Predicting DNA-Binding Proteins through Advanced Deep Transfer Learning Techniques

Jun Yan [1] and Hongjie Wu [1]

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

Abstract. DNA-binding proteins (DBPs) are crucial in gene-related life activities. Traditional methods for DBP prediction are labor-intensive and costly. We present a novel method using deep transfer learning to predict DBPs efficiently. Our approach extracts sequence and PSSM features, employs transfer learning algorithms to construct datasets, and uses an attention mechanism-equipped neural network for prediction.

Keywords: DNA-binding proteins.

1 Introduction

DBPs play a key role in DNA replication, transcription, regulation, and other cellular processes. Traditional experimental methods for DBP identification are resource-intensive. Computational methods offer a more efficient alternative. We focus on developing a computational approach using deep transfer learning to predict DBPs accurately.

Protein is very important for the human body. Some of these proteins can interact with DNA and are called DNA-binding proteins (DBPs). These are very important for gene-related life activities. For example, in DNA replication and repair functions, origins of replication sites [1] is the location where genomic DNA replication begins, and is important for the study of the DNA replication process. In Mathematical Biosciences and Engineering Volume 19, Issue 8, 7719-7736.transcription and regulatory functions, RNA is an important molecule in the cell. Messenger RNA passes genetic information to DNA and acts as a template for protein synthesis, while only 2% of RNA molecules in proteins act as templates, the rest being a molecule called MicroRNA, which plays an important regulatory role in biological processes. Identifying molecules of MicroRNA [2] helps to understand the whole regulatory process, while some other functions are single-stranded DNA binding and separation functions, chromatin formation functions and cell development functions [3,4]. In addition, research into drug target proteins [5,6] and DNA expression genetics are also quite popular, as drug target proteins are closely related to human diseases, while DNA expression genetics include

DNA N4-methylcytosine [7,8], histone modification, RNA interference, etc. The main study in this paper is DNA binding proteins. Identification of DBPs can help us better understand how proteins interact with DNA, thus promoting the development of life science.

Although the traditional method based on biological experiments can obtain high-precision results, it needs large quantities of time and human effort. In addition, with the advent of the postgenome era, Web-lab methods cannot keep up with the growth rate of protein sequences. By contrast, computational approach reduces the resources and manpower required and enables simple and efficient identification of DBPs from many protein sequences. Thus, for the development of bioinformatics, the use of computational methods to predict DBPs is of great value.

In the past decade, machine learning based algorithms are already getting a lot of attention, and researchers have also proposed several research algorithms. In general, DNA-binding proteins can be identified by two computational methods, one based on structure and the other on sequence. Gao et al. [9] proposed a knowledge-based method called DBD-Hunter. This method uses protein structural alignment and statistical potential energy assessment to predict DBPs. Nimrod et al. [10] used the 3D structure of proteins to predict DBPs. They used a random forest classifier to determine whether a protein was a DBP based on features obtained from the protein's evolutionary profile. Zhao et al. [11] Identification of DBPs proteins using 3D structures generated based on HHblits [12]. However, structure-based approaches rely on predicted or natural 3D protein structures, and obtaining these structures is difficult. As a result, many sequence-based methods have been developed. Kumar et al. [13]developed a random forest approach called DNA-Prot to identify DBPs from protein sequences. Liu et al. [14] developed a predictor called iDNAPro-PseAAC, which relies only on protein sequence

information. They applied PseAAC [15,16] to support vector machines to identify DBPs. Wei et al. [17]used the features extracted from the local PSE-PSSM (pseudo location-specific scoring matrix) in combination with a random forest classifier and to identify DBPs. Mishra et al. [18] proposed a method called StackDPPred, which uses features extracted from PSSM and residue-specific contact energy to help train a stacking-based machine learning method that can effectively predict DNA-binding proteins.

Nanni et al. [19] in order to build an optimal and most general classification system for DNA-binding proteins, features were experimentally extracted from proteins and trained and evaluated in a separate support vector machine, while the matrix of proteins was fine-tuned using convolutional neural networks with different parameter settings, and the decisions were fused with the support vector machine using weights and rules for predicting DBPs. In recent years, deep learning has proven to be very effective in image and natural language processing. Therefore, researchers gradually began to apply deep learning in bioinformatics. Deep learning methods need only to input raw data and do not need to manually extract features, as does machine learning. For example, Qu et al. [20] used a combination of LSTM and CNN and extracted features from protein sequences to predict DBPs.Shadab et al. [21] proposed two methods, DeepDBP-ANN and DeepDBP-CNN, by using deep

2 Method

Our method combines transfer learning with deep learning. We used two transfer learning algorithms, DDC and TrAdaBoost, to expand the dataset and improve prediction

performance. An attention mechanism was integrated into the deep neural network to enhance prediction accuracy.

2.1 Transfer Learning Framework

We utilized transfer learning to extract related datasets and train our model. The framework includes sequence and PSSM feature extraction followed by deep learning model training using an attention mechanism.

We applied DDC to reduce the distribution distance between source and target domains. TrAdaBoost was used for data migration, adjusting weights during iterations to focus on misclassified samples.

An attention mechanism was added to improve model efficiency and accuracy, allowing the model to focus on critical features similarly to human focus.

We used one-hot coding for sequence representation and PSSM for evolutionary information capture.

In the experiments of this study, the main approach to prediction was to use a conjunction of transfer learning and deep learning. First, the transfer learning algorithm was used to extract the data set S, which was related to the target sample, but not completely distributed based on sample similarity.

Then the sequence and PSSM [25] features of data set S were extracted, in a deep network with an attention mechanism, the features are input and trained.

In the deep learning part of this method, the sequence and PSSM features were entered into LSTM [26] and CNN [27] respectively. In subsequent improvements, ResNet [28] was used to replace CNN, and better results were obtained. The final prediction results of these two parts also need to go through the fully connected layer. Figure 1 shows an overall prediction framework, mainly based on the DBP [29,30] prediction framework of deep transfer learning

2.2 Train

Our model was trained using the Adam optimizer in PyTorch with cross-entropy loss over 40 epochs.

We used Accuracy, MCC, Sensitivity, and Specificity as our evaluation metrics.

We used the PDB186 dataset for testing and compared our method with other existing methods. Our method showed superior performance.

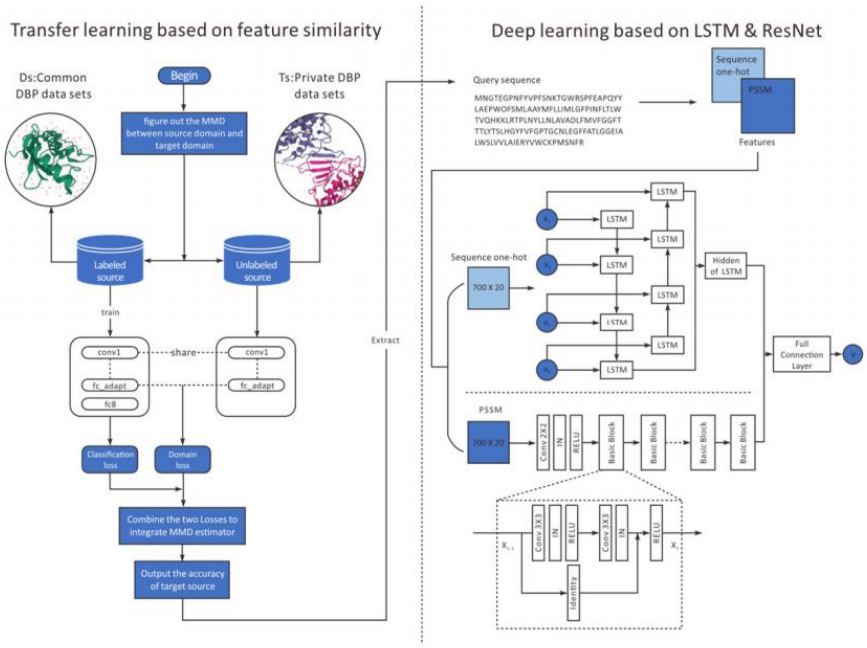

Fig. 1. Architecture.

We compared various neural network models and found that deeper models like ResNet improved performance.

Our transfer learning approach showed significant improvement in model performance with fewer labeled samples. Our deep transfer learning method outperformed traditional machine learning methods, demonstrating better prediction accuracy and robustness.

3 Results

Our deep transfer learning approach for DBP prediction is efficient and accurate. It overcomes the limitations of traditional methods by reducing the need for extensive resources. Future work will focus on improving prediction accuracy by addressing noisy samples.

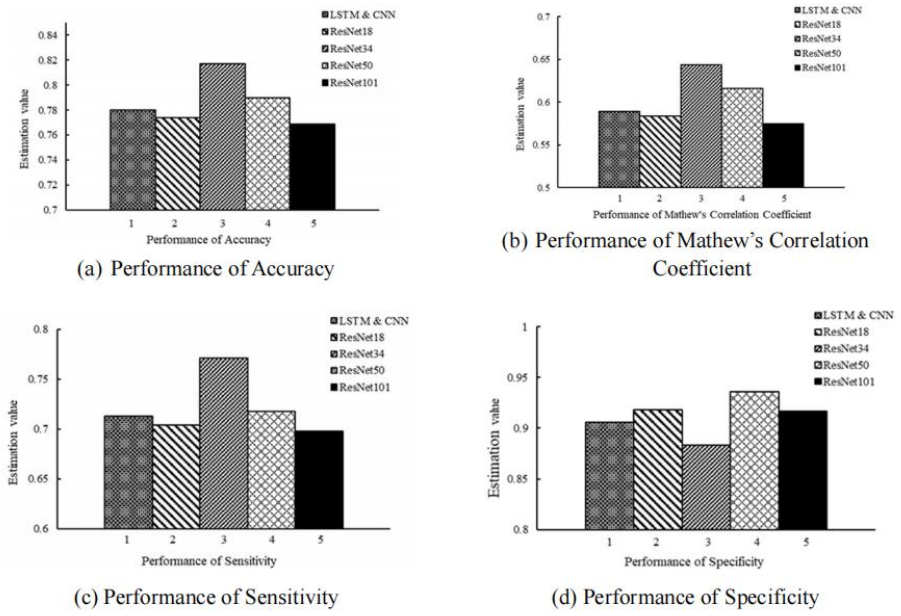

Fig. 2. Results.

The integration of multiscale CNNs and graph representations for drug molecules and protein sequences, respectively, yielded superior results in DTA prediction. Our model offers a promising approach for accelerating drug discovery.

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

L. Wei, W. He, A. Malik, R. Su, L. Cui, B. Manavalan, Computational prediction and interpretation of cell-specific replication origin sites from multiple eukaryotes by exploiting stacking framework, *Briefings Bioinf.*, **22** (2021). https://doi.org/10.1093/bib/bbaa275
2. L. Wei, M. Liao, Y. Gao, R. Ji, Z. He, Q. Zou, Improved and promising identification of human MicroRNAs by incorporating a high-quality negative set, *IEEE/ACM Trans. Comput. Biol. Bioinf.*,
11 (2014), 192–201. https://doi.org/10.1109/TCBB.2013.146
3. D. H. Ohlendorf, W. F. Anderson, R. G. Fisher, Y. Takeda, B.W. Matthews, The molecular basis
of DNA-protein recognition inferred from the structure of cro repressor, *Nature*, **298** (1982), 718–
23. https://doi.org/10.1038/298718a0

4. W. H. Hudson, E. A. Ortlund, The structure, function and evolution of proteins that bind DNA and
RNA, *Nat. Rev. Mol. Cell Biol.*, **15** (2014), 749–760. https://doi.org/10.1038/nrm3884
5. Y. Ding, J. Tang, F. Guo, Q. Zou, Identification of drug-target interactions via multiple kernel based triple collaborative matrix factorization, *Briefings Bioinf.*, **23** (2022), bbab582. https://doi.org/10.1093/bib/bbab582
6. Y. Ding, J. Tang, F. Guo, Identification of drug–target interactions via dual laplacian regularized
least squares with multiple kernel fusion, *Knowl.-Based Syst.*, **204** (2020), 106254. https://doi.org/10.1016/j.knosys.2020.106254
7. Y. Ding, P. Tiwari, Q. Zou, F. Guo, H. M. Pandey, C-loss based Higher-order Fuzzy Inference Systems for identifying DNA N4-methylcytosine Sites, *IEEE Trans. Fuzzy Syst.*, 2022. https://doi.org/10.1109/TFUZZ.2022.3159103
8. Y. Ding, W. He, J. Tang, Q. Zou, F. Guo, Laplacian regularized sparse representation based classifier for identifying DNA N4-methylcytosine Sites via L2,1/2-matrix norm, *IEEE/ACM Trans. Comput. Biol. Bioinf.*, 2021. https://doi.org/10.1109/TCBB.2021.3133309
9. M. Gao, J. Skolnick, DBD-Hunter: a knowledge-based method for the prediction of DNA-protein
interactions, *Nucleic Acids Res.*, **36** (2008), 3978–3992. https://doi.org/10.1093/nar/gkn332
10. G. Nimrod, M. Schushan, A. Szilagyi, C. Leslie, N. Ben-Tal, iDBPs: a web server for the identification of DNA binding proteins,
Bioinformatics,
26 (2010), 692–693.
https://doi.org/10.1093/bioinformatics/btq019
11. H. Zhao, J. Wang, Y. Zhou, Y. Yang, Predicting DNA-binding proteins and binding residues by
complex structure prediction and application to human proteome, PLoS One, (2014), e96694. https://doi.org/10.1371/journal.pone.0096694
12. M. Remmert, A. Biegert, A. Hauser, J. Soding, HHblits: lightning-fast iterative protein sequence
searching by HMM-HMM alignment, Nat. Methods, 9 (2011), 173–175. https://doi.org/10.1038/nmeth.1818
13. K. K. Kumar, G. Pugalenthi, P. N. Suganthan, DNA-Prot: identification of DNA binding proteins
from protein sequence information using random forest, J. Biomol. Struct. Dyn., 26 (2009), 679–686. https://doi.org/10.1080/07391102.2009.10507281
14. B. Liu, S. Wang, X. Wang, DNA binding protein identification by combining pseudo amino acid
composition and profile-based protein representation, Sci. Rep., 5 (2015), 15479. https://doi.org/10.1038/srep15479
15. K. C. Chou, Some remarks on protein attribute prediction and pseudo amino acid composition, J.
Theor. Biol., 273 (2011), 236–247. https://doi.org/10.1016/j.jtbi.2010.12.024
16. K. C. Chou, Prediction of protein cellular attributes using pseudo-amino acid composition, Proteins, 43 (2001), 246–255. https://doi.org/10.1002/prot.1035
17. L. Wei, J. Tang, Q. Zou, Local-DPP: an improved DNA-binding protein prediction method by
exploring local evolutionary information, Inf. Sci., 384 (2017), 135–144. https://doi.org/10.1016/j.ins.2016.06.026

18. A. Mishra, P. Pokhrel, M. T. Hoque, StackDPPred: a stacking based prediction of DNA-binding
protein from sequence, Bioinformatics, 35 (2019), 433–441.
https://doi.org/10.1093/bioinformatics/bty653
19. L. Nanni, S. Brahnam, Robust ensemble of handcrafted and learned approaches for DNA-binding
proteins, Appl. Comput. Inf., 2021. https://doi.org/10.1108/ACI-03-2021-0051
20. Y. H. Qu, H. Yu, X. J. Gong, J. H. Xu, H. S. Lee, On the prediction of DNA-binding proteins only
from primary sequences: a deep learning approach, PLoS One, (2017), e0188129.
https://doi.org/10.1371/journal.pone.0188129
21. S. Shadab, T. A. Khan, N. A. Neezi, S. Adilina, S. Shatabda, DeepDBP: deep neural networks for
identification of DNA-binding proteins, Inf. Med. Unlocked, 19 (2020), 100318.
https://doi.org/10.1016/j.imu.2020.100318
22. S. Ahmad, A. Sarai, PSSM-based prediction of DNA binding sites in proteins, BMC Bioinf., 6
(2005), 33. https://doi.org/10.1186/1471-2105-6-33
23. J. Zhang, Q. Chen, B. Liu, DeepDRBP-2L: a new genome annotation predictor for identifying DNA-binding proteins and RNA-binding proteins using convolutional neural network and long short-term memory, IEEE/ACM Trans. Comput. Biol. Bioinf., 18 (2021), 1451–1463.
https://doi.org/10.1109/TCBB.2019.2952338
24. J. Zhang, Q. Chen, B. Liu, iDRBP_MMC: identifying DNA-binding proteins and RNA-binding
proteins based on multi-label learning model and motif-based convolutional neural network, J. Mol. Biol., 432 (2020), 5860–5875. https://doi.org/10.1016/j.jmb.2020.09.008
25. G. Li, X. Du, X. Li, L. Zou, G. Zhang, Z. Wu, Prediction of DNA binding proteins using local
features and long-term dependencies with primary sequences based on deep learning, PeerJ, 9 (2021), e11262. https://doi.org/10.7717/peerj.11262
26. K. Greff, R. K. Srivastava, J. Koutnik, B. R. Steunebrink, J. Schmidhuber, LSTM: a search space
odyssey, IEEE Trans. Neural Networks Learn. Syst., 28 (2017), 2222–2232.
https://doi.org/10.1109/TNNLS.2016.2582924
27. T. Roska, L. O. Chua, The CNN universal machine: an analogic array computer, IEEE Trans. Circuits Syst. II, 40 (1993), 163–173. https://doi.org/10.1109/82.222815
28. C. Szegedy, S. Ioffe, V. Vanhoucke, A. A. Alemi, Inception-v4, inception-resnet and the impact of residual connections on learning, in Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, (2017), 4278–4284. Available from: https://dl.acm.org/doi/10.5555/3298023.3298188.
29. B. Liu, J. Xu, X. Lan, R. Xu, J. Zhou, X. Wang, et al., iDNA-Prot|dis: identifying DNA-binding proteins by incorporating amino acid distance-pairs and reduced alphabet profile into the general pseudo amino acid composition, PLoS One, (2014), e106691. https://doi.org/10.1371/journal.pone.0106691
30. Y. Wang, Y. Ding, F. Guo, L. Wei, J. Tang, Improved detection of DNA-binding proteins via
compression technology on PSSM information, PLoS One, (2017), e0185587.
https://doi.org/10.1371/journal.pone.0185587

31. R. Caruana, A. Niculescu-Mizil, An empirical comparison of supervised learning algorithms, in
Proceedings of the 23rd International Conference on Machine Learning, (2006), 161–168.
https://doi.org/10.1145/1143844.1143865
32. K. Weiss, T. M. Khoshgoftaar, D. Wang, A survey of transfer learning, J. Big Data, 3 (2016), 9.
https://doi.org/10.1186/s40537-016-0043-6
33. S.J. Pan, Q. Yang, A survey on transfer learning, IEEE Trans. Knowl. Data Eng., 22 (2010), 1345–
1359. https://doi.org/10.1109/TKDE.2009.191
34. M. Oquab, L. Bottou, I. Laptev, J. Sivic, Learning and transferring mid-level image representations using convolutional neural networks, in 2014 IEEE Conference on Computer Vision and Pattern Recognition, (2014), 1717–1724. https://doi.org/10.1109/CVPR.2014.222
35. W. Dai, Q. Yang, G. Xue, Y. Yu, Boosting for transfer learning, Machine Learning, in Proceedings
of the 24th International Conference on Machine Learning, (2007), 193–200.
https://doi.org/10.1145/1273496.1273521
36. S. R. Bowman, L. Vilnis, O. Vinyals, A. M. Dai, R. Jozefowicz, S. Bengio, Generating sentences
from a continuous space, in Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning, (2016), 10–21. https://doi.org/10.18653/v1/K16-1002
37. E. Tzeng, J. Hoffman, N. Zhang, K. Saenko, T. Darrell, Deep domain confusion: Maximizing for
domain invariance, preprint, arXiv:1412.3474.
38. H. Yan, Y. Ding, P. Li, Q. Wang, Y. Xu, W. Zuo, Mind the class weight bias: weighted maximum mean discrepancy for unsupervised domain adaptation, in 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), (2017), 945–954. https://doi.org/10.1109/CVPR.2017.107
39. W. Qin, X. Cui, C. A. Yuan, X. Qin, L. Shang, Z. K. Huang, et al., Flower species recognition system combining object detection and attention mechanism, in International Conference on Intelligent Computing, Springer, 2019. https://doi.org/10.1007/978-3-030-26766-7_1
40. K. Cho, B. V. Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, et al., Learning
phrase representations using RNN encoder-decoder for statistical machine translation, in Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), (2014), 1724–1734. https://doi.org/10.3115/v1/D14-1179
41. T. Mikolov, S. Kombrink, L. Burget, J. Černocký, S. Khudanpur, Extensions of recurrent neural
network language model, in 2011 IEEE International Conference on Acoustics, Speech and Signal
Processing (ICASSP), (2011), 5528–5531. https://doi.org/10.1109/ICASSP.2011.5947611
42. L. Wei, C. Zhou, H. Chen, J. Song, R. Su, ACPred-FL: a sequence-based predictor using effective
feature representation to improve the prediction of anti-cancer peptides, Bioinformatics, 34 (2018),
4007–4016. https://doi.org/10.1093/bioinformatics/bty451
43. Y. Ding, J. Tang, F. Guo, Protein crystallization identification via fuzzy model on linear neighborhood representation, IEEE/ACM Trans. Comput. Biol. Bioinf., 18 (2021), 1986–1995. https://doi.org/10.1109/TCBB.2019.2954826

44. Y. Ding, J. Tang, F. Guo, Human protein subcellular localization identification via fuzzy model
on kernelized neighborhood representation, Appl. Soft Comput., 96 (2020), 106596.
https://doi.org/10.1016/j.asoc.2020.106596
45. S. K. Knapp, Accelerate FPGA macros with one-hot approach, Electron. Des., 1990.
46. J. Soding, Protein homology detection by HMM-HMM comparison, Bioinformatics, 21 (2005),
951–960. https://doi.org/10.1093/bioinformatics/bti125
47. K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), (2016), 770–778.
https://doi.org/10.1109/CVPR.2016.90
48. V. Nair, G. E. Hinton, Rectified linear units improve restricted boltzmann machines, in Proceedings of the 27th International Conference on International Conference on Machine Learning, (2010), 807–814. Available from: https://dl.acm.org/doi/10.5555/3104322.3104425.
49. A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, et al., Automatic differentiation
in pytorch, 2017. Available from: https://paperswithcode.com/paper/automatic-differentiation-inpytorch.
50. D. P. Kingma, J. Ba, Adam: a method for stochastic optimization, CoRR, 2015. Available from:
https://www.semanticscholar.org/paper/Adam%3A-A-Method-for-Stochastic-OptimizationK-ingma-Ba/a6cb366736791bcccc5c8639de5a8f9636bf87e8.
51. W. Lou, X. Wang, F. Chen, Y. Chen, B. Jiang, H. Zhang, Sequence based prediction of DNAbinding proteins based on hybrid feature selection using random forest and Gaussian naive Bayes,
PLoS One, (2014), e86703. https://doi.org/10.1371/journal.pone.0086703
52. P. W. Rose, A. Prlic, C. Bi, W. F. Bluhm, C. H. Christie, S. Dutta, et al., The RCSB Protein Data
Bank: views of structural biology for basic and applied research and education, Nucleic Acids Res.,
43 (2015), D345–D356. https://doi.org/10.1093/nar/gku1214
53. X. Du, Y. Diao, H. Liu, S. Li, MsDBP: Exploring DNA-binding proteins by integrating multiscale
sequence information via Chou's five-step rule, J. Proteome Res., 18 (2019), 3119–3132.
https://doi.org/10.1021/acs.jproteome.9b00226