2024 IEEE International Conference on Bioinformatics and Biomedicine (BIBM) | 979-8-3503-8622-6/24/$31.00 ©2024 IEEE | DOI: 10.1109/BIBM62325.2024.10822462

Integrating Biological Language Processing and Memory Attention Model for Protein-DNA Binding Residue Prediction

Shixuan Guan, Xiucai Ye*, Tetsuya Sakurai
Department of Computer Science, University of Tsukuba, Tsukuba 3058577, Japan
*Corresponding author, Email: yexiucai@cs.tsukuba.ac.jp

***Abstract*—Proteins are among the most important substances in the human body, and identifying protein-DNA binding sites is crucial for studying their interactions. Although traditional wet-lab methods can accurately identify these sites, they are time-consuming, labor-intensive, and expensive, making it challenging to keep pace with the rapid increase in protein sequence data. In this study, we propose the Memory Attention-Based Protein-DNA Binding Sites Prediction (MAPDB) model, which leverages multi-head Memory Attention for predicting protein-DNA binding sites. Our model employs a pre-trained embedding module to generate numerical representations of protein sequences, followed by a feature extraction module that uses Memory Attention to capture both intra-sequence and inter-sequence relationships. Extensive experiments on five benchmark datasets show that our model outperforms other state-of-the-art methods, especially in improving MCC scores. These results indicate that MAPDB effectively captures complex relationships within protein sequences, leading to more accurate predictions of protein-DNA binding residues.**

Keywords—Protein-DNA Binding Sites, Bioinformatics, Attention Mechanism, Protein Language Model

I. INTRODUCTION

Proteins are essential components of cells, involved in various critical biological activities and processes within organisms. However, proteins do not function independently; they operate through interactions with other substances [1]. For example, proteins can bind with metal ions to carry out specific biological functions. The binding of Cu^{2+} ions with proteins can induce oxidative modification of aldose reductase, while the binding of Mn^{2+} ions with proteins can lead to the formation of an oxygen-evolving complex related to photosynthesis [2]. Proteins also interact with other biological molecules, such as DNA and RNA, which are fundamental to cellular functions. These interactions play crucial roles in gene transcription, replication, translation, and regulation of cell metabolism, as well as in the replication and transmission of genetic information and the transport of essential substances [3, 4]. Among these interactions, the interaction between proteins and DNA is particularly significant. It is essential for recognizing transcription sites, DNA transcription, and DNA splicing. Moreover, since protein-DNA binding depends on specific residues within the proteins, accurately identifying these binding residues is of great significance for understanding the mechanisms of protein-DNA interactions and for designing novel drugs [5].

Traditionally, protein-DNA binding residues are identified through wet-lab methods such as Fast ChIP, electrophoretic mobility shift assays, and X-ray crystallography. These methods provide highly accurate identification of protein-DNA binding residues and form the basis for understanding protein-DNA interactions and conducting subsequent analyses. However, traditional wet-lab methods are labor-intensive, costly, and often time-consuming [6]. Consequently, researchers have increasingly focused on computational approaches—particularly the widely popular machine learning and deep learning methods in recent years—to identify protein binding residues. These computational methods approach the identification of protein binding residues as a binary classification problem for each residue and then train models on benchmark datasets to distinguish protein binding residues.

To date, numerous methods using traditional machine learning or deep learning have been proposed to identify protein binding residues. Among traditional machine learning methods, Hu et al. introduced TargetDNA, which utilizes a linear kernel alignment algorithm to weight and combine the evolutionary information of proteins with predicted solvent accessibility, employing a support vector machine (SVM) as a classifier to predict protein-DNA binding residues [7]. To address the issue of imbalanced positive and negative samples in protein-DNA binding residue prediction, Zhu et al. proposed an SVM algorithm based on ensemble hyperplane distances [8]. Ding et al. used a graph-regularized k-local hyperplane distance nearest neighbor algorithm to tackle sample imbalance issues in protein-nucleotide classification [9]. In the realm of deep learning methods, PredDBR is a sequence-based approach that uses three features—predicted secondary structure, position-specific frequency matrix, and predicted ligand-binding residue probability—as input features and employs a convolutional neural network (CNN) for feature extraction, resulting in more accurate protein-DNA binding residue predictions [10]. DeepCSeqSite utilizes seven features extracted from protein sequences as inputs and is constructed using a multi-stage convolutional neural network [11].

However, most existing methods typically use a sliding window technique to pre-extract features of protein residues, which means that only a portion of the residue sequence is

This work was supported by JST SPRING (Grant Number JPMJSP2124 and Grant Number JPMJPF2017), and JSPS KAKENHI (Grant Numbers JP23H03411 and JP22K12144).

observed at a time [12]. Consequently, it is challenging for these methods to observe an entire protein sequence in one go. To address this issue, our previous work leveraged protein sequence features—including the Position-Specific Scoring Matrix (PSSM) and predicted secondary structure—and employed a deep learning model to hierarchically extract residue correlations across the entire protein sequence [13]. These extracted features were then used to predict binding residues. The drawback of this method is that the sequence features used are difficult to obtain, and the prediction accuracy still requires improvement.

As described earlier, with the deepening exploration of deep learning in bioinformatics, an increasing number of studies are investigating the potential of deep learning for predicting protein binding residues, leading to improved predictive performance. However, existing deep learning methods often fail to fully capture inter-residue relationships and tend to rely on manually crafted features of protein residues. This approach contradicts the inherent feature-learning capability of deep learning, preventing these models from reaching their full potential. Consequently, the predictive performance of current deep learning models remains suboptimal. With advancements in natural language processing (NLP) and the development of pre-trained models such as GPT [14] and BERT [15], researchers can now accurately represent input sequences by pre-training on large datasets. Drawing inspiration from these NLP models, we attempt to apply biological language processing, where protein sequences are analogized as natural text sequences and residues as tokens. To this end, we propose the Memory Attention-Based Protein-DNA Binding Sites Prediction (MAPDB) model. First, we use a pre-trained model to generate numerical representations of protein sequences. Additionally, to capture relationships both within each protein sequence and between different sequences in the dataset, we employ a Memory Attention (MA) deep learning model. This model can take the numerical representation of an entire protein sequence as input and extract inter-residue relationships within the same sequence. Furthermore, Memory Attention constructs a "knowledge base" to capture relationships across different protein sequences. Experimental results demonstrate that our model predicts protein-DNA binding residues more accurately than other state-of-the-art methods.

II. Methodology

The overall architecture and workflow of the MAPDB model are illustrated in Fig. 1, highlighting its three primary modules: (A) the Pre-trained Embedding Module, (B) the Feature Extraction Module, and (C) the Classification Module. The workflow is as follows: Each protein sequence in the dataset is processed through the Pre-trained Embedding Module to obtain its numerical representation. Then, Module B (as shown in Fig. 1B) further extracts features from the numerical representation of each protein. Since the multi-head Memory Attention model operates on the feature dimensions of individual residues, it imposes no requirements on the length of the input protein sequence. This enables us to input an entire protein sequence at once, without needing to use a sliding window technique to standardize the sequence length. Finally, the features extracted by Module B are passed to Module C, where a fully connected layer is used to classify and predict protein-DNA binding residues. Additionally, the model parameters in Module B and Module C are updated using backpropagation. However, the model parameters in Module A are frozen and not updated via backpropagation. The detailed description of each module will be provided in the following sections.

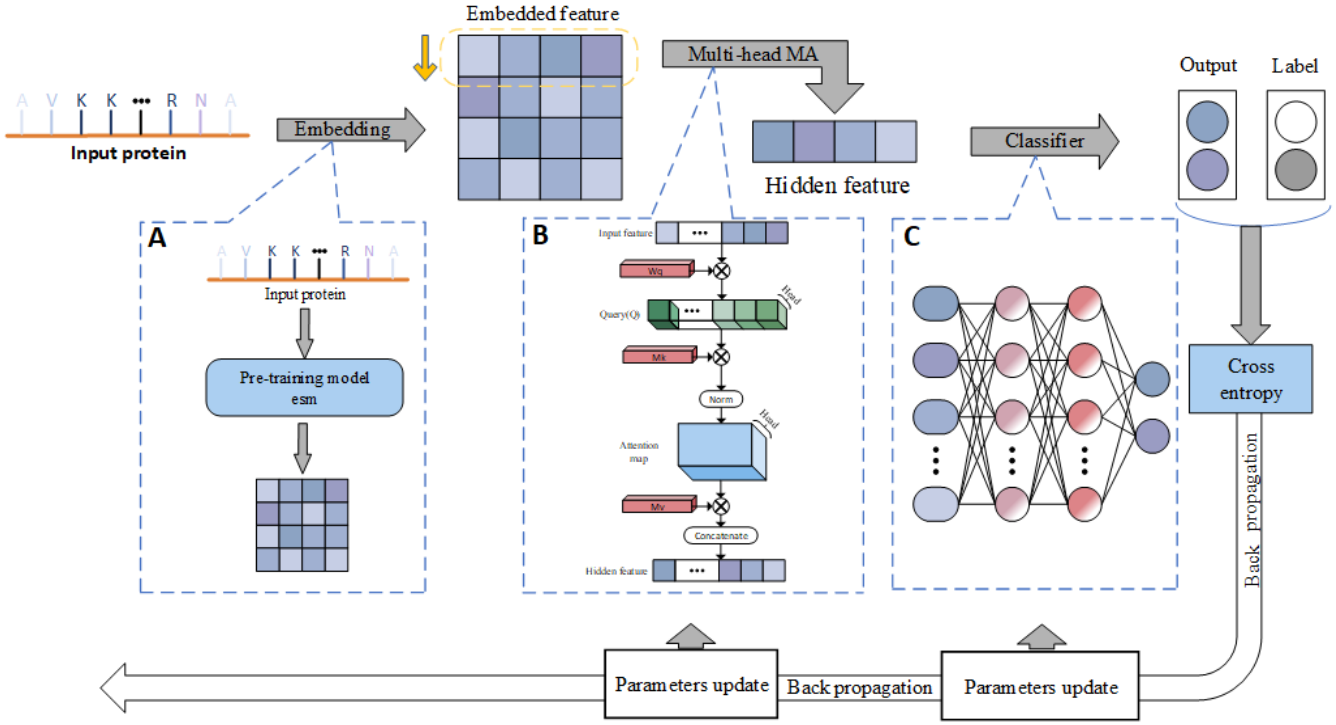

Fig. 1. The Overall Framework of the MAPDB Model

A. Pre-trained Embedding Module

To effectively represent protein sequences, we utilize the ESM pre-trained model to generate numerical embedding representations for each protein sequence [16]. The ESM model employs the Transformer encoder [17], treating each amino acid as a token in a manner similar to natural language processing, and completes training accordingly. The ESM model is pre-trained on the UniParc dataset [18], which contains 250 million protein sequences and 86 billion amino acid tokens. The resulting ESM-1b model consists of 33 Transformer encoding layers and has approximately 650 million parameters. The ESM model accepts an entire protein sequence as input. And if the sequence length is L, it outputs a feature representation of dimension L*1280, where each residue is represented by a 1280-dimensional feature vector. Since the ESM model treats each residue in a protein as a token and protein-DNA binding site prediction is also a residue-level classification task, the residue features obtained from the ESM model are highly suitable for predicting protein-DNA binding sites.

B. Feature Extraction Module

After obtaining the protein embedding representations, further feature extraction is required to derive the final features of the protein sequence. Here, we introduce the Self-Attention Mechanism and the Memory Attention approach proposed in this study.

(1) Self-Attention Mechanisms

The self-attention mechanism provides an effective way to capture the contextual information of the input sequence through the triplet of (Q-query, K-key, V-value) [19]. In the self-attention mechanism, the query Q, key K, and value V are all obtained from the same sequence through spatial mapping, with the formula as follows:

$$Q=W_qS \tag{1}$$

$$K=W_kS \tag{2}$$

$$V=W_vS \tag{3}$$

Where $S \in \mathbb{R}^{d\times N}$ represents the original sequence, with N and d denoting the sequence length and the embedding dimension of each token, respectively. $W_q \in \mathbb{R}^{q\times d}, W_k \in \mathbb{R}^{k\times d}, W_v \in \mathbb{R}^{v\times d}$ are the transformation matrices corresponding to Q, K, and V, respectively, where q, k, and v represent the dimensions of Q, K, and V. In general, $q = k = v$.

After obtaining Q, K, and V, Q and K can be multiplied to obtain the self-attention matrix A:

$$A=\text{softmax}\left(QK^T\right) \tag{4}$$

where $A \in \mathbb{R}^{N\times N}$ is the attention matrix composed of attention scores $a_{i,j}$, with $a_{i,j}$ indicating the strength of the association between the i-th token and the j-th token. Then, by multiplying V with the self-attention matrix A, the final output can be obtained:

$$Y=AV \tag{5}$$

Although the self-attention mechanism can capture the internal correlations within the input sequence, it only functions within the input sequence and cannot link different sequence inputs together. To extract associations across all data in the dataset, we propose a Memory Attention.

(2) Memory Attention

Given that the self-attention mechanism can only find correlations within the input sequence, we propose Memory Attention. It can construct a knowledge base to store the associations between all the sequences that have been input.Specifically, the attention mechanism functions through the interaction of the (Q-query, K-key, V-value) triplet. Here, we change the way Q, K, and V are generated. For Q-query, we still use the spatial mapping of the input sequence:

$$Q=W_qS \tag{6}$$

As for K-key and V-value, instead of using the input sequence as in the self-attention mechanism, we construct a global knowledge base to store key information from all sequences in the dataset. There are two specific implementation methods: (1) storing global information in the neurons of an MLP, and (2) constructing a global vector for K-key and V-value, respectively. By using these methods to build the global knowledge base, we can obtain the final output of the input sequence through MA. The specific implementation processes of the two methods are detailed in Algorithm 1 and Algorithm 2.

(3) Multi-head Memory Attention

Multi-head MA extends the basic MA by employing multiple independent MA mechanisms to obtain multiple independent spatial representations. This approach allows the model to focus on different positions, capturing relationships between different locations and across different layers.

Similar to the multi-head self-attention mechanism, MA can also be divided across different input dimensions, forming a multi-head MA. Using multi-head MA can enhance model performance. On one hand, it allows the model to focus on different positions, thereby capturing relationships between different locations. Additionally, it helps avoid potential negative impacts that might arise from relying on a single knowledge base. Multi-head MA can be calculated as follows:

$$\text{MultiHead}\ (Q, K, V)= \text{Concat}\ (\text{head}_1, \text{head}_2, \ldots, \text{head}_H)W_s \tag{7}$$

Algorithm 1 Pseudo-code for method (1).

```
# X - Input
# W_k-The K-key that stores global information, W_k = torch.nn.Linear().
# W_v-The V-value that stores global information, W_v=torch.nn.Linear()

Q = torch.nn.Linear(X)
attn = W_k(Q)
attn_score = softmax(attn)
Y = W_v(attn_score)
```

Algorithm 2 Pseudo-code for method (2).

```
# X — Input
# K —The K-key that stores global information, K = torch.nn.Parameter()
# V—The V-value that stores global information, V = torch.nn.Parameter()

Q = torch.nn.Linear(X)
attn = torch.bmm(Q, K)
attn_score = softmax(attn)
Y = torch.bmm(attn_score, V)
```

Where W_s is the spatial mapping function, ensuring that the dimensions of the input sequence remain consistent before and after multi-head MA. The calculation of $head_i$ differs depending on the implementation of MA. For Algorithm 1:

$$\text{head}_\text{i} = \text{MA}\ (\text{QW}_\text{q}^\text{i}, \text{W}_\text{k}, \text{W}_\text{v}) \tag{8}$$

Here, $W_q^i \in \mathbb{R}^{d \times d'}$ maps Q to different dimensions. The $head_i$ is then calculated using the shared W_k and W_v through the Algorithm 1. As for Algorithm 2:

$$\text{head}_\text{i} = \text{MA}\ (\text{QW}_\text{q}^\text{i}, \text{KW}_\text{k}^\text{i}, \text{VW}_\text{v}^\text{i}) \tag{9}$$

Here, W_q^i, W_k^i, W_v^i are the mapping matrices for the i-th head. The $head_i$ is then computed through the Algorithm 2.

III. Result

A. Dataset

In this study, we used five benchmark datasets—PDNA-543 [7], PDNA-41 [7], PDNA-335 [20], PDNA-52 [20], and PDNA-316 [21]—for the prediction of protein-DNA binding sites. To comprehensively evaluate the classification performance of our model, we conducted experiments on all five datasets. Cross-validation was performed on the PDNA-543, PDNA-335, and PDNA-316 datasets. Additionally, we trained the model on the PDNA-543 benchmark dataset and performed independent testing on the PDNA-41 dataset. Similarly, we trained the model on the PDNA-335 benchmark dataset and conducted independent testing on the PDNA-52 dataset. To prevent the influence of homologous sequences on the experimental results, all sequences were de-redundantized using CD-HIT [22]. In the PDNA-543 and PDNA-41 datasets, the similarity between any two protein sequences does not exceed 30%. In the PDNA-335 and PDNA-52 datasets, the similarity between any two protein sequences does not exceed 40%, and in the PDNA-316 dataset, the similarity does not exceed 30%. Table I shows the detailed information for each dataset.

TABLE I. DETAILS OF THE EXPERIMENTAL DATASETS

Dataset	No. Sequences[1]	Experiment Type	No. positive[2]	No. negative[3]	R_{DNABR}[4] (%)
PDNA-543	543	Ten-fold Cross-Validation	9549	134995	7.07
PDNA-316	316		5606	67109	8.35
PDNA-335	335	Five-fold Cross-Validation	6443	70768	9.11
PDNA-52	52	Independent Testing	967	10621	6.03
PDNA-41	41		734	14021	5.24

[1]: Number of Protein Sequences

[2]: Number of Positive Samples (Protein-DNA Binding Residues)

[3]: Number of Negative Samples (Non-Protein-DNA Binding Residues)

[4]: Positive-to-Negative Sample Ratio

B. Evaluation Metrics

For binary classification problems, the most commonly used evaluation metrics are Matthews Correlation Coefficient (MCC), Accuracy (ACC), Specificity (SP), and Sensitivity (SN). In this study, we use these four metrics to assess the performance of our model. Their calculations are as follows:

$$\text{MCC}=\frac{\text{TP}\times\text{TN-FP}\times\text{FN}}{\sqrt{(\text{TP+FP})(\text{TP+FN})(\text{TN+FP})(\text{TN+FN})}} \tag{10}$$

$$\text{ACC}=\frac{\text{TP+TN}}{\text{TP+FN+TN+FP}} \tag{11}$$

$$\text{SP}=\frac{\text{TP}}{\text{TP+FN}} \tag{12}$$

$$\text{SN}=\frac{\text{TP}}{\text{TN+FP}} \tag{13}$$

Where TP, TN, FP, and FN represent the numbers of true positive, true negative, false positive, and false negative samples predicted by the model, respectively. Since protein-DNA residue binding site prediction is an extremely imbalanced task, MCC is a more reliable indicator of the model's predictive quality. Both MCC and ACC evaluate the overall performance of the model. SP measures the proportion of correctly predicted non-binding residues, while SN measures the proportion of correctly predicted binding residues. For all the above metrics, higher values indicate better model performance.

C. Ablation Study

In the self-attention mechanism, the multi-head self-attention mechanism can capture multiple spatial features compared to the single-head self-attention mechanism. This often leads to better experimental results with multi-head attention [23]. Similarly, to verify whether multi-head MA has a positive impact on protein-DNA binding site prediction compared to single-head MA, we conducted experiments on three cross-validation datasets. The experimental results are shown in Table II. From Table II, it is evident that the training performance using multi-head MA outperforms single-head MA in all aspects. Specifically, we compared the experimental results of the single-head MA model with the 4-head MA model. The results on the three datasets show a significant improvement in MCC, along with noticeable improvements in ACC, SN, and SP. On the PDNA-543 dataset, the MCC of the 4-head MA model reached 0.4341, a 2.42% improvement compared to the single-head model (0.4099). Similarly, the multi-head MA model's MCC on the PDNA-316 dataset (0.4913) and the PDNA-335 dataset (0.4583) improved by 2.06% and 3.37%, respectively, compared to the single-head model. Therefore, using a multi-head MA model can effectively enhance predictive performance, with MCC improving by over 2% across all tested datasets.

TABLE II. EXPERIMENTAL COMPARISON BETWEEN MULTI-HEAD MA AND SINGLE-HEAD MA

Dataset	Num. of head	MCC	ACC	SN	SP
PDNA-543	1	0.4099	93.54	54.79	95.36
	4	0.4341	93.86	57.06	95.66
PDNA-316	1	0.4707	93.09	58.78	95.24
	4	0.4913	93.24	59.35	95.54
PDNA-335	1	0.4246	91.37	48.45	95.07
	4	0.4583	91.73	51.40	95.41

TABLE III. EXPERIMENTAL COMPARISON BETWEEN MULTI-HEAD MA AND MULTI-HEAD SELF-ATTENTION MECHANISM

Dataset	Method	MCC	ACC	SN	SP
PDNA-543	Self-attention	0.4142	93.46	53.71	95.46
	Memory attention	0.4341	93.86	57.06	95.66
PDNA-316	Self-attention	0.4783	94.91	54.12	96.95
	Memory attention	0.4913	93.24	59.35	95.54
PDNA-335	Self-attention	0.4454	91.74	50.89	95.22
	Memory attention	0.4583	91.73	51.40	95.41

MA is a crucial part of this study. To explore its effectiveness, we compare it with the traditional self-attention mechanism model. In these experiments, we set the number of heads for both the multi-head MA model and the multi-head self-attention mechanism model to 4. The number of heads in the self-attention mechanism model must be a divisor of 1280 due to dimensional requirements, so it is set to 4 for the comparative experiments. The results on the three cross-validation datasets are shown in Table III. The MA method generally outperforms the self-attention mechanism in SN and MCC, particularly on the PDNA-543 and PDNA-316 datasets, where sensitivity shows a significant improvement. This indicates that MA is better at identifying positive samples. Although the differences in ACC and SP are minimal, MA has a slight edge in specificity. Therefore, the MA method performs better in balancing overall performance while enhancing positive sample identification.

D. Comparative Study

To numerically evaluate the predictive performance of MAPDB, we compared it against other state-of-the-art methods across five datasets: PDNA-543, PDNA-41, PDNA-335, PDNA-52, and PDNA-316, as described in the dataset section. For the PDNA-543 dataset, we conducted ten-fold cross-validation and compared the results with TFDSite [13] and PredDBR [10], both of which have also been evaluated on this dataset. On the PDNA-316 dataset, we performed ten-fold cross-validation and compared the results with DISIS [24], DP-Bind [25], MetaDBSite [21], BindN-rf [26], and PredDBR, all of which have been evaluated on this dataset. For the PDNA-335 dataset, we conducted five-fold cross-validation and compared the results with TargetS [20], EC-RUS [27], and PredDBR. On the PDNA-52 dataset, we first trained the model on the PDNA-335 dataset and then performed independent testing on the PDNA-52 dataset. The results were compared and analyzed with TargetS, DNAPred, COACH [28], and PredDBR. For the PDNA-41 dataset, we first trained the model on the PDNA-543 dataset and then conducted independent testing. The results were compared with DP-Bind, DNABind, COACH, TFDSite, and PredDBR.

The experimental details and results are shown in Table IV. The reason for using different comparison methods across datasets is the unavailability of certain methods' online servers or downloadable code, which limits the experimental results to specific datasets.As shown in Table IV, our method outperforms other existing predictors on 4 out of the 5 datasets, with only slightly lower performance on the PDNA-52 dataset. Specifically, compared to the next best predictor, PredDBR, our

TABLE IV. COMPARATIVE EXPERIMENTS WITH OTHER METHODS

Dataset	Method	MCC	ACC	SN	SP
PDNA-543	TFD-Site	0.352	92.83	45.20	95.38
	PredDBR	0.415	91.43	45.35	95.50
	Ours	0.4341	93.86	57.06	95.66
PDNA-316	DISIS	0.250	92.00	19.00	98.00
	DP-Bind	0.290	78.00	69.00	79.00
	MetaDBSite	0.320	77.00	77.00	77.00
	BindN-rf	0.320	82.00	67.00	83.00
	PredDBR	0.489	92.30	53.08	95.82
	Ours	0.4913	93.24	59.35	95.54
PDNA-335	TargetS	0.362	89.90	41.70	94.50
	EC-RUS	0.378	92.60	48.70	95.10
	PredDBR	0.390	90.96	42.59	95.34
	Ours	0.4583	91.73	51.40	95.41
PDNA-52[1]	TargetS	0.377	93.30	41.30	96.50
	DNAPred	0.405	92.50	51.80	94.90
	COACH	0.420	91.55	59.91	93.45
	PredDBR	0.451	93.46	53.85	95.83
	Ours	0.4140	92.43	38.25	97.13
PDNA-41[2]	DP-Bind	0.241	81.40	61.72	82.43
	DNABind	0.264	79.78	70.16	80.28
	COACH	0.352	92.67	46.19	95.10
	TFDSite	0.357	94.87	47.57	96.44
	PredDBR	0.359	93.93	39.10	96.79
	Ours	0.3925	94.75	46.50	96.78

[1]: Trained on the PDNA-335 dataset and then independently tested on the PDNA-52 dataset.

[2]: Trained on the PDNA-543 dataset and then independently tested on the PDNA-41 dataset.

model achieved an average MCC improvement of 3% across all datasets. Notably, on the PDNA-335 and PDNA-41 datasets, the MCC of our method (PDNA-335: 0.4583, PDNA-41: 0.3925) was 6.83% and 3.35% higher, respectively, than that of the next best method, PredDBR (PDNA-335: 0.390, PDNA-41: 0.359).

IV. Conclusion

The MAPDB model introduced in this study provides a novel approach to predicting protein-DNA binding sites by integrating the strengths of pre-trained embeddings and Memory Attention. This approach effectively captures both local and global relationships within and between protein sequences, resulting in improved predictive accuracy. The ability of the model to process entire protein sequences without the need for truncation or padding simplifies preprocessing and enhances predictive power. Our experiments on five benchmark datasets show that MAPDB outperforms existing methods on four of the five datasets, with significant improvements in MCC and overall prediction accuracy. These findings underscore the potential of Memory Attention in capturing complex sequence relationships and suggest that MAPDB could become a valuable tool for protein-DNA interaction studies. Future work will focus on further enhancing the model's accuracy and exploring its applicability to other types of protein interactions.

References

[1] Bondos, Sarah E., A. Keith Dunker, and Vladimir N. Uversky. "Intrinsically disordered proteins play diverse roles in cell signaling." *Cell Communication and Signaling,* vol. 20, no. 20, 2022.

[2] Ono, Taka-aki. "Metallo-radical hypothesis for photoassembly of (Mn) 4-cluster of photosynthetic oxygen evolving complex." *Biochimica et Biophysica Acta (BBA)-Bioenergetics,* vol. 1503, no. 1, pp. 40-51, 2001.

[3] Tianyuan Liu, Huiyuan Qiao, Zixu Wang, Xinyan Yang, Xianrun Pan, Yu Yang, Xiucai Ye, Tetsuya Sakurai, Hao Lin, and Yang Zhang. "CodLncScape Provides a Self‐Enriching Framework for the Systematic Collection and Exploration of Coding LncRNAs." *Advanced Science*, vol. 11, no. 22, pp.2400009, 2024.

[4] Sabari, Benjamin R., Alessandra Dall'Agnese, and Richard A. Young. "Biomolecular condensates in the nucleus." *Trends in biochemical sciences,* vol. 45, no. 11, pp. 961-977, 2020.

[5] Kun Wu, Xiulong Yang, Zixu Wang, Na Li, Jialu Zhang, and Lizhuang Liu. "Data-balanced transformer for accelerated ionizable lipid nanoparticles screening in mRNA delivery." *Briefings in Bioinformatics*, vol. 25, no. 3, pp. bbae186, 2024.

[6] Shixuan Guan, Yuqing Qian, Tengsheng Jiang, Min Jiang, Yijie Ding, and Hongjie Wu. "MV-H-RKM: A Multiple View-Based Hypergraph Regularized Restricted Kernel Machine for Predicting DNA-Binding Proteins." *IEEE/ACM Transactions on Computational Biology and Bioinformatics,* vol. 20, no. 2, pp. 1246-1256, 2022.

[7] Hu Jun, Yang Li, Ming Zhang, Xibei Yang, Hong-Bin Shen, and Dong-Jun Yu. "Predicting protein-DNA binding residues by weightedly combining sequence-based features and boosting multiple SVMs." *IEEE/ACM transactions on computational biology and bioinformatics,* vol. 14, no. 6, pp. 1389-1398, 2016.

[8] Yi-Heng Zhu, Jun Hu, Xiao-Ning Song, and Dong-Jun Yu. "DNAPred: accurate identification of DNA-binding sites from protein sequence by ensembled hyperplane-distance-based support vector machines." *Journal of chemical information and modeling,* vol. 59, no. 6, pp. 3057-3071, 2019.

[9] Yijie Ding, Chao Yang, Jijun Tang, and Fei Guo. "Identification of protein-nucleotide binding residues via graph regularized k-local hyperplane distance nearest neighbor model." *Applied Intelligence,* vol. 52, pp. 6598-6612, 2022.

[10] Jun Hu, Yang-Song Bai, Lin-Lin Zheng, Ning-Xia Jia, Dong-Jun Yu, and Gui-Jun Zhang. "Protein-DNA binding residue prediction via bagging strategy and sequence-based cube-format feature." *IEEE/ACM transactions on computational biology and bioinformatics,* vol. 19, no. 6, pp. 3635-3645, 2021.

[11] Yifeng Cui, Qiwen Dong, Daocheng Hong, and Xikun Wang. "Predicting protein-ligand binding residues with deep convolutional neural networks." *BMC bioinformatics,* vol. 20, no. 93, pp. 1-12, 2019.

[12] Junnan Li, Jiang Li, Xiaoping Wang , Xin Zhan, and Zhigang Zeng. "A Domain Generalization and Residual Network-Based Emotion Recognition from Physiological Signals." *Cyborg and Bionic Systems,* vol. 5, p. 0074, 2024.

[13] Shixuan Guan, Quan Zou, Hongjie Wu, and Yijie Ding Shixuan. "Protein-dna binding residues prediction using a deep learning model with hierarchical feature extraction." *IEEE/ACM Transactions on Computational Biology and Bioinformatics,* vol. 20, no. 5, pp. 2619-2628, 2022.

[14] Floridi, Luciano, and Massimo Chiriatti. "GPT-3: Its nature, scope, limits, and consequences." *Minds and Machines,* vol. 30, pp. 681-694, 2020.

[15] Devlin, Jacob. "Bert: Pre-training of deep bidirectional transformers for language understanding." *arXiv preprint arXiv:1810.04805*, 2018.

[16] Alexander Rives, Joshua Meier, Tom Sercu, Siddharth Goyal, Zeming Lin, Jason Liu, et al. "Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences." *Proceedings of the National Academy of Sciences,* vol. 118, no. 15, p. e2016239118, 2021.

[17] Vaswani, Ashish. "Attention is all you need." *arXiv preprint arXiv:1706.03762,* 2017.

[18] Rolf Apweiler, Amos Bairoch, Cathy H. Wu, Winona C. Barker, Brigitte Boeckmann, Serenella Ferro, et al. "UniProt: the universal protein knowledgebase." *Nucleic acids research,* vol. 32, no. suppl_1, pp. D115-D119, 2004.

[19] Vaswani, Ashish. "Tensor2tensor for neural machine translation." *arXiv preprint arXiv:1803.07416*, 2018.

[20] Dong-Jun Yu, Jun Hu, Jing Yang, Hong-Bin Shen, Jinhui Tang, and Jing-Yu Yang. "Designing template-free predictor for targeting protein-ligand binding sites with classifier ensemble and spatial clustering." *IEEE/ACM transactions on computational biology and bioinformatics,* vol. 10, no. 4, pp. 994-1008, 2013.

[21] Jingna Si, Zengming Zhang, Biaoyang Lin, Michael Schroeder, and Bingding Huang. "MetaDBSite: a meta approach to improve protein DNA-binding sites prediction." *BMC systems biology,* vol. 5, pp. 1-7, 2011.

[22] Limin Fu, Beifang Niu, Zhengwei Zhu, Sitao Wu, and Weizhong Li. "CD-HIT: accelerated for clustering the next-generation sequencing data." *Bioinformatics,* vol. 28, no. 23, pp. 3150-3152, 2012.

[23] Zhouhan Lin, Minwei Feng, Cicero Nogueira dos Santos, Mo Yu, Bing Xiang, Bowen Zhou, et al. "A structured self-attentive sentence embedding." *arXiv preprint arXiv:1703.03130*, 2017.

[24] Ofran, Yanay, Venkatesh Mysore, and Burkhard Rost. "Prediction of DNA-binding residues from sequence." *Bioinformatics,* vol. 23, no. 13, pp. i347-i353, 2007.

[25] Hwang, Seungwoo, Zhenkun Gou, and Igor 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.

[26] Wang, Liangjiang, Mary Qu Yang, and Jack Y. Yang. "Prediction of DNA-binding residues from protein sequence information using random forests." *Bmc Genomics,* vol. 10, pp. 1-9, 2009.

[27] Yijie Ding, Jijun Tang, and Fei Guo. "Identification of protein–ligand binding sites by sequence information and ensemble classifier." *Journal of Chemical Information and Modeling,* vol. 57, no. 12, pp. 3149-3161, 2017.

[28] Jianyi Yang, Ambrish Roy, and Yang Zhang. "Protein–ligand binding site recognition using complementary binding-specific substructure comparison and sequence profile alignment." *Bioinformatics,* vol. 29, no. 20, pp. 2588-2595, 2013.