Briefings in Functional Genomics, 2023, **22**, 401–410
https://doi.org/10.1093/bfgp/elad016
Advance access publication date 8 May 2023
Protocol Article

Editor's Choice

RBPLight: a computational tool for discovery of plant-specific RNA-binding proteins using light gradient boosting machine and ensemble of evolutionary features

Upendra K. Pradhan, Prabina K. Meher*, Sanchita Naha, Soumen Pal, Sagar Gupta, Ajit Gupta and Rajender Parsad

*Corresponding author: Division of Statistical Genetics, ICAR-Indian Agricultural Statistics Research Institute, PUSA, New Delhi 110012, India.
Email: prabina.meher@icar.gov.in

Abstract

RNA-binding proteins (RBPs) are essential for post-transcriptional gene regulation in eukaryotes, including splicing control, mRNA transport and decay. Thus, accurate identification of RBPs is important to understand gene expression and regulation of cell state. In order to detect RBPs, a number of computational models have been developed. These methods made use of datasets from several eukaryotic species, specifically from mice and humans. Although some models have been tested on *Arabidopsis*, these techniques fall short of correctly identifying RBPs for other plant species. Therefore, the development of a powerful computational model for identifying plant-specific RBPs is needed. In this study, we presented a novel computational model for locating RBPs in plants. Five deep learning models and ten shallow learning algorithms were utilized for prediction with 20 sequence-derived and 20 evolutionary feature sets. The highest repeated five-fold cross-validation accuracy, 91.24% AU-ROC and 91.91% AU-PRC, was achieved by light gradient boosting machine. While evaluated using an independent dataset, the developed approach achieved 94.00% AU-ROC and 94.50% AU-PRC. The proposed model achieved significantly higher accuracy for predicting plant-specific RBPs as compared to the currently available state-of-art RBP prediction models. Despite the fact that certain models have already been trained and assessed on the model organism *Arabidopsis*, this is the first comprehensive computer model for the discovery of plant-specific RBPs. The web server RBPLight was also developed, which is publicly accessible at https://iasri-sg.icar.gov.in/rbplight/, for the convenience of researchers to identify RBPs in plants.

Keywords: RNA-binding proteins; deep learning; shallow learning; computational model; evolutionary feature

Introduction

RNA-protein interactions are involved in a wide range of biological activities connected to the gene regulation. Proteins that interact with RNAs are referred to as RNA-binding proteins (RBPs), a diverse class of proteins that contain one or more RNA binding domains in addition to other catalytic or functional domains. In plants, more than 1800 potential RBPs have been discovered, with over 800 of those being enriched in *Arabidopsis* [1, 2]. RBPs

Upendra Kumar Pradhan is presently working as a scientist in the division of Statistical Genetics at ICAR-Indian Agricultural Statistics Research Institute, New Delhi, India. His research interests are focused on bioinformatics, computational biology, biostatistics, machine learning, deep learning and remote sensing modelling.

Prabina Kumar Meher is currently working as a senior scientist at ICAR-Indian Agricultural Statistics Research Institute, New Delhi, India. His research interests are focused on application of machine learning, computational biology, statistical genetics and genomics. Dr. Meher is the recipient of Lal Bahadur Shastri outstanding Young Scientist Award from Indian Council of Agricultural Research (ICAR), Ministry of Agriculture and Farmer Welfare, Government of India. He is an Associate Fellow of National Academy of Agricultural Sciences (NAAS), India. He is also a Selected Member of National Academy Science, India (NASI).

Sanchita Naha is presently working as a scientist in the division of Computer Applications, ICAR-Indian Agricultural Statistics Research Institute (IASRI), New Delhi, India. Her research interests are focused on Recommender Systems, Artificial Intelligence and Machine Learning. Besides this, she has expertise in Web Development, Mobile Application Development, and Design and Development of Databases.

Soumen Pal worked as a scientist from 2010 to 2015, Scientist (Senior Scale) from 2016 to 2020 at ICAR-Indian Agricultural Statistics Research Institute (IASRI), New Delhi, India. Currently, he is working as a senior scientist in the division of Computer Applications, ICAR-IASRI, New Delhi, India. His research interests are focused on Statistical Modeling and ICT in the domain of agriculture and allied sectors. He has designed System Architecture for many National level e-Governance Applications for ICAR. Besides this, he has expertise in Web Development and Design and Development of Databases.

Sagar Gupta received his MSc degree in Bioinformatics from the University of Allahabad, Prayagraj, India, in 2020. He is a PhD research scholar in Bioinformatics at CSIR-Institute of Himalayan Bioresource Technology (CSIR-IHBT), India. His research interests are focused on computational genomics.

Ajit Gupta is currently working as PME Cell Incharge and Head(A), Division of Statistical Genetics at ICAR-Indian Agricultural Statistics Research Institute, New Delhi, India. He received his MSc and PhD degrees in Mathematical Statistics from Agra University, Agra, and Bundelkhand University, Jhansi, respectively. His research work got Peer-Recognition from Karnataka state Govt., India. His research expertise includes Ecological modelling, geostatistics and machine learning.

Rajender Parsad is the Director of ICAR-Indian Agricultural Statistics Research Institute, New Delhi, India. His research interests are focused on Statistics, Design of Experiments, Sample Surveys and Bioinformatics. He is a recipient of the National Award in Statistics for Young Statistician in honor of Prof. C.R. Rao in recognition of outstanding contribution in the field of Statistics from Ministry of Statistics and Programme Implementation, Govt. of India. He was formerly an ICAR National Fellow. He is an elected Fellow of National Academy of Agricultural Sciences and Indian Society of Agricultural Statistics. He is also an elected member of International Statistical Institute, Netherlands.

Received: December 5, 2022. **Revised:** April 12, 2023. **Accepted:** April 21, 2023

play important roles for the growth and development of plants, including genome organisation, stress response, immune response, mRNA processing and post-transcriptional gene regulation [2–7].

The RNA-protein interaction has primarily been analysed in wet lab experiments. In particular, RNA-protein interactions in plants were discovered using *in vitro* techniques like gel shift assay, mutant and knockout screening, nucleic acid-binding assay and other traditional genetics and cell biological methods [8–10]. However, the wet-experiment methods are time-consuming and costly. In other words, RBP experimental studies can be guided by developing computational methods for high-throughput prediction of RBPs. Thus, development of computational techniques that can precisely detect RBPs in plants is the need of the hour.

Several attempts have been made to develop effective computational models for predicting RBPs in eukaryotes. The two main categories of these techniques are machine learning- and template-based techniques. A query protein's similarity to a template RBP or RNA-binding domain (RBD) is measured using the template-based approaches to identify RBPs. Methods in this category include SPalign [11], SPOT-stru [12], SPOT-seq [12], SPOT-Seq-RNA [13] and APRICOT [14]. However, the predefined RBDs cannot be found in almost half of the experimentally identified RBPs [15–16]. Also, proteins with the presence of the RBDs may not always match to RBPs [17]. Therefore, in these two situations, the template-based methods might not work. On the other hand, machine learning-based methods train models using annotated training datasets that contain both RBPs and non-RBPs. Machine learning approaches have been more common in recent years due to their advantage in handling high-dimensional features generated from sequences or structures.

In the existing RBP prediction methodologies, both sequence-derived features and features derived from the protein's 3D structure have been used. The models such as BindUp [18], NucleicNet [19] and NAbind [20] have used 3D structural features to improve the prediction accuracy. However, the majority of the algorithms currently in use have been trained using features extracted from sequence data, as it is simpler to obtain sequence data than 3D structure data. In order to discriminate between RBPs and non-RBPs, Kumar *et al.* (2011) developed a method called RNApred in which binding residues and position-specific scoring matrix (PSSM) profiles were incorporated into the support vector machine (SVM) prediction algorithm [21]. Two distinct approaches based on the Random Forest model have been proposed by Ma *et al.* [22–23], where the two approaches differ in terms of features. In the first approach [22], physicochemical features, evolutionary features and amino acid compositional features were used, whereas in the second approach [23], conjoint triad, binding propensity, non-binding propensity and physicochemical attributes were used. Zhang and Liu (2017) developed RBPPred, a novel sequence-based approach that predicts RBPs using SVM and incorporates physiochemical and evolutionary features derived from PSSM profiles [24]. In another study, Wang *et al.* [25] presented a hierarchical ensemble learning model to integrate three levels of information and suggested a computational predictor called iDRBP-EL to discover DNA-binding proteins (DBPs) and RNA-binding proteins [25]. In another study, Wang *et al.* [26] further suggested a novel feature representation approach for representing protein sequences known as PSSM and Position-Specific Frequency Matrix (PSFM) Cross Transformation (PPCT). Based on the PPCT features and Random Forest method, the authors presented a new computational predictor named IDRBP-PPCT to identify DBPs, RBPs and DRBPs (both DBPs and RBPs).

Along with shallow machine learning models like SVM and Random Forest, deep learning models have also been utilized for the prediction of RBPs. A method called Deep-RBPPred was developed by Zheng *et al.* [27] by employing the protein feature of RBPPred and convolutional neural network (CNN). Zhang *et al.* [28] developed a multi-label learning model known as iDRBP_MMC based on the motif-based CNN to address the cross-prediction issue and enhance the predictive performance of DBPs and RBPs. In addition, Zhang *et al.* [29] also developed the DeepDRBP-2L, a two-level predictor for predicting RBPs by fusing CNN with Long Short-Term Memory (LSTM). In another study, Zhang *et al.* [30] proposed the PreRBP-TL model to detect species-specific RBPs based on transfer learning. In this model, weights were initially set up using pre-training on a sizable RBP dataset, and were then improved using transfer learning on a smaller RBP dataset that was specialised to a single species. In a similar vein, Peng *et al.* [31] proposed the RBP-TSTL technique, which integrates the learning from the annotated pre-training RBPs dataset with the feature embedding produced by a self-supervised pre-trained model.

The majority of the machine learning-based models mentioned above have been developed using RBP sequence data from a wide range of eukaryotic species, yielding models that are generalized in nature. However, RBPs are specific to distinct species as well as to lineage-specific families [1, 32]. Thus, it may not be possible to predict plant-specific RBPs with higher accuracy using the current generic models. Even though some of the models have been tested on model plants, these methods predict the RBPs with a poor degree of accuracy for other plant species. Despite significant advancements in RBP prediction, plant-based model development is mostly ignored. Thus, there is a need to develop computational method for prediction of plant-specific RBPs. In the present study, we proposed a novel computational tool called RBPLight to predict plant-specific RBPs. The devised method took use of machine learning algorithms for prediction purpose, where ensemble of evolutionary features was utilized as input in machine learning algorithms.

Materials and methods

Retrieval and processing of sequence data

The plant RBP sequences were collected from CISBP-RNA [33] and UniProtKB (accessed on 16/07/2022) [34] databases. The CISBP-RNA database includes experimentally validated RBP sequences. The UniProtKB database consists of two sections: reviewed and unreviewed. The reviewed section comprises manually annotated protein sequences with information extracted from literature, whereas the unreviewed section comprises protein sequences associated with computationally generated annotation (https://www.uniprot.org/help/uniprotkb). For the current investigation, we retrieved the RBP and non-RBP sequences from the reviewed section of the UniProtKB, based on gene ontology (GO) terms. More clearly, the RBP sequences were defined as the protein sequences annotated with the GO term 'RNA-binding' (GO: 0003723), whereas non-RBP sequences were defined as proteins without the annotation. Similar approach has also been adopted in earlier studies [21, 24–31] for retrieving RBP and non-RBP sequences from UniProtKB database. A total of 16,453 RBP sequences, including 13,162 sequences from CISBP-RNA and 3,291 sequences from UniProtKB, were obtained for 36 distinct plant species. On the other side, a total of 19,251 non-RBP sequences were retrieved from the UniProtKB database. Protein sequences with non-standard residues (B, J, O, U, X and Z) and less than 50 amino acids in length were removed. To eliminate the homologous bias in the prediction accuracy that may arise from including redundant or highly similar sequences, homology reduction was applied to both RBP and non-RBP datasets. Homology reduction in protein dataset refers to the process of reducing the redundancy by removing highly similar protein sequences while retaining the diversity of the dataset. The CD-HIT [35] approach was used to remove sequences from each data set that shared >40% of their sequence identity with any other sequences. The CD-HIT command is provided in the supplementary file. After removing redundancy, 6,921 non-RBP sequences and 2,696 RBP sequences were obtained. Out of the 2,696 RBPs, 200 sequences were kept aside in order to be utilised as an independent positive test set, and the remaining 2,496 RBPs were used as the positive training set. To prevent prediction bias toward the non-RBP class that had a larger number of observations, a balanced dataset of 2,496 RBP and 2,496 non-RBP sequences was considered. The 2,496 non-RBP sequences were selected randomly from the 6,921 non-RBP sequences. To generate the positive independent test set, we once more retrieved the plant RBP sequences from the UniProtKB database (accessed on 9 October 22). A total of 575 RBP sequences were found for 35 distinct plant species. After removing the protein sequences with irregular residues and length of less than 50 amino acids, a non-redundant dataset of 343 RBP sequences was obtained. Therefore, a total of 543 RBP sequences (343 + 200) were used for the positive independent test dataset. In order to have a fair prediction, 543 randomly selected non-RBP sequences were taken into consideration from the remaining 4,425 non-RBP sequences (after using 2,496 out of 6,921 non-RBP sequences for the training set). To put it simply, the independent dataset was created by combining 543 non-RBP sequences with 543 RBP sequences.

Generation of numeric features from sequence data

The development of sequence-based RBP prediction requires numerical representation of the RBP sequence, as machine learning algorithms cannot accept the sequence data directly. The numerical representation has a significant impact on credibility of the prediction model in terms of prediction accuracy. Both sequence-based features and PSSM-based evolutionary features were used in the current study. Specifically, we considered 20 different sequence-based feature sets (Supplementary Table S1) and 20 feature sets obtained from PSSM profile (Supplementary Table S2). The *protr* R-package [36], *ftrCOOL* R-Package [37], *Peptides* R-packages [38] and *iFeature* Python module [39] were used to generate the sequence-based features, whereas the *PSSMCOOL* R-package [40] and *POSSUM* standalone toolkit [41] were utilized to implement all functionalities for PSSM-based feature descriptors. In Supplementary Data, a concise description of each sequence and PSSM-based feature set is provided with the required citations.

Prediction algorithms

The existing research on RBP prediction have used both shallow learning and deep learning algorithms. In the present study, we evaluated the accuracy of ten different shallow learning techniques, including SVM [42], extreme gradient boosting (XGBoost) [43], Random Forests (RFs) [44], light gradient boosting machine (LightGBM) [45], multi-layer perceptrons (MLP) [46], Bagging [47], adaptive boosting (AdaBoost) [48], stochastic gradient descent (SGD) [49], NaiveBayes [50] and gradient tree boosting (GBDT) [51], using both sequence-derived and PSSM-derived features. In addition, five deep learning models were also used, including one-dimensional convolutional neural networks (CNN_1D) [52], attention-based convolutional neural networks (ABCNN) [53], long short-term memory (LSTM) [54], bidirectional LSTM (Bi-LSTM) [55] and AutoEncoder (AE) [56]. The AE was not used for classification in this study. Rather, the feature representation of the input data acquired through the usage of the AE supervised learning model was employed as input in the deep neural network (DNN) for classification. The DNN along with AE features was denoted as AE_DNN in this study. The SVM, RF, XGBoost, AdaBoost, NaiveBayes, LightGBM, MLP, Bagging, SGD and GBDT algorithms, respectively, were implemented using the R-packages *e1071*, *randomForest*, *xgboost*, *adabag*, *fastNaiveBayes*, *lightgbm*, *RSNNS*, *ipred*, *sgd* and *gbm*. With the help of the *PyTorch* and *TensorFlow* libraries of Python, deep learning models were executed. Supplementary Table S3 provides information on the software used to implement the learning models and the parameter setup.

Cross-validation and performance metrics

A repeated five-fold cross-validation approach was used to assess the performance of the classification models, and the experiment was repeated 100 times. Each RBP and non-RBP dataset was randomly separated into five subgroups of equal size in order to perform the five-fold cross-validation [57]. In each fold of the cross-validation, one randomly selected subset from the RBP and non-RBP classes was used as a test set, and the remaining four subsets from both classes were pooled to serve as a training set. Distinct training and test sets were used five times during the five-fold classification process. The performance metrics were calculated by taking average of the accuracy over all five test sets and 100 replications. Figure 1 shows the methodological flow diagram outlining each phase of the proposed computational model. The five-fold cross validation and different performance metrics used to evaluate the effectiveness of the prediction models are shown in Figure 2.

Result

Prediction analysis of shallow learning models

The accuracy of 10 shallow learning algorithms was examined with 20 sequence-derived and 20 evolutionary feature sets using 50% of the observations of the entire dataset. When compared to features obtained from sequences, evolutionary features were shown to be more accurate. Among the 10 algorithms, LightGBM and XGBoost had the highest accuracy for both kinds of feature sets. Among the sequence-derived feature sets, the CKSAAP feature set with the LightGBM approach had the highest AU-ROC (85.50%) and AU-PRC (85.40%) values (Figure 3). In other words, CKSAAP was found to be the only sequence-derived feature set with >85% accuracy. The accuracy, on the other hand, was <85% for the remaining algorithms and sequence-based feature set combinations. Out of 20 PSSM-derived feature sets, only seven feature sets such as AADP_PSSM, AATP_PSSM, TPC_PSSM, DP_PSSM, PSE_PSSM, Kbigram_PSSM and Trigram_PSSM showed ≥90% AU-ROC and AU-PRC (Figure 3). Specifically, the Trigram_PSSM feature set and LightGBM method achieved the highest AU-ROC (91.9%) and AU-PRC (92.3%). Similar accuracies for LightGBM were also obtained for other six PSSM-derived feature sets (Figure 3).

Performance analysis of deep learning models

The performance of five cutting-edge deep learning models, including CNN_1D, LSTM, Bi-LSTM, AE_DNN and ABCNN, was then evaluated using the seven PSSM-derived feature and one

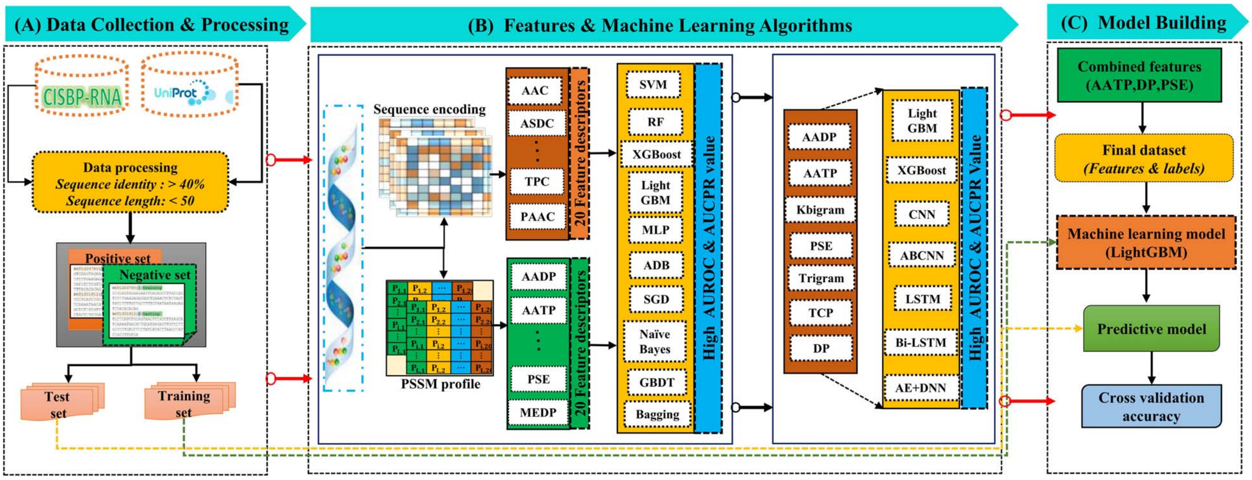

Figure 1: Illustration of the brief outline of the proposed approach. The diagram depicts the overall design of the entire computational strategies followed to develop the RBP prediction model. (A) Retrieval of RBP and non-RBP sequences from the Uniprot and CISBP-RNA database and processing of sequence data; (B) sequence- and PSSM-derived feature generation and selection of most important features descriptor and MLA, based on AU-ROC and AU-PRC; (C) model building using different machine learning techniques and assessment of cross-validation accuracy.

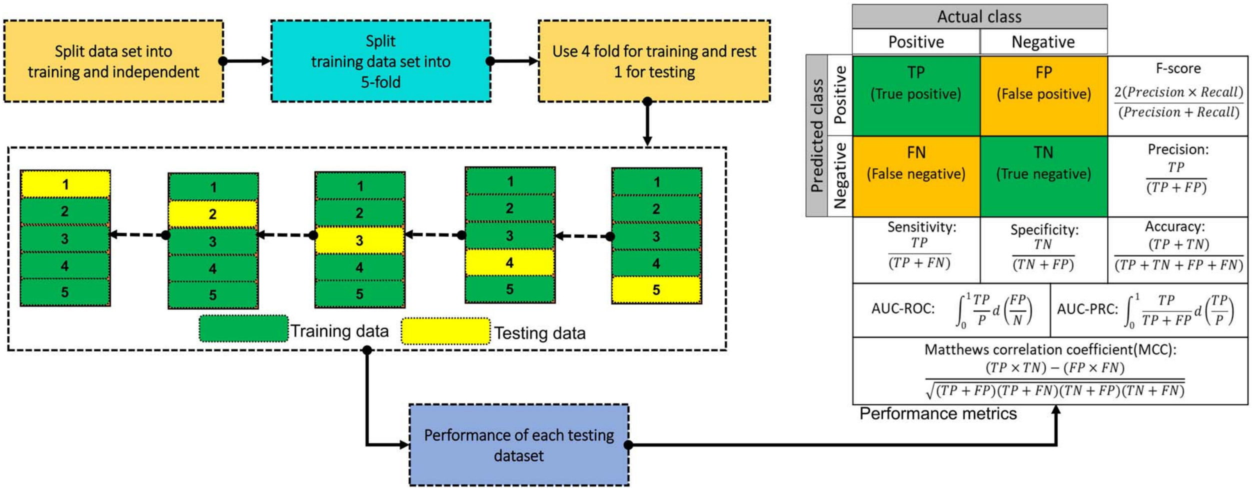
Split data set into training and independent
Split training data set into 5-fold
Use 4 fold for training and rest 1 for testing
1
2
3
4
5
Training data
Testing data
Performance of each testing dataset
Actual class
Positive
Negative
Predicted class
Positive
Negative
TP (True positive)
FP (False positive)
FN (False negative)
TN (True negative)
F-score $\frac{2(Precision \times Recall)}{(Precision + Recall)}$
Precision: $\frac{TP}{(TP+FP)}$
Sensitivity: $\frac{TP}{(TP+FN)}$
Specificity: $\frac{TN}{(TN+FP)}$
Accuracy: $\frac{(TP+TN)}{(TP+TN+FP+FN)}$
AUC-ROC: $\int_0^1 \frac{TP}{P} d\left(\frac{FP}{N}\right)$
AUC-PRC: $\int_0^1 \frac{TP}{TP+FP} d\left(\frac{TP}{P}\right)$
Matthews correlation coefficient(MCC): $\frac{(TP \times TN) - (FP \times FN)}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}$
Performance metrics

Figure 2: Illustration of the five-fold cross validation approach and different performance metrics used for evaluating the performance of learning algorithms.

sequence-derived feature sets in which higher accuracy was discovered using shallow learning algorithms. The performance of the deep learning models was also compared with LightGBM and XGBoost, as these two algorithms were found to be the top two performers among shallow learning algorithms. For all the feature sets, the LightGBM and XGBoost achieved higher accuracy as compared to the five deep learning models. In comparison to other deep learning models, CNN_1D achieved greater accuracy for four PSSM-derived feature sets, including PSE_PSSM (AU-ROC: 90.04%, AU-PRC: 89.94%), AATP_PSSM (AU-ROC: 87.16%, AU-PRC: 87.14%), AADP_PSSM (AU-ROC: 85.07%, AU-PRC: 86.38%) and Kbigram_PSSM (AU-ROC: 87.07%, AU-PRC: 87.36%) (Figure 4). Similar to how AE_DNN obtained higher accuracy for the DP_PSSM feature set, LSTM did so for the TPC_PSSM (AU-ROC: 87.15%, AU-PRC: 87.08%) and Trigram_PSSM (AU-ROC: 89.13%, AU-PRC: 87.75%) feature sets, respectively (Figure 4). The LightGBM and XGBoost with PSSM-based features outperformed other plausible combinations of learning algorithms and feature sets (Figure 4). Consequently, LightGBM and XGBoost along with the PSSM-based features were taken into consideration for subsequent analysis.

Prediction analysis with feature combination

To reduce computational complexity, the Trigram_PSSM feature set was ignored because it includes 8,000 features, which was far more than other PSSM-based feature sets. AATP_PSSM was chosen over AADP_PSSM because, it had a marginally higher accuracy despite having the same number of features (420). Similarly, TPC_PSSM was chosen over Kibigram_PSSM due to the same number of features (400), but slightly higher accuracy. In total, four feature sets such as AATP_PSSM, DP_PSSM, PSE_PSSM and TPC_PSSM were taken into account for further analysis. The four feature sets were combined in different conceivable ways for the prediction analysis (Table 1). Similar accuracy was obtained across feature combinations by XGBoost and LightGBM. For both learning algorithms, AU-ROC and AU-PRC, respectively, were obtained >90% and > 91% for all feature combinations. The AATP_PSSM+DP_PSSM+PSE_PSSM and AATP_PSSM+DP_PSSM+PSE_PSSM+TPC_PSSM feature sets performed slightly better than other feature combinations (Table 1). LightGBM achieved AU-ROC of 91.24 ± 0.212 and AU-PRC of 91.91 ± 0.210 for the feature combination AATP_PSSM+DP_PSSM +PSE_PSSM, which was slightly higher than that of XGBoost

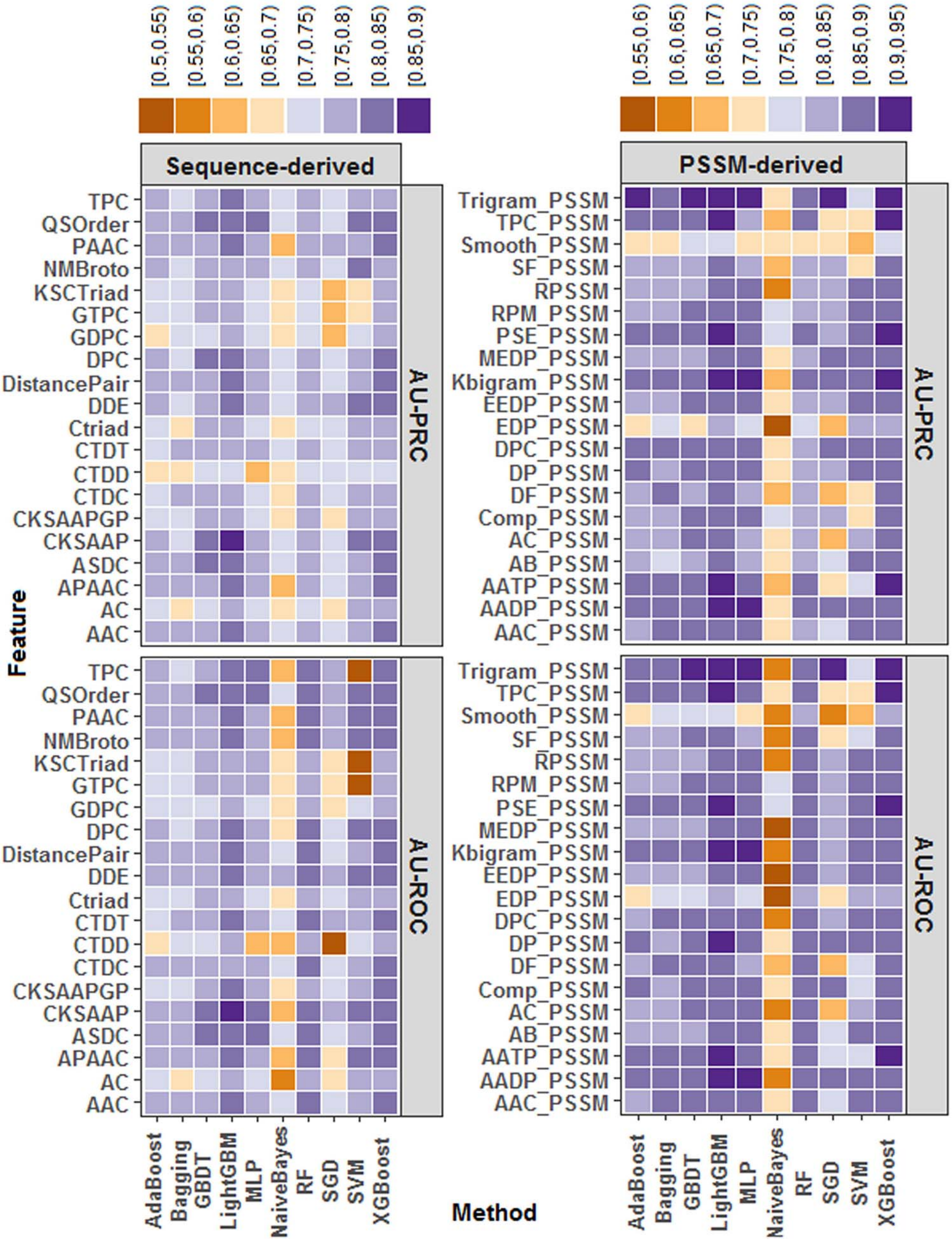

Figure 3: Heat maps of the AU-ROC and AU-PRC for predicting plant-specific RBPs employing 10 shallow learning algorithms coupled with 20 sequence-derived feature sets and 20 PSSM-derived feature sets.

(AU-ROC: 91.07 ± 0.151, AU-PRC: 91.84 ± 0.153). Also, the estimates of AU-ROC (91.21 ± 0.165) and AU-PRC (91.92 ± 0.167) of LightGBM were marginally higher than those of XGBoost (91.10 ± 0.163, 91.86 ± 0.141) for the feature set AATP_PSSM+DP_PSSM+PSE_PSSM+TPC_PSSM. When both feature sets were compared, nearly the same accuracy was found. As a result, the feature set with 400 less features, AATP_PSSM+DP_PSSM+PSE_PSSM, was chosen to develop the final prediction model.

Prediction with independent test set

The performance of the developed model was further evaluated using an independent dataset consisting of 543 RBP and 543 non-RBP sequences. Using the LightGBM learning technique, the prediction model was trained with the AATP_PSSM+DP_PSSM+PSE_PSSM features derived from the training dataset of 2,496 RBP and 2,496 non-RBP sequences. For the test dataset, AU-PRC of 94.00% and AU-ROC of 94.50% were achieved (Figure 5). This suggests the robustness and effectiveness of the developed model for predicting the plant-specific RBPs with higher accuracy.

Performance of existing tools on plant dataset

Since the plant dataset have been used in the training model of the existing tools, the performance of the developed approach (trained with 2,496 RBPs and 2,496 non-RBPs) and the existing tools were computed and compared based on the accuracy of an independent dataset that comprises 543 RBPs and 543 non-RBPs of plants. The performance of 10 state-of-art existing models and the proposed approach is shown in Table 2. Among the existing tools, RBPPred had the best accuracy (79.10%), whereas Deep-RBPPred and iDRBP-ECHF were shown to have the highest sensitivity (85.08%) and specificity (79.56%), respectively. The first model developed for the prediction of RBPs, RNApred, had the

Table 1. Performance of XGBoost and LightGBM with different combination of PSSM-derived features

Feature combination	XGBoost		LightGBM	
	AU-ROC±SE	AU-PRC ± SE	AU-ROC±SE	AU-PRC ± SE
F1 + F2	91.08 ± 0.214	91.84 ± 0.209	91.23 ± 0.166	91.91 ± 0.186
F1 + F3	90.55 ± 0.185	91.30 ± 0.194	90.55 ± 0.196	91.24 ± 0.192
F1 + F4	90.35 ± 0.172	91.13 ± 0.180	90.42 ± 0.183	91.08 ± 0.195
F2 + F3	90.61 ± 0.183	91.28 ± 0.187	90.57 ± 0.193	91.21 ± 0.221
F2 + F4	91.03 ± 0.161	91.78 ± 0.183	91.18 ± 0.170	91.89 ± 0.169
F3 + F4	90.54 ± 0.192	91.29 ± 0.184	90.54 ± 0.202	91.24 ± 0.195
F1 + F2 + F3	**91.07 ± 0.151**	**91.84 ± 0.153**	**91.24 ± 0.212**	**91.92 ± 0.210**
F1 + F2 + F4	91.06 ± 0.198	91.82 ± 0.178	91.23 ± 0.166	91.91 ± 0.186
F1 + F3 + F4	90.50 ± 0.189	91.27 ± 0.170	90.55 ± 0.196	91.24 ± 0.192
F2 + F3 + F4	91.11 ± 0.177	91.83 ± 0.144	91.23 ± 0.211	91.91 ± 0.212
F1 + F2 + F3 + F4	**91.10 ± 0.163**	**91.86 ± 0.141**	**91.24 ± 0.165**	**91.92 ± 0.167**

F1: AATP_PSSM, F2: DP_PSSM, F3: PSE_PSSM, F4: TPC_PSSM

Table 2. Performance of existing RBP prediction tools on plant dataset

Method	Sensitivity	Specificity	Precision	Accuracy	F1 Score	MCC
RNApred	79.74	15.29	48.49	47.51	64.82	15.14
RBPPred	84.16	74.03	76.42	**79.10**	80.11	58.50
Deep-RBPPred	85.08	42.73	59.77	63.90	70.21	30.70
iDRBP_MMC	36.10	76.61	60.68	56.35	45.27	13.90
DeepDRBP-2L	74.95	74.59	74.68	74.77	74.82	49.54
iDRBP-EL	55.06	70.17	64.86	62.62	59.56	25.52
iDRBP-ECHF	25.97	79.56	55.95	52.76	35.47	6.54
IDRBP-PPCT	63.54	69.24	67.38	66.39	65.40	32.83
PreRBP-TL[a]	77.16	76.61	76.74	**76.89**	76.95	53.78
RBP-TSTL[a]	83.43	54.51	64.71	68.97	72.44	39.22
Our approach	**86.74**	**86.74**	**86.74**	**86.74**	**86.74**	**73.48**

[a] *A. thaliana* model; bold font denotes a higher value. MCC, Matthews correlation coefficient

lowest accuracy (47.51%). With the same dataset, the accuracy of the proposed model was determined to be 86.74%, which was ~7% higher than the top-performing model (RBPPred) from the pool of existing models. Additionally, it was observed that the MCC and F1-score of the proposed model were, respectively, 19% and 7% higher than that of RBPPred. The AU-ROC and the AU-PRC of the proposed approach was also found to be much higher than that of best performing two existing models, i.e. RBPPred and PreRBP-TL (Figure 5). To further assess the accuracy of the top two existing models and the proposed approach on the most recent RBP dataset, 871 RBP sequences were retrieved from the UniProtKB on 21/02/2023. It was made sure that these sequences weren't present in the positive sets of the training and independent test sets. After eliminating non-standard amino acids and sequences with more than 40% sequence similarity to other sequences, the remaining 360 sequences were used for prediction analysis. It was observed that 91.38, 64.17 and 68.33 percentage of the sequences were correctly predicted by RBPLight, RBPPred and PreRBP-TL, respectively.

Species-specific RBP prediction

For *Arabidopsis*, rice, tomato, soybean and sorghum, respectively, 198, 198, 29, 24 and 18 experimentally validated RBPs were collected from the UniProtKB on 12 October 2022 in order to further verify the effectiveness of the developed model. Additionally, it was made sure that these sequences weren't in the training positive set. The proposed model was used to predict these sequences, and it was found that for *Arabidopsis*, rice, tomato, soybean and sorghum, respectively, 87.37, 89.39, 89.65, 91.66 and 100 percentage of the sequences were correctly predicted.

Prediction server RBPLight

An online prediction tool called RBPLight (https://iasri-sg.icar.gov.in/rbplight/) was developed based on the proposed model for the prediction of RBPs in plants. The user has to supply the protein sequences in FASTA format, excluding non-standard residues. The probability with which each sequence was predicted as RBP or non-RBP by LightGBM learning algorithms are displayed in a tabular manner. For making prediction using a larger size dataset, the link to download the source code is available at the server site.

Discussion

RBPs are necessary for several biological processes related to gene regulation in plants. Consequently, the discovery of RBPs has significant theoretical and practical implications for plant proteomics and genomics research [7, 58–59]. Identification of RBPs within proteomes is a difficult process because of the variety of RNA properties and the existence of inherently disordered regions [17]. A number of computational methods have been developed for the purpose of identifying RBPs, the majority of which focus on human data while a few models have been evaluated on *Arabidopsis thaliana*. On their own test datasets, every existing model was said to have performed well. However, in our evaluation using the most recent RBP collection, which includes experimentally validated plant RBPs, the existing models performed poorly. Despite tremendous progress in this domain, no

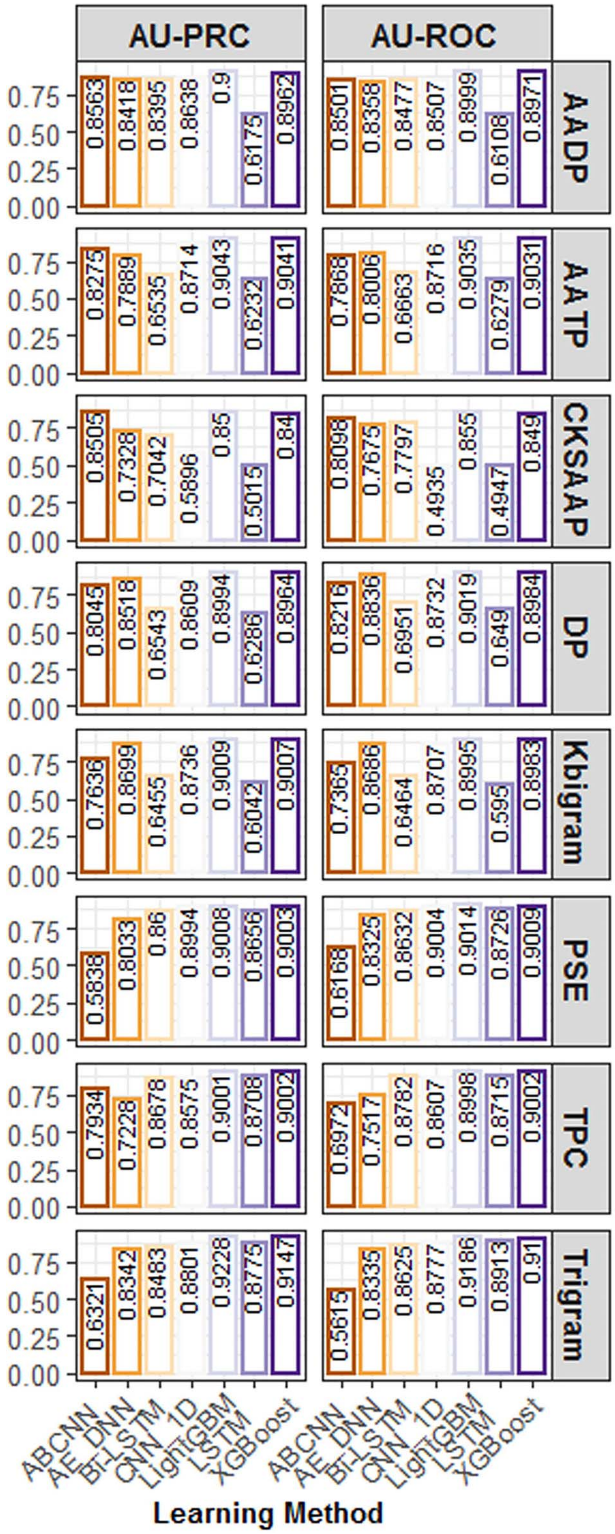

Figure 4: Bar diagrams of the AU-ROC and AU-PRC for prediction of RBPs using five deep learning and two shallow learning models with seven selected PSSM-derived feature sets and one sequence-derived feature set.

method for identifying RBP specific to plants have been developed. In this paper, we introduced a novel computational model called RBPLight that makes use of evolutionary feature information as input to predict plant-specific RBPs from protein sequences using light gradient boosting machine (LightGBM).

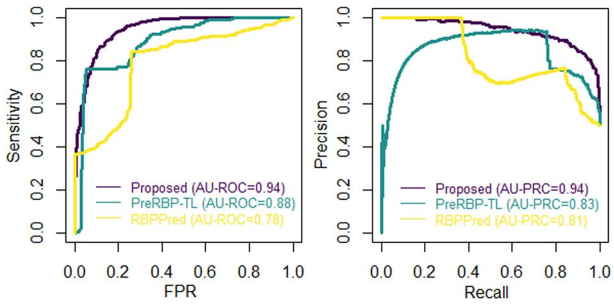

Figure 5: ROC and PR curves for predicting independent dataset of plant RBPs.

Both sequence-derived features and PSSM-based evolutionary features were utilized in the current study due to their successful implementation in earlier RBP prediction models [21, 24, 27]. In contrast to previous studies that included only a few sequence-derived and evolutionary features, we conducted a thorough assessment which included 20 different types of both sequence-derived and PSSM-derived feature sets. It was found that evolutionary features based on PSSM had higher prediction accuracy than features obtained from sequence. Earlier RBP prediction models have also used PSSM-derived features including PSSM-400 [21, 24, 27, 60, 61], BLOSUM62 [25–26, 28–30] and PSSM-TPC [62]. However, past research has not examined the feature sets used in the current work, such as AATP_PSSM, DP_PSSM and PSE_PSSM. Furthermore, the top three feature sets (420 for AATP _PSSM, 240 for DP_PSSM and 40 for PSE_PSSM), which were used for the final prediction, were combined to yield a total of 700 features. Since the number of observations (2496 RBPs and 2496 non-RBPs) was substantially more than the number of features (700) employed, there is a very less chance that the model will be poorly trained. Thus, no feature selection strategy was employed in the present study.

Structure-based features have been shown to improve RBP prediction accuracy. However, we have used only the sequence-derived feature in this investigation. This is because the number of experimentally solved protein structures is less, whereas sequence-based models have access to much more data that can help improve accuracy by training the model with a larger dataset. Besides, predicting a protein's structure is a more computationally difficult task than extracting sequence-derived features, and hence sequence-based models are much easier to train and deploy than structure-based models.

Five deep learning models and ten shallow learning models were assessed for prediction in the current study. With very few exceptions, it was observed that shallow learning models outperformed deep learning methods across feature sets. In particular, XGBoost and LightGBM outperformed the other learning algorithms in terms of accuracy. When comparing the accuracy of these two algorithms, LightGBM came out slightly ahead of XGBoost. The key distinction between the two algorithms is that trees grow leaf-wise in LightGBM whereas trees develop depth-wise in XGBoost [63]. One of the likely causes of LightGBM's higher accuracy in comparison to XGBoost is that it follows a leaf wise split strategy rather than a level-wise split approach. Although the LightGBM model has been used for prediction in other areas of computational biology, such as protein–protein interactions [64], protein-ATP binding residues [65] and DNA-binding residue prediction [66], RBP prediction has not yet been explored using this learning algorithm.

To further assess the reliability of the developed model, we compared the performance of RBPLight with ten other state-of-the-art methods using an independent test dataset. The proposed method was found to outperform the compared methods in terms of accuracy. To put it another way, the existing methods were less accurate at predicting RBPs specific to plants than they are for other eukaryotic species like human and mouse. RBPs are well known to be highly tissue- and species-lineage-specific [1, 32], in contrast to the existing models, which were developed based on protein sequences from a range of eukaryotic species and are therefore less accurate for plant-specific RBPs. Despite having been trained on the model organism *Arabidopsis*, models like PreRBP-TL [30] and RBP-TSTL [31] show poor accuracy when predicting the RBPs of other plant species. This suggests that building an effective RBP prediction model for plants may not be possible by focusing exclusively on the model organism of the plant. Hence, we included RBPs from over 36 plant species in the current study.

The performance of the proposed computational model was also evaluated using an independent test dataset in addition to cross-validation analysis, in order to demonstrate its robustness and generalization ability. It was found that the overall accuracy using the independent dataset was comparable with the accuracy obtained using cross-validation. This shows that the accuracy of the model was not overestimated or underestimated. In addition to independent validation, the effectiveness of the developed model was evaluated for predicting experimentally validated RBPs for five distinct species, including *Arabidopsis*, rice, tomato, soybean and sorghum. The higher accuracy for species-specific prediction endorses up the dependability and generalizability of the proposed model in respect of predicting RBPs for different plant species.

Conclusion

The proposed method RBPLight offers a substantially improvement in the prediction accuracy for plant-specific RBPs when compared with the existing approaches. Due to encouraging results, the RBPLight can be effectively used for large-scale annotation of plant-specific proteins by utilising only sequence information. For predicting plant-specific RBPs, we have developed an online prediction tool RBPLight (https://iasri-sg.icar.gov.in/rbplight/). It is anticipated that the proposed approach will supplement the existing models and experimental techniques for identifying plant-specific RBPs.

Key Points

- Proposed a novel computational method, RBPLight for identifying plant-specific RNA-binding proteins (RBPs).
- The RBPLight achieved high accuracy and outperformed several existing tools for RBP identification in different plant species.
- Species-specific RBP identification using experimentally validated RBP sequences confirmed the reliability and generalized predictive ability of RBPLight.
- RBPLight is also available in the form of an online prediction tool which is freely accessible at https://iasri-sg.icar.gov.in/rbplight/.
- The proposed approach is expected to supplant the existing tools and methodologies for recognizing plant-specific RBPs.

Supplementary data

Supplementary data mentioned in the text are available to subscribers in *Briefings in Functional Genomics* online.

Funding

This work was funded by ICAR-Indian Agricultural Statistics Research Institute, PUSA, New Delhi, India (Grant no. AGEDIAS-RISIL202101700188).

Conflict of Interest statement

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

All the datasets used in the present study are available at https://iasri-sg.icar.gov.in/rbplight/dataset.php.

References

1. Marondedze C, Thomas L, Serrano NL, *et al.* The RNA-binding protein repertoire of Arabidopsis thaliana. *Sci Rep* 2016;**6**(1):29766.
2. Marondedze C. The increasing diversity and complexity of the RNA-binding protein repertoire in plants. *Proc R Soc B: Biol Sci* 2020;**287**(1935):20201397.
3. Woloshen V, Huang S, Li X. RNA-binding proteins in plant immunity. *J Pathog* 2011;**2011**:1–11.
4. Huh SU, Paek K-H. Plant RNA binding proteins for control of RNA virus infection. *Front Physiol* 2013;**4**:397.
5. Lee K, Kang H. Emerging roles of RNA-binding proteins in plant growth, development, and stress responses. *Mol Cells* 2016;**39**:179–85.
6. Dedow LK, Bailey-Serres J. Searching for a match: structure, function and application of sequence-specific RNA-binding proteins. *Plant Cell Physiol* 2019;**60**:1927–38.
7. Muthusamy M, Kim J-H, Kim JA, *et al.* Plant RNA binding proteins as critical modulators in drought, high salinity, heat, and cold stress responses: an updated overview. *Int J Mol Sci* 2021;**22**:6731.
8. Vermel M, Guermann B, Delage L, *et al.* A family of RRM-type RNA-binding proteins specific to plant mitochondria. *Proc Natl Acad Sci USA* 2002;**99**:5866–71.
9. Staiger D, Zecca L, Wieczorek Kirk DA, *et al.* The circadian clock regulated RNA-binding protein AtGRP7 autoregulates its expression by influencing alternative splicing of its own pre-mRNA. *Plant J* 2003;**33**:361–71.
10. Lee J, Lee I. Regulation and function of SOC1, a flowering pathway integrator. *J Exp Bot* 2010;**61**:2247–54.
11. Yang Y, Zhan J, Zhao H, *et al.* A new size-independent score for pairwise protein structure alignment and its application to structure classification and nucleic-acid binding prediction. *Proteins* 2012;**80**:2080–8.
12. Zhao H, Yang Y, Zhou Y. Highly accurate and high-resolution function prediction of RNA binding proteins by fold recognition and binding affinity prediction. *RNA Biol* 2011;**8**:988–96.
13. Yang Y, Zhao H, Wang J, *et al.* SPOT-Seq-RNA: predicting protein-RNA complex structure and RNA-binding function by fold recognition and binding affinity prediction. *Methods Mol Biol* 2014;**1137**:119–30.

14. Sharan M, Förstner KU, Eulalio A, *et al.* APRICOT: an integrated computational pipeline for the sequence-based identification and characterization of RNA-binding proteins. *Nucleic Acids Res* 2017;**45**:e96.
15. Beckmann BM, Horos R, Fischer B, *et al.* The RNA-binding proteomes from yeast to man harbour conserved enigmRBPs. *Nat Commun* 2015;**6**:10127.
16. Van Nostrand EL, Freese P, Pratt GA, *et al.* A large-scale binding and functional map of human RNA-binding proteins. *Nature* 2020;**583**:711–9.
17. Hentze MW, Castello A, Schwarzl T, *et al.* A brave new world of RNA-binding proteins. *Nat Rev Mol Cell Biol* 2018;**19**: 327–41.
18. Paz I, Kligun E, Bengad B, *et al.* BindUP: a web server for non-homology-based prediction of DNA and RNA binding proteins. *Nucleic Acids Res* 2016;**44**:W568–74.
19. Lam JH, Li Y, Zhu L, *et al.* A deep learning framework to predict binding preference of RNA constituents on protein surface. *Nat Commun* 2019;**10**:4941.
20. Shazman S, Mandel-Gutfreund Y. Classifying RNA-binding proteins based on electrostatic properties. *PLoS Comput Biol* 2008;**4**:e1000146.
21. Kumar M, Gromiha MM, Raghava GPS. SVM based prediction of RNA-binding proteins using binding residues and evolutionary information. *J Mol Recognit* 2011;**24**:303–13.
22. Ma X, Guo J, Sun X. Sequence-based prediction of RNA-binding proteins using Random Forest with minimum redundancy maximum relevance feature selection. *Biomed Res Int* 2015;**2015**:425810.
23. Ma X, Guo J, Xiao K, *et al.* PRBP: prediction of RNA-binding proteins using a Random Forest algorithm combined with an RNA-binding residue predictor. *IEEE/ACM Trans Comput Biol Bioinform* 2015;**12**:1385–93.
24. Zhang X, Liu S. RBPPred: predicting RNA-binding proteins from sequence using SVM. *Bioinformatics* 2017;**33**:854–62.
25. Wang N, Zhang J, Liu B. iDRBP-EL: identifying DNA- and RNA-binding proteins based on hierarchical ensemble learning. *IEEE/ACM Trans Comput Biol Bioinform* 2021;**20**(1):432–41.
26. Wang N, Zhang J, Liu B. IDRBP-PPCT: identifying nucleic acid-binding proteins based on position-specific score matrix and position-specific frequency matrix cross transformation. *IEEE/ACM Trans Comput Biol Bioinform* 2021;**19**(4):2284–93.
27. 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.
28. Zhang J, Chen Q, Liu B. 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* 2020;**432**:5860–75.
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.
30. Zhang J, Yan K, Chen Q, *et al.* PreRBP-TL: prediction of species-specific RNA-binding proteins based on transfer learning. *Bioinformatics* 2022;**38**:2135–43.
31. Peng X, Wang X, Guo Y, *et al.* RBP-TSTL is a two-stage transfer learning framework for genome-scale prediction of RNA-binding proteins. *Brief Bioinform* 2022;**23**:bbac215.
32. Nagarajan R, Gromiha MM. Prediction of RNA binding residues: an extensive analysis based on structure and function to select the best predictor. *PloS One* 2014;**9**:e91140.
33. Ray D, Kazan H, Cook KB, *et al.* A compendium of RNA-binding motifs for decoding gene regulation. *Nature* 2013;**499**:172–7.
34. The UniProt Consortium. UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Res* 2021;**49**:D480–9.
35. Huang Y, Niu B, Gao Y, *et al.* CD-HIT Suite: a web server for clustering and comparing biological sequences. *Bioinformatics* 2010;**26**:680–2.
36. Xiao N, Cao D-S, Zhu M-F, *et al.* protr/ProtrWeb: R package and web server for generating various numerical representation schemes of protein sequences. *Bioinformatics* 2015;**31**:1857–9.
37. Amerifar S, Norouzi M, Ghandi M. A tool for feature extraction from biological sequences. *Brief Bioinform* 2022;**23**:bbac108.
38. Osorio D, Rondón-Villarreal P, Torres R. Peptides: a package for data mining of antimicrobial peptides. *The R Journal* 2015;**7**:4.
39. Chen Z, Zhao P, Li F, *et al.* iFeature: a Python package and web server for features extraction and selection from protein and peptide sequences. *Bioinformatics* 2018;**34**:2499–502.
40. Mohammadi A, Zahiri J, Mohammadi S, *et al.* PSSMCOOL: a comprehensive R package for generating evolutionary-based descriptors of protein sequences from PSSM profiles. *Biol Methods Protoc* 2022;**7**:bpac008.
41. Wang J, Yang B, Revote J, *et al.* POSSUM: a bioinformatics toolkit for generating numerical sequence feature descriptors based on PSSM profiles. *Bioinformatics* 2017;**33**:2756–8.
42. Vapnik V. Pattern recognition using generalized portrait method. *Autom Remote Control* 1963;**24**:774–80.
43. Chen T, Guestrin C. XGBoost: a scalable tree boosting system. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. San Francisco, CA, USA: ACM, 2016; 785–94.
44. Breiman L. Random Forests. *Mach Learn* 2001;**45**:5–32.
45. Ke G, Meng Q, Finley T, *et al.* LightGBM: a highly efficient gradient boosting decision tree. *Adv Neural Inf Process Syst* 2017;**30**: 3149–57.
46. Wang SC. *Interdisciplinary Computing in Java Programming*. Boston, MA: Springer, 2003, 3–15.
47. Breiman L. Bagging predictors. *Mach Learn* 1996;**24**:123–40.
48. Freund Y, Schapire RE. *A Short Introduction to Boosting*, Vol. **14**. Weinheim: Wiley-VCH Verlag.
49. Duchi J, Hazan E, Singer Y. Adaptive subgradient methods for online learning and stochastic optimization. *J Mach Learn Res* 2011;**12**:2121–59.
50. McCallum A, Nigam K. A comparison of event models for naive Bayes text classification. In: *AAAI Conference on Artificial Intelligence*. Madison, Wisconsin: AAAI, 1998; 41–48.
51. Friedman JH. Greedy function approximation: a gradient boosting machine. *Ann Stat* 2001;**29**:1189–232.
52. Kim Y. Convolutional neural networks for sentence classification. In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Doha, Qatar: ACL, 2014; 1746–51.
53. Yin W, Ebert S, Schütze H. Attention-based convolutional neural network for machine comprehension. In: *Proceedings of the Workshop on Human-Computer Question Answering*. San Diego, California: ACL, 2016; 15–21.
54. Hochreiter S, Schmidhuber J. Long short-term memory. *Neural Comput* 1997;**9**:1735–80.
55. Brahma S. Improved sentence modeling using suffix bidirectional LSTM. arXiv: Learning, 2018. https://arXiv.org/1805.07340.
56. Liou C-Y, Cheng W-C, Liou J-W, *et al.* Autoencoder for words. *Neurocomputing* 2014;**139**:84–96.
57. Jiang G, Wang W. Error estimation based on variance analysis of k-fold cross-validation. *Pattern Recogn* 2017;**69**:94–106.

58. Burjoski V, Reddy ASN. The landscape of RNA-protein interactions in plants: approaches and current status. *Int J Mol Sci* 2021;**22**:2845.
59. Haroon M, Afzal R, Zafar MM, *et al.* Ribonomics approaches to identify RBPome in plants and other eukaryotes: current progress and future prospects. *Int J Mol Sci* 2022;**23**:5923.
60. Sun X, Jin T, Chen C, *et al.* RBPro-RF: use Chou's 5-steps rule to predict RNA-binding proteins via Random Forest with elastic net. *Chemom Intel Lab Syst* 2020;**197**:103919.
61. Mishra A, Khanal R, Kabir WU, *et al.* AIRBP: accurate identification of RNA-binding proteins using machine learning techniques. *Artif Intell Med* 2021;**113**:102034.
62. Wei Q, Zhang Q, Gao H, *et al.* DEEPStack-RBP: accurate identification of RNA-binding proteins based on autoencoder feature selection and deep stacking ensemble classifier. *Knowl Based Syst* 2022;**256**:109875.
63. Bentéjac C, Csörgő A, Martínez-Muñoz G. A comparative analysis of gradient boosting algorithms. *Artif Intell Rev* 2021;**54**:1937–67.
64. Sharma A, Singh B. AE-LGBM: sequence-based novel approach to detect interacting protein pairs via ensemble of autoencoder and LightGBM. *Comput Biol Med* 2020;**125**:103964.
65. Song J, Liu G, Jiang J, *et al.* Prediction of protein-ATP binding residues based on ensemble of deep convolutional neural networks and LightGBM algorithm. *Int J Mol Sci* 2021; **22**:939.
66. Deng L, Pan J, Xu X, *et al.* PDRLGB: precise DNA-binding residue prediction using a light gradient boosting machine. *BMC Bioinform* 2018;**19**:522.