2024 5th International Conference on Computer, Big Data and Artificial Intelligence (ICCBD+AI) | 979-8-3315-3399-1/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICCBD-AI65562.2024.00009

Prediction of Protein-DNA Binding Sites Based on GraphSAGE

1st Ruiyan Huang
School of Information Engineering
Jingdezhen Ceramic University
Jingdezhen, China
huangruiyan477@163.com

2nd Xian Chen
School of Information Engineering
Jingdezhen Ceramic University
Jingdezhen, China
2514043452@qq.com

3rd Wangren Qiu
School of Information Engineering
Jingdezhen Ceramic University
Jingdezhen, China
qiuone@163.com

4th Weizhong Lin*
School of Information Engineering
Jingdezhen Ceramic University
Jingdezhen, China
linweizhong@jcu.edu.cn

***Abstract*—Accurately pinpointing the locations where proteins bind to DNA is crucial for the discovery of novel drug targets and the development of targeted therapeutic strategies. In this study, a graph neural network architecture called GraphSAGE has notably improved the precision of predicting protein-DNA binding sites. This model represents protein sequences as a graph, with amino acid residues serving as nodes, and employs a pre-trained, large-scale protein language model as ESM to derive residues' features. The effectiveness of the model was rigorously evaluated using ten-fold cross-validation on benchmark datasets, including PDNA-316, PDNA-335, and PDNA-543, yielding Matthews Correlation Coefficients (MCC) of 0.741, 0.750, and 0.735, respectively. These results demonstrate considerable improvements over the existing state-of-the-art methods, with respective increments of 0.207, 0.185, and 0.174. Furthermore, the methodology is not only confined to binding site prediction but is also extensible to applications in protein functional annotation, drug discovery, and pharmaceutical design. The entire codebase for this project can be accessed without restriction through the subsequent URL: https://github.com/primrosehry/iProtDNA-SAGE.**

Keywords-protein-DNA binding site; graph neural network; imbalanced dataset learning

I. INTRODUCTION

The precise prediction of protein-DNA binding sites is crucial for understanding the mechanisms of protein-DNA interactions and may provide key information for drug development and therapeutic strategies[1]. However, prevalent empirical techniques tend to be expensive and laborious. Therefore, developing fast and accurate computational prediction algorithms is particularly important. Predictive models are expected to accelerate research processes, reduce costs, and play a broad role in the biomedical field.

This prediction task also faces many challenges. Firstly, data on protein-DNA binding sites is relatively scarce, making it difficult to obtain representative training data. Secondly, compared to other bioinformatics tasks, predicting protein-DNA binding sites requires consideration of more factors, such as the physicochemical properties of protein and DNA sequences and the interactions between sequences[2].

In previous studies, a range of approaches have been adopted for pinpointing sites of protein-DNA interaction, encompassing the analysis of protein sequences, phylogenetic insights, secondary structural examination, and propensities for binding of specific residues. Yu et al.[3] proposed the TargetS model, which integrates these features to construct discriminative features for prediction. Hu et al.[4] developed the TargetDNA method which utilizes primary sequence information, evolutionary data, and predicted solvent accessibility to recognize protein-DNA binding residues. Ding et al.[5] used protein sequences and discrete cosine transformations to extract features and combined relative solvent accessibility information to improve accuracy. Shen et al.[6] introduced MLAB method, which harnesses local evolutionary information from primary sequence data classified DNA-protein binding sites without 3D information. PredDBR[7], introduced by Hu and colleagues, detects residues that interact with DNA through the analysis of PSFM, PSS, and ligand-binding residue prediction probabilities as features, when the spatial configuration of the protein is not accessible. Wang et al.[8] introduced iDRNA-ITF on sequence data, which utilizes inductive and transfer framework to incorporate the functional characteristics of residues. This approach aids in the recognition of residues that interact with DNA and RNA.

Dealing with imbalanced datasets poses a significant challenge in forecasting protein-DNA interaction regions. Zhu et al.[3] designed the E-HDSVM algorithm, using a two-stage imbalanced learning approach to develop DNAPred, resulting in substantial improvements in the precision of determining protein-DNA interaction regions. Song et al.[9] crafted a refinement method harnessing the interaction likelihood of focal residues and their surroundings to rectify the skewed outcomes from the classifier, thereby accurately foreseeing the particular residues involved in protein-DNA interfaces.

Neural networks' progress has led to a growing adoption of deep learning models for this particular task. Cao and colleagues recommended employing deep learning models with convolutional layers to enhance accuracy in predicting protein-DNA binding sites[10]. In 2019, Nguyen et al.[11] created the iProDNA-CapsNet, which utilizes a position-specific scoring matrix (PSMM) and capsule neural networks (CapsNets). Hendrix et al.[12] designed and tested a deep learning framework called DeepDISE using the 3D coordinates and surface atom

979-8-3315-3399-1/24/$31.00 ©2024 IEEE
DOI 10.1109/ICCBD-AI65562.2024.00009

types of proteins to successfully predict DNA binding residues. Guan et al.[13] proposed a seq2seq model utilizing transformers to extract sequence features and improving identifying performance of protein-DNA binding residues. Yuan et al.[14] introduced GraphSite, a novel approach for identifying DNA-binding residues building upon AlphaFold2's protein structure prediction capabilities. However, proteins, as biological macromolecules, have complex topological structures, and CNNs are mainly used to process data in Euclidean space, which is not convenient for processing non-Euclidean protein topological structures[15].

This study proposes an innovative approach that adopts graph neural networks to forecast protein-DNA binding sites. The model first converts protein sequences into graph-structured data, then uses graph inductive representation learning models to extract features of nodes (amino acids) in the graph, and finally uses MLP to predict binding sites. The model's efficacy has been confirmed through extensive testing across various datasets.

II. METHODS

A. Dataset Description

Assessing the model's forecasting accuracy involved the use of five distinct datasets for protein-DNA interfaces: 543[4], 335, 316[3], 52[6], and 41[4]. Detailed statistics of these datasets are presented in TABLE I. , the datasets 543, 335, and 316 were utilized in ten-part cross-validation assessments, whereas 41 and 52 served for standalone verification exercises.

The PDNA-543 collection encompasses 543 proteins that bind to DNA, while the PDNA-41 set contains 41 of these protein arrangements. In the combined set of these two datasets, the homology between any two sequences does not exceed 30%.

The PDNA-335 dataset is composed of 335 sequences of DNA-binding proteins, and the PDNA-52 dataset contains 52 sequences. Within the PDNA-335 dataset, no sequence has a pairwise identify exceeding 40% with any sequence in PDNA-52 dataset.

Lastly, the PDNA-316 set, assembled by researchers including Si[16], consists of 316 DNA-binding protein sequences, with no more than 30% sequence similarity among any pair within the compilation.

TABLE I. SUMMARY STATISTICS FOR THE FIVE BENCHMARK DATASETS

Dataset	No. of Sequences	No. of Positive samples	No. of Negative samples	Imbalance Ratio
PDNA-543	543	9549	134995	14.14
PDNA-335	335	6461	71320	11.04
PDNA-316	316	5609	67109	11.96
PDNA-52	52	973	16225	16.68
PDNA-41	41	734	14021	19.10

B. Framework of iProtDNA-SAGE

The iProtDNA-SAGE model discussed in this paper employs GraphSAGE[16] for the identification of binding regions involving proteins and DNA. The comprehensive framework of this model is depicted in Figure 1. This process typically involves the subsequent steps:

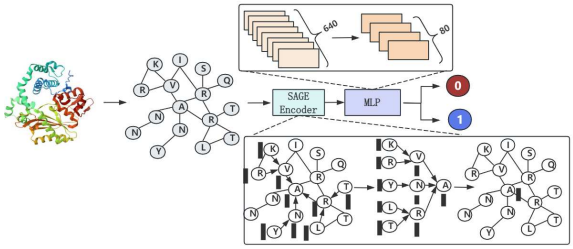

Figure 1. The framework of iProtDNA-SAGE.

1) Sequence to Graph Conversion: Protein sequences are converted into graph structures with each amino acid corresponding to a node. Features of the amino acid residue are extracted using the pre-trained ESM2 model[3], assigning each amino acid a feature vector of 5120 dimensions. The depiction of the protein's adjacency matrix is outlined below:

a) First, obtain the corresponding PDB file from the Protein Data Bank, based on the protein sequence ID in the dataset.

b) Extract the three-dimensional coordinates of the Cα atoms of each residue from the PDB file. Then, represent the coordinates of the Cα atoms as the spatial positions of the residue molecules[17].

c) Calculate the distance between the Cα atoms of each residue pair within the residue sequence and generate a distance matrix. For example, for a residue sequence of length N, an N×N distance matrix will be generated. The matrix is a symmetric, with entries at position (i, j) and (j, i) representing the spatial separation of the Cα atoms from residues i and j, respectively.

d) Select a distance threshold; when the distance between two residues is less than this threshold, it is considered that there is an interaction relationship between these two residues[18]. Then, set the values in the distance matrix that are less than the distance threshold to 1, and the rest to 0, while setting the main diagonal of the matrix to 0, thus obtaining the adjacency matrix of the protein graph.

2) Feature Extraction: Informatin pertaining to the protein's structure and function is captured by extracting features from its graph representation. These features may include the amino acid's properties, the connectivity of nodes, and the overall topology of the graph.

3) Graph Convolutional Networks (GCNs): Graph convolutional layers, such as the SAGEConv layer mentioned earlier, are applied to learn node representations. These layers integrate information from the surrounding nodes of a given node to refine its feature vector, allowing the model to capture the local structure around each amino acid.

4) Training and Prediction): the MLP is employed for the identification of sites where proteins interact with DNA. After passing through the SAGEConv layer, the model obtains a set of feature vectors representing the protein-DNA complex, which are then input into the MLP for further processing and classification.

C. Evaluation Indexes

To ensure a comprehensive and fair evaluation of the model's performance, five metrics are utilized: Acc, Sen, Spe, Pre, and the MCC. The calculations for these metrics are detailed below:

$$\mathrm{Sen} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{1}$$

$$\mathrm{Spe} = \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FP}} \tag{2}$$

$$\mathrm{Acc} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{FN} + \mathrm{TN} + \mathrm{FP}} \tag{3}$$

$$\mathrm{Pre} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \tag{4}$$

$$\mathrm{MCC} = \frac{\mathrm{TP} \cdot \mathrm{TN} - \mathrm{FP} \cdot \mathrm{FN}}{\sqrt{(\mathrm{TP} + \mathrm{FN})(\mathrm{TP} + \mathrm{FP})(\mathrm{TN} + \mathrm{FN})(\mathrm{TN} + \mathrm{FP})}} \tag{5}$$

In this context, TP refers to the number of instances correctly identified as positive, while TN represents the number of instances correctly identified as negative; FP refers to the number of instances incorrectly identified as positive rather than negative; FN pertains to the number of instances incorrectly identified as negative rather than positive.

Specifically, the present study predicts a binary classification problem with class imbalance. Therefore, the MCC holds particular significance as it comprehensively considers all aspects of the confusion matrix, thereby providing a more balanced measure that is independent of the class distribution. This makes it an excellent choice for evaluating models on imbalanced datasets.

III. Results and Discussions

A. Distance threshold between amino acids

The distance threshold between amino acids is a crucial parameter in constructing the protein graph. To determine the optimal threshold, we set different values on the datasets 543, 316, and 335, conducting cross-validation across ten partitions. The outcomes are presented in TABLE II to IV.

The experimental result reveals that the MCC metric achieves its optimal value across all three datasets when the distance threshold is fixed at 8Å. As the distance threshold increases, the MCC tends to decrease. A higher threshold results in more edges being generated in the graph, which, while allowing the target residues to aggregate information from more adjacent residues, may also introduce more noise. This could avoid the model quickly memorizing the dataset throughout the ten-fold cross-validation training process.

B. Comparison with Other Methods

1) Comparison and Analysis of 10-Fold Cross-Validation on the Training Set

On the 316 dataset, iProDNA-SAGE was subjected to a ten-fold cross-validation process, with results compared against models like DBS-PRED[3], BindN[3], DNABindR[19], DP-Bind[20], BindN-rf[21], TargetDNA[4], EC-RUS[5], DNAPred, PredDBR[7], and ULDNA[3]. The results are presented in TABLE V.

TABLE II. TEN-FOLD CROSS VALIDATION RESULTS OF IPROTDNA-SAGE ON PDNA-543

Threshold(Å)	Sen	Spe	Acc	Pre	MCC
8	**0.802**	**0.978**	**0.966**	**0.719**	**0.741**
10	0.783	0.977	0.964	0.709	0.726
12	0.784	0.975	0.963	0.692	0.717
15	0.752	**0.978**	0.963	0.710	0.711
18	0.742	**0.978**	0.963	0.708	0.705

a. The parameters for the FocalLoss loss function are set to α=0.85 and γ=2.25.

TABLE III. TEN-FOLD CROSS VALIDATION RESULTS OF IPROTDNA-SAGE ON PDNA-316

Threshold(Å)	Sen	Spe	Acc	Pre	MCC
8	0.779	0.976	**0.961**	**0.733**	**0.735**
10	**0.784**	0.974	0.959	0.714	0.726
12	0.772	0.973	0.958	0.709	0.717
15	0.736	**0.977**	0.959	0.729	0.710
18	0.745	0.975	0.958	0.716	0.707

b. The parameters for the FocalLoss loss function are set to α=1 and γ=2.

TABLE IV. TEN-FOLD CROSS VALIDATION RESULTS OF IPROTDNA-SAGE ON PDNA-335

Threshold(Å)	Sen	Spe	Acc	Pre	MCC
8	**0.813**	**0.973**	**0.960**	**0.733**	**0.750**
10	0.810	0.969	0.956	0.706	0.732
12	0.811	0.969	0.956	0.704	0.732
15	0.798	0.970	0.957	0.713	0.731
18	0.778	0.972	0.956	0.714	0.721

c. The parameters for the FocalLoss loss function are set to α=1 and γ=2.

TABLE V. COMPARISON OF TEN-FOLD CROSS-VALIDATION RESULTS OF IPROTDNA-SAGE WITH OTHER METHODS ON PDNA-316

Method	Sen	Spe	Acc	MCC
DBS-PRED	0.530	0.760	0.750	0.170
BindN	0.540	0.800	0.780	0.210
DNABindR	0.660	0.740	0.730	0.230
DP-Bind	0.690	0.790	0.780	0.290
BindN-rf	0.670	0.830	0.820	0.320
TargetDNA	0.430	0.950	0.910	0.375
EC-RUS(WSRC)	0.511	0.950	0.916	0.439
DNAPred	0.521	0.951	0.918	0.452
PredDBR	0.561	0.953	0.921	0.497
ULDNA	0.676	0.950	0.929	0.561
iProtDNA-SAGE	**0.780**	**0.976**	**0.961**	**0.735**

On the PDNA-543 dataset, iProDNA-SAGE was subjected to ten-fold cross-validation, and its performance was evaluated against other models such as iProDNA-CapsNet[11], TargetDNA, Hierarchical Feature[13], DNAPred, EC-RUS, PredDBR, and ULDNA. The results can be found in TABLE VI.

TABLE VI. COMPARISON OF TEN-FOLD CROSS-VALIDATION RESULTS OF IPROTDNA-SAGE WITH OTHER METHODS ON PDNA-543

Method	Sen	Spe	Acc	MCC
iProDNA-CapsNet	0.642	0.850	0.837	0.313
TargetDNA	0.406	0.950	0.914	0.339
Hierarchical Feature	0.452	0.954	0.928	0.352
DNAPred	0.449	0.950	0.917	0.373
EC-RUS (WSRC)	0.476	0.949	0.918	0.392
PredDBR	0.454	0.955	0.914	0.415
ULDNA	0.668	0.950	0.931	0.534
iProtDNA-SAGE	**0.802**	**0.978**	**0.966**	**0.741**

On the PDNA-335 dataset, iProDNA-SAGE underwent ten-fold cross-validation, with performance evaluated against PredDBR, TargetS, EC-RUS, and

DNAPred models. Details of these findings are presented in TABLE VII.

TABLE VII. COMPARISON OF TEN-FOLD CROSS-VALIDATION RESULTS OF IPROTDNA-SAGE WITH OTHER METHODS ON PDNA-335

Method	Sen	Spe	Acc	MCC
EC-RUS	0.4870	0.951	0.926	0.378
TargetS	0.4170	0.945	0.899	0.362
DNAPred	0.5430	0.917	0.886	0.390
PredDBR	0.4259	0.953	0.910	0.390
ULDNA	0.6760	0.948	0.925	0.565
iProtDNA-SAGE	**0.8130**	**0.973**	**0.960**	**0.750**

Tables V through VII demonstrate that iProDNA-SAGE significantly outperforms other prediction models on the datasets 316, 543, and 335. Notably, its MCC metric exceeds that of the highest-performing ULDNA model by approximately 0.18.

2) Comparison and Analysis on Independent Test Sets

To expand the evaluation of iProDNA-SAGE's accuracy, this study conducted standalone tests on 41 test cases using the 543 dataset as the training collection, and on 52 test cases using the 335 dataset as the training data. The model was then compared with the two current top-performing predictors, PredDBR and ULDNA, in comparative experiments, with the results presented in TABLE VIII.

TABLE VIII. COMPARISON OF INDEPENDENT TESTING OF IPROTDNA-SAGE ON PDNA-41 AND PDNA-52

DataSet	Method	Sen	Spe	Acc	MCC
PDNA-41	PreDBR	0.391	0.968	0.939	0.359
	ULDNA	**0.556**	0.970	0.950	**0.499**
	iProtDNA-SAGE	0.492	**0.978**	**0.954**	0.489
PDNA-52	PreDBR	0.539	**0.958**	**0.935**	0.451
	ULDNA	**0.704**	0.944	0.931	**0.517**
	iProtDNA-SAGE	0.679	0.946	0.931	0.505

As shown in TABLE VIII. , iProDNA-SAGE achieved excellent performance on both independent test sets. It demonstrated substantial enhancement compared to PredDBR, achieving an MCC boost of 0.13 for the 41 dataset and a 0.05 rise for the 52 dataset. While iProDNA-SAGE's performance is slightly behind the current best predictor, ULDNA, the difference is minimal. However, iProDNA-SAGE demonstrated superior performance in the ten-fold cross-validation across the three training sets, with remarkably consistent results across all datasets.

These series of comparative experimental results indicate that iProDNA-SAGE exhibits stable performance across different datasets, achieving excellent results in both cross-validation and independent testing.

IV. CONCLUSION

This paper introduces an approach leveraging both sequence and structural data to forecast regions where proteins interact with DNA. It begins by creating a protein graph, with nodes representing the protein's residues, with connections between nodes determined by a truncated distance threshold. Subsequently, graph neural networks are employed to generate feature representations for the nodes, followed by a classifier for node classification.

Additionally, ESM2, which is trained beforehand on protein sequences, is employed to derive characteristics for the amino acid units. These features are subsequently merged with the GraphSAGE approach to construct an encoder that refines the node embeddings. To boost the accuracy of the protein-DNA interaction site forecaster on skewed datasets, the FocalLoss function is implemented as the objective function. Experiments conducted across five benchmark datasets and comparisons with various existing algorithms demonstrate that iProtDNA-SAGE achieves significant improvements in identifying interaction regions between proteins and DNA, effectively validating the model's strengths.

ACKNOWLEDGMENT

We are grateful for the financial backing from the National Natural Science Foundation of China (Grant No. 62162032) and the Education Department of Jiangxi Province's technology projects (GJJ2201040), which was pivotal for accomplishing this study. Their contributions have significantly enhanced our work, and we are thankful for the opportunities provided by these institutions.

REFERENCES

[1] Zhang L, Liu T. PDNAPred: Interpretable prediction of protein-DNA binding sites based on pre-trained protein language models. Int J Biol Macromol. 2024;281 (Pt 2):136147.

[2] Daanial Khan Y, Alkhalifah T, Alturise F, et al. DeepDBS: Identification of DNA-binding sites in protein sequences by using deep representations and random forest. Methods. 2024; 231:26-36.

[3] Zhu Y, Yu D J. ULDNA: Integrating Unsupervised Multi-Source Language Models with LSTM-Attention Net-work for Protein-DNA Binding Site Prediction[J]. bioRxiv, 2023: 2023.05. 30.542787.

[4] Hu J, Li Y, Zhang M, et al. Predicting protein-DNA bind-ing residues by weightedly combining sequence-based features and boosting multiple SVMs[J]. IEEE/ACM transactions on computational biology and bioinformatics, 2016, 14(6): 1389-1398.

[5] Ding Y, Tang J, Guo F. Identification of protein–ligand binding sites by sequence information and ensemble classifier[J]. Journal of Chemical Information and Mod-eling, 2017, 57(12): 3149-3161.

[6] Shen C, Ding Y, Tang J, et al. Identification of DNA–protein binding sites through multi-scale local average blocks on sequence information[J]. Molecules, 2017, 22(12): 2079.

[7] Hu J, Bai Y S, Zheng L L, et al. Protein-dna binding res-idue prediction via bagging strategy and sequence-based cube-format feature[J]. IEEE/ACM transactions on com-putational biology and bioinformatics, 2021, 19(6): 3635-3645.

[8] Wang N, Yan K, Zhang J, et al. iDRNA-ITF: identifying DNA-and RNA-binding residues in proteins based on induction and transfer framework[J]. Briefings in Bioin-formatics, 2022, 23(4): bbac236.

[9] Song J, Liu G, Jiang J. A novel prediction method for protein DNA-binding residues based on neighboring residue correlations[J]. Biotechnology & Biotechnologi-cal Equipment, 2022, 36(1): 865-877.

[10] Cao Z, Zhang S. Simple tricks of convolutional neural network architectures improve DNA–protein binding prediction[J]. Bioinformatics, 2019, 35(11): 1837-1843.

[11] Nguyen B P, Nguyen Q H, Doan-Ngoc G N, et al. iProDNA-CapsNet: identifying protein-DNA binding residues using capsule neural networks[J]. BMC bioin-formatics, 2019, 20: 1-12.

[12] Hendrix S G, Chang K Y, Ryu Z, et al. DeepDISE: DNA binding site prediction using a deep learning method[J]. International Journal of Molecular Sciences, 2021, 22(11): 5510.

[13] Guan S, Zou Q, Wu H, et al. Protein-dna binding residues prediction using a deep learning model with hierarchical feature extraction[J]. IEEE/ACM Transactions on Com-putational Biology and Bioinformatics, 2022.

[14] Shi W, Singha M, Pu L, et al. Graphsite: ligand binding site classification with deep graph learning[J]. Biomolecules, 2022, 12(8): 1053

[15] Xia B B, Wang J. Protein modeling and design based on deep learning. Chinese Journal of Biotechnology, 2021, 37(11): 3863-3879.

[16] Hamilton W, Ying Z, Leskovec J. Inductive representation learning on large graphs[J]. Advances in neural infor-mation processing systems, 2017, 30.

[17] Atilgan A R, Akan P, Baysal C. Small-world communica-tion of residues and significance for protein dynamics[J]. Biophysical journal, 2004, 86(1): 85-91.

[18] Greene L H, Higman V A. Uncovering network systems within protein structures[J]. Journal of molecular biology, 2003, 334(4): 781-791.

[19] Yan C, Terribilini M, Wu F, et al. Predicting DNA-binding sites of proteins from amino acid sequence[J]. BMC bio-informatics, 2006, 7: 1-10.

[20] Hwang S, Gou Z, Kuznetsov I B. DP-Bind: a web server for sequence-based prediction of DNA-binding residues in DNA-binding proteins[J]. Bioinformatics, 2007, 23(5): 634-636.

[21] Wang L, Yang M Q, Yang J Y. Prediction of DNA-binding residues from protein sequence information using random forests[J]. Bmc Genomics, 2009, 10: 1-9.