Briefings in Bioinformatics, 2024, **25(4)**, bbae285
https://doi.org/10.1093/bib/bbae285
Problem Solving Protocol

Improved prediction of DNA and RNA binding proteins with deep learning models

Siwen Wu and Jun-tao Guo*

Department of Bioinformatics and Genomics, University of North Carolina at Charlotte, Charlotte, NC 28223, United States
*Corresponding author. Department of Bioinformatics and Genomics, University of North Carolina at Charlotte, Charlotte, NC, 28223, United States.
Email: jguo4@charlotte.edu

Abstract

Nucleic acid-binding proteins (NABPs), including DNA-binding proteins (DBPs) and RNA-binding proteins (RBPs), play important roles in essential biological processes. To facilitate functional annotation and accurate prediction of different types of NABPs, many machine learning-based computational approaches have been developed. However, the datasets used for training and testing as well as the prediction scopes in these studies have limited their applications. In this paper, we developed new strategies to overcome these limitations by generating more accurate and robust datasets and developing deep learning-based methods including both hierarchical and multi-class approaches to predict the types of NABPs for any given protein. The deep learning models employ two layers of convolutional neural network and one layer of long short-term memory. Our approaches outperform existing DBP and RBP predictors with a balanced prediction between DBPs and RBPs, and are more practically useful in identifying novel NABPs. The multi-class approach greatly improves the prediction accuracy of DBPs and RBPs, especially for the DBPs with ~12% improvement. Moreover, we explored the prediction accuracy of single-stranded DNA binding proteins and their effect on the overall prediction accuracy of NABP predictions.

Keywords: RNA binding protein (RBP); single-stranded DNA binding protein (SSB); double-stranded DNA binding protein (DSB); convolutional neural network (CNN); long short-term memory (LSTM); multi-class model

Introduction

DNA-binding proteins (DBPs) and RNA-binding proteins (RBPs) are two different types of nucleic acid-binding proteins (NABPs), which play crucial roles in many biological processes, such as DNA replication, transcriptional regulation, alternative splicing and translation [1–5]. DBPs include double-stranded DNA binding proteins (DSBs) and single-stranded DNA binding proteins (SSBs). While DSBs are mainly involved in transcriptional regulation, DNA cleavage and chromosome packaging, SSBs participate in DNA recombination, replication and repair, and serve as key players in the maintenance of genomic stability [6–8]. Although experimental methods can be used to identify the functions of some proteins, it is time-consuming and expensive. In addition, there are a large number of uncharacterized proteins in the protein sequence database [9–12], making it impossible to characterize and annotate each of them by experimental methods. Computational methods, on the other hand, can complement the experimental approaches by efficiently predicting the functional categories of the unannotated proteins and help narrow the number down for experimental validations.

A number of computational methods have been developed so far to predict nucleic acid binding proteins from sequences, especially for DBP predictions [13–27]. Recently, advanced machine learning approaches, such as deep learning, become popular in bioinformatics research and have been applied to predict DBPs and RBPs [26–33]. However, almost all of these methods were trained to predict only either DBPs or RBPs and therefore may have limited their applications due to the similarities between DBPs and RBPs. For example, DNAbinder [18], DPP-PseAAC [23], PlDBPred [30], and DBPMod [31], using support vector machine (SVM), random forest (RF), adaptive boosting (ADB), and light gradient boosting (LGB) methods, were developed for predicting DBPs. The SVM-based RNAPred [34], convolutional neural network (CNN)-based DeepRBPPred [28], LGB-based RBPLight [32] and CNN-based RBProkCNN [33] are RBPs predictors. For the DBP predictors, DBPs were used as the positive datasets with non-DBPs (including both non-NABPs and RBPs) as the negative datasets for model training. Similar approach for dataset construction and model training was adopted for the RBP predictors. Therefore, DBP predictors may predict DBPs with relatively high accuracy, but tend to incorrectly predict many RBPs as DBPs, and similarly for the RBP predictors. Recently, Zhang *et al.* [29] developed DeepDRBP-2 L for prediction of both RBPs and DBPs using CNN and long short-term memory (LSTM). While the

Siwen Wu is currently a postdoc in the Department of Bioinformatics and Genomics at the University of North Carolina at Charlotte. She received her Ph.D. degree in Bioinformatics and Computational Biology in 2023 from UNC Charlotte.
Jun-tao Guo is a professor in the Department of Bioinformatics and Genomics at the University of North Carolina at Charlotte. His research interest is in structural bioinformatics with a special focus on protein-nucleic acid interactions.

Received: March 5, 2024. **Revised:** May 20, 2024. **Accepted:** May 31, 2024

prediction accuracy of DBPs by DeepDRBP-2 L is very good, the prediction accuracy of RBPs is relatively low. The annotation of the datasets based only on GO terms may play a role in the low prediction accuracy of RBPs [29]. For example, protein Q9VPT8 is annotated as both an RBP and a DBP, a dual function protein, in the Swiss-Prot database [35]. However, this protein was classified as an RBP in their dataset since the GO term for this protein only has GO:0003723 (RNA binding), not GO:0003677 (DNA binding). As demonstrated previously by Zaitzeff *et al*. [36] accurate datasets are essential for developing better prediction models. Adding to the complexity, there are two types of DBPs, SSBs that bind single-stranded DNA and DSBs that bind double-stranded DNA while RBPs bind to diverse types of secondary and tertiary RNA structures besides single-stranded RNA [37–39].

We adopted three strategies to address these issues. Firstly, we generated new datasets with a more restricted keyword-based selection method for selecting the NABPs. Secondly, we developed a hierarchical approach using two layers of CNN and one layer of LSTM. For our hierarchical approach, the first step is the prediction of non-NABPs/NABPs followed by the prediction of DBPs/RBPs. Thirdly, even though each of the steps in the hierarchical approach can achieve >80% prediction accuracy, the actual prediction accuracy for DBPs and RBPs for any given protein is lower since it is contingent on the first step (non-NABPs/NABPs) performance. As such, we developed a multi-class deep learning model that predicts non-NABPs, DBPs, and RBPs simultaneously. Results show that our hierarchical approach outperforms the existing DBPs and RBPs prediction tools with balanced prediction accuracy between DBPs and RBPs. The multi-class approach can predict DBPs and RBPs more accurately when compared with the overall accuracy from the hierarchical approach for any given protein, especially for prediction of DBPs, which improved dramatically from 64.7% to 76.6%.

In this study, we also investigated the prediction accuracy of SSBs and their effect on the overall prediction performance of the DBPs and RBPs. To our knowledge, this is the first time that SSBs are explicitly investigated as part of DBPs and RBPs prediction, which can provide guidance in developing models for predicting novel SSBs.

Materials and methods

Datasets

We downloaded a total of 484,143 proteins with GO term annotations in Swiss-Prot from the UniProt database [35] and removed the redundancy of the proteins using a sequence identity cutoff of 0.4 with CD-HIT v4.8.1 [40], which resulted in a non-redundant (NR) dataset of 65,076 proteins. Similar to previous studies, we selected proteins with length between 40 and 1500 amino acids, which represent 96.5% (62,797) of the initial NR dataset (Fig. 1A). The types of NABPs were then defined as shown in Table 1. More specifically, if the protein file contains all the four keywords 'DNA', 'binding', 'single', and 'strand', it is annotated as an SSB; if the protein file contains all the four keywords 'DNA', 'binding', 'double', and 'strand', or any other four types of descriptions related to transcription factors as shown in Table 1, it is considered as a DSB; if the protein has keyword 'RNA-binding' or 'RNA binding', it is defined as an RBP; and finally if it does not contain any of the keywords 'DNA', 'RNA', 'nucleic acid', and 'nucleotide' in its keywords description line, does not contain keyword 'binding' and GO terms 'GO0003676', 'GO0003677', 'GO0003723', 'GO0003697', and 'GO0003690' in all its descriptions, it is considered as a non-NABP. Combing GO terms with keywords for non-NABP dataset generation is to maximize the removal of potential NABPs from the non-NABP dataset. The above selection process resulted in 561 SSBs, 4520 DSBs, 5836 RBPs, and 12,899 non-NABPs (Fig. 1B).

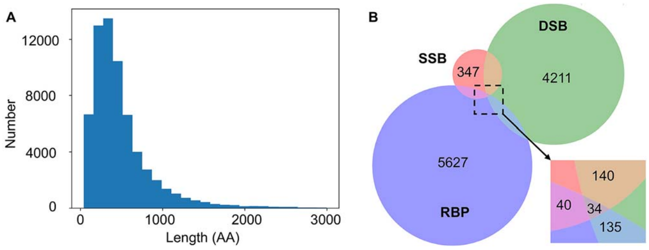

Figure 1. Datasets used in this study. A. The length distribution of the non-redundant dataset. B. Venn diagram showing the numbers of different types of NABPs, including RBPs, DSBs, SSBs and proteins with capabilities of binding more than one types of nucleic acids (zoomed-in square diagram).

Table 1. Keywords used for generating each dataset

Types	Keywords
SSB	'DNA' + 'binding' + 'single' + 'strand'
DSB	'DNA' + 'binding' + 'double' + 'strand' 'DNA-binding transcription activator activity' 'DNA-binding transcription factor activity' 'DNA-binding transcription repressor activity' 'sequence-specific DNA binding'
RBP	'RNA-binding' 'RNA binding'
non-NABP	does not contain: 'DNA', 'RNA', 'nucleic acid', 'nucleotide' in keywords and 'binding' in all description and GO:0003676, GO:0003677, GO:0003723, GO:0003697, GO:0003690 in all description

It is well known that some NABPs can bind different types of nucleic acids. For example, some proteins can bind both DNA and RNA [29]. In our dataset, there are 135 proteins that can bind to both RNA and dsDNA, 40 proteins that are annotated as RBPs and SSBs, 140 proteins that are capable of binding dsDNA and ssDNA, and 34 proteins can bind all three types of nucleic acids (Fig. 1B). The NABPs that bind only one type of nucleic acids consist of 347 SSBs, 4211 DSBs, and 5627 RBPs (Fig. 1B). In this study, we only used NABPs with distinct binding type annotations, SSBs, DSBs, RBPs, and non-NABPs, for model training and testing.

Position-specific scoring matrix calculation

The position-specific scoring matrix (PSSM) of each protein was used to train, validate, and test our machine learning models. To calculate the PSSM for each protein, psi-blast (2.11.0+) [41] was used with an e-value cutoff of 0.001 and three iterations against the uniref90 datasets [42]. For each protein, the initial PSSM, a $L*20$ matrix (L: length of the protein), was first transformed as a $20*L$ matrix. Since the fully connected layer of CNN model requires the inputs to be of the same length, zero was added after the original PSSM to make all the proteins having the same length ($20*1500$).

CNN model building

For the deep learning model, a four-layer architecture, which consists of one CNN layer, one dropout layer, one max-pooling layer, and a second dropout layer, was repeated once (Fig. 2A).

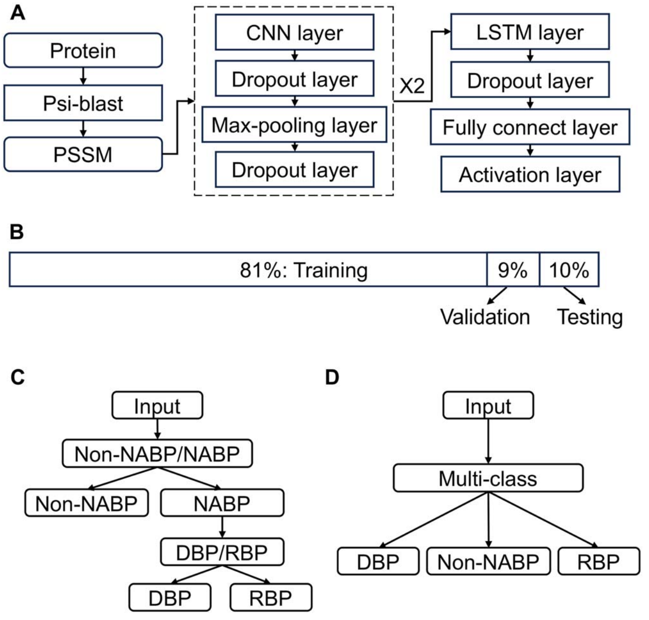

Figure 2. Deep learning models. A. Flowchart of the CNN model. B. Schematic representation of the10-fold cross validation of the training, validation and testing datasets. C. Flowchart of the hierarchical approach. D. Flowchart of the multi-class approach.

The process is completed with four more layers, consisting of an LSTM layer, one dropout layer, one fully connected layer and an activation layer (Fig. 2A). The CNN model was trained using the training and validation datasets with a 10-fold cross validation strategy for 100 epochs and the model with the highest validation accuracy was saved and tested with the testing datasets (Fig. 2B). Since the datasets were randomly selected, the CNN model was run for 500 rounds with each random selection to calculate the mean and standard deviation (SD) of the prediction accuracy. In each approach, we tried different set of parameters for training, and the parameters with the best performance were selected for the final models.

Hierarchical prediction approach

We first developed a hierarchical approach for predicting non-NABPs, DBPs, and RBPs using the CNN model (Fig. 2C). The first step is the prediction between non-NABPs and NABPs. In this step, all the SSBs, DSBs, and RBPs in our datasets were combined as the NABPs dataset (positive dataset), and 79% of non-NABPs were randomly selected as the negative dataset to match the number of NABPs positive dataset. The CNN models were trained with the following parameters: number of filters = 128, kernel size = 37, pooling size = 4, dropout rate = 0.2, LSTM output size = 50, batch size = 128.

The second step of the hierarchical approach is the prediction between DBPs and RBPs. In this step, all the SSBs and DSBs in the datasets were pooled together as the DBP dataset (positive dataset), and 81% of the RBPs were randomly selected as the RBP dataset (negative dataset) in order to have the same size of the DBP dataset. The parameters used for this step are: number of filters = 64, kernel size = 29, pooling size = 6, dropout rate = 0.2, LSTM output size = 80, batch size = 16.

Multi-class approach

The multi-class approach predicts non-NABPs, DBPs, and RBPs simultaneously using the same CNN model architecture (Fig. 2D). To make the dataset sizes comparable, all the SSBs and DSBs were pooled as the DBP dataset, 81% of the RBPs were randomly selected as the RBP dataset, and 35% of the non-NABPs were randomly selected as the non-NABP dataset for each of the 500 rounds. The parameters for the multi-class approach are as follows: number of filters = 64, kernel size = 37, pooling size = 4, dropout rate = 0.2, LSTM output size = 60, batch size = 32.

Prediction without SSBs

To investigate the effect of SSBs on the prediction accuracy for both the hierarchical and multi-class approaches, the SSBs were removed from the DBP dataset and the number of RBPs and non-NABPs were adjusted accordingly to make sure that each dataset has the same number of proteins. The same CNN model architecture and parameters for the hierarchical and multi-class approaches were applied, respectively.

Evaluation metrics

Four metrics were used to evaluate the performance of each predictor: Accuracy (*ACC*), Recall (*REC*), Precision (*PRE*), and F1-Score (*F1*) (Equations 1–4). ACC calculates the number of correctly predicted cases over the total cases. *REC* represents number of correctly predicted positive samples over the total of positive samples. *PRE* is the ratio of the truly predicted positive samples over the total positive predictions. *F1* is a measure of recall and precision, which reflects the overall performance.

$$ACC = \frac{TP + TN}{TP + FN + TN + FP} \tag{1}$$

$$REC = \frac{TP}{TP + FN} \tag{2}$$

$$PRE = \frac{TP}{TP + FP} \tag{3}$$

$$F1 = \frac{2 * REC * PRE}{REC + PRE} \tag{4}$$

where TP, TN, FP, and FN represent true positive, true negative, false positive and false negative, respectively.

Results

Prediction of non-NABPs, DBPs, and RBPs with the hierarchical approach

In both non-NABP/NABP and DBP/RBP prediction of the hierarchical approach, the mean validation accuracy and the mean testing accuracy are comparable with similar standard deviations, suggesting that there is no overfitting of the models (Fig. 3 and see online supplementary material for a colour version of Table S1). For non-NABP/NABP prediction, the overall accuracy in validation is 82.3% while in testing is 81.2% (Fig. 3A). In the second step for DBP/RBP prediction, the overall validation and testing accuracy are 82.8% and 81.4%, respectively (Fig. 3B). When comparing the performance of the positive and negative datasets separately, we found that the testing accuracy for NABP (83.4%) is higher than that for non-NABP (79.1%) in the first step (Fig. 3A). In the second step, the testing accuracy for RBP (83.1%) is better than that for DBP (79.7%) (Fig. 3B).

Prediction accuracy of SSBs and DSBs in the DBP dataset

Due to the structural differences between ssDNA and dsDNA, the embedding signal of their binding proteins may be different. To investigate if there are any differences in performance between

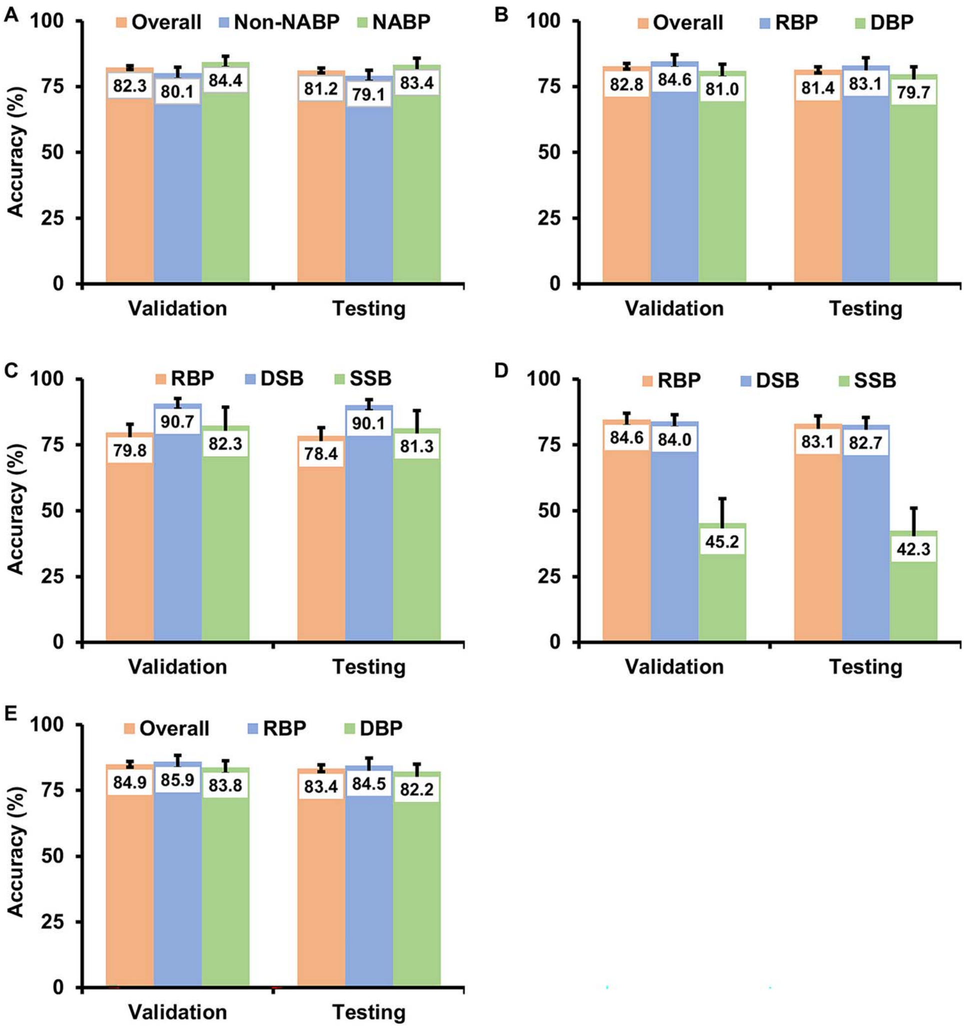

Figure 3. Accuracy for the validation and testing datasets of the hierarchical approach. A. Overall prediction accuracy and accuracy of non-NABPs and NABPs from the non-NABP/NABP prediction model. B. Overall prediction accuracy and accuracy of RBPs and DBPs from the DBP/RBP prediction model. C. Prediction accuracy of RBPs, DSBs and SSBs from the non-NABP/NABP prediction model. D. Prediction accuracy of RBPs, DSBs and SSBs from the DBP/RBP prediction model. E. Overall prediction accuracy and accuracy of RBPs and DBPs from the DBP/RBP prediction model without SSBs included.

SSBs and DSBs, we compared the prediction accuracy for SSBs and DSBs of DBPs separately. Similar validation and testing accuracy of each dataset, RBP, DSB, and SSB, were found in both non-NABP/NABP and DBP/RBP prediction steps, indicating no bias or overfitting towards any of the datasets from the prediction models (Figs. 3C and D, see online supplementary material for a colour version of Table S2). In the non-NABP/NABP prediction step, the highest testing accuracy was achieved for DSBs (90.1%), while the testing accuracy for SSBs and RBPs are 81.3% and 78.4%, respectively (Fig. 3C, see online supplementary material for a colour version of Table S2). However, in the DBP/RBP prediction step, testing accuracy of SSBs (42.3%) is much lower than those of the DSBs (82.7%) and RBPs (83.1%) (Fig. 3D, see online supplementary material for a colour version of Table S2). In other words, in discriminating non-NABPs and NABPs, SSBs achieved similar prediction accuracy to those of RBPs and DSBs. The high prediction accuracy for SSBs in the first step is not surprising since the sequence features from SSBs and non-NABPs are probably very different. The low prediction accuracy for SSBs in the second step could be a combined result of the small SSB dataset when compared to the DSB and RBP datasets and the difficulty of differentiating SSBs from RBPs since they both have single stranded components.

To investigate the effect of SSBs on the overall prediction accuracy, the SSBs were removed from the DBP dataset for the DBP/RBP prediction step. The overall testing accuracy increased from 81.4% to 83.4% with a larger improvement for the DBPs, from 79.7% to 82.2% while the prediction accuracy for RBPs also improves from 83.1% to 84.5% (Fig. 3E, see online supplementary material for a colour version of Table S1).

Comparison with existing predictors

We compared our DBP/RBP prediction model with several published DBP, RBP, or DBP/RBP predictors as reported in previous studies [18, 23, 28–30, 32, 34]. Since in these DBP or RBP predictions, the negative datasets used are either non-DBP (for DBP predictors) or non-RBP (for RBP predictors), which include both non-NABPs and either RBPs (for DBP predictors) or DBPs (for RBP predictors), for the purpose of fair comparisons, we constructed the negative datasets with 455 proteins by combining non-NABPs with either DBPs (for RBP predictors) or RBPs (for RBP predictors) with a 1:1 ratio. The datasets used for each predictor are summarized in, see online supplementary material for a colour version of, Tables S3–S5.

Table 2. Comparison of our DBP/RBP predictor with the existing predictors (bold numbers represent the highest values in each comparison metric)

Method	PredictionTypes	Methods used	REC	PRE	F1	Neg. Set ACC	Total ACC
DNAbinder	DBP	SVM	0.72	0.50	0.59	0.27	0.49
DPP-PseAAC	DBP	RF	0.47	0.52	0.49	0.57	0.52
PlDBPred	DBP	ADB	0.65	0.75	0.70	0.78	0.71
RNAPred	RBP	SVM	0.82	0.57	0.67	0.38	0.60
DeepRBPPred (balance)	RBP	CNN	**0.88**	0.55	0.68	0.28	0.58
DeepRBPPred (unbalance)	RBP	CNN	0.76	0.55	0.64	0.38	0.57
RBPLight	RBP	LGB	0.72	0.69	0.70	0.67	0.70
DeepDRBP-2 L	DBP/RBP	CNN + LSTM	**0.88**	0.76	**0.81**	0.72	0.80
Hierarchical DBP/RBP	DBP/RBP	CNN + LSTM	0.81	**0.81**	**0.81**	**0.82**	**0.81**

As shown in Table 2, our predictor has the highest overall accuracy of 81%, 1% more than the second highest program DeepDRBP-2 L (80%) while the prediction accuracy for the other predictor's ranges from 49% to 71%. We would like to note that the prediction accuracy reported here for other predictors are generally higher than the published accuracy of the predictors, probably because our datasets are less ambiguous in terms of functional annotations [18, 23, 28, 29, 34]. Not surprisingly, the prediction accuracies for PlDBPred [30] and RBPLight [32] reported here are lower than the published accuracy since both programs were developed and trained for predicting plant DBPs and RBPs.

Not only does our predictor have the highest overall accuracy, but our model also demonstrated a nice balance between recall (0.81) and precision (0.81) with a negative set accuracy of 0.82. While most of the DBP and RBP predictors have relatively good recall (0.65–0.88) except for DPP-PseAAC (0.47), they typically have lower precision values and lower accuracy from the negative testing datasets, suggesting that the prediction is biased toward the positive dataset (Table 2). For example, DeepRBPPred-balanced has the highest recall (0.88) but with a low precision of 0.55, indicating a high number of false positive predictions. PlDBPred, RBPLight, and DeepDRBP-2 L are programs with relatively good precision (0.75, 0.69, and 0.76, respectively) and accuracy of negative testing datasets (0.78, 0.67, and 0.72, respectively) (Table 2).

Prediction of non-NABPs, DBPs, and RBPs using a multi-class approach

In the hierarchical approach, while the overall prediction accuracy is high from both the non-NABP/NABP step and the DBP/RBP step, the actual prediction accuracy for any given protein to be a DBP or RBP is the combined result of the two steps, ~81.2%∗81.4% = 66.1% (Fig. 3A and B, see online supplementary material for a colour version of Table S1). Here we develop a new multi-class approach for predicting the non-NABPs, DBPs and RBPs simultaneously. The results show an overall testing accuracy of 72.7%, and 71.2% for non-NABPs, 70.3% for RBPs, and 76.6% for DBPs (Fig. 4A, see online supplementary material for a colour version of Table S6). These data revealed that, while the prediction accuracy of non-NABP in the multi-class approach (71.2%) is lower than that in the hierarchical approach (79.1%), the prediction accuracy of DBP improved dramatically from 64.7% (81.2%∗79.7%) to 76.6% and the prediction accuracy of RBP improved from 67.5% (81.2%∗83.1%) to 70.3%. For the purpose of identifying novel DBPs and RBPs, better prediction accuracy for DBPs and RBPs is more important than that for non-NABPs.

Similar to the hierarchical approach, we found that the testing prediction accuracy for SSB (39.4%) is much lower than DSB (79.6%) (Fig. 4B, see online supplementary material for a colour version of Table S2). After removing SSBs from the DBP dataset in the multi-class approach, the overall testing accuracy improved from 72.7% to 73.9% (Fig. 4C, see online supplementary material

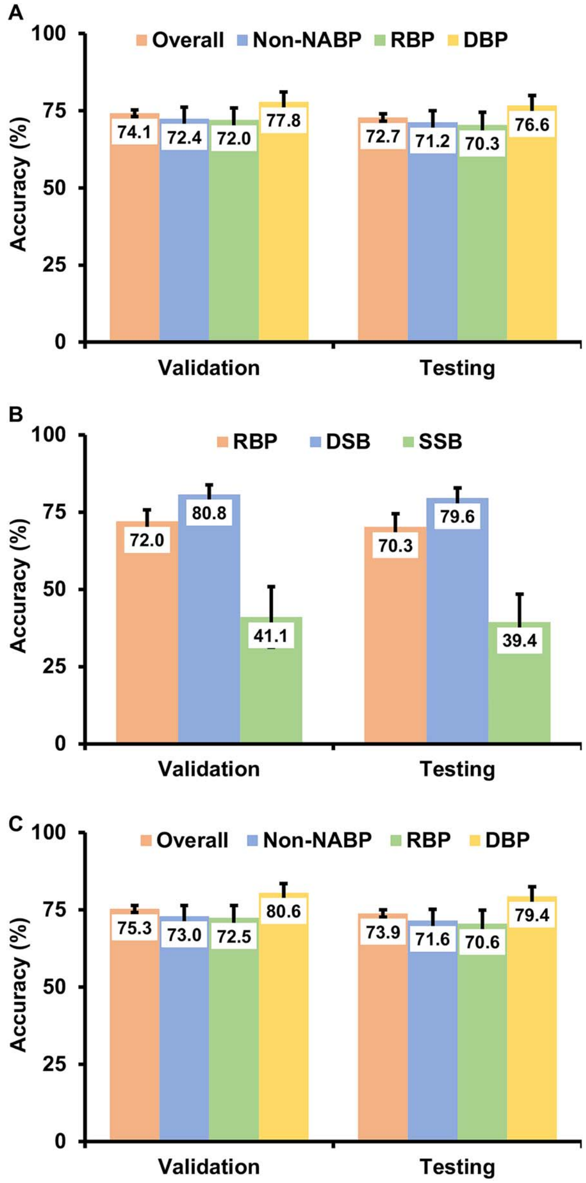

Figure 4. Prediction accuracy of the multi-class approach. A. Prediction accuracy of the overall, non-NABPs, RBPs and DBPs. B. Accuracy of RBPs, DSBs and SSBs. C. Accuracy of the overall, non-NABPs, RBPs and DBPs without SSBs included.

for a colour version of Table S6). While there is a minimal increase of prediction accuracy for RBPs, 70.3% to 70.6%, without SSBs in the DBP dataset, the accuracy for DBP increased about 3% (76.6% to 79.4%) (Fig. 4C, see online supplementary material for a colour version of Table S6).

Discussion

NABPs, including DBPs and RBPs, play essential roles in biological processes. It is well known that proteins with unknown functions, termed 'the dark matter of the sequence universe', represent a significant fraction of known sequences (20%–50%) [9]. It is practically impossible to experimentally characterize whether each of them is a DBP or RBP [9–12]. Therefore, it is of particular importance to develop efficient computational methods for predicting new DNA or RNA binding proteins. Not only can it expand the landscape of nucleic acid binding proteins, it can also narrow down the cases to a reasonable number for follow-up experimental validations. There are a number of existing computational methods for predicting DBPs and RBPs. However, the predictors generally have limited their applications in RBP or DBP annotations for any given protein because of the setups of the training and testing sets. In addition, the datasets used by these methods are not robust enough [43].

In this paper, we compiled new non-redundant datasets for non-NABPs, DBPs, RBPs, DSBs, and SSBs, with more accurate annotations. With the advancement of deep learning models, we developed a hierarchical and a multi-class prediction model, based on CNN and LSTM, for more accurate DBP and RBP predictions. We demonstrated that our DBP/RBP predictor outperforms other DBP and/or RBP prediction tools with a great balance between DBPs and RBPs prediction accuracy [18, 23, 28, 29, 34]. More importantly, our multi-class predictor shows a much better accuracy for predicting DBPs or RBPs for any given protein when compared with the hierarchical approach. In addition, for the first time (to our knowledge), we explicitly included annotated SSBs as part of the DBP dataset and investigated the prediction accuracy of SSBs and the effect of SSBs on the overall prediction accuracy of DBPs and RBPs. The approach is practically useful since the models developed in this study can be used to do predictions for any given protein such as non-NABP, DBP or RBP. While we employed the up-to-date Swiss-Prot database with all the reviewed entries and the data is big enough for applying deep learning methods, more data and novel features will certainly be beneficial for better learning and better prediction in the future.

To investigate if there are any patterns from the wrongly predicted proteins, we developed a strategy to compile the correct predictions and incorrect predictions. Since we carried out 500-rounds of tests due to random dataset selection and randomly splitting the datasets for training and testing, the proteins that are correctly predicted at least 75% of the times are grouped as the correctly predicted set while the wrongly predicted set consists of the ones having percentage falls under 25%. The numbers of the correctly and wrongly predicted proteins from each model are shown in, see online supplementary material for a colour version of, Table S7. We found that the wrongly predicted proteins tend to be longer than the correctly predicted proteins with statistically significant differences (see online supplementary material for a colour version of Fig. S1). We then investigated the GO function enrichment of the wrongly predicted proteins (see online supplementary material for a colour version of Fig. S2) [44]. The results provide a couple of clues with respect to why some proteins are not predicted correctly. For the incorrectly predicted non-NABPs, one enriched function is heterocyclic compound binding and organic cyclic compound binding (GO:1901363) (see online supplementary material for a colour version of Fig. S2 A and E). The child terms of GO:1901363 include nucleobase binding, suggesting the dataset selection process can be more refined in the future. Another common feature is from the wrongly predicted RBPs that are enriched in functions involved in interacting with double-stranded RNA (GO:0003725) (see online supplementary material for a colour version of Fig. S2 D and G). Since double-stranded RNA has some similarity to double stranded DNA, they may have similar binding signals to DBPs and can be predicted incorrectly as DBPs.

In this study, we also demonstrated that SSBs have much lower prediction accuracy (~40%) than DSBs and RBPs in both the hierarchical and multi-class models. In other words, more than half of the SSBs are incorrectly predicted as RBPs. This has important implications in developing methods for SSB predictions of proteins with unknown functions. Several programs have been developed to classify DSBs from SSBs [43, 45–47]. While the performance is decent for classifying DSB and SSB in these studies, their usefulness is limited in applications for annotating SSBs for any given protein since these methods and tools take DBPs as input. However, as we demonstrated in this study from the DBP/RBP prediction step, when RBPs are involved in the dataset, the prediction accuracy for SSBs is much lower. Therefore, even though high classification accuracy for DBPs can be achieved, the overall prediction accuracy for SSB is low when RBPs are in the mix. The low SSB prediction accuracy could be combined results of two factors, a small SSB dataset and insufficient feature representations. New strategies need to be explored for more accurate prediction of SSBs for proteins with unknown functions.

Key Points

- Robust datasets including separate SSB annotations were generated for training and testing the NABP, DBP, and RBP predictors.
- Our hierarchical deep learning model outperforms existing DBP and RBP predictors with a balanced prediction accuracy between DBPs and RBPs.
- Our multi-class deep learning model shows dramatic improvement for DBP and RBP predictions, especially for predicting DBPs (~12% improvement).
- For the first time, we investigated the prediction accuracy of SSBs and their effect on the overall prediction accuracy of NABPs, RBPs, and DBPs.

Supplementary data

Supplementary data is available at *Briefings in Bioformatics* online.

Funding

This work was supported by the National Institutes of Health [R15GM132846 to J.G]; and by the National Science Foundation [DBI-2051491 to J.G].

Data availability

The datasets, source code and user notes are available at GitHub https://github.com/unccguolab/Prediction-of-DNA-and-RNA-binding-proteins.

References

1. Hudson WH, Ortlund EA. The structure, function and evolution of proteins that bind DNA and RNA. *Nat Rev Mol Cell Biol* 2014;**15**: 749–60. https://doi.org/10.1038/nrm3884
2. Luscombe NM, Austin SE, Berman HM. *et al.* An overview of the structures of protein-DNA complexes. *Genome Biol* 2000;**1**:REVIEWS001
3. Glisovic T, Bachorik JL, Yong J. *et al.* RNA-binding proteins and post-transcriptional gene regulation. *FEBS Lett* 2008;**582**: 1977–86. https://doi.org/10.1016/j.febslet.2008.03.004
4. Gerstberger S, Hafner M, Tuschl T. A census of human RNA-binding proteins. *Nat Rev Genet* 2014;**15**:829–45. https://doi.org/10.1038/nrg3813
5. Schleif R. DNA binding by proteins. *Science* 1988;**241**:1182–7. https://doi.org/10.1126/science.2842864
6. Corona RI, Guo JT. Statistical analysis of structural determinants for protein-DNA-binding specificity. *Proteins* 2016;**84**:1147–61. https://doi.org/10.1002/prot.25061
7. Lin M, Malik FK, Guo JT. A comparative study of protein-ssDNA interactions. *NAR Genom Bioinform* 2021;**3**:lqab006. https://doi.org/10.1093/nargab/lqab006
8. Guo JT, Malik F. Single-stranded DNA binding proteins and their identification using machine learning-based approaches. *Biomolecules* 2022;**12**:1187. https://doi.org/10.3390/biom12091187.
9. Levitt M. Nature of the protein universe. *Proc Natl Acad Sci U S A* 2009;**106**:11079–84. https://doi.org/10.1073/pnas.0905029106
10. Galperin MY, Koonin EV. 'Conserved hypothetical' proteins: prioritization of targets for experimental study. *Nucleic Acids Res* 2004;**32**:5452–63. https://doi.org/10.1093/nar/gkh885
11. Shumilin IA, Cymborowski M, Chertihin O. *et al.* Identification of unknown protein function using metabolite cocktail screening. *Structure* 2012;**20**:1715–25. https://doi.org/10.1016/j.str.2012.07.016
12. Ellens KW, Christian N, Singh C. *et al.* Confronting the catalytic dark matter encoded by sequenced genomes. *Nucleic Acids Res* 2017;**45**:11495–514. https://doi.org/10.1093/nar/gkx937
13. Adilina S, Farid DM, Shatabda S. Effective DNA binding protein prediction by using key features via Chou's general PseAAC. *J Theor Biol* 2019;**460**:64–78. https://doi.org/10.1016/j.jtbi.2018.10.027
14. Ali F, Ahmed S, Swati ZNK. *et al.* DP-BINDER: machine learning model for prediction of DNA-binding proteins by fusing evolutionary and physicochemical information. *J Comput Aided Mol Des* 2019;**33**:645–58. https://doi.org/10.1007/s10822-019-00207-x
15. Chowdhury SY, Shatabda S, Dehzangi A. iDNAProt-ES: identification of DNA-binding proteins using evolutionary and structural features. *Sci Rep* 2017;**7**:14938. https://doi.org/10.1038/s41598-017-14945-1
16. Zaman R, Chowdhury SY, Rashid MA. *et al.* HMMBinder: DNA-binding protein prediction using HMM profile based features. *Biomed Res Int* 2017;**2017**:4590609
17. Du X, Diao Y, Liu H. *et al.* MsDBP: exploring DNA-binding proteins by integrating multiscale sequence information via Chou's five-step rule. *J Proteome Res* 2019;**18**:3119–32. https://doi.org/10.1021/acs.jproteome.9b00226
18. Kumar M, Gromiha MM, Raghava GP. Identification of DNA-binding proteins using support vector machines and evolutionary profiles. *BMC Bioinformatics* 2007;**8**:463. https://doi.org/10.1186/1471-2105-8-463
19. Xu R, Zhou J, Liu B. *et al.* enDNA-Prot: identification of DNA-binding proteins by applying ensemble learning. *Biomed Res Int* 2014;**2014**:294279
20. Lou W, Wang X, Chen F. *et al.* Sequence based prediction of DNA-binding proteins based on hybrid feature selection using random forest and Gaussian naïve Bayes. *PloS One* 2014;**9**:e86703. https://doi.org/10.1371/journal.pone.0086703.
21. Mishra A, Pokhrel P, Hoque MT. StackDPPred: a stacking based prediction of DNA-binding protein from sequence. *Bioinformatics* 2019;**35**:433–41. https://doi.org/10.1093/bioinformatics/bty653
22. Motion GB, Howden AJM, Huitema E. *et al.* DNA-binding protein prediction using plant specific support vector machines: validation and application of a new genome annotation tool. *Nucleic Acids Res* 2015;**43**:e158. https://doi.org/10.1093/nar/gkv805.
23. Rahman MS, Shatabda S, Saha S. *et al.* DPP-PseAAC: a DNA-binding protein prediction model using Chou's general PseAAC. *J Theor Biol* 2018;**452**:22–34. https://doi.org/10.1016/j.jtbi.2018.05.006
24. Wang J, Zheng H, Yang Y. *et al.* PredDBP-stack: prediction of DNA-binding proteins from HMM profiles using a stacked ensemble method. *Biomed Res Int* 2020;**2020**:7297631
25. Zhang X, Liu S. RBPPred: predicting RNA-binding proteins from sequence using SVM. *Bioinformatics* 2017;**33**:854–62. https://doi.org/10.1093/bioinformatics/btw730
26. Hu S, Ma R, Wang H. An improved deep learning method for predicting DNA-binding proteins based on contextual features in amino acid sequences. *PloS One* 2019;**14**:e0225317. https://doi.org/10.1371/journal.pone.0225317.
27. Qu YH, Yu H, Gong XJ. *et al.* On the prediction of DNA-binding proteins only from primary sequences: a deep learning approach. *PloS One* 2017;**12**:e0188129. https://doi.org/10.1371/journal.pone.0188129.
28. Zheng J, Zhang X, Zhao X. *et al.* Deep-RBPPred: predicting RNA binding proteins in the proteome scale based on deep learning. *Sci Rep* 2018;**8**:15264. https://doi.org/10.1038/s41598-018-33654-x
29. Zhang J, Chen Q, Liu B. 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 Bioinform* 2021;**18**: 1451–63. https://doi.org/10.1109/TCBB.2019.2952338
30. Pradhan UK, Meher PK, Naha S. *et al.* PlDBPred: a novel computational model for discovery of DNA binding proteins in plants. *Brief Bioinform* 2023;**24**:bbac483. https://doi.org/10.1093/bib/bbac483
31. Pradhan UK, Meher PK, Naha S. *et al.* DBPMod: a supervised learning model for computational recognition of DNA-binding proteins in model organisms. *Brief Funct Genomics* 2023;elad039. https://doi.org/10.1093/bfgp/elad039
32. Pradhan UK, Meher PK, Naha S. *et al.* RBPLight: a computational tool for discovery of plant-specific RNA-binding proteins using light gradient boosting machine and ensemble of evolutionary features. *Brief Funct Genomics* 2023;**22**:401–10. https://doi.org/10.1093/bfgp/elad016
33. Pradhan UK, Naha S, das R. *et al.* RBProkCNN: deep learning on appropriate contextual evolutionary information for RNA binding protein discovery in prokaryotes. *Comput Struct Biotechnol J* 2024;**23**:1631–40. https://doi.org/10.1016/j.csbj.2024.04.034

34. Kumar M, Gromiha MM, Raghava GP. SVM based prediction of RNA-binding proteins using binding residues and evolutionary information. *J Mol Recognit* 2011;**24**:303–13. https://doi.org/10.1002/jmr.1061
35. UniProt C. UniProt: the universal protein knowledgebase in 2023. *Nucleic Acids Res* 2023;**51**:D523–31
36. Zaitzeff A, Leiby N, Motta FC. *et al.* Improved datasets and evaluation methods for the automatic prediction of DNA-binding proteins. *Bioinformatics* 2021;**38**:44–51. https://doi.org/10.1093/bioinformatics/btab603
37. Mortimer SA, Kidwell MA, Doudna JA. Insights into RNA structure and function from genome-wide studies. *Nat Rev Genet* 2014;**15**:469–79. https://doi.org/10.1038/nrg3681
38. Ganser LR, Kelly ML, Herschlag D. *et al.* The roles of structural dynamics in the cellular functions of RNAs. *Nat Rev Mol Cell Biol* 2019;**20**:474–89. https://doi.org/10.1038/s41580-019-0136-0
39. Zhang J, Fei Y, Sun L. *et al.* Advances and opportunities in RNA structure experimental determination and computational modeling. *Nat Methods* 2022;**19**:1193–207. https://doi.org/10.1038/s41592-022-01623-y
40. Li W, Godzik A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* 2006;**22**:1658–9. https://doi.org/10.1093/bioinformatics/btl158
41. McGinnis S, Madden TL. BLAST: at the core of a powerful and diverse set of sequence analysis tools. *Nucleic Acids Res* 2004;**32**:W20–5. https://doi.org/10.1093/nar/gkh435
42. UniProt C. UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Res* 2021;**49**:D480–9
43. Wang W, Sun L, Zhang S. *et al.* Analysis and prediction of single-stranded and double-stranded DNA binding proteins based on protein sequences. *BMC Bioinformatics* 2017;**18**:300. https://doi.org/10.1186/s12859-017-1715-8
44. Ashburner M, Ball CA, Blake JA. *et al.* Gene ontology: tool for the unification of biology. The gene ontology consortium. *Nat Genet* 2000;**25**:25–9. https://doi.org/10.1038/75556
45. Ali F, Arif M, Khan ZU. *et al.* SDBP-Pred: prediction of single-stranded and double-stranded DNA-binding proteins by extending consensus sequence and K-segmentation strategies into PSSM. *Anal Biochem* 2020;**589**:113494. https://doi.org/10.1016/j.ab.2019.113494
46. Sharma R, Kumar S, Tsunoda T. *et al.* Single-stranded and double-stranded DNA-binding protein prediction using HMM profiles. *Anal Biochem* 2021;**612**:113954. https://doi.org/10.1016/j.ab.2020.113954
47. Tan C, Wang T, Yang W. *et al.* PredPSD: a gradient tree boosting approach for single-stranded and double-stranded DNA binding protein prediction. *Molecules* 2019;**25**:98. https://doi.org/10.3390/molecules25010098