Advancing Identification of DNA-Protein Binding Residues Using Deep Learning Techniques

Haipeng Zhao [1] and Hongjie Wu [1]

[1] Suzhou University of Science and Technology
hongjie.wu@qq.com

Abstract. Accurate identification of DNA-protein binding sites is vital for understanding biological processes and facilitating drug discovery. This study introduces a novel method that integrates a Transformer encoder with Bi-directional Long Short-Term Memory (BiLSTM) to predict DNA-protein binding residues effectively. The method enriches protein representation by combining evolutionary information from the position-specific scoring matrix (PSSM) with spatial information from predicted secondary structures. Experimental results demonstrate the method's competitiveness, achieving 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.

1 Introduction

DNA-protein interactions are critical for biological processes like transcription and DNA repair. Identifying binding sites is essential for understanding gene regulation and disease mechanisms and for drug design. Traditional experimental methods are costly and time-consuming. Computational methods offer a more efficient alternative.

Given the importance of protein-DNA binding, many wet-lab methods have been developed to identify protein-DNA binding residues. These methods include X-ray crystallography [6], Fast ChIP [7], and electrophoretic mobility shift assays (EMSAs) [8,9]. Although wet-lab methods can yield precise identification outcomes, they are expensive and labor intensive. Moreover, they cannot keep up with the growth rate of protein sequences in the post-genomic era [10]. Therefore, there is a need to develop an efficient and convenient computation-based method for identifying protein-DNA binding residues. With advancements in computer theory, a number of computational methods have emerged for this purpose. These methods can be broadly categorized into three types: sequence-based, structure-based, and hybrid methods [11].

Bioinformatics research primarily focuses on sequence-based methods, which pose a significant challenge. Predicting protein-DNA binding residues using only sequence-based features may have poor performance due to the limited information contained in protein sequences. However, the number of protein sequences is increasing day by day, research in this area is still focused on utilizing sequence features. In the past decade, several sequence-based methods have been proposed. These include BindN [12],

ProteDNA [13], DP-Bind [14], BindN+ [15], MetaDBSite [16], TargetDNA [17], DNABind [18], DNAPred [19] and PredDBR [20], among others. In BindN, they utilized three types of protein sequence features: hydrophobicity, side chain pKa value, and molecular mass of amino acids. These features were inputted into a support vector machine (SVM) to accurately predict protein-DNA binding residues. In DP-Bind, they utilized evolutionary information obtained from protein sequences, specifically the position-specific scoring matrix (PSSM) [21]. To enhance the recognition accuracy of protein-DNA binding residues, three conventional machine learning techniques were combined: penalized logistic regression, SVM, and kernel logistic regression. In TargetDNA, they used two protein sequence features, solvent accessibility and evolutionary information, and made use of an undersampling technique to divide the raw data into multiple sub-datasets and applied multiple SVMs for ensemble learning to predict protein-DNA binding residues.

Structure-based methods utilize either natural or predicted 3D structure information of proteins. This is because the 3D structure of a protein contains a large amount of information and the structure of a protein determines the function of the protein to some extent. Consequently, utilizing protein structure information for predicting protein-DNA binding residues often yields better performance than sequencebased methods. Common structure-based methods include: DBD-Hunter [22], DNABINDPROT [23], DR_bind [24], PreDs [25], etc. All these methods mentioned above use only the structure information of the protein and ignore the information that may be contained in the protein sequence that may be helpful in predicting the protein-DNA binding residues. To enhance prediction accuracy, hybrid methods integrate both sequence and structure information. Some common hybrid methods include: TargetATP [26], COACH [27], TargetS [28], SVMPred [29] and NsitePred [30], etc. In DR_bind, the model predicts protein-DNA binding residues by utilizing evolutionary, geometric and electrostatic properties to describe the protein structure. In COACH, they designed an algorithm named TM-SITE to infer binding sites from homologous structural templates and also an algorithm named S-SITE for sequence.

2 Method

The study uses the PDNA-543 and PDNA-41 datasets, enriching protein features by combining PSSM evolutionary information with secondary structure predictions. The model architecture includes a Transformer encoder, BiLSTM, and a convolutional feature extraction module, followed by a multilayer perceptron (MLP) decoder for residue classification.

PSSM features were generated using PSI-BLAST, and secondary structure predictions were made using PSIPRED. These features were combined to form a comprehensive protein representation.

The model integrates a Transformer encoder and BiLSTM to capture long-range dependencies and local residue features. A convolutional layer processes the encoded protein feature matrix, and an MLP decoder generates the binding pattern.

2.1 Framework

The PDNA-543 and PDNA-41 datasets were utilized, with the former used for training and the latter for testing the model's generalization performance.

2.2 Train

The model was evaluated using the DUD-E dataset and the Human dataset, which are standard benchmarks for DTI prediction.

The model was trained using binary cross-entropy loss and the Adam optimizer. Evaluation metrics included MCC, SP, SN, and ACC.

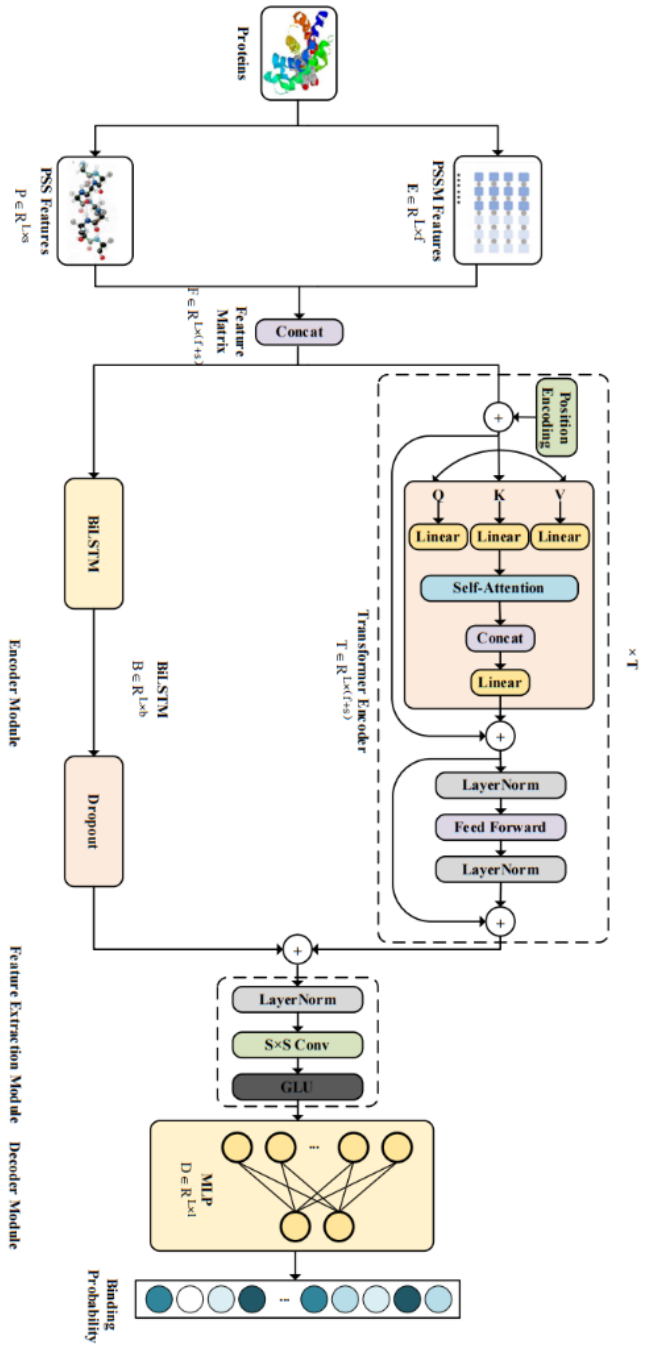

Fig. 1. Architecture.

3 Results

The proposed method demonstrated improved performance over existing classifiers, with significant improvements in MCC, SP, SN, and ACC on the PDNA-41 dataset. The combination of Transformer encoder and BiLSTM effectively captured both global and local residue features.

The study presents a robust method for identifying DNA-protein binding residues using deep learning. The method's effectiveness lies in its ability to capture long-range dependencies and local features, offering a user-friendly approach that requires only protein sequences as input. Future work will explore incorporating three-dimensional structural information and graph neural networks for further enhancements.

In this study, we propose an encoder-decoder model to predict protein-DNA binding sites. To represent a protein sequence, we use two sequence-based features, the evolutionary feature PSSM and the predicted secondary structure, respectively. Unlike current state-of-the-art methods, our model enables end to end prediction of an entire protein sequence without the need for feature pre-extraction for each residue using a sliding window technique, which demonstrates the ease of use of our model. Comparing with previous methods, our model achieves respectable performance on the PDNA-41 test set (MCC:0.343, SP:96.37%, SN:46.34%, ACC:94.79%), which proves the effectiveness of our model.

While our method has made some progress and can handle variable length protein sequences, it also limits our model to one protein input at a time. Therefore, we will further try more models for the problem of inconsistent protein sequence lengths. Given the success of graph neural networks in bioinformatics, we will try to employ graph structures to represent protein sequences to identify DNA binding residues. In addition, the features used in this work could be improved. With the great achievements in the field of protein structure prediction in recent years, we can use the predicted structural information to aid in this task.

Acknowledgments. This research was supported by the National Natural Science Foundation of China and other institutions.

Disclosure of Interests. The authors declare no conflict of interest.

References

Dobson C M. Chemical space and biology[J]. Nature, 2004, 432(7019): 824-828.11
Gao M, Skolnick J. The distribution of ligand-binding pockets around protein-protein interfaces suggests a general mechanism for pocket formation[J]. Proceedings of the National Academy of Sciences, 2012, 109(10): 3784-3789.
Zhao J, Cao Y, Zhang L. Exploring the computational methods for protein-ligand binding site prediction[J]. Computational and structural biotechnology journal, 2020, 18: 417-426.
Ofran Y, Mysore V, Rost B. Prediction of DNA-binding residues from sequence[J]. Bioin

formatics, 2007, 23(13): i347-i353.
5. Jones S, Van Heyningen P, Berman H M, et al. Protein-DNA interactions: a structural analysis[J]. Journal of molecular biology, 1999, 287(5): 877-896.
6. Smyth M S, Martin J H J. x Ray crystallography[J]. Molecular Pathology, 2000, 53(1): 8.
7. Nelson J D, Denisenko O, Bomsztyk K. Protocol for the fast chromatin immunoprecipitation (ChIP) method[J]. Nature protocols, 2006, 1(1): 179-185.
8. Heffler MA, Walters RD, Kugel J F. Using electrophoretic mobility shift assays to measure equilibrium dissociation constants: GAL4(p53 binding DNA as a model system[J]. Biochemistry and Molecular Biology Education, 2012, 40(6): 383-387 .
9. Hellman L M, Fried M G. Electrophoretic mobility shift assay (EMSA) for detecting protein±nucleic acid interactions[J]. Nature protocols, 2007, 2(8): 1849-1861.
10. Vajda S, Guarnieri F. Characterization of protein-ligand interaction sites using experimental and computational methods[J]. Current Opinion in Drug Discovery and Development, 2006, 9(3): 354.
11. Ding Y, Yang C, Tang J, et al. Identification of protein-nucleotide binding residues via graph regularized k-local hyperplane distance nearest neighbor model[J]. Applied Intelligence, 2022: 1-15.
12. Wang L, Brown S J. BindN: a web-based tool for efficient prediction of DNA and RNA binding sites in amino acid sequences[J]. Nucleic acids research, 2006, 34(suppl_2): W243-W248.
13. Chu W Y, Huang Y F, Huang C C, et al. ProteDNA: a sequence-based predictor of sequence-specific DNA-binding residues in transcription factors[J]. Nucleic acids research, 2009, 37(suppl_2): W396-W401.
14. 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.
15. Wang L, Huang C, Yang M Q, et al. BindN+ for accurate prediction of DNA and RNA binding residues from protein sequence features[J]. BMC Systems Biology, 2010, 4: 1-9.
16. Si J, Zhang Z, Lin B, et al. MetaDBSite: a meta approach to improve protein DNA-binding
sites prediction[J]. BMC systems biology, 2011, 5(1): 1-7.
17. Hu J, Li Y, Zhang M, et al. Predicting protein-DNA binding 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.
18. Liu R, Hu J. DNABind: A hybrid algorithm for structure(based prediction of DNA(binding residues by combining machine learning(and template(based approaches[J]. PROTEINS: structure, Function, and Bioinformatics, 2013, 81(11): 1885-1899.
19. Zhu Y H, Hu J, Song X N, et al. DNAPred: accurate identification of DNA-binding sites
from protein sequence by ensembled hyperplane-distance-based support vector ma chines[J]. Journal of chemical information and modeling, 2019, 59(6): 3057-3071.

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

Reference

[1] Chou KC, Elrod DW. Prediction of membrane protein types and subcellular locations. Proteins: Structure, Function, and Bioinformatics 1999; 34(1), 137-153. https://doi.org/10.1002/(SICI)1097-0134(19990101)34:1<137::AID-PROT11>3.0.CO;2-O.
[2] Cai YD, Zhou GP, Chou KC. Support vector machines for predicting membrane protein types by using functional domain composition. Biophysical journal 2003; 84(5): 3257-3263. https://doi.org/10.1016/S0006-3495(03)70050-2.
[3] Cai YD, Chou KC. Predicting membrane protein type by functional domain composition and pseudo-amino acid composition. Journal of Theoretical Biology 2006; 238(2): 395-400. https://doi.org/10.1016/j.jtbi.2005.05.035.
[4] Chou KC, Shen HB. MemType-2L: a web server for predicting membrane proteins and their types by incorporating evolution information through Pse-PSSM. Biochemical and biophysicalresearch communications 2007; 360(2): 339-345. https://doi.org/10.1016/j.bbrc.2007.06.027.
[5] Liu H, Yang J, Wang M, Xue L, Chou KC. Using Fourier spectrum analysis and pseudo amino acid composition for prediction of membrane protein types. The Protein Journal 2005; 24(6):385-389. https://doi.org/10.1007/s10930-005-7592-4.
[6] Shen H, Chou KC. Using optimized evidence-theoretic K-nearest neighbor classifier and pseudo-amino acid composition to predict membrane protein types. Biochemical and biophysical research communications 2005; 334(1): 288-292. https://doi.org/10.1016/j.bbrc.2005.06.087.
[7] Shen HB, Yang J, Chou KC. Fuzzy KNN for predicting membrane protein types from pseudo-amino acid composition. Journal of theoretical biology 2006; 240(1): 9-13. https://doi.org/10.1016/j.jtbi.2005.08.016.
[8] Wang M, Yang J, Liu GP, Xu ZJ, Chou KC. Weighted-support vector machines for predicting membrane protein types based on pseudo-amino acid composition. Protein Engineering Design and Selection 2004; 17(6): 509-516.
[9] Wang M, Yang J, Liu GP, Xu ZJ, Chou KC. Weighted-support vector machines for predicting membrane protein types based on pseudo-amino acid composition. Protein Engineering Design and Selection 2004; 17(6): 509-516.
[10] Liu H, Wang M, Chou KC. Low-frequency Fourier spectrum for predicting membrane protein types. Biochemical and biophysical research communications 2005; 336(3): 737-739. https://doi.org/10.1016/j.bbrc.2005.08.160.
[11] Wang SQ, Yang J, Chou KC. Using stacked generalization to predict membrane protein types based on pseudo-amino acid composition. Journal of theoretical biology 2006; 242(4): 941-946. https://doi.org/10.1016/j.jtbi.2006.05.006.
[12] Chen YK, Li KB. Predicting membrane protein types by incorporating protein topology, domains, signal peptides, and physicochemical properties into the general form of Chou's pseudo amino acid composition. Journal of Theoretical Biology 2013; 318: 1-12. https://doi.org/10.1016/j.jtbi.2012.10.033.
[13] Han GS, Yu ZG, Anh V. A two-stage SVM method to predict membrane protein types by incorporating amino acid classifications and physicochemical properties into

a general form of Chou's PseAAC. Journal of Theoretical Biology 2014; 344: 31-39. https://doi.org/10.1016/j.jtbi.2013.11.017.
[14] Hayat M, Khan A. Predicting membrane protein types by fusing composite protein sequence features into pseudo amino acid composition. Journal of theoretical biology 2011; 271(1): 10-17. https://doi.org/10.1016/j.jtbi.2010.11.017.
[15] Hayat M, Khan A, Yeasin M. Prediction of membrane proteins using split amino acid and ensemble classification. Amino acids 2012; 42(6): 2447-2460. https://doi.org/10.1007/s00726-011-1053-5.
[16] Rezaei MA, Abdolmaleki P, Karami Z, Asadabadi EB, Sherafat MA, Abrishami-Moghaddam, H, Forouzanfar M. Prediction of membrane protein types by means of wavelet analysis and cascaded neural networks. Journal of theoretical biology 2008; 254(4): 817-820. https://doi.org/10.1016/j.jtbi.2008.07.012.
[17] Shen Y, Tang J, Guo F. Identification of protein subcellular localization via integrating evolutionary and physicochemical information into Chou's general PseAAC. Journal of Theoretical Biology 2019; 462: 230-239. https://doi.org/10.1016/j.jtbi.2018.11.012.
[18] Wang Y, Ding Y, Guo F, Wei L, Tang J. Improved detection of DNA-binding proteins via compression technology on PSSM information. PloS one 2017; 12(9): e0185587. https://doi.org/10.1371/journal.pone.0185587.
[19] Shen C, Ding Y, Tang J, Xu X, Guo F. An ameliorated prediction of drug–target interactions based on multi-scale discrete wavelet transform and network features. International journal of molecular sciences 2017; 18(8): 1781. https://doi.org/10.3390/ijms18081781.
[20] Ahmed N, Natarajan T, Rao KR. Discrete cosine transform. IEEE transactions on Computers 1974; 100(1): 90-93. https://doi.org/10.1109/T-C.1974.223784.
[21] Ding Y, Tang J, Guo F. Identification of protein–protein interactions via a novel matrix-based sequence representation model with amino acid contact information. International journal of molecular sciences 2016; 17(10): 1623. https://doi.org/10.3390/ijms17101623.
[22] Boeckmann B, Bairoch A, Apweiler R, Blatter MC, Estreicher A, Gasteiger E, Schneider M. The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. Nucleic acids research 2003; 31(1): 365-370.
[23] Li W, Godzik A, Cd-hit: a fast program for clustering and comparing large sets of protein or
nucleotide sequences, Bioinformatics 2006; 22(13): 1658-1659.
[24] Limin Fu, Beifang Niu, Zhengwei Zhu, Sitao Wu and Weizhong Li, CD-HIT: accelerated for clustering the next generation sequencing data. Bioinformatics 2012; 28(23): 3150-3152. https://doi.org/10.1093/bioinformatics/bts565.
[25] cheol Jeong J, Lin X, Chen XW. On position-specific scoring matrix for protein function prediction. IEEE/ACM transactions on computational biology and bioinformatics 2010; 8(2): 308-315. https://doi.org/10.1109/TCBB.2010.93.
[26] Nanni L, Brahnam S, Lumini A. Wavelet images and Chou's pseudo amino acid composition for protein classification. Amino Acids 2012; 43(2): 657-665. https://doi.org/10.1007/s00726-011-1114-9.
[27] Zhou D, Huang J, Schölkopf B. Learning with hypergraphs: Clustering, classification, and embedding. Advances in neural information processing systems 2006; 19.

[28] Huang Y, Liu Q, Metaxas D. Video object segmentation by hypergraph cut. In 2009 IEEE conference on computer vision and pattern recognition 2009; 1738-1745. https://doi.org/10.1109/CVPR.2009.5206795.
[29] Huang Y, Liu Q, Zhang S, Metaxas DN. Image retrieval via probabilistic hypergraph ranking. In 2010 IEEE computer society conference on computer vision and pattern recognition 2010; 3376-3383. https://doi.org/10.1109/CVPR.2010.5540012.
[30] Gao Y, Wang M, Zha ZJ, Shen J, Li X, Wu X. Visual-textual joint relevance learning for tag-based social image search. IEEE Transactions on Image Processing 2012; 22(1): 363-376. https://doi.org/10.1109/TIP.2012.2202676.
[31] Hwang T, Tian Z, Kuangy R, Kocher JP. Learning on weighted hypergraphs to integrate protein interactions and gene expressions for cancer outcome prediction. In 2008 Eighth IEEE International Conference on Data Mining 2008; 293-302. https://doi.org/10.1109/ICDM.2008.37.
[32] Gao Y, Wang M, Tao D, Ji R, Dai Q. 3-D object retrieval and recognition with hypergraph analysis. IEEE Transactions on Image Processing 2012; 21(9): 4290-4303. https://doi.org/10.1109/TIP.2012.2199502.
[33] Feng Y, You H, Zhang Z, Ji R, Gao Y. Hypergraph neural networks. In Proceedings of the AAAI conference on artificial intelligence 2019; 33(1): 3558-3565. https://doi.org/10.1609/aaai.v33i01.33013558.
[34] Henaff M, Bruna J, LeCun Y. Deep convolutional networks on graph-structured data. arXiv preprint arXiv: 1506.05163 2015. https://doi.org/10.48550/arXiv.1506.05163.
[35] Defferrard M, Bresson X, Vandergheynst P. Convolutional neural networks on graphs with fast localized spectral filtering. Advances in neural information processing systems 2016; 29.
[36] Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R. Dropout: a simple way to prevent neural networks from overfitting. The journal of machine learning research 2014; 15(1): 1929-1958.
[37] Kingma DP, Ba J. Adam. A method for stochastic optimization. arXiv preprint arXiv:1412.6980 2014..
[38] Alhamdoosh M, Wang D. Fast decorrelated neural network ensembles with random weights. Information Sciences 2014; 264: 104-117. https://doi.org/10.1016/j.ins.2013.12.016.
[39] Chou KC. Prediction of protein cellular attributes using pseudo‐amino acid composition. Proteins: Structure, Function, and Bioinformatics 2001; 43(3): 246-255. https://doi.org/10.1002/prot.1035.
[40] Wang L, Yuan Z, Chen X, Zhou Z. The prediction of membrane protein types with NPE. IEICE Electronics Express 2010; 7(6): 397-402. https://doi.org/10.1587/elex.7.397.
[41] Shen HB, Chou KC. Using ensemble classifier to identify membrane protein types. Amino acids 2007; 32(4): 483-488. https://doi.org/10.1007/s00726-006-0439-2.