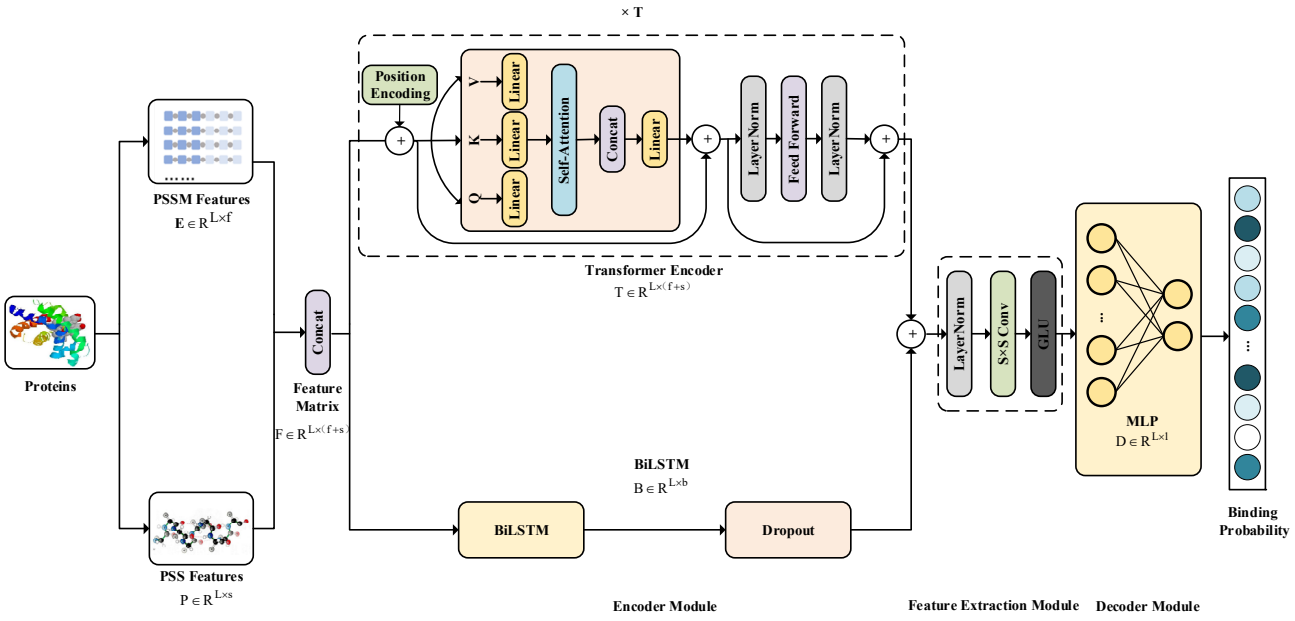

Figure 1. The overall structure of the proposed approach. The workflow of the model is: Initially, concatenate the PSSM information and predicted protein secondary structure information to form residue feature vectors. Subsequently, input these vectors separately into Transformer encoder and BiLSTM to learn the dependencies between residues. Combine the obtained results. Then, process the encoded protein feature matrix using convolutional layers to obtain local residue features. Finally, employ a multilayer perceptron (MLP) as the decoder to generate the DNA binding pattern.

2.3.1. Transformer encoder

The transformer exhibits outstanding performance when processing sequential data. In this work, we consider proteins as sequences, with each amino acid regarded as a sequence element. With the use of a Transformer encoder, we can process the protein sequence to capture the attention relationships between residues. Figure 1 illustrates the composition of the Transformer encoder. The acquisition of residue attention between different positions is achieved through a self-attention mechanism, and the specific calculation method is as follows:

$$\text{Attention}(\text{Q}, \text{K}, \text{V}) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)\text{V} \tag{2}$$

where Q, K and V are derived from the original input X through linear transformations, d_k is a scaling factor designed to prevent the numerical values of the results from becoming too large. Its value

depends on the number of heads and the dimensionality of the K matrix. Single-head attention mechanism can only capture attention relationships between residues in a single dimension, whereas multi-head attention mechanism can obtain different types of interactions across multiple dimensions. The calculation method of the multi-head attention mechanism is:

$$head_i = Attention\left(XW_i^Q, XW_i^K, XW_i^V\right) \tag{3}$$

$$\text{MultiHead}(\text{Q}, \text{K}, \text{V}) = \text{Concat}(head_1, \ldots, head_h)W^0 \tag{4}$$

where X represents the original input, while W_i^Q, W_i^K and W_i^V denote the matrices of linear transformation coefficients.

2.3.2. BiLSTM

Long Short-Term Memory (LSTM) can learn the dependencies between residues in a protein sequence through a gating mechanism. The specific computational process is as follows:

First, generate a numerical value between 0 and 1 based on the forget gate. This operation takes the current input x_t and the hidden state from the previous time step h_{t-1} to determine the degree of forgetting information. The sigmoid function is employed to remap the results, as follows:

$$f_t = \sigma\left(W_f[h_{t-1}, x_t] + b_f\right) \tag{5}$$

The next step, in the input gate, the previous time step's hidden state h_{t-1} and the current time step's input x_t are used as computational parameters to transform the input values into a positive number less than 1, representing the proportion of new information to be added. Then, the tanh function is employed to calculate the candidate value for the new information. i.e.,

$$i_t = \sigma(W_i[h_{t-1}, x_t] + b_i) \tag{6}$$

$$j_t = \tanh\left(W_j[h_{t-1}, x_t] + b_j\right) \tag{7}$$

where W_i, b_i, W_j and b_j are trainable parameters. At the current moment, we utilize the cell state C_{t-1} from the preceding time step and certain parameters to compute the new cell state C_t. i.e.,

$$C_t = f_t * C_{t-1} + i_t * j_t \tag{8}$$

Ultimately, the calculation of the output gate O_t is carried out to regulate the cell state's value. The final cell's hidden state is represented as h_t. The specific computation is as follows:

$$O_t = \sigma(W_o[h_{t-1}, x_t] + b_o) \tag{9}$$

$$h_t = O_t * \tanh(C_t) \tag{10}$$

LSTM can learn information only before the target position. In order to allow the algorithm model to access future context information just like it does with past context information, BiLSTM is designed with two separate LSTM hidden layers that operate in opposite directions. In this structure, information from both before and after the target residue can be learned at the output layer. i.e.,

$$\overrightarrow{h_t} = \overrightarrow{LSTM}\left(x_t, \overrightarrow{h_{t-1}}\right) \tag{11}$$

$$\overleftarrow{h_t} = \overleftarrow{LSTM}\left(x_t, \overleftarrow{h_{t-1}}\right) \tag{12}$$

$$H_t = w_t\overrightarrow{h_t} + v_t\overleftarrow{h_t} + b_t \tag{13}$$

where w_t matches with the weights of the forward hidden layer, while v_t corresponds to the weights of the backward hidden layer. H_t contains information from both, making it richer in information compared to the original input x_t.

2.3.3. Feature Extraction Module

The binding residues within the sequence of DNA-binding proteins are partially scattered, with some being adjacent. Furthermore, a significant portion of non-binding residues is also adjacent. Additionally, the properties of adjacent residues in a protein sequence exhibit a degree of similarity to each other. Therefore, we considered the features of the target residue and its adjacent residues on both the left and right sides. Given the superior performance of convolutional neural networks in extracting local information from feature maps, convolutional operations are employed in this module to process protein feature matrices. We adjusted the size of the convolutional kernel to S*S, with a stride set to 1. Such parameter settings precisely capture information about the target residue in the protein sequence and the (S-1)/2 residues on each side. Here, S is a variable value, and a larger S value implies the consideration of more neighboring residue information. Additionally, we investigated the performance variations associated with different convolutional kernel sizes, as elaborated in Section 3.2.

2.3.4. Decoder Module

In this module, we use a decoder based on MLP to obtain the predicted probabilities of each residue. The main function of MLP is to generate prediction results based on the encoded features. Through the combination of multiple hidden layers and nonlinear activation functions, MLP can learn the complex mapping relationship between encoded features and output predictions:

$$0 = \sigma(W_{MLP}[\mathrm{k}] + b_{MLP}) \tag{14}$$

where k is the output processed by the Feature Extraction Module. W_{MLP} is a collection of matrices, with each matrix corresponding to the connections between neuron layers, and b_{MLP} is a vector, where each element corresponds to a neuron. The σ function transforms the raw output to obtain probability values within the range of 0 to 1.

2.4. Training and evaluation

2.4.1. Loss function

The loss is a metric used to quantify the disparity between the model's predicted outcome and the actual labels. Throughout the training process of deep learning methods, our objective is to minimize

this loss. For the problem addressed in this study, binary cross-entropy loss is effective in measuring the disparity between predicted values and actual labels:

$$\mathcal{L}(y, P) = -\sum_{i=1}^{L}(y_i \log(P_i) + (1 + y_i) \log(1 - P_i)) + \frac{\lambda}{2}\left|\left|\Theta\right|\right|_2^2 \quad (15)$$

where y_i and P_i are the true labels and predicted values, respectively, for the i-th residue in the DNA-binding protein sequences. The symbol λ represents the regularization parameter controlling the complexity of the model. Θ is a vector or matrix that contains weights and biases. Moreover, to find better classification boundaries and improve classification accuracy, the Adam optimizer is used in this work.

2.4.2. Hyper-parameter-tuning

There are multiple modules included in our work, and each module contains many parameters. The variation of parameter values in the modules can lead to certain differences in model performance. We conducted a parameter search to explore the optimal value for performance.

1) **The number of layers of the Transformer encoder:** A deeper encoder can capture more complex semantic and dependency relationships in DNA-binding protein sequences, thereby improving the modeling ability of the model for input data. However, too many layers will result in more computational resources and time. Therefore, when choosing the number of layers for a Transformer encoder, a proper balance needs to be struck. We explored layer values ranging from 1 to 5 and discovered that the model achieved its optimal performance when employing 2 layers.
2) **The number of attention heads:** Multi-head attention can capture more diverse feature representations at different levels. Each attention head can focus on different aspects of the input sequence, thus extracting a variety of features. After experimenting with different numbers of attention heads (1, 2, 4 and 8), the results demonstrated that employing 4 attention heads led to the most favorable outcomes.
3) **The number of hidden units in BiLSTM:** The quantity of hidden units affects the capacity and representational power of the model, thereby influencing its learning. By altering this parameter across {16, 32, 64, 128, 256, 512}, it was observed that the model achieved its peak performance when equipped with 256 units.

2.4.3. Evaluation measurements

In DNA-protein binding residue identification samples, the majority of residues are non-binding, with only a minority being binding residues. This presents an imbalanced classification problem and implies that the more robust Matthews Correlation Coefficient (MCC) can better assess the classifier's performance because it takes into account various classification scenarios. In addition to this, Accuracy (ACC) is indicative of the proportion of accurately identified residues to the overall residue count. Specificity (SP) and Sensitivity (SN) are used to measure the classifier's performance in identifying non-binding and binding residues, respectively, with the calculations as follows:

$$\mathrm{SN} = \frac{TP}{TP+FN} \quad (16)$$

$$\mathrm{SP} = \frac{TP}{TN+FP} \tag{17}$$

$$\mathrm{ACC} = \frac{TP+TN}{TP+TN+FN+FP} \tag{18}$$

$$\mathrm{MCC} = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FN)\times(TP+FP)\times(TN+FN)\times(TN+FP)}} \tag{19}$$

where TP, TN, FP and FN stand for true positives, true negatives, false positives and false negatives, respectively.

3. Results

3.1. BiLSTM improves model performance

To demonstrate the beneficial impact of BiLSTM on model performance, models with different architectures were subjected to ten-fold cross-validation on the PDNA-543 dataset. The protein amino acid features utilized in the three sets of experiments are identical. The results of the three sets of experiments are clearly presented in Table 2. We can observe that models based on the Transformer encoder exhibit commendable performance, with MCC, SP, SN and ACC values of 0.336, 95.23%, 44.16% and 92.79% respectively. Its performance significantly outperformed the BiLSTM architecture model. This result aligns with our expectations since the Transformer encoder, which utilizes self-attention mechanisms, takes into account all positions within the sequence, whereas the step-by-step processing of BiLSTM fails to capture long-range dependencies. The combined model architecture approach showed improvements over the Transformer encoder architecture model in terms of MCC, SN and ACC, with increases of 0.005, 2.87 percentage points and 0.27 percentage points, respectively. Figure 2 illustrates the receiver operating characteristic (ROC) curves for three different models. In most cases, our model's curve lies above the curves of both the BiLSTM and Transformer models. The results indicate that the method that combines the advantages of Transformer encoder and BiLSTM can achieve better performance in DNA-protein binding residue identification task.

Table 2. The performance of models with different architectures in ten-fold cross-validation on PDNA-543.

Network	MCC	SP (%)	SN (%)	ACC (%)
BiLSTM	0.315	85.78	41.46	83.61
Transformer	0.336	95.23	44.16	92.79
Transformer +BiLSTM	0.341	95.16	47.03	93.06

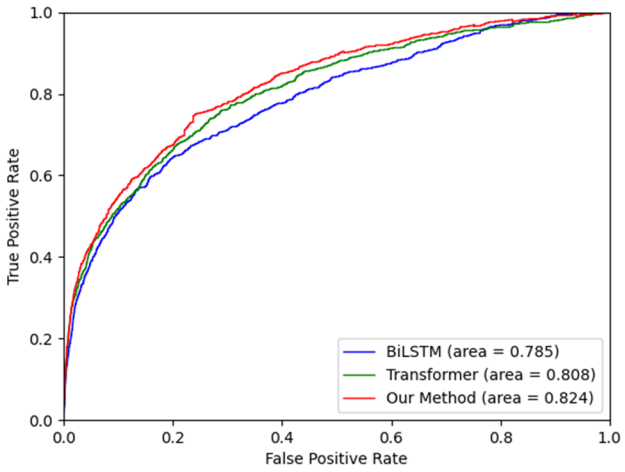

Figure 2. The ROC curves of models with different architectures in ten-fold cross-validation on PDNA-543.

3.2. Performance comparison for different convolutional kernel sizes

In this work, after encoding the protein sequences with the Encoder Module, we utilized convolutional kernels of size S*S to process the features of the target residue along with its (S-1)/2 neighboring residues on both sides. We observed variations in the model's performance when different values of S were chosen. In order to identify the optimal model parameters for achieving the best results, we conducted ten-fold cross-validation experiments on PDNA-543, altering the S value in the range of 1 to 7, incrementing it by 2 each time. Table 3 displays the varying performance of the model with different values of S. When S is set to 1, it implies that neighboring residues of the target residue are not considered. At this point, the model has already exhibited favorable performance. When we increase the value of S to 3, the model achieves its optimal performance, with an MCC value of 0.341 and a SN value of 47.03%. As we further increase the value of S, the model's MCC, ACC and SN noticeably decrease. Throughout the entire process, the SP value remains relatively stable. This indicates that, after processing by the encoder, considering the target residue and its adjacent residues on both sides has minimal impact on the identification of negative samples. However, within a certain range, taking these residues into account contributes to improving the model's identification performance for positive samples. Here, we utilize a combination of BiLSTM and Transformer encoder as Encoder Module architecture.

Table 3. Performance comparison under different values of S.

Value of S	MCC	SP (%)	SN (%)	ACC (%)
1	0.338	95.30	43.38	92.67
3	0.341	95.16	47.03	93.06
5	0.339	95.18	46.10	92.99
7	0.334	95.23	43.95	92.78

3.3. Comparison of overall performance with previous methods

When independently tested on PDNA-41 against the currently prominent methods, our approach demonstrated favorable performance with MCC, SP, SN and ACC scores of 0.349, 96.50%, 44.03% and 94.59%, respectively. These methods encompass several sequence-based approaches, namely BindN [31], ProteDNA [32], DP-Bind [33], BindN+ [34], MetaDBSite [35], TargetDNA and DNAPred, as well as a structure-based method called DNABind. The comparison results are shown in Table 4. The SP value of the ProteDNA method is 99.84%, slightly higher than our method, but its SN value is approximately one-tenth of our method's. This implies that our method exhibits significant improvement in positive class classification, and a higher MCC value suggests that our approach is more balanced and robust. DNAPred (with FPR set approximately at 5%) reached an SN value of 44.7%, but there was only a marginal improvement of 1.52% compared to our method. Our method demonstrated superior experimental outcomes, with notable enhancements of 3.56% in MCC, 1.69% in SP and 2.37% in ACC.

Table 4. Performance comparison of our approach versus previous methods on PDNA-41.

Method	MCC	SP (%)	SN (%)	ACC (%)
BindN	0.143	80.90	45.64	79.15
ProteDNA	0.160	99.84	4.77	95.11
BindN+(SP ≈ 95%)	0.178	95.11	24.11	91.58
BindN+(SP ≈ 85%)	0.213	85.41	50.81	83.69
MetaDBSite	0.221	93.35	34.20	90.41
DP-Bind	0.241	82.43	61.72	81.40
DNABind	0.264	80.28	70.16	79.78
TargetDNA (SN ≈ SP)	0.269	85.79	60.22	84.52
TargetDNA (SN ≈ 95%)	0.300	93.27	45.50	90.89
DNAPred (FPR ≈ 5%)	0.337	94.9	44.7	92.4
Our method	0.349	96.50	44.03	94.59

3.4. Case study

To visualize the superiority of the proposed method, we selected a protein-DNA complex, 4XR0_A (PDB ID = 4XR0, Chain: A), from the PDNA-41 dataset. We employed the trained model to predict this instance. Figure 3 presents a comparison of the DNA binding site prediction results between our proposed method and TargetDNA (SP $\approx$ 95%). This protein chain comprises a total of 308 residues, including 34 DNA-binding residues. Our method identified 19 TP instances, 15 FN instances and 27 FP instances. Moreover, the prediction results for TargetDNA include 15 TP instances, 19 FN instances and 26 FP instances. From Figure 3, it can be observed that our method, as compared to TargetDNA, exhibits minimal changes in the red and gray regions, while the green region significantly increases and the blue region notably decreases. It can be inferred that our method shows almost no difference compared to TargetDNA in identifying negative samples, while there is a significant improvement in its ability to recognize positive samples.

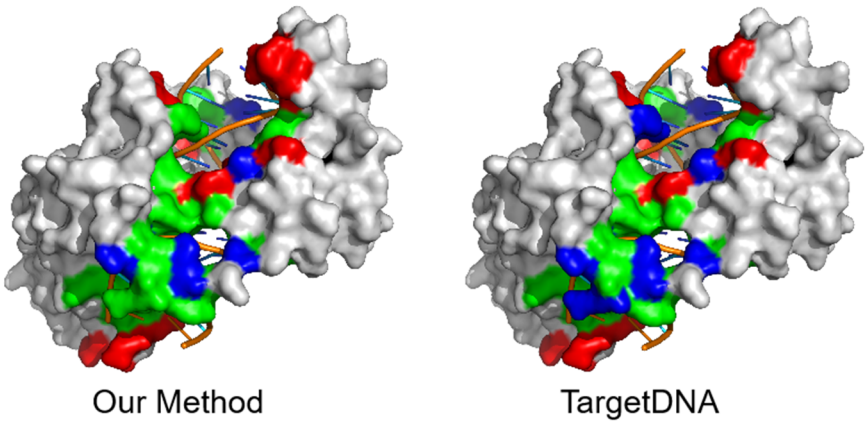

Figure 3. Visualization of an instance (PDB ID: 4XR0, Chain A) from the PDNA-41 dataset predicted by our method and TargetDNA (SP $\approx$ 95%). Green, red, yellow and gray represent TP, FP, FN and TN, respectively.

4. Conclusions and discussion

In this research project, a method for identifying DNA-protein binding residues is introduced, which takes into account the interconnections between global residues and the local feature information of target residues. This method can effectively capture features at both a broader, global scale and a more focused, local level. When representing protein features, we not only utilize the protein's evolutionary information through PSSM but also leverage the protein's spatial information, namely its secondary structure, to enrich its feature representation. Our method is simple and user-friendly, requiring only the input of protein sequences of arbitrary lengths, without the need for feature preprocessing at the individual residue level, to obtain results. Our method achieved favorable performance on the PDNA-41 dataset, with an MCC of 0.349, SP of 96.50%, SN of 44.03% and ACC of 94.59%. Compared to previous work, it is a classifier that excels in both positive and negative class classification, offering a more robust and balanced performance.

Considering that the interaction between proteins and DNA often involves specific domains or conformations, this information can be reflected in the three-dimensional structure of proteins. Three-dimensional structural information will be considered for further advancements in our research. Additionally, considering the successful applications of GNN in many bioinformatics domains, we will attempt to represent amino acids as graph nodes, with the attention level between amino acids serving as the edge weights. This transformation aims to convert the problem into a graph node classification problem.

Furthermore, we are attentive to the crucial role of the interactions between long non-coding RNA (lncRNA) and microRNA (miRNA) in various biological processes such as cell metabolism [36,37] and gene regulation [38,39]. Research in this area will provide valuable insights into gene markers associated with COVID-19 and diabetes. This has sparked significant research interest for us. Currently, many researchers have achieved promising results in this field [40,41]. For example, Wang et al., aiming to overcome the limitations of graph convolutional network (GCN) in predicting potential relationships between lncRNA and miRNA, proposed GCNCRF to infer the relationship between

lncRNA and miRNA. This method begins by constructing a heterogeneous network based on known interaction information, similarity networks and the LncRNA/miRNA feature matrix in the database. Subsequently, the initial embeddings are obtained and updated using a GCN with a conditional random field. Ultimately, predictive scores are generated. Zhang et al., aiming to address the gap in high-precision, high-performance computational models, proposed NDALMA based on distance analysis. This method is constructed through two types of similarity networks, including similarity networks of lncRNAs and miRNAs, as well as a Gaussian interaction profile kernel similarity network. Subsequently, a distance analysis is performed on the integrated network. The preliminary work of the researchers has provided us with a solid theoretical framework and technical paradigms. We hope to draw inspiration from their achievements and make significant contributions to the forefront of biomedical research, building on their success and achieving more exciting progress.

Use of AI tools declaration

The authors declare they have not used Artificial Intelligence (AI) tools in the creation of this article.

Acknowledgments

This paper is supported by the National Natural Science Foundation of China (62372318, 62073231, 62176175), National Research Project (2020YFC2006602)，Provincial Key Laboratory for Computer Information Processing Technology，Soochow University (KJS2166)，Opening Topic Fund of Big Data Intelligent Engineering Laboratory of Jiangsu Province (SDGC2157).

Conflict of interest

The authors declare that there are no conflicts of interest.

References

1. V. Charoensawan, D. Wilson, S. A. Teichmann, Genomic repertoires of DNA-binding transcription factors across the tree of life, *Nucleic Acids Res.*, **38** (2010), 7364–7377. https://doi.org/10.1093/nar/gkq617
2. J. Si, R. Zhao, R. Wu, An overview of the prediction of protein DNA-binding sites, *Int. J. Mol. Sci.*, **16** (2015), 5194–5215. https://doi.org/10.3390/ijms16035194
3. K. A. Aeling, N. R. Steffen, M. Johnson, G. W. Hatfield, R. H. Lathrop, D. F. Senear, DNA deformation energy as an indirect recognition mechanism in protein-DNA interactions, *IEEE/ACM Trans. Comput. Biol. Bioinf.*, **4** (2007), 117–125. https://doi.org/10.1109/TCBB.2007.1000
4. M. Ljungman, Activation of DNA damage signaling, *Mutat. Res. Fundam. Mol. Mech. Mutagen.*, **577** (2005), 203–216. https://doi.org/10.1016/j.mrfmmm.2005.02.014
5. G. Zhu, S. Cansiz, M. You, L. Qiu, D. Han, L. Zhang, et al., Nuclease-resistant synthetic drug-DNA adducts: Programmable drug-DNA conjugation for targeted anticancer drug delivery, *NPG Asia Mater.*, **7** (2015). https://doi.org/10.1038/am.2015.19

6. S. Peled, O. Leiderman, R. Charar, G. Efroni, Y. Shav-Tal, Y. Ofran, De-novo protein function prediction using DNA binding and RNA binding proteins as a test case, *Nat. Commun.*, **7** (2016), 13424. https://doi.org/10.1038/ncomms13424
7. C. J. Jeffery, Current successes and remaining challenges in protein function prediction, *Front. Bioinf.*, **3** (2023). https://doi.org/10.3389/fbinf.2023.1222182
8. C. P. Ponting, J. Schultz, F. Milpetz, P. Bork, SMART: Identification and annotation of domains from signalling and extracellular protein sequences, *Nucleic Acids Res.*, **27** (1999), 229–232. https://doi.org/10.1093/nar/27.1.229
9. N. M. Luscombe, R. A. Laskowski, J. M. Thornton, Amino acid–base interactions: A three-dimensional analysis of protein–DNA interactions at an atomic level, *Nucleic Acids Res.*, **29** (2001), 2860–2874. https://doi.org/10.1093/nar/29.13.2860
10. Y. Mandel-Gutfreund, H. Margalit, Quantitative parameters for amino acid-base interaction: Implications for prediction of protein-DNA binding sites, *Nucleic Acids Res.*, **26** (1998), 2306–2312. https://doi.org/10.1093/nar/26.10.2306
11. Y. H. Zhu, J. Hu, X. N. Song, D. Yu, DNAPred: Accurate identification of DNA-binding sites from protein sequence by ensembled hyperplane-distance-based support vector machines, *J. Chem. Inf. Model.*, **59** (2019), 3057–3071. https://doi.org/10.1021/acs.jcim.8b00749
12. X. Ma, J. Guo, H. D. Liu, J. Xie, X. Sun, Sequence-based prediction of DNA-binding residues in proteins with conservation and correlation information, *IEEE/ACM Trans. Comput. Biol. Bioinf.*, **9** (2012), 1766–1775. https://doi.org/10.1109/TCBB.2012.106
13. J. Yan, L. Kurgan, DRNApred, fast sequence-based method that accurately predicts and discriminates DNA-and RNA-binding residues, *Nucleic Acids Res.*, **45** (2017). https://doi.org/10.1093/nar/gkx059
14. L. Wang, M. Q. Yang, J. Y. Yang, Prediction of DNA-binding residues from protein sequence information using random forests, *BMC Genomics*, **10** (2009). https://doi.org/10.1186/1471-2164-10-S1-S1
15. H. A. Maghawry, M. G. M. Mostafa, T. F. Gharib, A new protein structure representation for efficient protein function prediction, *J. Comput. Biol.*, **21** (2014), 936–946. https://doi.org/10.1089/cmb.2014.0137
16. Y. Xia, C. Q. Xia, X. Pan, H. Shen, GraphBind: Protein structural context embedded rules learned by hierarchical graph neural networks for recognizing nucleic-acid-binding residues, *Nucleic Acids Res.*, **49** (2021). https://doi.org/10.1093/nar/gkab044
17. H. Zhou, D. Ren, H. Xia, M. Fan, X. Yang, H. Huang, Ast-gnn: An attention-based spatio-temporal graph neural network for interaction-aware pedestrian trajectory prediction, *Neurocomputing*, **445** (2021), 298–308. https://doi.org/10.1016/j.neucom.2021.03.024
18. R. Liu, J. Hu, DNABind: A hybrid algorithm for structure-based prediction of DNA-binding residues by combining machine learning-and template-based approaches, *Proteins Struct. Funct. Bioinf.*, **81** (2013), 1885–1899. https://doi.org/10.1002/prot.24330
19. S. Jones, H. P. Shanahan, H. M. Berman, J. M. Thornton, Using electrostatic potentials to predict DNA-binding sites on DNA-binding proteins, *Nucleic Acids Res.*, **31** (2003), 7189–7198. https://doi.org/10.1093/nar/gkg922
20. Y. Tsuchiya, K. Kinoshita, 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 Struct. Funct. Bioinf.*, **55** (2004), 885–894. https://doi.org/10.1002/prot.20111

21. T. Wang, J. Sun, Q. Zhao, Investigating cardiotoxicity related with hERG channel blockers using molecular fingerprints and graph attention mechanism, *Comput. Biol. Med.*, **153** (2023), 106464. https://doi.org/10.1016/j.compbiomed.2022.106464
22. Z. Chen, L. Zhang, J. Sun, R. Meng, S. Yin, Q. Zhao, DCAMCP: A deep learning model based on capsule network and attention mechanism for molecular carcinogenicity prediction, *J. Cell. Mol. Med.*, **27** (2023), 3117–3126. https://doi.org/10.1111/jcmm.17889
23. R. Meng, S. Yin, J. Sun, H. Hu, Q. Zhao, scAAGA: Single cell data analysis framework using asymmetric autoencoder with gene attention, *Comput. Biol. Med.*, **165** (2023), 107414. https://doi.org/10.1016/j.compbiomed.2023.107414
24. J. Hu, Y. Li, M. Zhang, X. Yang, H. Shen, D. Yu, Predicting protein-DNA binding residues by weightedly combining sequence-based features and boosting multiple SVMs, *IEEE/ACM Trans. Comput. Biol. Bioinf.*, **14** (2017), 1389–1398. https://doi.org/10.1109/TCBB.2016.2616469
25. W. Li, A. Godzik, Cd-hit: A fast program for clustering and comparing large sets of protein or nucleotide sequences, *Bioinformatics*, **22** (2006), 1658–1659. https://doi.org/10.1093/bioinformatics/btl158
26. J. cheol Jeong, X. Lin, X. W. Chen, On position-specific scoring matrix for protein function prediction, *IEEE/ACM Trans. Comput. Biol. Bioinf.*, **8** (2010), 308–315. https://doi.org/10.1109/TCBB.2010.93
27. J. Zahiri, O. Yaghoubi, M. Mohammad-Noori, R. Ebrahimpour, A. Masoudi-Nejad, PPIevo: Protein–protein interaction prediction from PSSM based evolutionary information, *Genomics*, **102** (2013), 237–242. https://doi.org/10.1016/j.ygeno.2013.05.006
28. S. F. Altschul, T. L. Madden, A. A. Schäffer, J. Zhang, Z. Zhang, W. Miller, et al., Gapped BLAST and PSI-BLAST: A new generation of protein database search programs, *Nucleic Acids Res.*, **25** (1997), 3389–3402. https://doi.org/10.1093/nar/25.17.3389
29. The UniProt Consortium, UniProt: A worldwide hub of protein knowledge, *Nucleic Acids Res.*, **47** (2019), D506–D515. https://doi.org/10.1093/nar/gky1049
30. L. J. McGuffin, K. Bryson, D. T. Jones, The PSIPRED protein structure prediction server, *Bioinformatics*, **16** (2000), 404–405. https://doi.org/10.1093/bioinformatics/16.4.404
31. L. Wang, S. J. Brown, BindN: A web-based tool for efficient prediction of DNA and RNA binding sites in amino acid sequences, *Nucleic Acids Res.*, **34** (2006), W243–W248. https://doi.org/10.1093/nar/gkl298
32. W. Y. Chu, Y. F. Huang, C. C. Huang, Y. Cheng, C. Huang, Y. Oyang, ProteDNA: A sequence-based predictor of sequence-specific DNA-binding residues in transcription factors, *Nucleic Acids Res.*, **37** (2009), W396–W401. https://doi.org/10.1093/nar/gkp449
33. S. Hwang, Z. Gou, I. B. Kuznetsov, DP-Bind: A web server for sequence-based prediction of DNA-binding residues in DNA-binding proteins, *Bioinformatics*, **23** (2007), 634–636. https://doi.org/10.1093/bioinformatics/btl672
34. L. Wang, C. Huang, M. Q. Yang, J. Y. Yang, BindN+ for accurate prediction of DNA and RNA-binding residues from protein sequence features, *BMC Syst. Biol.*, **4** (2010). https://doi.org/10.1186/1752-0509-4-S1-S3
35. J. Si, Z. Zhang, B. Lin, M. Schroeder, B. Huang, MetaDBSite: A meta approach to improve protein DNA-binding sites prediction, *BMC Syst. Biol.*, **5** (2011). https://doi.org/https://doi.org/10.1186/1752-0509-5-S1-S7

36. J. Li, H. Tian, J. Yang, Z. Gong, Long noncoding RNAs regulate cell growth, proliferation, and apoptosis, *DNA Cell Biol.*, **35** (2016), 459–470. https://doi.org/10.1089/dna.2015.3187
37. M. D. Paraskevopoulou, A. G. Hatzigeorgiou, Analyzing miRNA–lncRNA interactions, in *Long Non-coding RNAs: Methods and Protocols*, Humana press, (2016), 271–286. https://doi.org/10.1007/978-1-4939-3378-5_21
38. J. C. R. Fernandes, S. M. Acuña, J. I. Aoki, L. M. Floeter-Winter, S. M. Muxel, Long non-coding RNAs in the regulation of gene expression: Physiology and disease, *Non-coding RNA*, **5** (2019), 17. https://doi.org/10.3390/ncrna5010017
39. X. Li, C. Q. Zhong, R. Wu, X. Xu, Z. Yang, S. Cai, et al., RIP1-dependent linear and nonlinear recruitments of caspase-8 and RIP3 respectively to necrosome specify distinct cell death outcomes, *Protein Cell*, **12** (2021), 858–876. https://doi.org/10.1007/s13238-020-00810-x
40. W. Wang, L. Zhang, J. Sun, Q. Zhao, J. Shuai, Predicting the potential human lncRNA–miRNA interactions based on graph convolution network with conditional random field, *Briefings Bioinf.*, **23** (2022), bbac463. https://doi.org/10.1093/bib/bbac463
41. L. Zhang, P. Yang, H. Feng, Q. Zhao, H. Liu, Using network distance analysis to predict lncRNA–miRNA interactions, *Interdiscip. Sci.: Comput. Life Sci.*, **13** (2021), 535–545. https://doi.org/10.1007/s12539-021-00458-z

AIMS Press