OXFORD

Briefings in Bioinformatics, 2023, **24(1)**, 1–11
https://doi.org/10.1093/bib/bbac538
Problem Solving Protocol

Prediction of RNA-interacting residues in a protein using CNN and evolutionary profile

Sumeet Patiyal, Anjali Dhall, Khushboo Bajaj, Harshita Sahu and Gajendra P.S. Raghava

Corresponding author. G.P.S. Raghava, Department of Computational Biology, Indraprastha Institute of Information Technology, Delhi, Okhla Industrial Estate, Phase III, (Near Govind Puri Metro Station), New Delhi, 110020, India. Office: A-302 (R&D Block). Tel.: +91-11-26907444; Fax: 911126907420; E-mail: raghava@iiitd.ac.in; Website: http://webs.iiitd.edu.in/raghava/

Abstract

This paper describes a method Pprint2, which is an improved version of Pprint developed for predicting RNA-interacting residues in a protein. Training and independent/validation datasets used in this study comprises of 545 and 161 non-redundant RNA-binding proteins, respectively. All models were trained on training dataset and evaluated on the validation dataset. The preliminary analysis reveals that positively charged amino acids such as H, R and K, are more prominent in the RNA-interacting residues. Initially, machine learning based models have been developed using binary profile and obtain maximum area under curve (AUC) 0.68 on validation dataset. The performance of this model improved significantly from AUC 0.68 to 0.76, when evolutionary profile is used instead of binary profile. The performance of our evolutionary profile-based model improved further from AUC 0.76 to 0.82, when convolutional neural network has been used for developing model. Our final model based on convolutional neural network using evolutionary information achieved AUC 0.82 with Matthews correlation coefficient of 0.49 on the validation dataset. Our best model outperforms existing methods when evaluated on the independent/validation dataset. A user-friendly standalone software and web-based server named 'Pprint2' has been developed for predicting RNA-interacting residues (https://webs.iiitd.edu.in/raghava/pprint2 and https://github.com/raghavagps/pprint2).

Keywords: RNA-interacting residues, binary profile, evolutionary profile, convolutional neural network, machine learning techniques

Introduction

Proteins and RNA are the most crucial biological components of life, whereas RNA forms the essential part of ribosome, spliceosome and performs diverse roles within cell [1]. The RNA-protein interactions are necessary for several biological functions such as gene expression regulation, viral assembly & replication, posttranscriptional modification and protein synthesis [2–7]. Recent studies reveal that RNA-protein interactions shows major involvement in developing human cancers and neurological disorders such as amyotrophic lateral sclerosis and Alzheimer's [8–12]. These interactions also play a very crucial role in various infectious and genetic disorders [13–15]. In order to understand the functions and mechanisms of any biological process, it is necessary to get information regarding the RNA-protein interaction residues. In addition to that, RNA-protein interactions play major role in cellular homeostasis and malfunctioning in these interactions may lead to abnormal cellular functions and diseases [16–19]. Therefore, identification of the RNA-interacting residues can help with biotechnological manipulation. With the better understanding of the RNA-protein interacting residues one can design the RNA-based therapy to treat several RNA associated disorders [20–23]. The advancements in experimental techniques such as X-ray crystallography and nuclear magnetic resonance (NMR) several structures of protein-RNA-interacting residues have been discovered and reported in Protein Data Bank [24]. However these experimental methods are very cost expensive and time-consuming. Whereas, computational approaches based on sequence information are very viable and cost-efficient in the probable detection of RNA-binding residues. Identification of RNA-interacting residues/sites in proteins is a highly significant and complex processes in molecular biology, and it has drawn many researchers to work on it for several years in order to provide the best of the best research in this field.

In the last few years a wide range of computational algorithms has been developed for the prediction and identification of RNA-interacting proteins and residues. Broadly, these methods can be divided into two categories, sequence-based and structure-based methods [25–31]. In case of structure based methods, RNA-interacting residues are identified from structure of RNA-protein complex. Unfortunately, due to limitation of experimental

Sumeet Patiyal is currently working as PhD in Computational Biology from Department of Computational Biology, Indraprastha Institute of Information Technology, New Delhi, India.
Anjali Dhall is currently working as PhD in Computational Biology from Department of Computational Biology, Indraprastha Institute of Information Technology, New Delhi, India.
Khushboo Bajaj is currently working as MTech in Computer Science and Engineering from Department of Computer Science and Engineering, Indraprastha Institute of Information Technology, New Delhi, India.
Harshita Sahu is currently working as MTech in Computer Science and Engineering from Department of Computer Science and Engineering, Indraprastha Institute of Information Technology, New Delhi, India.
Gajendra P.S. Raghava is currently working as professor and Head of Department of Computational Biology, Indraprastha Institute of Information Technology, New Delhi, India.
Received: June 8, 2022. **Revised:** September 28, 2022. **Accepted:** November 8, 2022

techniques it is not possible to determine structure of RNA-binding protein. In the last two decades; numerous sequence based methods have been developed where RNA-interacting residues are identified from sequence only. Kumar *et al.* [32] developed a sequence based model using machine learning techniques and evolutionary information. Song *et al.* [33] developed two approaches for predicting RNA-protein interaction residues in protein sequences; first is sequence-based method and second is feature-based method. In addition, PredRBR integrated huge number of sequence, structure-based features for the prediction of protein-RNA-binding affinity [34]. Another computational method named RPiRLS [35] predicts RNA-protein interactions using the structural information. Recently developed methods such as ProNA2020 and hybridNAP predict using sequence information [36]. List of all the available tools is provided in Table 1.

In this study, we have made a systematic attempt to construct machine learning and deep learning based models using the sequence information. We compute physiochemical properties, binary profile and position-specific scoring matrix or evolutionary profile-based features. In order to develop prediction models, we used a wide range of machine learning algorithms such as decision tree, extra-tree classifier, random forest, naive Bayes, gradient boosting, logistic regression and k-nearest neighbor. In addition, models have been developed using deep learning particularly convolutional neural network. In order to provide uniform benchmarking, all existing methods have been evaluated on an independent or validation dataset. We have integrated our best models in the webserver and standalone package for the prediction of RNA-interacting or non-interacting residues in a protein sequence. It has been shown that our convolutional neural network based model using evolutionary profile better than existing methods.

Material and methods

Dataset collection and pre-processing

In the current study, we have compiled the datasets from the recently published paper HybridNAP [48] and ProNA2020 [36], which consist of 1057 and 360 protein sequences that are annotated as RNA-binding proteins. In HybridNAP, the proteins were collected from BioLiP database [52] that provides annotations of biologically relevant protein–ligand interactions for complexes extracted from Protein Data Bank (PDB). Whereas, in ProNA2020, RNA-protein-binding dataset was collected from the Protein-RNA Interface Database (PRIDB) [53], which is a comprehensive repository of protein-RNA interfaces extracted from the protein-RNA complexes available in the PDB. We have used CD-HIT software with the standards of 30% sequence identity to take care the redundancy. In order to compare the performance of the proposed method with HybridNAP and ProNA2020, we have used their training dataset to train the model and independent/validation dataset to evaluate the final model, after removing the redundancy by implementing CD-HIT software. No protein in validation dataset has >30% similarity with any protein in training dataset. Eventually, we were left with 545 (18 559 RNA-interacting and 171 879 non-interacting residues) protein sequences in the training dataset and 161 (6966 RNA-interacting and 44 349 non-interacting residues) sequences in the validation dataset. After that, we generated the overlapping patterns for each sequence and with length 17, as done in previous studies [32, 54]. The central i.e. 9th residue was taken as the representative of whole pattern, such as if the 9th residue is RNA-interacting the pattern is assigned as RNA-interacting otherwise non-interacting. To handle the terminal residues, eight 'X' residues were added to both terminals of the sequences before generating the patterns, so that each residue gets the chance to be the central residue. Figure 1 represents the complete workflow adapted in this study.

Amino acid composition

Amino acid composition (AAC) of residues in a protein sequence was computed using Pfeature standalone version [55]. For AAC, it calculates the percent composition and generates a fixed length vector of 20 amino acid residues as shown in Equation (1).

$$AAC_i = \frac{R_i}{L} \tag{1}$$

where AAC_i is amino acid composition of residue type i; R_i and L number of residues of type i and length of sequence.

Feature generation

Binary profile

The binary profile is generated with the help of Pfeature [55], here we computed binary profile. Where, each amino acid residue represented with a fixed vector size i.e. 21, for example A is represented as 1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0; in which 20 are natural amino acids and one is for dummy variable, whereas X is denoted as 0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1. Hence, each pattern is denoted by a fixed length vector of size 357 (17 * 21).

Physicochemical properties profile

We have implemented Pfeature [55] to calculate physicochemical properties profile using binary profile module. In this approach, each amino acid is represented with a vector of size 25, where each element denotes the particular physicochemical property, such as amino acid 'A' is represented as 0,0,1,0,1,1,0,0,0,0,1,1,0,0,0,0,1,0,0,1,0,0,1,1,0; where 'X' is denoted by zero vector of length 25. '1' denotes the presence and '0' denoted the absence of a particular property. Since, the size of each pattern is 17, hence the resulting length for each vector is 425 (17 * 25).

Evolutionary profile

In order to compute the evolutionary information of residues we generate Position-Specific Scoring Matrix (PSSM) profiles. The PSSM profile was created using PSI-BLAST, which searches each sequence against the Swiss-Prot database. PSI-BLAST was conducted with three iterations along with an e-value of $1e^{-3}$. As shown in Equation (2), we further normalized the profiles, where each pattern is represented as a vector of length 357 in the final matrix for each sequence, which is of dimension $N \times 21$, where N is the length of the protein sequence.

$$\mathrm{PSSM}_N = \frac{1}{1+e^{-x}} \tag{2}$$

where, x is the PSSM score and PSSM_N is the normalized value.

Model building

In order to classify RNA-interacting and non-interaction residues in a protein sequence, we have implemented various classifiers. These machine learning classifiers implemented in python library Scikit-learn [56]. To develop prediction models, we implement various classifiers such as decision tree (DT), random forest (RF),

Table 1. List of computational resources for the prediction of RNA-interacting residues

Name	Year	Description (Weblink)	Working status
BindN [37]	2006	SVM based method for predicting DNA and RNA binding sites (http://bioinformatics.ksu.edu/bindn/)	No
RNABindR [38]	2007	Distance cut-off based prediction (http://bindr.gdcb.iastate.edu/RNABindR)	Yes
Pprint [32]	2008	SVM and PSSM profile-based prediction method (https://webs.iiitd.edu.in/raghava/pprint/)	Yes
PRINTR [39]	2008	SVM and PSSM profile-based prediction method (http://210.42.106.80/printr/)	No
BindN+ [40]	2010	SVM based DNA or RNA-binding site prediction using PSSM profile (http://bioinfo.ggc.org/bindn+/)	No
PRBR [41]	2011	RF based model developed using hybrid feature (http://www.cbi.seu.edu.cn/PRBR/)	No
RPISeq [42]	2011	Sequence information prediction method for RNA-protein interaction (http://pridb.gdcb.iastate.edu/RPISeq/)	Yes
RNABindRPlus [43]	2014	Machine Learning and Sequence Homology-Based Methods (http://ailab-projects2.ist.psu.edu/RNABindRPlus/)	Yes
DR_bind [44]	2014	Evolutionary conserved structural and energetic features based prediction (https://drbind.limlab.ibms.sinica.edu.tw/)	Yes
RBScore & Nbench [45]	2015	Scoring scheme based linking of feature values with nucleic acid-binding probabilities (http://ahsoka.u-strasbg.fr/rbscorenbench/)	No
SNBRFinder [46]	2015	Prediction of nucleic acids binding residues using hybrid algorithm (http://ibi.hzau.edu.cn/SNBRFinder)	No
PredRBR [34]	2016	Structure based prediction method (http://denglab.org/PredRBR/)	Yes
DRNAPred [47]	2017	Fast sequence-based method that accurately predicts RNA-interacting residues (http://biomine.cs.vcu.edu/servers/DRNApred/)	Yes
HybridNAP [48]	2017	RNA-binding residue prediction method (http://biomine.cs.vcu.edu/servers/hybridNAP/)	Yes
RPI-Bind [29]	2017	Structure-based method for accurate identification of RNA-protein-binding sites (http://ctsb.is.wfubmc.edu/publications/RPI-Bind-Pred.php)	No
iDeepS [27]	2018	Deep convolutional and neural networks based prediction method (https://github.com/xypan1232/iDeepS)	Yes
RPiRLS [35]	2018	Quantitative matrix based predictions of RNA interaction (http://bmc.med.stu.edu.cn/RPiRLS)	No
SVMnuc & NucBind [49]	2019	Support vector machine-based ab-initio method (https://yanglab.nankai.edu.cn/NucBind/)	Yes
PRIME-3D2D [31]	2020	Structure based RNA-binding residue prediction method (http://www.rnabinding.com/PRIME-3D2D/)	Yes
ProNA2020 [36]	2020	Standard neural networks based method (www.predictprotein.org)	Yes
NCBRPred [50]	2021	Multi-label learning framework method (http://bliulab.net/NCBRPred/)	Yes
PST-PRNA [30]	2022	Protein surface topology method (http://www.zpliulab.cn/PSTPRNA)	Yes
iDRNA-ITF [51]	2022	Induction & transfer framework based method (http://bliulab.net/iDRNA-ITF/)	Yes

logistic regression (LR), eXtreme gradient boosting (XGB), Gaussian Naive Bayes (GNB), extra-tree classifier (ET), K-nearest neighbor and one-dimensional convolutional neural network (1D-CNN).

Five-fold cross-validation

To train, test and validate the prediction models, we employed 5-fold cross-validation and an external validation as implemented in previous studies [57, 58]. Five-fold cross-validation was employed only on the training dataset. In which the data were split into five parts, where four of which were used to train the model and the fifth set was utilized for testing. This process is iterated five times, in which each set is used in training and testing the models. We have hyper-tuned the parameters using grid-search approach for traditional machine learning and deep-learning technique. For 1D-CNN model, the parameters were hyper-tuned at three level: layers (filters in each convolution layer, convolution filters' size, number of dense layers and number of neurons in each layer); functions (optimizers, loss function and activation function) and rates (learning and dropout).

Performance evaluation

In this work, we examined the performance of the model using sensitivity, specificity, $f1$ score, accuracy, area under the receiver operating characteristic (AUROC) and Matthews correlation coefficient (MCC). These parameters belong to threshold dependent (i.e. sensitivity, specificity, accuracy and MCC) and independent category (AUROC) to evaluate our models. These parameters were

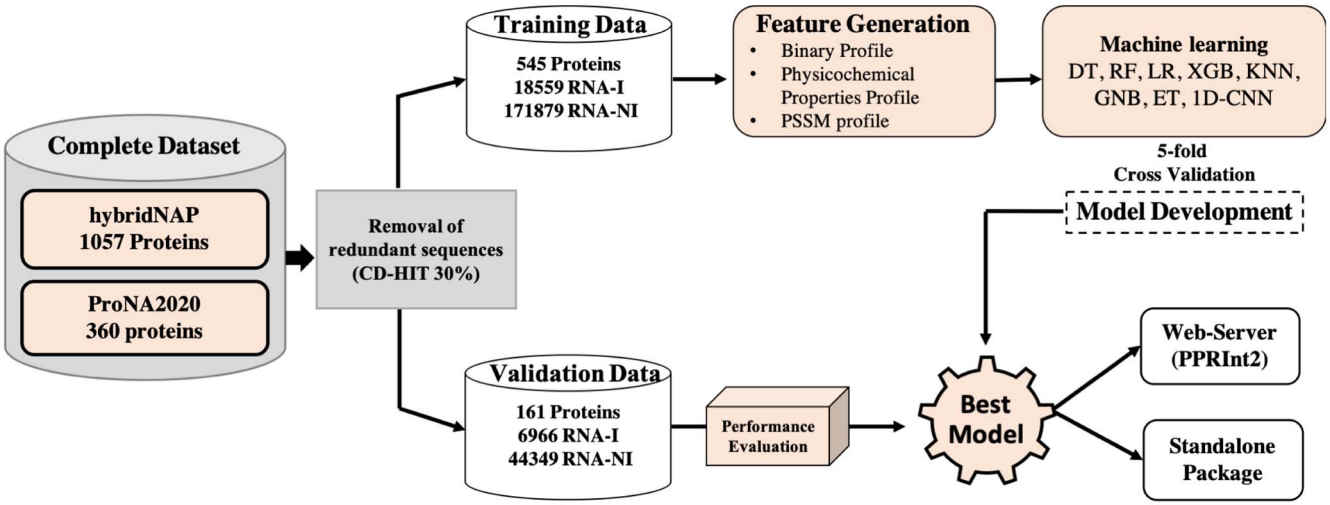
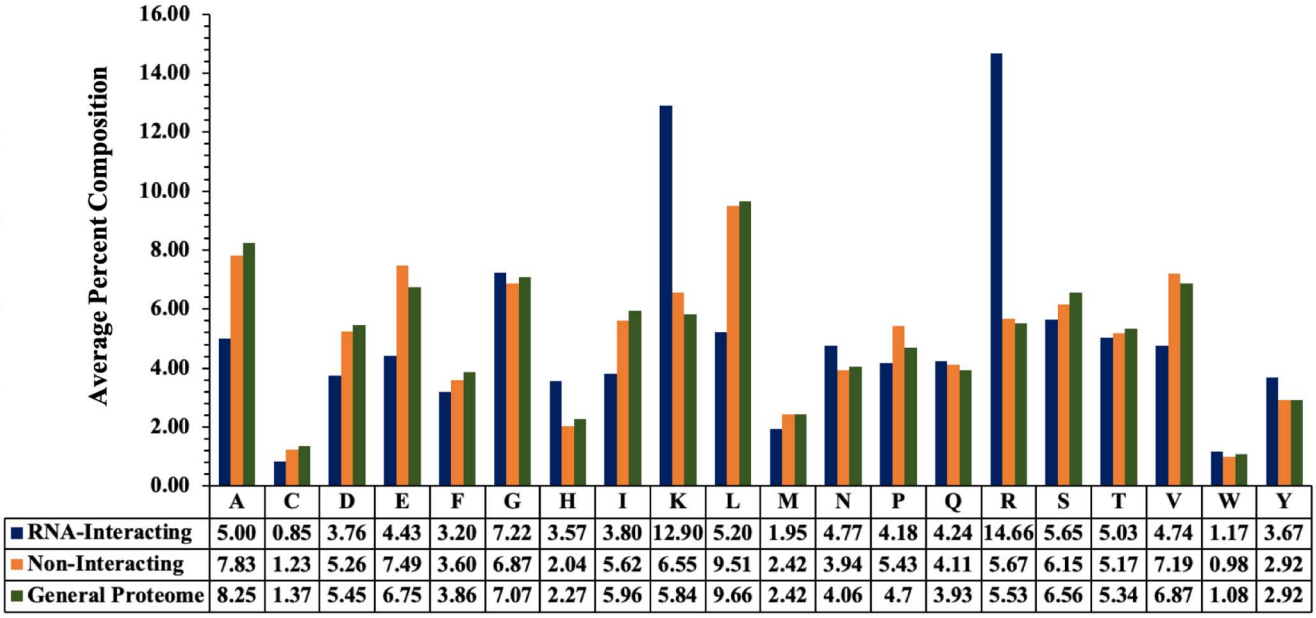


Figure 1. Complete workflow of the study including the dataset creation, feature generation, model development and evaluation.



	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
RNA-Interacting	5.00	0.85	3.76	4.43	3.20	7.22	3.57	3.80	12.90	5.20	1.95	4.77	4.18	4.24	14.66	5.65	5.03	4.74	1.17	3.67
Non-Interacting	7.83	1.23	5.26	7.49	3.60	6.87	2.04	5.62	6.55	9.51	2.42	3.94	5.43	4.11	5.67	6.15	5.17	7.19	0.98	2.92
General Proteome	8.25	1.37	5.45	6.75	3.86	7.07	2.27	5.96	5.84	9.66	2.42	4.06	4.7	3.93	5.53	6.56	5.34	6.87	1.08	2.92

Figure 2. Average percent composition of each residue in RNA-interacting, non-interacting residues and general proteome.

computed using the following Equations (3–7).

$$\text{Sensitivity} = \frac{TP}{TP + FN} * 100 \quad (3)$$

$$\text{Specificity} = \frac{TN}{TN + FP} * 100 \quad (4)$$

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + TN + FN} * 100 \quad (5)$$

$$F1\text{-score} = \frac{2TP}{FP + FN} \quad (6)$$

$$\text{MCC} = \frac{(TP * TN) - (FP * FN)}{\sqrt{(TP + FP)\,(TP + FN)\,(TN + FP)\,(TN + FN)}} \quad (7)$$

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

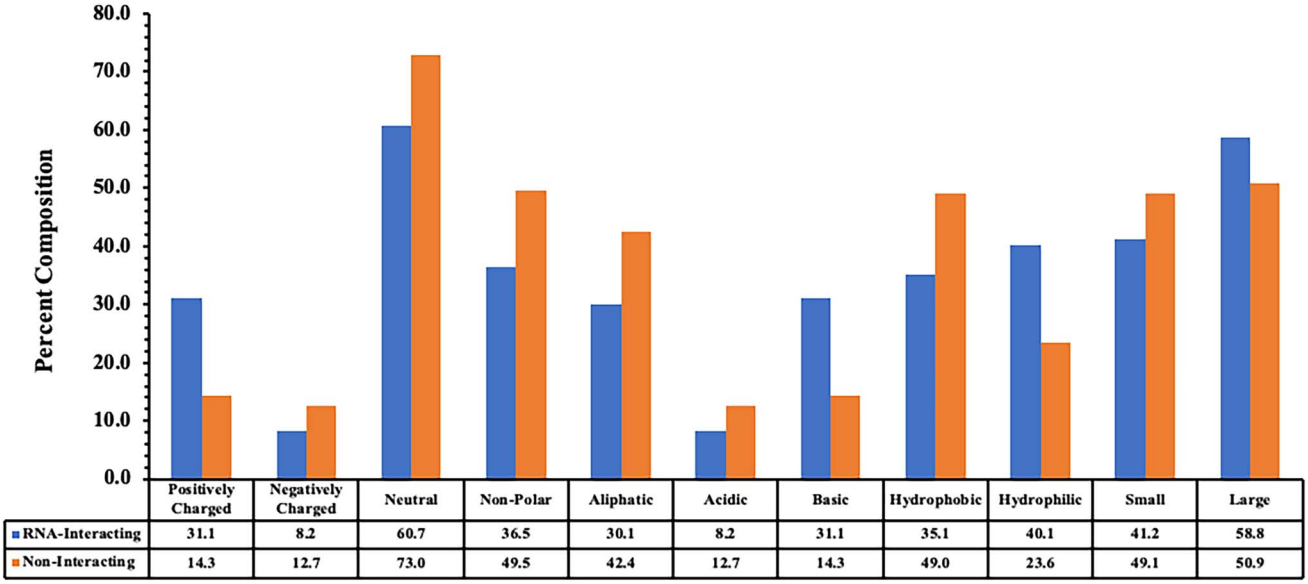
Results

Compositional analysis

In order to understand the distribution of amino acid residues in the RNA-interacting versus non-interacting versus general proteome, we have calculated the percent amino acid composition. For percent amino acid composition of general proteome, we have used the values from the Swiss-Prot database available at https://web.expasy.org/protscale/pscale/A.A.Swiss-Prot.html. As exhibited by Figure 2, positively charged residue H, K and R are abundant in RNA-interacting residues as compare to non-interacting or general proteome, where A, D, E, I, L and V are rich in non-interacting and general proteome.

Physiochemical property analysis

In order to explore the nature of amino acids involved in the RNA-interaction, we computed physiochemical properties-based percent composition of residues involved in RNA-interaction.



	Positively Charged	Negatively Charged	Neutral	Non-Polar	Aliphatic	Acidic	Basic	Hydrophobic	Hydrophilic	Small	Large
RNA-Interacting	31.1	8.2	60.7	36.5	30.1	8.2	31.1	35.1	40.1	41.2	58.8
Non-Interacting	14.3	12.7	73.0	49.5	42.4	12.7	14.3	49.0	23.6	49.1	50.9

Figure 3. Percent composition based on physicochemical properties of residues involved in RNA-interaction.

Table 2. The performance binary profile based on models developed using different classifiers

Classifier	Training dataset							Validation dataset						
	Sens	Spec	Acc	AUC	F1	K	MCC	Sens	Spec	Acc	AUC	F1	K	MCC
DT	19.14	88.93	82.13	0.54	0.17	0.07	0.07	15.94	89.07	79.14	0.53	0.17	0.05	0.05
RF	66.57	67.53	67.43	0.73	0.29	0.16	0.21	55.77	68.54	66.81	0.67	0.31	0.15	0.18
LR	67.20	68.16	68.07	0.74	0.29	0.16	0.22	55.79	70.37	68.39	0.68	0.32	0.16	0.19
XGB	66.88	67.54	67.48	0.74	0.29	0.16	0.21	55.86	69.65	67.78	0.68	0.32	0.16	0.19
KNN	60.32	63.75	63.41	0.64	0.24	0.10	0.15	52.60	64.66	63.02	0.60	0.28	0.10	0.12
GNB	66.63	65.14	65.28	0.71	0.27	0.14	0.19	56.30	67.11	65.64	0.65	0.31	0.14	0.17
ET	68.25	65.72	65.97	0.73	0.28	0.15	0.21	58.38	66.64	65.52	0.67	0.32	0.15	0.18
1D-CNN	73.91	73.76	73.78	0.81	0.36	0.24	0.31	50.66	74.63	71.38	0.68	0.33	0.17	0.19

DT: decision tree; RF: random forest; LR: logistic regression; XGB: eXtreme gradient boosting; KNN: K-nearest neighbour; GNB: Gaussian Naïve Bayes; ET: extra tree; 1D-CNN: one-dimensional convolutional neural network; Sens: sensitivity; Spec: specificity; Acc: accuracy; AUC: area under the receiver operating characteristic curve; K: Kappa; MCC: Matthews correlation coefficient.

As shown by Figure 3, positively charged, basic and hydrophilic residues are abundant in the RNA-binding sites as RNA has a negative backbone. On the other hand, negatively charged, neutral charged, acidic, hydrophobic, small, non-polar and aliphatic residues are scare in RNA-interacting sites.

Two sample logo

In the past, number of studies have shown that the interaction property of a residue is influenced by its neighboring residues. In this study, we have created the patterns of length 17, where central i.e. 9th residue is the representative of whole pattern and based on that a pattern is assigned as interacting or non-interacting. Figure 4 represents the preference of residues as each position in RNA-interacting and non-interacting patterns, and it shows that residue R is highly preferred at interacting site followed by K and H, flanked by positively charged residues R and K. On the contrary, non-interacting sites are preferred by residue L, followed by E, A and V, where these residues are flanked by residues E and L.

Performance of binary profile

We have used binary profile as input features to train and evaluate the prediction models by implementing various machine learning classifiers. We used five-fold cross-validation technique to build prediction models on training datasets. As shown in Table 2, in case of machine learning techniques both logistic regression and XGB achieve maximum AUC 0.74. In case of deep learning, 1D-CNN based model obtained AUC 0.81, which is highest. In order obtain unbiased performance of models we evaluate these models on validation/independent dataset. In case of machine learning techniques, we achieved maximum AUC 0.68 for both logistic regression and XGB based models. In case of 1D-CNN performance of model decrease drastically from 0.81 to 0.68 when tested on validation/independent dataset instead of training dataset. In summary, both 1D-CNN and machine learning based models achieve same performance on validation dataset.

Performance of physicochemical properties profile

In addition, models have been developed using physicochemical properties profile, which represents each amino acid with vector of length 25. In case of machine learning techniques, we got maximum AUC 0.74 and 0.68 on training and independent/validation dataset, respectively (Table 3).

In case of deep learning 1D-CNN, we achieved maximum AUC 0.79 and 0.68 on training and validation dataset, respectively.

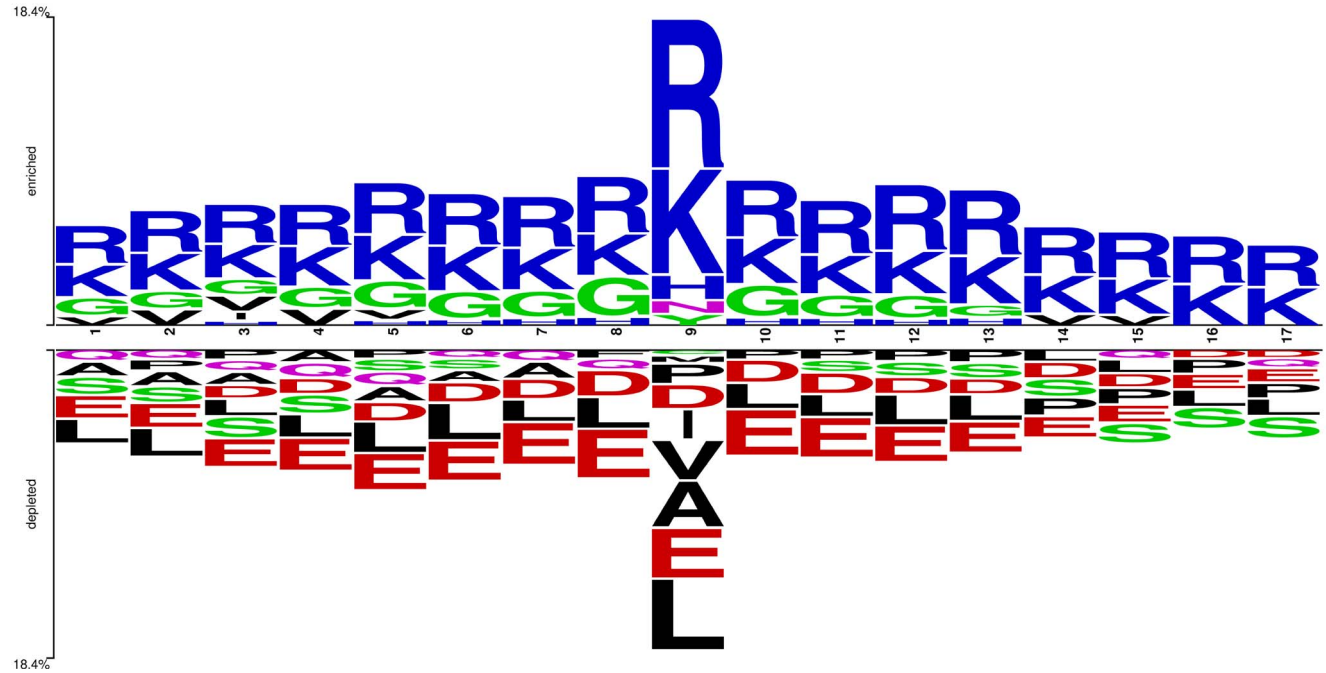


Figure 4. Two sample logo exhibits the preference of residues at each position in RNA-interacting and non-interacting patterns.

Table 3. Performance measures for models developed by implementing various classifiers using physicochemical properties profile as the input feature

Classifier	Training dataset							Validation dataset						
	Sens	Spec	Acc	AUC	F1	K	MCC	Sens	Spec	Acc	AUC	F1	K	MCC
DT	16.15	91.03	83.73	0.54	0.16	0.07	0.07	14.00	91.46	80.95	0.53	0.17	0.06	0.06
RF	67.74	64.56	64.87	0.72	0.27	0.14	0.20	56.88	66.37	65.08	0.66	0.31	0.14	0.17
LR	67.30	68.19	68.10	0.74	0.29	0.16	0.22	55.70	70.48	68.47	0.68	0.32	0.16	0.19
XGB	65.41	68.60	68.29	0.73	0.29	0.16	0.21	54.22	70.50	68.29	0.68	0.32	0.16	0.18
KN	62.43	63.32	63.23	0.65	0.25	0.11	0.16	53.22	63.81	62.37	0.60	0.28	0.10	0.12
GNB	65.44	66.15	66.08	0.71	0.27	0.14	0.19	54.19	68.58	66.63	0.66	0.31	0.14	0.16
ET	65.81	67.25	67.11	0.72	0.28	0.15	0.20	54.65	68.74	66.83	0.66	0.31	0.14	0.17
1D-CNN	69.84	69.60	69.62	0.77	0.31	0.19	0.25	56.42	71.14	69.14	0.69	0.33	0.17	0.20

DT: decision tree; RF: random forest; LR: logistic regression; XGB: eXtreme gradient boosting; KNN: K-nearest neighbour; GNB: Gaussian Naïve Bayes; ET: extra tree; 1D-CNN: one-dimensional convolutional neural network; Sens: sensitivity; Spec: specificity; Acc: accuracy; AUC: area under the receiver operating characteristic curve; K: Kappa; MCC: Matthews correlation coefficient.

Table 4. Performance of various classifiers using PSSM profile as input feature for training and validation dataset

Classifier	Training dataset							Validation dataset						
	Sens	Spec	Acc	AUC	F1	K	MCC	Sens	Spec	Acc	AUC	F1	K	MCC
DT	30.54 ± 1.56	92.02 ± 0.21	86.03 ± 0.31	0.61 ± 0.008	0.29 ± 0.018	0.22 ± 0.016	0.22 ± 0.016	22.08	92.42	82.87	0.57	0.26	0.17	0.17
RF	74.35 ± 1.79	70.63 ± 1.52	71.01 ± 1.20	0.80 ± 0.008	0.33 ± 0.012	0.21 ± 0.009	0.28 ± 0.009	62.26	70.90	69.73	0.73	0.36	0.20	0.24
LR	71.93 ± 1.69	71.63 ± 0.47	71.66 ± 0.29	0.79 ± 0.007	0.33 ± 0.14	0.21 ± 0.011	0.28 ± 0.013	61.71	74.12	72.44	0.75	0.38	0.23	0.27
XGB	74.23 ± 1.33	73.65 ± 0.61	73.71 ± 0.42	0.82 ± 0.006	0.36 ± 0.012	0.24 ± 0.008	0.31 ± 0.009	61.47	75.83	73.88	0.76	0.39	0.25	0.28
KN	74.19 ± 1.91	66.17 ± 0.79	66.96 ± 0.64	0.75 ± 0.012	0.30 ± 0.015	0.18 ± 0.012	0.25 ± 0.014	64.36	66.71	66.39	0.69	0.34	0.18	0.22
GNB	67.64 ± 1.70	68.82 ± 1.06	68.70 ± 0.80	0.75 ± 0.007	0.30 ± 0.010	0.17 ± 0.005	0.23 ± 0.007	54.78	73.28	70.77	0.70	0.34	0.18	0.21
ET	74.28 ± 1.87	70.24 ± 1.79	70.64 ± 1.44	0.80 ± 0.007	0.33 ± 0.012	0.21 ± 0.010	0.28 ± 0.009	62.82	70.12	69.13	0.72	0.36	0.20	0.24
1D-CNN	83.08 ± 1.29	83.37 ± 1.34	83.34 ± 1.36	0.91 ± 0.008	0.49 ± 0.025	0.41 ± 0.023	0.47 ± 0.021	80.19	82.35	82.05	0.82	0.55	0.45	0.49

DT: decision tree; RF: random forest; LR: logistic regression; XGB: eXtreme gradient boosting; KNN: K-nearest neighbour; GNB: Gaussian Naïve Bayes; ET: extra tree; 1D-CNN: one-dimensional convolutional neural network; Sens: sensitivity; Spec: specificity; Acc: accuracy; AUC: area under the receiver operating characteristic curve; K: Kappa; MCC: Matthews correlation coefficient.

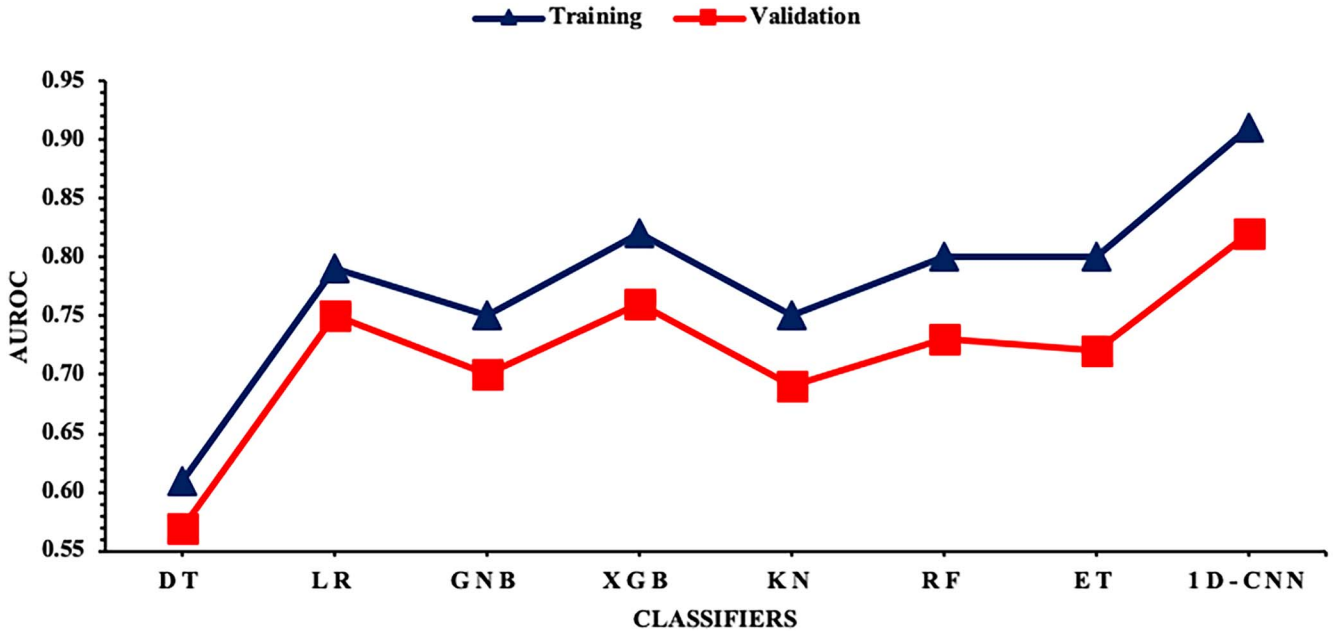


Figure 5. Difference in the AUROC of various models developed using PSSM profile.

These results indicate that models based on physicochemical properties and binary profile have nearly same precision.

Performance of evolutionary profile

It has been shown in the previous studies that evolutionary information provides more information then single sequence. Thus, in this study PSSM profiles were generated for proteins to capture evolutionary information. Machine learning based models have been developed using PSSM profile for predicting RNA-interacting residues. As shown in Table 4, almost all classifies developed using PSSM profile perform better then binary profile-based models. In case of machine learning, XGB based model obtain AUC 0.82 and 0.76 on training and independent/validation dataset, respectively. It means the performance increases from 0.68 to 0.76 on the validation dataset, when we used PSSM profile in place of binary profile. It shows importance of PSSM profile in predicting RNA-interacting residues. Our 1D-CNN based model obtained AUC 0.91 and 0.82 on training and validation datasets respectively. In summary, 1D-CNN achieved maximum AUC 0.82 on validation dataset.

We have also compared the difference between the AUC of models built using PSSM profile, to understand the overfitting in the models by comparing their performances in training and validation dataset as shown in Figure 5. The difference in the AUC for training and validation dataset showed that DT model is least overfitted, whereas the performance of the deep-learning model exhibits that the resulting model is quite overfitted as compare to the other models.

Comparison with existing methods

In order to justify a newly developed method, it is essential to compare its performance with the existing methods. The existing methods have been trained and evaluated on different datasets as these methods have been developed over the years. In order to provide unbiased evaluation of methods, one should evaluate performance of all methods on a common dataset. In this paper, evaluate the performance of the existing methods on independent/validation dataset. The performance measures

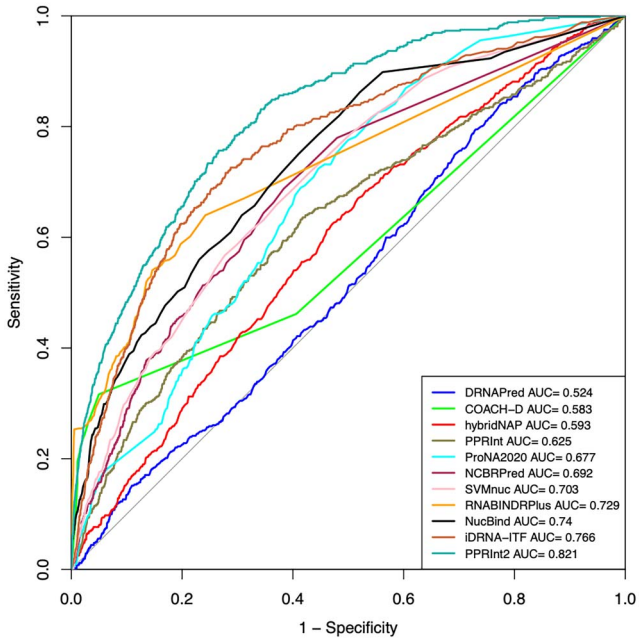


Figure 6. ROC curve shows the performance of all methods on validation dataset in term of area under curve (AUC).

taken into consideration were sensitivity, specificity, AUC, accuracy, F1, Kappa and MCC. As shown in Table 5 and Figure 6, Pprint2 perform better than existing methods with AUC of 0.82, whereas iDRNA-ITF performed second best among the existing methods with AUC 0.77 (See Table 5).

Service to scientific community

In order to facilitate scientific community, we developed a web server and standalone package called Pprint2, which is available at https://webs.iiitd.edu.in/raghava/pprint2/. Our web server allows the users to predict RNA-interacting residues in a protein sequence pasted/uploaded by user. It allows users to select any

Table 5. Performance of existing methods on the validation dataset

Methods	Sensitivity	Specificity	Accuracy	AUC	F1	K	MCC
DRNAPred	45.40	51.87	51.59	0.52	0.08	0.00	0.01
COACH-D	46.17	59.37	58.80	0.58	0.09	0.01	0.02
HybridNAP	56.90	56.05	56.09	0.59	0.10	0.02	0.05
Pprint	60.15	60.29	60.28	0.63	0.12	0.04	0.08
ProNA2020	62.07	62.44	62.43	0.68	0.13	0.05	0.10
NCBRPred	37.93	86.03	83.95	0.69	0.17	0.11	0.14
SVMnuc	63.03	65.39	65.29	0.70	0.14	0.06	0.12
RNABindR-Plus	67.24	68.36	68.31	0.73	0.16	0.09	0.15
NucBind	65.71	66.92	66.87	0.74	0.15	0.08	0.14
iDRNA-ITF	71.65	71.83	71.82	0.77	0.18	0.11	0.19
Pprint2	80.19	82.34	82.05	0.82	0.55	0.45	0.49

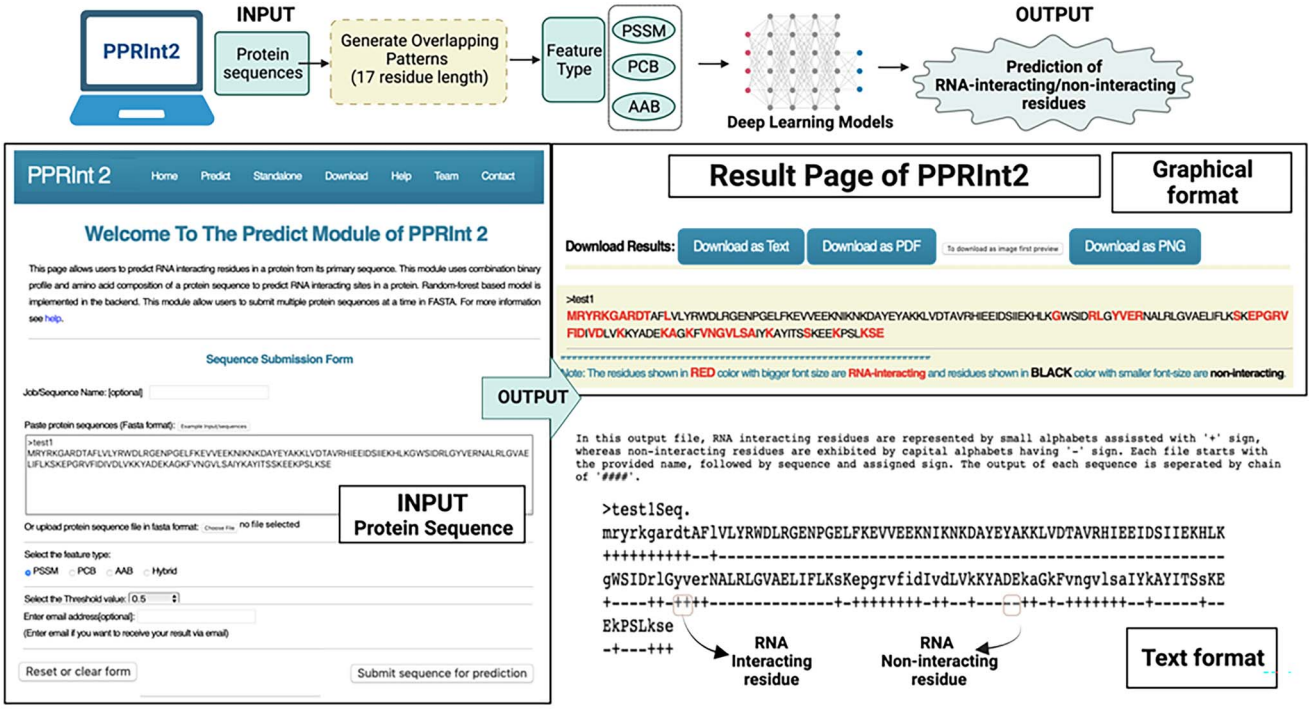


Figure 7. Complete usage of prediction module of webserver.

modules for prediction that include PSSM profile and binary profile. The resulting page displays the input sequence(s) with highlighting the interacting residues in red color as represented in Figure 7. In addition, server provides the facility to download the results in .txt, .pdf and .png format. In addition, we have also provided the python-based standalone package, which is available from https://github.com/raghavagps/pprint2.

Discussion and conclusion

The interaction between RNA and protein complexes is responsible for many fundamental processes such as splicing, translation, transport and silencing [59, 60]. Accurate identification of amino acid residues involving in the interaction with the RNA lead to the better understanding of the sequence specific mechanism of RNA-protein interaction [61, 62]. Correct identification of RNA-interacting residues in a protein is only possible if structure of RNA-protein complexes is available. Unfortunately, crystallization of all RNA-protein complexes is not possible due to number of limitation of experimental techniques like crystallography and NMR. In addition, experimental techniques are costly and time-consuming. In order to facilitate researcher in the field of RNA biology, many methods have been developed for predicting RNA-interacting residues [63–65]. One of the major limitation of previous methods was that they have been trained and evaluated on limited RNA-binding proteins. For example, in our previous method Pprint [32], we trained and test our models on only 86 RNA-binding proteins. Due to lack of sufficient data, a lenient cut-off level 70% has been used for removing redundant proteins. This is true for most of old methods where limited set of RNA-binding proteins were used for training and evaluation. In addition, proteins in dataset contain highs level of similarity with each other. Over the years structures of RNA-protein complexes have been grown drastically in PDB [62, 66]. Thus, there is a need to develop new method on large set of RNA-binding proteins whose structure is available in PDB. In addition, there is a need to create separate dataset for training and validation. In order to avoid any over optimization of machine learning models that proteins in

training and validation should not have minimum similarity. In this study, we make an attempt to train our models on large set of proteins. Initially, we obtained 1057 and 360 protein sequences from recently published articles hybridNAP [67] and proNA2020 [36]. In order to create dataset of redundant RNA-binding protein, we removing the redundant sequences using CD-HIT at 30%. This leads to a dataset of 545 proteins in training and 161 sequences in independent or validation dataset. This is one of the largest dataset used for training and validation. In addition, proteins in training and validation dataset have similarity 30% or less.

As shown in results section, in most of the cases performance of our machine learning based model on training and validation dataset is nearly same. This indicates that our models are not over optimized on training dataset. Our results indicate that performance of models improved significantly using evolutionary information. This support previous studies including our previous method Pprint that evolutionary information is important for prediction. The performance of models based on binary profile and physiochemical property is nearly same. It means over optimization is not a major challenge in case of machine learning techniques if models developed using cross-validation techniques. In case of 1D-CNN (deep learning), we observed high level of over optimization. Most of the 1D-CNN models have high performance on training dataset compare to training dataset. Thus cross-validation techniques are not sufficient for evaluating deep-learning models. These deep-learning models should be evaluated on validation dataset for unbiased evaluation.

Key Points

- Machine learning based models were developed using different profiles.
- PSSM profile of a protein was created to extract evolutionary information.
- PSSM profiles of proteins were generated using PSI-BLAST.
- Convolutional neural network based model was developed using PSSM profile.
- Webserver, Python- and Perl-based standalone package and GitHub is available.

Data availability

All the datasets generated in this study are available at 'Pprint2' web server https://webs.iiitd.edu.in/raghava/pprint2/dataset.php. BioRxiv DOI: https://doi.org/10.1101/2022.06.03.494705

Authors' contributions

SP and GPSR collected and processed the datasets. SP, KP, HS and GPSR implemented the algorithms and SP developed the prediction models. SP, AD and GPSR analysed the results. SP created the back-end of the web server the front-end user interface. SP, AD and GPSR penned the manuscript. GPSR conceived and coordinated the project. All authors have read and approved the final manuscript.

Acknowledgments

Authors are thankful to the Department of Bio-Technology (DBT) and Department of Science and Technology (DST-INSPIRE) for fellowships and the financial support and Department of Computational Biology, IIITD New Delhi for infrastructure and facilities.

Funding

The current work has received grant from the Department of Bio-Technology (DBT), Government of India, India (BT/PR40158/BTIS/137/24/2021).

References

1. Jones S, Daley DT, Luscombe NM, *et al.* Protein-RNA interactions: a structural analysis. *Nucleic Acids Res* 2001;**29**:943–54.
2. Turner M, Diaz-Munoz MD. RNA-binding proteins control gene expression and cell fate in the immune system. *Nat Immunol* 2018;**19**:120–9.
3. Lin B, Pang Z. Stability of methods for differential expression analysis of RNA-seq data. *BMC Genomics* 2019;**20**:35.
4. Pattnaik A, Palermo N, Sahoo BR, *et al.* Discovery of a non-nucleoside RNA polymerase inhibitor for blocking Zika virus replication through in silico screening. *Antiviral Res* 2018;**151**:78–86.
5. Payne JL, Khalid F, Wagner A. RNA-mediated gene regulation is less evolvable than transcriptional regulation. *Proc Natl Acad Sci U S A* 2018;**115**:E3481–90.
6. Standart N, Jackson RJ. Regulation of translation by specific protein/mRNA interactions. *Biochimie* 1994;**76**:867–79.
7. Gangloff S, Soustelle C, Fabre F. Homologous recombination is responsible for cell death in the absence of the Sgs1 and Srs2 helicases. *Nat Genet* 2000;**25**:192–4.
8. Carey KT, Wickramasinghe VO. Regulatory potential of the RNA processing machinery: implications for human disease. *Trends Genet* 2018;**34**:279–90.
9. Kwiatkowski TJ, Jr, Bosco DA, Leclerc AL, *et al.* Mutations in the FUS/TLS gene on chromosome 16 cause familial amyotrophic lateral sclerosis. *Science* 2009;**323**:1205–8.
10. Idda ML, Munk R, Abdelmohsen K, *et al.* Noncoding RNAs in Alzheimer's disease, Wiley Interdiscip Rev. *RNA* 2018;**9**(2):e1463.
11. Tsai MC, Spitale RC, Chang HY. Long intergenic noncoding RNAs: new links in cancer progression. *Cancer Res* 2011;**71**:3–7.
12. Zhou M, Zhao H, Wang X, *et al.* Analysis of long noncoding RNAs highlights region-specific altered expression patterns and diagnostic roles in Alzheimer's disease. *Brief Bioinform* 2019;**20**:598–608.
13. Gebauer F, Schwarzl T, Valcarcel J, *et al.* RNA-binding proteins in human genetic disease. *Nat Rev Genet* 2021;**22**:185–98.
14. Castello A, Fischer B, Hentze MW, *et al.* RNA-binding proteins in Mendelian disease. *Trends Genet* 2013;**29**:318–27.
15. Kapeli K, Martinez FJ, Yeo GW. Genetic mutations in RNA-binding proteins and their roles in ALS. *Hum Genet* 2017;**136**:1193–214.
16. Ramanathan M, Porter DF, Khavari PA. Methods to study RNA-protein interactions. *Nat Methods* 2019;**16**:225–34.
17. Allerson CR, Cazzola M, Rouault TA. Clinical severity and thermodynamic effects of iron-responsive element mutations in hereditary hyperferritinemia-cataract syndrome. *J Biol Chem* 1999;**274**:26439–47.
18. Batista PJ, Chang HY. Long noncoding RNAs: cellular address codes in development and disease. *Cell* 2013;**152**:1298–307.
19. Khalil AM, Rinn JL. RNA-protein interactions in human health and disease. *Semin Cell Dev Biol* 2011;**22**:359–65.
20. Guo P, Coban O, Snead NM, *et al.* Engineering RNA for targeted siRNA delivery and medical application. *Adv Drug Deliv Rev* 2010;**62**:650–66.

21. Schmidt N, Lareau CA, Keshishian H, *et al.* The SARS-CoV-2 RNA-protein interactome in infected human cells. *Nat Microbiol* 2021;**6**:339–53.
22. Yu AM, Choi YH, Tu MJ. RNA drugs and RNA targets for small molecules: principles, progress, and challenges. *Pharmacol Rev* 2020;**72**:862–98.
23. Kolinski M, Kaluzna E, Piwecka M. RNA-protein interactomes as invaluable resources to study RNA viruses: insights from SARS CoV-2 studies. *Wiley Interdiscip Rev RNA* 2022;e1727.
24. Berman HM, Westbrook J, Feng Z, *et al.* The Protein Data Bank. *Nucleic Acids Res* 2000;**28**:235–42.
25. Zhao H, Yang Y, Zhou Y. Structure-based prediction of RNA-binding domains and RNA-binding sites and application to structural genomics targets. *Nucleic Acids Res* 2011;**39**:3017–25.
26. Poursheikhali Asghari M, Abdolmaleki P. Prediction of RNA- and DNA-binding proteins using various machine learning classifiers. *Avicenna J Med Biotechnol* 2019;**11**:104–11.
27. Pan X, Rijnbeek P, Yan J, *et al.* Prediction of RNA-protein sequence and structure binding preferences using deep convolutional and recurrent neural networks. *BMC Genomics* 2018;**19**:511.
28. Sanchez de Groot N, Armaos A, Grana-Montes R, *et al.* RNA structure drives interaction with proteins. *Nat Commun* 2019;**10**:3246.
29. Luo J, Liu L, Venkateswaran S, *et al.* RPI-Bind: a structure-based method for accurate identification of RNA-protein binding sites. *Sci Rep* 2017;**7**:614.
30. Li P, Liu ZP. PST-PRNA: Prediction of RNA-binding sites using protein surface topography and deep learning. *Bioinformatics* 2022;**2162**:2168.
31. Xie J, Zheng J, Hong X, *et al.* PRIME-3D2D is a 3D2D model to predict binding sites of protein-RNA interaction. *Commun Biol* 2020;**3**:384.
32. Kumar M, Gromiha MM, Raghava GP. Prediction of RNA binding sites in a protein using SVM and PSSM profile. *Proteins* 2008;**71**:189–94.
33. Jiazhi Songa GL, Wanga R, Suna L, *et al.* A novel method for predicting RNA-interacting residues in proteins using a combination of feature-based and sequence template-based methods. *Biotechnol Biotechnolog Equip* 2019;**33**(1):1138–49.
34. Deng DLYTCFZCL. *PredRBR: Accurate Prediction of RNA-Binding Residues in proteins using Gradient Tree Boosting*. IEEE International Conference on Bioinformatics and Biomedicine (BIBM), 2016;**47**:52.
35. Shen WJ, Cui W, Chen D, *et al.* RPiRLS: quantitative predictions of RNA interacting with any protein of known sequence. *Molecules* 2018;**23**(3):**540**.
36. Qiu J, Bernhofer M, Heinzinger M, *et al.* ProNA2020 predicts protein-DNA, protein-RNA, and protein-protein binding proteins and residues from sequence. *J Mol Biol* 2020;**432**:2428–43.
37. Wang L, Brown SJ. BindN: a web-based tool for efficient prediction of DNA and RNA binding sites in amino acid sequences. *Nucleic Acids Res* 2006;**34**:W243–8.
38. Terribilini M, Sander JD, Lee JH, *et al.* RNABindR: a server for analyzing and predicting RNA-binding sites in proteins. *Nucleic Acids Res* 2007;**35**:W578–84.
39. Wang Y, Xue Z, Shen G, *et al.* PRINTR: prediction of RNA binding sites in proteins using SVM and profiles. *Amino Acids* 2008;**35**:295–302.
40. Wang L, Huang C, Yang MQ, *et al.* BindN+ for accurate prediction of DNA and RNA-binding residues from protein sequence features. *BMC Syst Biol* 2010;**4**(Suppl 1):S3.
41. Ma X, Guo J, Wu J, *et al.* Prediction of RNA-binding residues in proteins from primary sequence using an enriched random forest model with a novel hybrid feature. *Proteins* 2011;**79**:1230–9.
42. Muppirala UK, Honavar VG, Dobbs D. Predicting RNA-protein interactions using only sequence information. *BMC Bioinform* 2011;**12**:489.
43. Walia RR, Xue LC, Wilkins K, *et al.* RNABindRPlus: a predictor that combines machine learning and sequence homology-based methods to improve the reliability of predicted RNA-binding residues in proteins. *PLoS One* 2014;**9**:e97725.
44. Chen YC, Wright JD, Lim C. DR_bind: a web server for predicting DNA-binding residues from the protein structure based on electrostatics, evolution and geometry. *Nucleic Acids Res* 2012;**40**:W249–56.
45. Miao Z, Westhof E. RBscore& NBench: a high-level web server for nucleic acid binding residues prediction with a large-scale benchmarking database. *Nucleic Acids Res* 2016;**44**:W562–7.
46. Yang X, Wang J, Sun J, *et al.* SNBRFinder: a sequence-based hybrid algorithm for enhanced prediction of nucleic acid-binding residues. *PLoS One* 2015;**10**:e0133260.
47. Yan J, Kurgan L. DRNApred, fast sequence-based method that accurately predicts and discriminates DNA- and RNA-binding residues. *Nucleic Acids Res* 2017;**45**: e84.
48. Zhang J, Ma Z, Kurgan L. Comprehensive review and empirical analysis of hallmarks of DNA-, RNA- and protein-binding residues in protein chains. *Brief Bioinform* 2019;**20**:1250–68.
49. Su H, Liu M, Sun S, *et al.* Improving the prediction of protein-nucleic acids binding residues via multiple sequence profiles and the consensus of complementary methods. *Bioinformatics* 2019;**35**:930–6.
50. Zhang J, Chen Q, Liu B. NCBRPred: predicting nucleic acid binding residues in proteins based on multilabel learning. *Brief Bioinform* 2021;**22**(5):bbaa397.
51. Wang N, Yan K, Zhang J, *et al.* iDRNA-ITF: identifying DNA- and RNA-binding residues in proteins based on induction and transfer framework. *Brief Bioinform* 2022;**23**(4):bbac236.
52. Yang J, Roy A, Zhang Y. BioLiP: a semi-manually curated database for biologically relevant ligand-protein interactions. *Nucleic Acids Res* 2013;**41**:D1096–103.
53. Lewis BA, Walia RR, Terribilini M, *et al.* PRIDB: a protein-RNA interface database. *Nucleic Acids Res* 2011;**39**:D277–82.
54. Patiyal S, Dhall A, Raghava GPS. A deep learning-based method for the prediction of DNA interacting residues in a protein. *Brief Bioinform* 2022;**23**(5).
55. Pande A, Patiyal S, Lathwal A, *et al.* Computing wide range of protein/peptide features from their sequence and structure. *Journal of Computational Biology*. 2022.
56. Pedregosa F, Varoquaux G, Gramfort A, *et al.* Scikit-learn: machine learning in Python. *J Mach Learn Res* 2012;**12**:2825–30.
57. Dhall A, Patiyal S, Sharma N, *et al.* Computer-aided prediction and design of IL-6 inducing peptides: IL-6 plays a crucial role in COVID-19. *Brief Bioinform* 2021;**22**:936–45.
58. Dhall A, Patiyal S, Raghava GPS. HLAncPred: a method for predicting promiscuous non-classical HLA binding sites. *Brief Bioinform* 2022;**23**(5):bbac192.
59. Cozzolino F, Iacobucci I, Monaco V, *et al.* Protein-DNA/RNA interactions: an overview of investigation methods in the -Omics era. *J Proteome Res* 2021;**20**:3018–30.
60. Re A, Joshi T, Kulberkyte E, *et al.* RNA-protein interactions: an overview. *Methods Mol Biol* 2014;**1097**:491–521.
61. Jain DS, Gupte SR, Aduri R. A data driven model for predicting RNA-protein interactions based on gradient boosting machine. *Sci Rep* 2018;**8**:9552.
62. Chen Y, Varani G. Engineering RNA-binding proteins for biology. *FEBS J* 2013;**280**:3734–54.

63. Chen W, Zhang SW, Cheng YM, *et al*. Identification of protein-RNA interaction sites using the information of spatial adjacent residues. *Proteome Sci* 2011;**9**(Suppl 1):S16.
64. Xiong D, Zeng J, Gong H. RBRIdent: an algorithm for improved identification of RNA-binding residues in proteins from primary sequences. *Proteins* 2015;**83**:1068–77.
65. Chen YC, Sargsyan K, Wright JD, *et al*. Identifying RNA-binding residues based on evolutionary conserved structural and energetic features. *Nucleic Acids Res* 2014;**42**: e15.
66. Velankar S, Burley SK, Kurisu G, *et al*. The Protein Data Bank archive. *Methods Mol Biol* 2021;**2305**:3–21.
67. Amirkhani A, Kolahdoozi M, Wang C, *et al*. Prediction of DNA-binding residues in local segments of protein sequences with fuzzy cognitive maps. *IEEE/ACM Trans Comput Biol Bioinform* 2020;**17**:1372–82.