**Supplementary Information**

**ESM-NBR: fast and accurate nucleic acid-binding residue prediction via protein language model feature representation and multi-task learning**

Wenwu Zeng1, Dafeng Lv 1, and Shaoliang Peng1,\*

1 College of Computer Science and Electronic Engineering, Hunan University, Changsha 410082, China;

\* Address correspondence to S.L. Peng at slpeng@hnu.edu.cn

**Supporting Texts**

**Text S1. Benchmark Datasets**

Two pairs of widely used mixed datasets of DBPs and RBPs are employed to evaluate the proposed methods fairly and comprehensively. The first one named YK17 is collected by Yan *et al*. [1] from the Protein Data Bank (PDB) database [2], which contains a training subset (denoted as YK17-Tr) and an independent test subset (denoted as YK17-Tst). The second one (denoted as DRNA-1314) is constructed by Xia *et al*. [3] from the BioLip database [4] which composed of four subsets, i.e., DNA-573\_Train, RNA-495\_Train, DNA-129\_Test, and RNA-117\_Test. In this study, DNA-573\_Train and RNA-495\_Train (or DNA-129\_Test and RNA-117\_Test) are combined as training data set abbreviated as DRNATr-1068 (or independent test set abbreviated as DRNATst-246) of multi-task model. In these two datasets, the sequence identities between the protein chains in the test set and the protein chains in the training set are less than 30% [1, 5]. The detailed compositions are listed in Table S1.

**Table S1.** Composition of the training and testing data sets

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Data set | Subset | No. protein | No. DBR *a* | No. RBR *a* | No. non-NBR *a* |
| YK17 | YK17-Tr | 488 | 7,764 | 4,684 | 90,594 |
| YK17-Tst | 82 | 955 | 807 | 17,119 |
| DRNA-1314 | DNA-573\_Train | 573 | 14,479 | 0 | 145,404 |
| RNA-495\_Train | 495 | 0 | 14,609 | 122,290 |
| DNA-129\_Test | 129 | 2,240 | 0 | 35,275 |
| RNA-117\_Test | 117 | 0 | 2,031 | 35,314 |

1. “DBR”, “RBR”, and “non-NBR” mean the residues binding to DNA, residues binding to RNA, and residues neither bind to nucleic acid residue nor are disordered residue, respectively.

**Text S2. Architecture of multi-task neural network**

Various neural networks like BiGRU, CNN, and graph transformer [6] have been utilized by researchers to identify nucleic acid-binding residue [7] [8] [9]. These networks are usually chosen to better match the input features. For example, in GraphSite [9], to capture spatial information, Yuan *et al.* employed a graph transformer model with predicted protein 3D structural from AlphaFold2 as feature, which takes the protein structural information into account. Here, considering the small training set and the long ESM2 features, we used the relatively lightweight BiLSTM and MLP as prediction model, which is suitable for capturing the long-term dependence of sequences and is not prone to overfitting. Concretely, for each input matrix of size  where and mean sequence length and feature dimension, respectively, we first feed it into two stacked BiLSTM layers whose size of hidden layer is set to 100; then, the output of each time step of second BiLSTM of length 200 is fed into the later three MLP layers shared by DNA and RNA; finally, the output of the shared layer is fed into two separate MLP blocks contained three linear layers, which are specifically designed to predict DNA- and RNA-binding residues, respectively. The final prediction results of DNA- and RNA-binding residues can be determined according to the outputs of two separate MLP blocks. Multi-task model is more space- and time-efficient than single-task model and is relatively less prone to overfitting due to the need to fit multiple labels simultaneously. The number of parameters need to be learned of the whole model are 872,404, 1,000,404, 1,128,404, 1,640,404, and 2,664,404 when lengths of input features are 320, 480, 640, 1280, and 2,560, respectively.

During the training phase of the model, the cross-entropy loss is used to calculated by follow functions:

where means the total number of residues in the training data set; and are the true DNA and RNA labels of the -th residue, respectively; and are the predicted probabilities that the target residue is predicted to be a positive sample; is the L2 regularization factor and is set to 0.0001; represents the total number of model parameters. Since the unknown residues in the disordered region do not participate in the loss calculation, their labels are set to -1. The Adam algorithm [10] is used to optimize the loss function with a learning rate of 0.0001. To prevent overfitting and enhance fitting ability, the dropout algorithm [11] with a dropout rate of 0.5 and RLUE function are applied to all BiLSTM and MLP layers, respectively. All the training process was done on a Tesla V100 with 16G of memory.

**Text S3. Evaluation Indexes**

MCC is a threshold-related index for the comprehensive assessment of unbalanced datasets; AUC is threshold-independent and indicates overall prediction performance of negative and positive samples. In contrast to AUC, AP mainly focuses on prediction performance of positive samples; The AURC applied on cross-prediction problem is used to indicate the proportion of native RNA-binding residue that are mistakenly predicted to be DNA-binding residue or native DNA-binding residue that are mistakenly predicted to be RNA-binding residue. Out of four indexes, only the smaller the AURC the better the model prediction performance. Besides the above four indexes, the Pearson correlation coefficient (PCC) and p-value in Student’s t-test are employed to indicate the linear correlation degree and the difference between ESM-NBR and other methods.

**Text S4. Performance comparison of models for multi-task and single-task**

There are shared and particular properties of DNA- and RNA-binding residues which should be able to be learnt simultaneously by the neural network. In this section, to investigate the effectiveness of multi-task network for nucleic acid-binding residues prediction by learning common and private knowledge, the prediction of multi-task and single-task models are performed on DRNATst-246 and YK17-Tst, respectively. Detail prediction results are shown in Table S2. By looking at the data on YK17-Tst, we can know that performance of multi-task model is outperforms that of single-task model both in terms of DNA- and RNA-binding residue prediction. Take DNA-binding residue for example, the MCC, AUC, and AP values of multi-task model are 0.391, 0.881, and 0.350, which are 7.41, 1.26, and 0.28 percent higher than those of single-task model respectively. In contrast, the predictive performance of the single-task model is better than that of the multi-task model across the board on the DRNATst-246. For example, the MCC, AUC, and AP values of single-task model on DNA-binding residue of DRNATst-246 are 0.474, 0.923, and 0.526, which are 4.17, 3.24, and 9.12 percent higher than those of multi-task model respectively. Completely opposite prediction performances on the two datasets caught our attention. By investigating the two training sets, i.e., DRNATr-1068 and YK17-Tr, we find that there are 12 proteins in YK17-Tr bind both DNA and RNA, containing 336 DNA-binding residues and 187 RNA-binding residues, whereas proteins in DRNATr-1068 bind only one of DNA and RNA. That is say, there more complementary information about DNA-binding and RNA-binding patterns in YK17-Tr, which is more suitable for multi-task model to learn. The *p*-value and PCC indexes also show also demonstrate the difference in prediction results between the two models, suggesting that both two tasks have learnt useful knowledge from each other to aid their own predictions compared to the single-task model. In addition, it is clear that the multi-task model is superior to the single-task model in both space and time, which means it can be more easily generalized to massive protein sequences, thus advancing the process of nucleic acid-protein interaction research.

**Table S2.** Performance comparisons of multi-task and single task models on DRNATst-246 and YK17-Tst over independent validation

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Dataset | Model |  | DNA | | | | |  | RNA | | | | |
|  | MCC | AUC | AP | *p*-value *b* | PCC *c* |  | MCC | AUC | AP | *p*-value | PCC |
| DRNATst-246 *a* | single-task |  | **0.474** | **0.923** | **0.526** | 1.36e-89 | 7.65e-01 |  | **0.284** | **0.824** | **0.247** *d* | 2.47e-92 | 6.31e-01 |
| multi-task |  | 0.455 | 0.894 | 0.482 | - | - |  | 0.219 | 0.800 | 0.186 | - | - |
| YK17-Tst | single-task |  | 0.364 | 0.871 | 0.349 | 3.62e-82 | 7.64e-01 |  | 0.235 | **0.793** | 0.232 | 7.63e-10 | 6.28e-01 |
| multi-task |  | **0.391** | **0.881** | **0.350** | - | - |  | **0.275** | 0.785 | **0.233** | - | - |

1. Since DRNATst-246 is divided into two subsets, i.e., DNA-129\_Test and RNA-117\_Test, and the single-task model can only target one task in either DNA or RNA, the experiments here are performed independently on the two subsets.
2. The *p*-values in Student’s *t*-test are calculated for the differences between single-task model and multi-task model using the probability that the target residue is predicted to be a positive sample.
3. The PCC are calculated for the linear correlation coefficient between single-task model and multi-task model using the probability that the target residue is predicted to be a positive sample.
4. Bolded font indicates the best result.

**Text S5. Case Studies**

To visualize the advantages of the ESM-NBR, one native DNA-binding protein chain (PDB ID: 4OWX1) and one native RNA-binding protein chain (PDB ID: 6FQ3\_A) are employed from YK17-Tst and RNA-117\_Test for case studies. Figure S1 demonstrates the prediction results of ESM-NBR and four control methods on these two cases. The native DNA/RNA-binding residues are highlighted using sphere. Note that the DNA binding with 4OWX1 is a single-stranded DNA (ssDNA). By looking at the Figure S1, it is intuitive that ESM-NBR has better prediction performance on both protein chains. In particularly, most native DNA-binding residues are correctly predicted by the ESM-NBR on 4OWX1. In contrast, iDRNA-ITF and DRNAPred only correctly predicted a small percentage of DNA-binding residues. Even less favorable are the predictions of NCBRPred and GraphBind, who failed to predict even a single DNA-binding residue correctly. On 6FQ3\_A, iDRNA-ITF and NCBRPred do not correctly predict any of the RNA-binding residues. GraphBind correctly predicted just a handful of RNA-binding residues. Although DRNAPred correctly identifies all positive samples, it also predicts a large number of negative samples as RNA-binding residues. On the other hand, the ESM-NBR shows a more accurate prediction and guarantees predictive performance for both positive and negative samples. The MCC of ESM-NBR on 6FQ3\_A is 0.384 which are significantly higher than that of iDRNA-ITF, DRNAPred, GraphBind, and NCBRPred. The experimental results show that ESM-NBR has unique advantages on these two proteins.

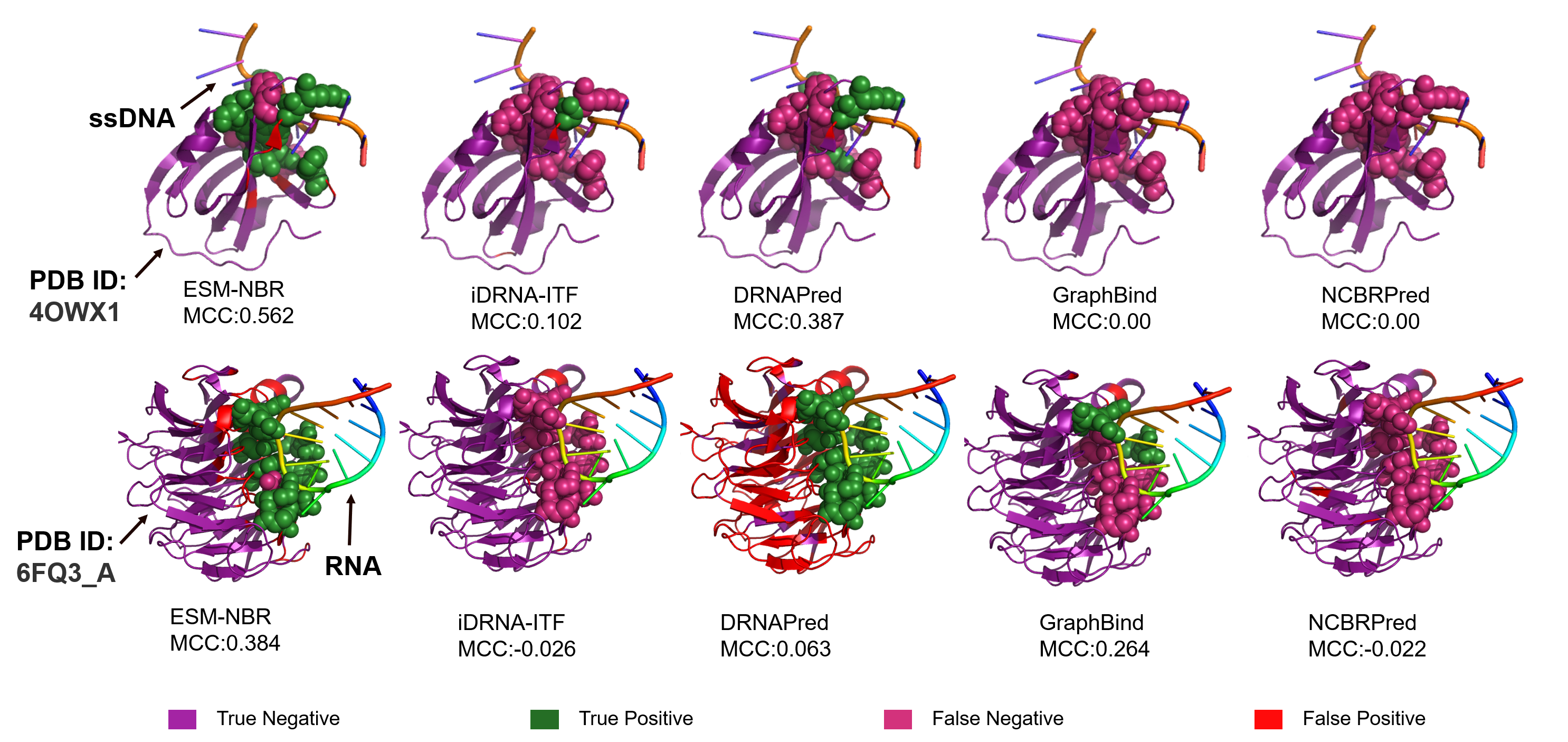


Fig. S1. Visualization of prediction results of ESM-NBR, iDRNA-ITF, DRNAPred, GraphBind, and NCBRPred on a DNA-binding protein chain (PDB ID: 4OWX1) and an RNA-binding protein chain (PDB ID: 6FQ3\_A). The figures are plotted using pymol [12]. The spheres mean native DNA/RNA-binding residues.

**REFERENCES**

[1] J. Yan, and L. Kurgan, “DRNApred, fast sequence-based method that accurately predicts and discriminates DNA- and RNA-binding residues,” *Nucleic Acids Research,* vol. 45, no. 10, pp. 16, Jun, 2017.

[2] J. L. Sussman, D. Lin, J. Jiang, N. O. Manning, J. Prilusky, O. Ritter, and E. E. J. A. C. S. D. B. C. Abola, “Protein Data Bank (PDB): database of three-dimensional structural information of biological macromolecules,” vol. 54, no. 6, pp. 1078-1084, 1998.

[3] Y. Xia, C. Q. Xia, X. Y. Pan, and H. B. Shen, “GraphBind: protein structural context embedded rules learned by hierarchical graph neural networks for recognizing nucleic-acid-binding residues,” *Nucleic Acids Research,* vol. 49, no. 9, pp. 17, May, 2021.

[4] J. Yang, A. Roy, and Y. Zhang, “BioLiP: a semi-manually curated database for biologically relevant ligand–protein interactions,” *Nucleic Acids Research,* vol. 41, no. D1, pp. D1096-D1103, 2012.

[5] N. Wang, K. Yan, J. Zhang, and B. Liu, “iDRNA-ITF: identifying DNA- and RNA-binding residues in proteins based on induction and transfer framework,” *Briefings in Bioinformatics,* vol. 23, no. 4, Jul, 2022.

[6] S. Yun, M. Jeong, R. Kim, J. Kang, and H. J. Kim, "Graph Transformer Networks," *Advances in Neural Information Processing Systems*, 2019.

[7] J. Hu, Y. S. Bai, L. L. Zheng, N. X. Jia, D. J. Yu, and G. J. 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, Nov-Dec, 2022.

[8] J. Zhang, Q. C. Chen, and B. Liu, “NCBRPred: predicting nucleic acid binding residues in proteins based on multilabel learning,” *Briefings in Bioinformatics,* vol. 22, no. 5, Sep, 2021.

[9] Q. Yuan, S. Chen, J. Rao, S. Zheng, H. Zhao, and Y. Yang, “AlphaFold2-aware protein–DNA binding site prediction using graph transformer,” *Briefings in Bioinformatics,* vol. 23, no. 2, 2022.

[10] D. P. Kingma, and J. Ba, “Adam: A method for stochastic optimization,” *arXiv preprint arXiv:1412.6980*, 2014.

[11] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, “Dropout: A Simple Way to Prevent Neural Networks from Overfitting,” *Journal of Machine Learning Research,* vol. 15, no. 1, pp. 1929-1958, 2014.

[12] S. G. Yuan, H. C. S. Chan, and Z. Q. Hu, “Using PyMOL as a platform for computational drug design,” *Wiley Interdisciplinary Reviews-Computational Molecular Science,* vol. 7, no. 2, Mar-Apr, 2017.