Computational and Structural Biotechnology Journal 20 (2022) 3195–3207

COMPUTATIONAL AND STRUCTURAL BIOTECHNOLOGY JOURNAL

journal homepage: www.elsevier.com/locate/csbj

A comparative analysis of machine learning classifiers for predicting protein-binding nucleotides in RNA sequences

Ankita Agarwal [a,b], Kunal Singh [b], Shri Kant [b], Ranjit Prasad Bahadur [b,*]

[a] School of Bio Science, Indian Institute of Technology Kharagpur, Kharagpur 721302, India
[b] Computational Structural Biology Laboratory, Department of Biotechnology, Indian Institute of Technology Kharagpur, Kharagpur 721302, India

ARTICLE INFO

Article history:
Received 23 March 2022
Received in revised form 14 June 2022
Accepted 14 June 2022
Available online 17 June 2022

Keywords:
RNA-protein interactions
Protein-binding nucleotides
Machine learning
Stratified cross validation
Random forest classifier

ABSTRACT

RNA-protein interactions play vital roles in driving the cellular machineries. Despite significant involvement in several biological processes, the underlying molecular mechanism of RNA-protein interactions is still elusive. This may be due to the experimental difficulties in solving co-crystallized RNA-protein complexes. Inherent flexibility of RNA molecules to adopt different conformations makes them functionally diverse. Their interactions with protein have implications in RNA disease biology. Thus, study of binding interfaces can provide a mechanistic insight of the molecular functioning and aberrations caused due to altered interactions. Moreover, high-throughput sequencing technologies have generated huge sequence data compared to available structural data of RNA-protein complexes. In such a scenario, efficient computational algorithms are required for identification of protein-binding interfaces of RNA in the absence of known structures. We have investigated several machine learning classifiers and various features derived from nucleotide sequences to identify protein-binding nucleotides in RNA. We achieve best performance with nucleotide-triplet and nucleotide-quartet feature-based random forest models. An overall accuracy of 84.8%, sensitivity of 83.2%, specificity of 86.1%, MCC of 0.70 and AUC of 0.93 is achieved. We have further implemented the developed models in a user-friendly webserver "Nucpred", which is freely accessible at "http://www.csb.iitkgp.ac.in/applications/Nucpred/index".

© 2022 The Author(s). Published by Elsevier B.V. on behalf of Research Network of Computational and Structural Biotechnology. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

1. Introduction

RNA-protein interaction is one of the most diverse cellular phenomena involved in several gene regulatory pathways and RNA metabolism, which eventually drives the cellular machineries [1–4]. Interactions of RNAs with partner proteins often guide the overall folding of biomolecular assemblies [5–7]. The complex network of interactions between RNA-binding proteins (RBPs) and their target RNA is highly specific and intricate [8,9]. Any misregulated interaction may lead to either gain-of-function or loss-of-function, thereby causing molecular and cellular defects. This in turn can result into various pathological conditions [10–13]. For example, RBP-RNA interactions are frequently involved in viral recognition and replication [14,15]. Defects in RNA-protein interactions are also annotated with several neurological disorders [16–18], genetic disorders [19] and various types of cancers [20–23]. Considering the diverse roles of RNA in RBP-RNA interactions [24], identification of nucleotides interacting with RBP becomes crucial to understand the functional implications of their recognition.

Experimental techniques including X-ray crystallography, NMR spectroscopy and Electron Microscopy have solved several structures of RBPs and their complexes. However, a huge gap exists between experimentally determined structures of RNA-protein complexes and those persist in nature. This gap is probably due to high cost and time associated with the experimental techniques along with their own limitations. In such a scenario, efficient computational methods can complement to overcome the inherent limitations of the experimental techniques. Combining known biological data with computer-aided algorithms can be an alternative technique to determine the binding interface from sequence information. Advancements in speed and efficiency in computational handling of big data led to the increasing use of Machine Learning (ML) techniques in various fields of physics, biology and medicine [25–28]. Several ML approaches have been used to solve many bioinformatics problem [29,30] including prediction of protein secondary structures [31], prediction of protein functional sites [32], modeling of gene regulatory networks [33], prediction of nucleic-acid binding proteins and residues from structure and sequence

* Corresponding author.
E-mail address: r.bahadur@bt.iitkgp.ac.in (R.P. Bahadur).

https://doi.org/10.1016/j.csbj.2022.06.036
2001-0370/© 2022 The Author(s). Published by Elsevier B.V. on behalf of Research Network of Computational and Structural Biotechnology.
This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

information [34–37] and ATP-binding sites in proteins [38]. All these methods utilize features derived from known data to provide probable solution for unknown data based on the models trained in ML algorithms. Predicting protein interacting sites in a given RNA sequence using ML is challenging due to the inherent complexity of RNA structures and their less diverse sequence pattern as compared to proteins. However, certain molecular properties of RNA sequence differ considerably between binding and non-binding regions in RNA-protein complexes. These distinctive molecular descriptors can be efficiently used in computational predictions of binding sites with feature selection, input data refinement and parameter optimization techniques of ML algorithms.

In this study, we have designed sequence-based features to train them in various ML classifiers to predict protein-binding nucleotides in a given RNA sequence. Efficiency of several algorithms including Random Forest (RF), XGBoost (XGB), Gradient-boosted trees (GBT), AdaBoost (ADB), Support Vector Machine (SVM), Naïve Bayes (NB), K-nearest neighbors (KNN) and Multilayer Perceptron (MLP) are also compared [39–42]. Several physicochemical features of RNA sequence including molecular mass, pKa and binary encoding along with nucleotide compositions including singlet, doublet, triplet and quartet are calculated. For each nucleotide (nt), feature vectors are generated using sliding window strategy. Performance of all the models are evaluated on a non-redundant dataset with various window size. We find, nucleotide-triplet composition (NC-triplet) based RF model performed best compared to all other classifiers with an optimized window-size of 23 nt. We have obtained 83.2% sensitivity, 86.1% specificity, 84.8% accuracy, 82.8% PPV, 0.7 MCC and 0.93 AUC using RF model with 10-fold repeated stratified cross validation.

2. Materials and methods

2.1. Training dataset preparation

A dataset of 180 protein-RNA complexes was curated for this study [43]. All the complexes with resolution better than 3 Å were obtained from the Protein Data Bank (PDB) [44]. We kept all the protein-RNA complexes with protein chains of at least 30 amino acids and RNA chains of at least 5 nt. To remove redundancy between the sequences, pairwise sequence alignment was performed for all the entries in the dataset using BLAST [45]. When protein or RNA component in any two complexes had more than 35% sequence identity, the one with better resolution was retained. Further, complexes with missing or non-standard nucleotides (N, T, X) were removed. The pipeline followed for dataset curation is schematically represented in supplementary Fig. S1. A few PDB structures contain multiple RNA chains that interact with protein. Considering all possible binary interactions, the final non-redundant training dataset – 'PB-RNA194' was curated to develop the training model. PB-RNA194 contains 194 RNA sequences including 33 tRNA, 70 ssRNA, 77 dsRNA and 14 ribosomal RNA (Table 1).

Table 1
Non-redundant training dataset of protein-binding RNAs.

RNA-type	No. of RNA chains	PDB IDs
ssRNA	70	1AV6_B, 1C9S_W, 1CVJ_M, 1G2E_B, 1JBS_C, 1JID_B, 1K8W_B, 1KNZ_W, 1KQ2_R, 1LNG_B, 1M5O_E, 1M8V_O, 1M8W_C, 1M8W_E, 1WPU_C, 1WSU_E, 1ZBH_E, 1ZH5_D, 2A8V_E, 2ANR_B, 2ASB_B, 2B3J_E, 2BX2_R, 2DB3_E, 2G4B_B, 2GIC_R, 2IX1_B, 2J0S_E, 2JEA_C, 2JLU_C, 2PY9_E, 2Q66_X, 2R7R_X, 2VNU_B, 2XGJ_C, 2XNR_C, 2XS2_B, 2XZO_D, 3AEV_C, 3BX2_C, 3D2S_E, 3I5X_B, 3IEV_D, 3K5Q_B, 3MDG_C, 3NMR_B, 3O8C_C, 3PF4_R, 3QJJ_Q, 3R2C_R, 3RC8_E, 3T5N_C, 4H5P_E, 4J1G_E, 4J7M_B, 4M59_C, 4M59_D, 4MDX_C, 4N2Q_B, 5AOR_C, 5DET_Q, 5EIM_C, 5ELH_R, 5ELK_R, 5ELR_B, 5ELS_I, 5EX7_B, 5GXH_B, 5I4A_D, 5LTA_E
dsRNA	77	1DI2_D, 1HQ1_B, 1MSW_R, 1N35_B, 1N35_C, 1OOA_C, 1R3E_C, 1R9F_B, 1R9F_C, 1SI3_B, 1WNE_B, 1WNE_C, 1YVP_E, 1YVP_F, 1YVP_H, 1ZBI_C, 2AZ0_C, 2AZ0_D, 2BGG_P, 2BGG_Q, 2EZ6_C, 2EZ6_D, 2F8S_C, 2GJW_E, 2GJW_F, 2GJW_H, 2GXB_E, 2GXB_F, 2OZB_C, 2PJP_B, 2QUX_C, 2R8S_R, 2XD0_G, 2Y8W_B, 2YKG_C, 2YKG_D, 2ZI0_C, 2ZKO_C, 2ZKO_D, 3A6P_D, 3A6P_E, 3BSN_P, 3BSN_T, 3BT7_C, 3DH3_F, 3EQT_C, 3EQT_D, 3FTE_C, 3FTE_D, 3IAB_R, 3KS8_E, 3KS8_F, 3MOJ_A, 3O3I_A, 3OIJ_C, 3RW6_H, 3SNP_C, 3ZC0_M, 4ATO_G, 4ERD_C, 4ERD_D, 4FVU_B, 4FVU_C, 4IG8_B, 4IG8_C, 4ILL_C, 4ILL_R, 4L8H_R, 4ZT0_B, 5AOX_C, 5ED1_B, 5ED1_C, 5F5F_B, 5F5H_C, 5ID6_G, 5TF6_B, 5WTK_B
tRNA	33	1ASY_R, 1B23_R, 1C0A_B, 1FFY_T, 1GAX_C, 1H3E_B, 1H4S_T, 1J1U_B, 1N78_C, 1QF6_B, 1QTQ_B, 1SER_T, 1U0B_A, 1VFG_D, 2AZX_C, 2BTE_B, 2CSX_C, 2DLC_Y, 2DRB_B, 2DU3_D, 2FK6_R, 2FMT_C, 2ZM5_C, 2ZZM_B, 3ADB_C, 3AMT_B, 3EPH_E, 3HL2_E, 3VJR_B, 4YCP_B, 4YVJ_C, 5HR7_D, 5T8Y_X
Ribosomal RNA	14	1DFU_M, 1DFU_N, 1FEU_B, 1FEU_C, 1G1X_D, 1G1X_E, 1I6U_C, 1MJI_C, 1MMS_C, 1MZP_B, 1S03_A, 1SDS_D, 2HW8_B, 5WTY_C

2.2. Classification of dataset into protein binding and non-binding nucleotides

To identify protein binding sites in RNA or RNA binding sites in proteins, distance-based methods have been used with cut-offs in the range between 3.0 Å and 6.0 Å in various prediction methods [46,47]. Prediction algorithms based on distance are highly dependent on the selected cut-off. Solvent accessible surface area (SASA) based criteria have also been used to determine the residues and nucleotides at protein-RNA interfaces [48]. In this study, we used SASA-based calculations to identify the nucleotides as binding or non-binding to partner protein. RNA chains were extracted from each protein-RNA complex in the training dataset. SASA was calculated using NACCESS [49], which implements the Lee and Richards algorithm [50] with default water probe of radius 1.4 Å. If any atom of a nucleotide loses SASA upon complexation, the corresponding nucleotides were marked as protein binding and the rest were labelled as non-binding nucleotides (supplementary file S1). Out of total 5414 nt present over all the sequences, 2411 were identified as protein binding and 3003 were identified as non-binding. Sequence length of each type of RNA chain and total number of interacting nucleotides in each chain are provided in supplementary Table S1.

2.3. Feature vector encoding with sliding-window strategy

For each RNA sequence in PB-RNA194 dataset, overlapping patterns of different window size (WS) starting from 5 nt to 31 nt were created. Each sliding window pattern was classified as positive if the central nucleotide of the pattern interacts with protein, or else, the pattern was classified as negative or non-interacting. To create a pattern for terminal nucleotides in a given sequence, "X" dummy nucleotides were added at both the termini of RNA, where $X = (WS - 1)/2$. For instance, for a window size of seven, three "X" were added at both the termini to create L patterns from sequence of length L. This sliding-window based strategy ensured that each nucleotide in the dataset occupies the central position once. It also

ensured that all the nucleotides in a given sequence and the effect of their neighbours were considered during training.

The overlapping window patterns, corresponding to each nt, were encoded by features derived from RNA sequence. Sequence length (SL) and nucleotide composition (NC) were calculated to encode global properties of each RNA sequence in the training set. Further, local features of the nucleotides including primary pKa value (pKa) and molecular mass (M) of each nucleotide of a given window was used as an attribute to represent the physicochemical nature of the nucleotides. Binary encoding (BE) representing the type of nucleotide at each position was used as another attribute to train the classifier, which would generate numerical values. Nucleotide compositions (NC-singlet, NC-doublet, NC-triplet and NC-quartet) were generated to represent the composition profile. For each window segment, comprising of four primary nucleotides (A, U, G and C) and one dummy nucleotide (X), NC-singlet (A, U, G, C, X), NC-doublet (AA, AU, AG, AC, AX, ..., CC), NC-triplet (AAA, AAC, AAG, . . ., CCC) and NC-quartet (AAAA, AAAU, AAAG,....., CCCC) features were represented by 5, 25, 125 and 625 numerical values, respectively. Feature vectors were generated for binary classification of the central nucleotide in each sequence window using each of these features individually and as an ensemble.

2.4. ML classifiers and hyperparameter tuning

ML approaches are the most cost-effective and time-efficient methods used extensively in various fields to develop prediction models. In this study, prediction models were trained to obtain the best performing model using experimentally validated data and various ML algorithms. All the algorithms discussed below were implemented using sklearn library from Scikit-learn [51] and python programming.

Naive Bayes (NB) classifiers are a collection of supervised classification algorithms based on Bayes' theorem and a common principle, where every pair of features being classified is independent of each other. Gaussian NB (GNB) and Multinomial NB (MNB) are two variants of NB classifier implemented in this study with default parameters. While GNB follows Gaussian distribution and supports continuous valued features, MNB supports discrete features. K-Nearest Neighbors classifier (KNN) is a simple supervised non-parametric ML algorithm. Its classification is based on clustering of similar data points (close proximity) into one class. In this study, KNN was implemented with tuning of hyperparameter k. Support Vector Machine (SVM) is another commonly used supervised algorithm. SVM uses an optimal hyperplane or decision boundaries to classify the data points depending on the choice of extreme vectors, called support vectors. In this study, we evaluated the prediction performance using both linear and RBF kernels of SVM. The other important hyperparameters, i.e., kernel coefficient (gamma) and regularization parameter (C) were tuned through grid-search cross-validation to get optimum performing SVM model. Multi-layer Perceptron (MLP) classifier is an artificial neural network (ANN) based supervised algorithm that trains by back-propagation. It has advantages of learning non-linear models based on optimization of the number of hidden layers between the input and output layers. The algorithm comprises of a regularization parameter (α) that can be varied to prevent overfitting of the data. In this study, all the parameters of MLP were set to default and α was optimized during training.

Random Forest (RF) is an ensemble learning method that builds multiple decision trees through random sampling of the dataset with replacement. Each of the individual decision trees gives a class prediction based on the attributes. The model select the final class based on the most voted one over all the individual trees. The advantage of RF is that it is faster than many classification algorithms and unlikely to overfit even if the number of trees is increased. In this study, RF was tuned by optimizing number of estimators and maximum depth for each tree. RF was implemented with bootstrap aggregation and the oob_score was set to true. AdaBoost (ADB), Gradient Boosted Trees (GBT) and Extreme Gradient Boosting (XGBoost) are other ensemble-based learning methods compared in this study. XGBoost is a revised version of gradient boosting decision tree algorithm, and is being increasingly used in ML predictions due to its higher execution speed and more accurate model performance compared to traditional gradient boosting. Several hyper-parameters of these ensemble methods were tuned to optimize the performance such as learning rate, number of estimators, maximum depth, minimum child weight, gamma, subsample, objective function and number of threads.

2.5. Performance evaluation of the classifier

Cross-validation is a widely used method to evaluate the performance of any prediction model. In our study, we used 10-fold cross-validation technique to compare the performance of all the classifiers on the benchmark dataset. In this technique, all the nucleotides were randomly divided into ten sets. Of these sets, nine were used for training and the remaining one was used for testing. This process was repeated ten times such that each set was used once for testing. The final performance was obtained by averaging the performance of all the ten sets. Further, to keep the positive and negative dataset balanced, we used repeated stratified cross validation with number of folds and repeats set to 10 such that in each of the ten validation sets, which is repeated ten times, the ratio of positive and negative samples in both the train and test dataset is preserved during each fold. To assess the performance of the classifier, we calculated commonly used metrics, which includes overall accuracy (ACC), specificity (SPE), sensitivity (SEN), positive predictive value (PPV), F-measure and Matthews correlation coefficient (MCC) [52].

In the equations (Table 2), TP, TN, FP and FN are true positives, true negatives, false positives and false negatives, respectively. ACC determines the amount of correct prediction of a class over the entire sample space. SEN defines the total number of correctly predicted positive instances over the total number of positive instances, while SPE defines the total number of correctly predicted negative instances over the total number of negative instances. PPV signifies the ratio of correctly predicted positive instances to the total positive predictions. The parameter MCC signifies the correlation between actual and predicted classifications. In addition to above-mentioned measures, a threshold-independent parameter, i.e., area under the curve (AUC) of the receiver operating characteristic (ROC) plots was computed for each variation of the prediction model. The AUC value obtained from ROC curves (plot of true positive rate against false positive rate) determines whether a classification model is perfect (AUC = 1) or random (AUC = 0.5).

Table 2
Performance Evaluation Metrics.

Sl. no.	Parameter	Expression
1	Accuracy (ACC)	$ACC = \frac{TP+TN}{TP+FN+TN+FP}$
2	Specificity (SPE)	$SPE = \frac{TN}{TN+FP}$
3	Sensitivity (SEN)	$SEN = \frac{TP}{TP+FN}$
4	Precision (PPV)	$PPV = \frac{TP}{TP+FP}$
5	F-measure (F-score)	$F-score = \frac{2\times PPV\times SEN}{PPV+SEN}$
6	Mathews-correlation coefficient (MCC)	$MCC = \frac{(TP)(TN)-(FP)(FN)}{\sqrt{(TP+FP)(TN+FP)(TP+FN)(TN+FN)}}$

2.6. Pipeline for model development and prediction

Given a RNA sequence in NCBI fasta format, sliding window patterns were generated from the query sequence. For each pattern, triplet nucleotide compositions were calculated and preprocessed in numerical format to generate feature vector for each nucleotide. The classification model was trained to give the predictions based on patterns of calculated features of the query sequence and previously trained RF structure. The final predictions were given in binary format as '1′ or '0′ for each nucleotide classified as binding or non-binding, respectively (Fig. 1).

3. Results

3.1. Dataset analysis and development of nucleotide features

Residue-nucleotide paired compositions at the protein-RNA interfaces are calculated to find out nucleotide preferences to any particular amino acid (Fig. 2). We do not find any strong preference of amino acids towards the four nucleotides. All the nucleotides show high preference for Arg and Lys and least preference for Cys, Tyr and Met. G and C show a slightly higher preference for Arg and Lys than A and U. The nucleotides in RNA-protein interfaces generally occur in contiguous stretches. Hence, to get an idea about the prevalence of certain stretches of nucleotides, we have calculated the occurrence of nucleotides at the interface in our training dataset. It is observed that almost 80% of the dataset constitute of one to six nucleotide stretches at the interface. Of these, stretches with three (13.3%), four (13.7%) and five nucleotides (14.9%) being the most common (Fig. 3). This implies that compared to single nucleotides (7.2%), doublets, triplets, quartets and pentets have greater preference at the interface and can be used for classification of binding and non-binding sites. To find out any distinguishing pattern, we have calculated nucleotide singlet, doublet and triplet compositions at the interface and non-interface regions (Fig. 4 A to C). For singlet nucleotide composition (Fig. 4A), all the four nucleotides show almost similar interface composition. Among the four, G shows the highest non-interface composition. Both A and U frequently occur at the interface than at the non-interface, while C does not show any such differences. The interface propensities of four nucleotides (Fig. S2, Table S2) also show that A and U have positive interface propensity, whereas G has strong negative interface propensity. However, C does not have any preference for both interface and non-interface. Among the nucleotide doublets (Fig. 4B), AA, AC and UA are slightly preferred at interface, whereas, AG, UC, CA, GU, GG, CG and CC are more preferred at non-interface. Some doublets, AU and UU do not show any preference for both interface and non-interface. Among the nucleotide triplets (Fig. 4C), clear interface preferences are observed for AAA, UUU, AUA, AUU, UAA, UAU, CAC and CUA. However, many triplets including AGA, AGU, AGG, AGC, UUC, UGG, UCG, UCC, GAU, GAG, GUU, GUC, GGU, GGG, CAG, CGA, CGG, CCG and CCC show clear non-interface preferences. These results indicate that NC-triplet can be used as a distinguishing feature for binary classification of binding and non-binding sites.

We have also analyzed the training data based on four different types of RNA present in the dataset. We have calculated the nucleotide compositions at interface and non-interface for each

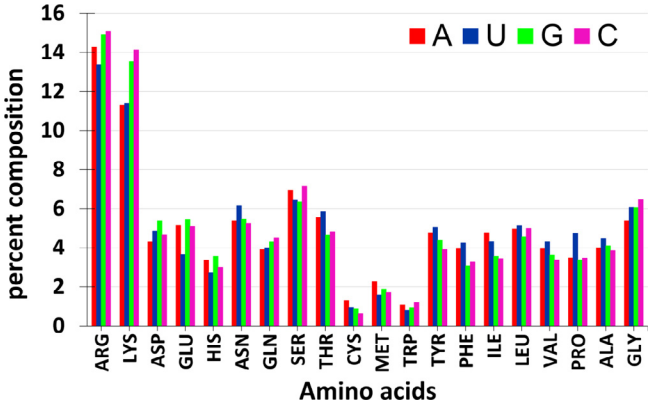

Fig. 2. Pairwise residue-nucleotide interface composition in training dataset.

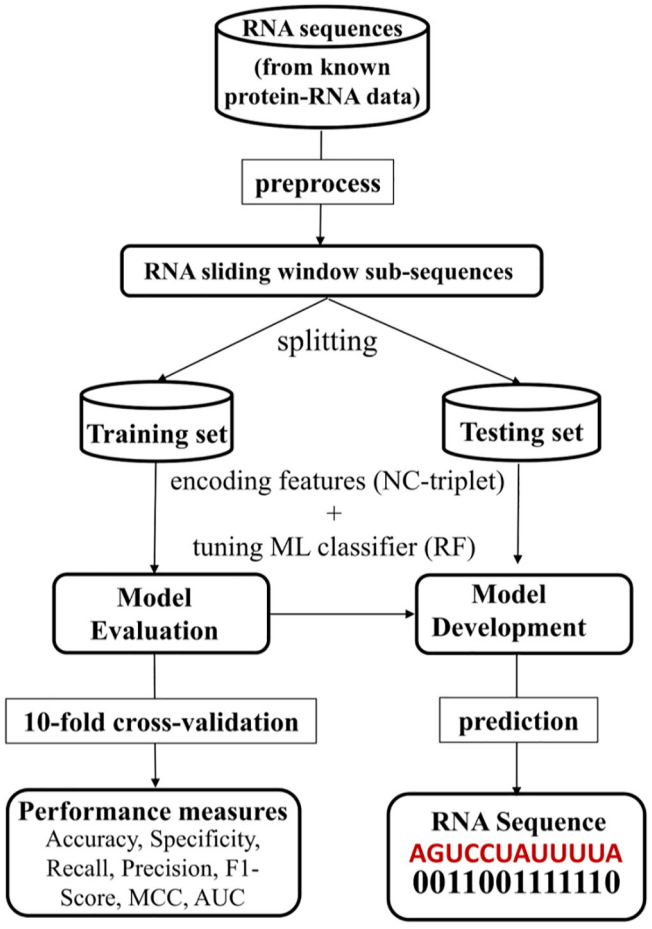

Fig. 1. Pipeline for development and prediction of classification model.

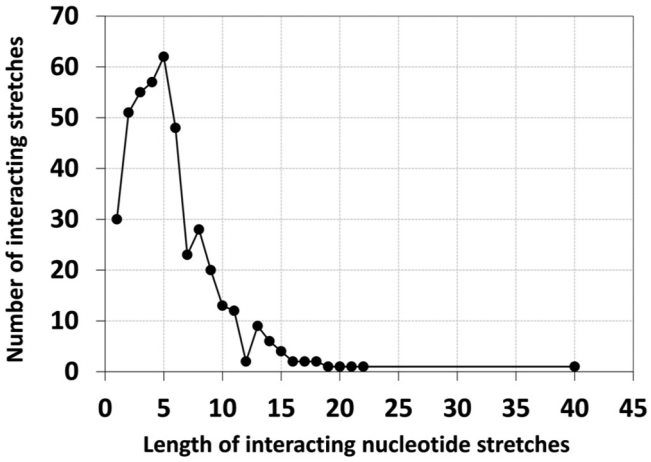

Fig. 3. Distribution of nucleotide stretches at the interface in training dataset.

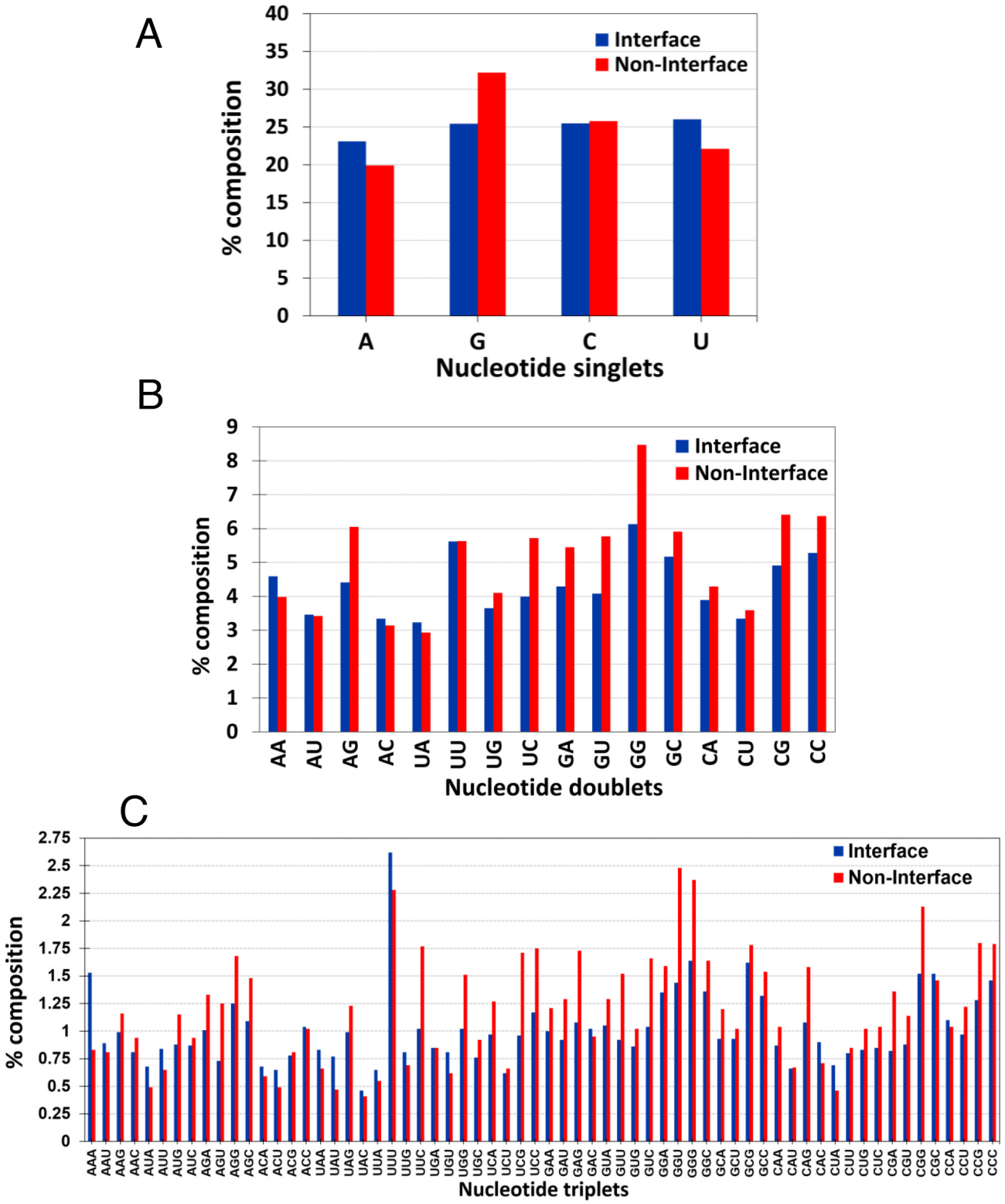

Fig. 4. Nucleotide compositions (A) NC-singlets, (B) NC-doublets and (C) NC-triplets at interface and non-interface regions in training dataset.

class of RNA to find out whether any nucleotide is preferred more at the interface in a particular RNA-type (Table S3). While ssRNA have higher composition of U, the other three types of RNA have higher composition of G. Further, the interfaces with tRNA is enriched with C, whereas higher occurrence of U and A are found at the interfaces with ssRNA. Higher prevalence of U is found at the interfaces with rRNA. In dsRNA, A is slightly favoured at the interface. In all the four RNA types, G is disfavoured at the interface with greater non-interface preference in ssRNA, dsRNA and ribosomal RNAs.

3.2. Performance evaluation of different RNA features and sliding window optimization

To validate the effect of feature selection on the prediction performance, we have evaluated the parametrically tuned RF model with different RNA features derived from the sequence at varying window size (WS). WS optimization is performed based on the AUC values obtained from ROC curves for each of the feature (Fig. 5 A to F). Table 3 provides the 10-fold CV metrics at optimum WS for each of the feature. Compared to individually trained features (Global, BE or Local), ensemble features (Global + Local and Global + Local + BE) performs better (Table 3). However, performance obtained with ensemble features is comparatively lower than individual NC-triplet and NC-quartet features. This is evident from large decrease in AUC value from 0.93 (for NC-triplet and NC-quartet) to 0.78 (for Global + Local) and 0.77 (for Global + Local + BE). MCC values are also reduced from 0.70 (for NC-triplet and NC-quartet) to 0.44 (for Global + Local) and 0.42 (for Global + Loc al + BE). Results of 10-fold CV at different WS for other local features of RNA, i.e., BE, NC-singlet and NC-doublet show that NC-doublet performs better than NC-singlet and BE (supplementary Tables S4, S5 and S6). With BE and NC-singlet, maximum AUC of 0.72 and 0.75 is obtained at WS of 27 nt and 29 nt, respectively. However, maximum AUC of 0.92 is obtained for NC-doublet at WS of 23 nt. The CV metrics for NC-triplet (Table 4) and NC-quartet features (Table 5) at variable WS show that both of these individual feature-based RF model provides similar performance at an optimum WS of 23 nt. On the other hand, ensemble feature (Global + Local + NC-triplet) achieves the highest AUC of 0.88, which is much lower than individual NC-triplet and NC-quartet features. Thus, NC-triplet and NC-quartet features provide the best results compared to all other features as well as the ensemble feature. Further, we have also evaluated the prediction performance on four different types of RNA sequences present in our dataset (Table 6). All the four types of RNA achieves reasonable performance with ribosomal RNA and tRNA obtaining slightly higher performance compared to other two types of RNA sequences.

3.3. Performance comparisions of different ML classifiers for NC-triplet model

Parameter tuning is one of the most important steps in building a good predictive model. Models for binary classification often involve prediction of a class based on certain threshold value of prediction probability. By default, RF gives prediction based on threshold of 0.5. However, the default threshold can be tuned to optimize the performance. To calculate the optimum threshold value for the developed model, we varied the threshold based on precision-recall optimization (Fig. S3). We have obtained optimum recall and precision of 0.85 and 0.83, respectively at threshold value of 0.5. Further, to optimize the learning process, an important hyperparameter of RF, the number of estimator, is also optimized. Based on sensitivity, n_estimators is set to 275 (Fig. S4). For SVM-RBF, kernel coefficient (gamma) and cost parameter (C) are also varied. Maximum accuracy and AUC is achieved with C value of 1.0 and gamma value of 0.1 (Fig. S5 A to D). The value of 'k', defining the number of neighbors, is tuned to 3 to achieve maximum accuracy and AUC in KNN (Fig. S6 A and B). We have implemented 10 different ML classifiers with optimized parameters and NC-triplet feature to compare the predictive performance of all these algorithms. Table 7 provides the prediction performance obtained from 10-fold CV for NC-triplet feature based model implemented using 10 ML classifiers. ROC curves for all the 10 classifiers at variable WS are provided in supplementary Fig. S7. ROC curves in Fig. 6 show the AUC values obtained for 10 classifiers with optimized parameters at window-size of 23 nt. The highest AUC of 0.93 is achieved with RF and XGBoost, followed by MLP (AUC = 0.92), SVM (AUC = 0.91) and KNN (AUC = 0.91). Naïve Bayes gives poor performance (AUC = 0.68) among all the classifiers. The performance of RF is the best with the highest AUC of 0.93, MCC of 0.7 and specificity of 87%. The other classifiers that give better performance are KNN (MCC = 0.68), RBF SVM (MCC = 0.67), MLP (MCC = 0.67) and XGBoost (MCC = 0.68). The performance of XGBoost is the closest to RF. The best performing classifiers are further used to train a voting ensemble classifier (with soft voting), and the predictive performance is measured. The ensemble classifier performs slightly better than a few other individual classifiers, except RF, which shows better performance than all the other individual classifiers as well as the voting ensemble classifier (Fig. 7).

3.4. Performance evaluation on other validation and test datasets

We have achieved maximum validation AUC of 0.93 (Fig. 8A) and accuracy of 84.72% (Table 4) with RF-based NC-triplet model at optimized WS of 23 nt in 10-fold repeated stratified CV on PB-RNA194 dataset. Fig. 8B shows the confusion matrix obtained for the same. The trained model is evaluated on another dataset RNA-208 [53] comprising of 208 RNA chains obtained from PRIDB database with RNA chain length greater than 10 nt. This dataset consists of 46,582 nt, of which, 10,198 are protein interacting according to the cutoff distance of 5.0 Å. The validation metrics achieved by RF model trained on 'PB-RNA194′ and tested on 'RNA-208′ (Table 8) shows poor performance with maximum MCC of 0.51. In RNA 208 dataset, a class imbalance is observed with binding class being the minority. Hence, we use random oversampling technique (ROS) to deal with class imbalance followed by model training and testing. Although, an increase in performance (MCC increase by 0.2) is observed, yet the maximum prediction performance achieved is still low. Probable reason for poor performance on RNA-208 could be the differences in length distribution of the two datasets. On plotting the histogram of RNA sequence lengths for both the datasets, we observe that the longest sequence in PB-RNA194 have 160 nt (Fig. S8A), while RNA-208 have few RNA sequences with length 1500 nt to 3000 nt (Fig. S8B). Thus, the trained predictor performs better for small and medium length RNA chains but provide poor performance on very long chains. To validate this claim, a subset of sequence is selected from RNA-208 with length below 500 nt (RNA-150). The model trained on PB-RNA194 is tested on this subset. An increase in the performance of the model in both accuracy (0.83 to 0.91) and AUC (0.87 to 0.95) is observed on this validation subset (Table 8). Further, a new dataset 'RNA-344′ is prepared by combining PB-RNA194 and RNA-150 datasets. SASA is used for assigning the nucleotides as binding or non-binding as described in the section 2.2. Out of 13,452 nt present in RNA-344, 4693 are protein binding and 8759 are non-binding. Maximum AUC of 0.94 is achieved with WS of 19 and above (Fig. S9). Performance metrics of RF model obtained after 10-fold CV on RNA-344 dataset for varying WS are provided in Table S7. We have obtained AUC much greater than 0.50 (random), indicating the feasibility of predicting the interface nucleotides using only RNA sequence as input information.

3.5. Prediction performance on independent test dataset

An independent test dataset (RNA30) comprising of 30 RNA sequences is prepared by taking all non-redundant RNA sequences of protein-RNA complexes deposited in PDB after January 2018. The filtered sequences have sequence similarity below 50% with all the sequences in training dataset. RNA30 is used as an independent test set to evaluate the prediction performance of the NC-triplet RF model trained on original data (PB-RNA194) to exclude the possibility of overfitting. The prediction performance on

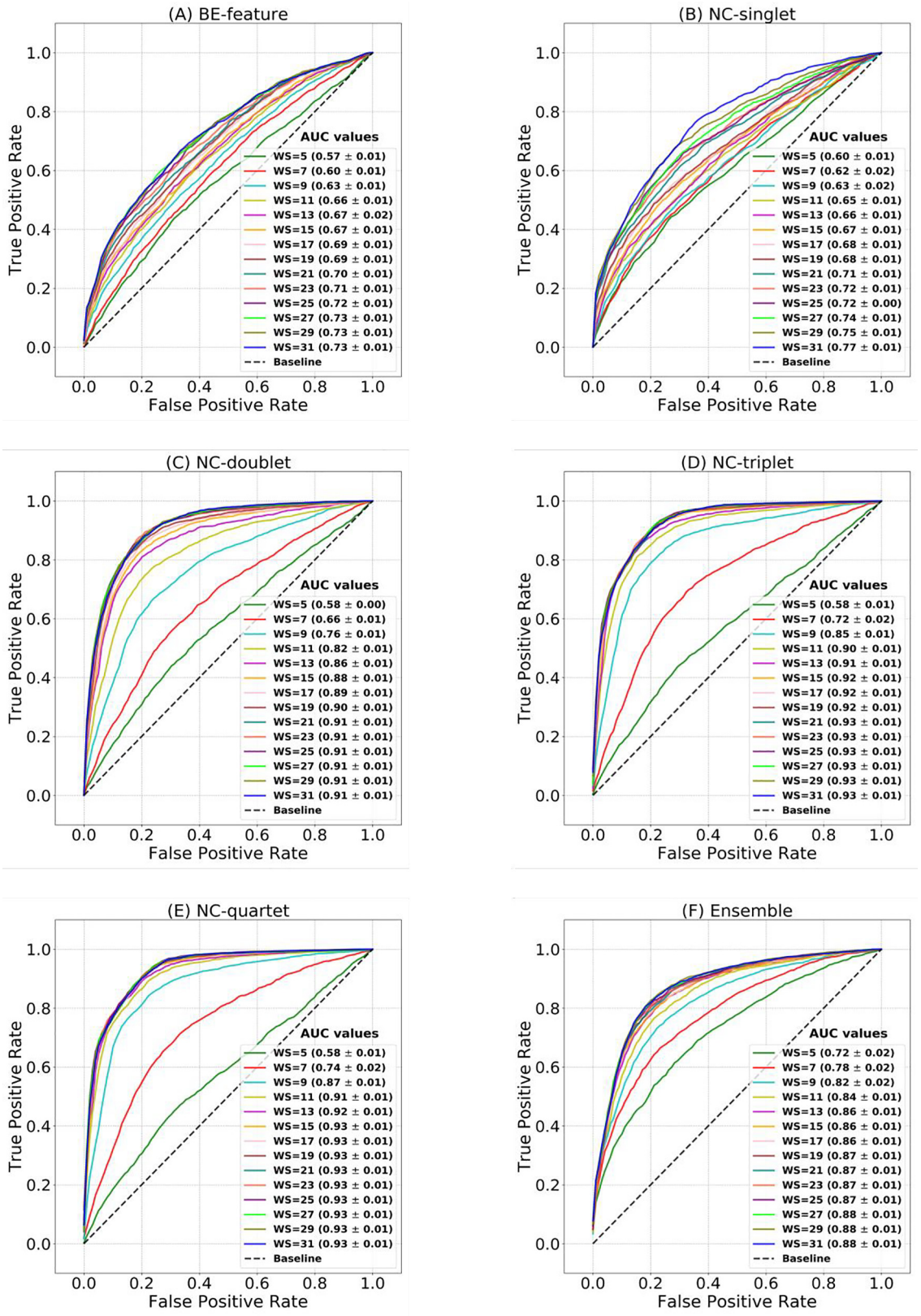

Fig. 5. ROC curves at variable window size (5 nt to 30 nt) for (A) BE-feature, (B) NC-singlet, (C) NC-doublet, (D) NC-triplet, (E) NC-quartet and (F) Ensemble-feature based RF models.

Table 3
10-fold CV results for different RNA features at optimized WS using RF.

Features	WS	ACC	SPE	SEN	F-score	PPV	MCC	AUC
Global (SL + NC)	–	0.68	0.77	0.57	0.61	0.67	0.35	0.73
Local (pKa)	29	0.68	0.85	0.46	0.55	0.71	0.34	0.72
Local (mass)	29	0.67	0.85	0.45	0.55	0.70	0.33	0.72
Local (pKa + mass)	27	0.67	0.85	0.45	0.55	0.71	0.34	0.72
Binary encoding (BE)	27	0.67	0.85	0.45	0.54	0.70	0.34	0.72
NC-singlet	29	0.70	0.73	0.66	0.66	0.67	0.40	0.75
NC-doublet	23	0.84	0.86	0.82	0.82	0.82	0.69	0.92
NC-triplet	23	0.85	0.86	0.84	0.83	0.83	**0.70**	**0.93**
NC-quartet	23	0.85	0.86	0.84	0.83	0.83	**0.70**	**0.93**
Global + Local + BE	27	0.71	0.84	0.42	0.56	0.74	0.42	0.77
Global + Local (SL + NC + pKa + mass)	23	0.72	0.84	0.58	0.65	0.75	0.44	0.78
Ensemble features (Global + Local + NC-triplet)	27	0.81	0.87	0.75	0.78	0.83	0.63	0.88

Best performing features are marked in bold considering AUC and MCC.

Table 4
10-fold CV results at varying WS for nucleotide triplet composition using RF.

WS	Accuracy	Specificity	Sensitivity	F-score	PPV	MCC	AUC
5	56.35	59.11	52.92	51.92	50.96	11.99	0.58
7	68.78	73.03	63.50	64.44	65.40	36.65	0.72
9	80.13	84.42	74.78	77.02	79.39	59.62	0.85
11	83.06	86.08	79.30	80.66	82.06	65.63	0.89
13	84.12	87.01	80.51	81.86	83.27	67.77	0.91
15	84.45	86.48	81.92	82.43	82.95	68.48	0.92
17	84.17	85.75	82.21	82.22	82.24	67.96	0.92
19	84.24	86.15	81.87	82.23	82.59	68.08	0.92
21	84.34	85.81	82.50	82.43	82.36	68.30	0.93
23	84.72	85.61	84.11	83.06	82.04	**69.75**	**0.93**
25	84.82	85.71	83.70	83.08	82.47	69.32	0.93
27	83.99	84.82	82.95	82.19	81.43	67.65	0.928
29	84.13	84.68	83.45	82.41	81.39	67.98	0.926
31	84.43	85.08	83.62	82.71	81.82	68.56	0.926

WS with optimum performance are marked in bold considering AUC and MCC.

Table 5
10-fold CV results at varying WS for nucleotide quartet composition using RF.

WS	Accuracy	Specificity	Sensitivity	F-score	PPV	MCC	AUC
5	56.83	59.84	53.09	52.28	51.49	12.90	0.58
7	69.40	73.69	64.12	65.14	66.18	37.95	0.74
9	81.23	85.25	76.23	78.35	78.35	61.88	0.87
11	83.34	85.95	80.09	81.07	82.07	66.21	0.91
13	84.21	86.31	81.58	82.15	82.72	68.00	0.92
15	84.80	86.78	82.33	82.83	83.33	69.20	0.92
17	84.72	86.55	82.46	82.78	83.11	69.06	0.93
19	85.00	86.38	83.28	83.18	83.08	69.65	0.93
21	84.91	86.11	83.41	83.12	82.83	69.48	0.93
23	85.04	86.11	83.70	83.29	82.87	**69.75**	**0.93**
25	85.02	86.08	83.70	83.27	82.84	69.71	0.93
27	84.41	85.15	83.49	82.67	81.86	68.52	0.93
29	84.48	84.92	83.95	82.82	81.71	68.70	0.93
31	84.10	84.98	82.99	82.29	81.61	67.87	0.93

WS with optimum performance are marked in bold considering AUC and MCC.

Table 6
Performance measures of NC-triplet RF model for different RNA-types in the dataset.

RNA-type	Accuracy	Specificity	Sensitivity	F-score	PPV	MCC
tRNA	0.975	0.978	0.970	0.960	0.960	0.940
ssRNA	0.930	0.910	0.950	0.940	0.930	0.860
dsRNA	0.950	0.930	0.960	0.950	0.930	0.890
rRNA	0.980	0.980	0.980	0.980	0.980	0.960

Table 7
10-fold CV comparisions for 10 ML algorithms at optimized WS using NC-triplet feature.

Sl.no.	Algorithm	Optimum WS	ACC	SPE	SEN	F-score	PPV	MCC	AUC
1.	GNB	27	0.63	0.82	0.40	0.49	0.64	0.24	0.69
2.	MNB	23	0.62	0.70	0.52	0.55	0.58	0.22	0.68
3.	**KNN**	17	0.84	0.85	0.83	0.82	0.82	0.68	0.91
4.	**MLP**	25	0.84	0.85	0.81	0.81	0.81	0.67	0.92
5.	Linear SVM	29	0.72	0.87	0.53	0.62	0.76	0.42	0.80
6.	**RBF SVM**	25	0.84	0.85	0.82	0.81	0.81	0.67	0.91
7.	ADB	29	0.70	0.77	0.61	0.64	0.68	0.40	0.79
8.	GBT	27	0.79	0.85	0.71	0.75	0.80	0.59	0.88
9.	**XGBoost**	27	0.84	0.85	0.83	0.82	0.82	0.68	0.93
10.	**RF**	23	0.85	0.87	0.84	0.83	0.83	0.70	0.93
11.	Voting Ensemble	25	0.84	0.86	0.82	0.82	0.82	0.67	0.92

Best performing ML classifiers are marked in bold. Voting ensemble is trained using five best performing classifiers marked in bold.

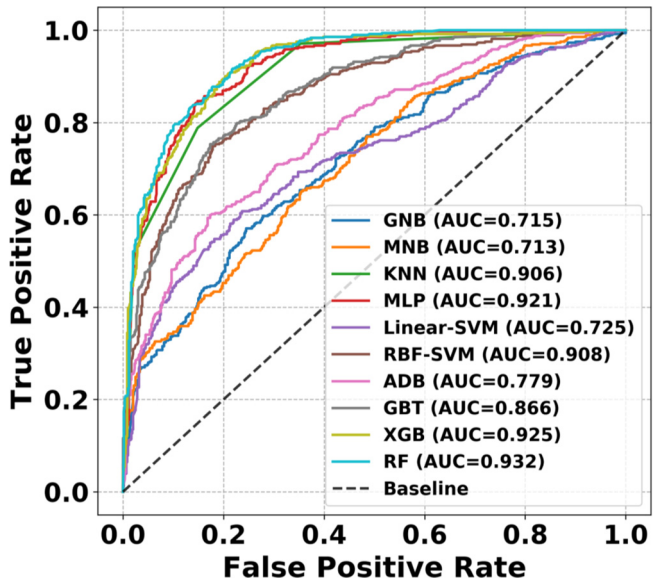

Fig. 6. ROC curves obtained for NC-triplet model at optimum window size of 23 nt for 10 different ML classifiers.

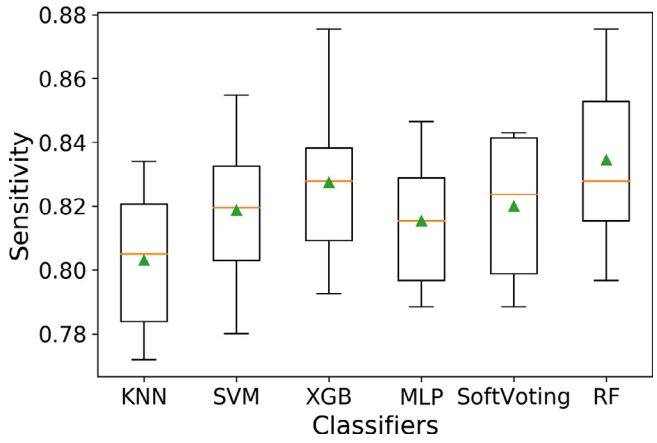

Fig. 7. Boxplot showing the CV performance of individual and voting ensemble classifiers based on sensitivity.

RNA30 test data is shown in Table 8. The results show that a reasonable performance is obtained on the test data with AUC of 0.89, MCC of 0.62 and accuracy of 0.83.

3.6. Prediction performance on a test case

The capability of the prediction model is tested on an independent case (PDB ID: 4JXZ). The RNA chain of the complex is a 71 nt tRNA interacting with glutaminyl-tRNA synthetase. The interface nucleotides in the complex are identified based on SASA calculations and are shown as red spheres (Fig. 9A). The predicted nucleotides are represented in red (TP), cyan (TN), yellow (FP) and purple (FN) spheres (Fig. 9B). The model correctly predicts 24 out of 34 true protein-binding nucleotides, achieving an overall accuracy of 83.2%, specificity of 97.1%, sensitivity of 67.7%, precision of 95.0%, F-score of 80.1% and MCC of 0.69.

3.7. Model implementation in web based tool

NC-triplet and NC-quartet feature-based models optimized with 10-fold CV with best performing RF predictor and other optimized predictors such as SVMRBF and KNN are implemented in a user-friendly web based tool "Nucpred". The input to the developed predictor is the protein-binding RNA sequence in fasta format. The web server will automatically generate NC-triplet feature profiles from the submitted sequence and use them as the input to the default trained RF classifier. The webserver prompts the user to either paste or upload their sequence and an option to select other classification models including RF-NC-quartet, SVMRBF and KNN apart from the default RF-NC-triplet model. The results provided by the webserver includes the raw probability scores and the annotations 1 or 0 to the nucleotides as RBP-binding or non-binding, respectively. The webserver is freely accessible from following link: "http://www.csb.iitkgp.ac.in/applications/Nucpred/index".

4. Discussion

Over the years, multitude of protein-coding and non-coding RNA have been discovered which interacts with proteins to perform various important biological functions. Nucleotides in RNA sequences specifically bind to their partner protein to carry out various cellular processes. For example, the specific recognition of GGU motif by the zinc finger domain in fused in sarcoma protein indicates the significance of protein binding nucleotides [54]. Hence, identification of the binding nucleotides can contribute to our understanding of the molecular basis of these interactions. Although, several methods have been developed to predict RNA-binding residues in protein sequences, yet the problem of predicting protein-binding nucleotides in RNA sequences has received little attention. Moreover, predicting protein-binding nucleotides is more challenging due to less diverse RNA sequence with only four nucleotides as compared to protein sequence with 20 amino acid residues.

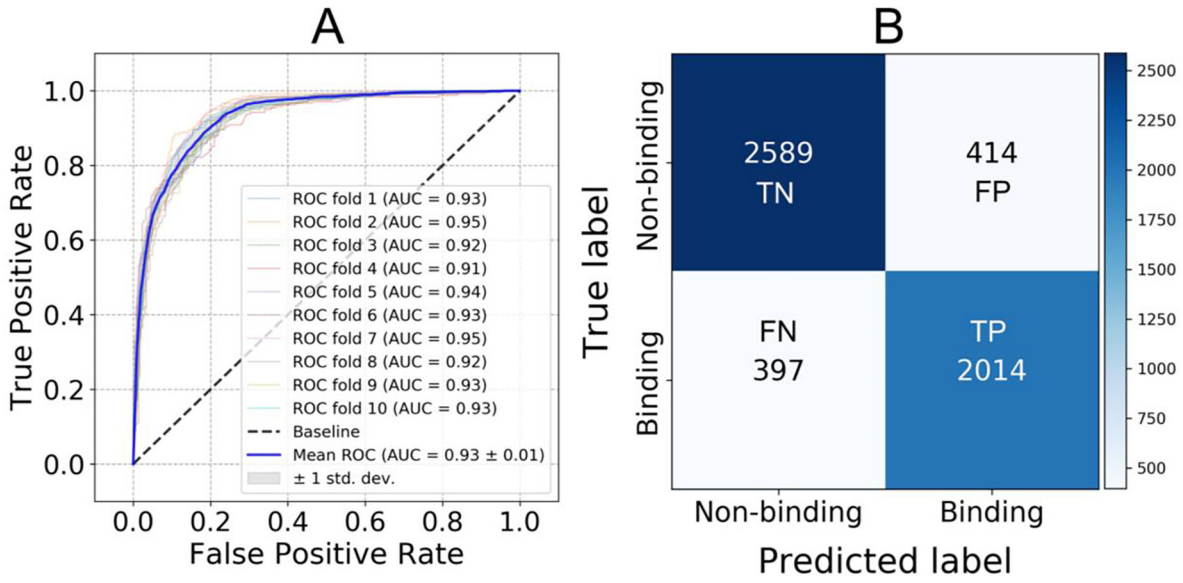

Fig. 8. (A) ROC curve and (B) Confusion matrix obtained with 10-fold CV for NC-triplet RF model at optimum window size of 23 nt.

Table 8
10-fold CV measures of NC-triplet RF model for different validation and test datasets.

Dataset	ACC	SPE	SEN	F-score	PPV	MCC	AUC
PB-RNA194	0.85	0.87	0.84	0.83	0.83	0.70	0.93
RNA-208	0.83	0.89	0.62	0.62	0.63	0.52	0.88
RNA-208 (ROS)	0.83	0.86	0.71	0.64	0.58	0.53	0.87
RNA-150	0.91	0.95	0.80	0.81	0.84	0.75	0.95
RNA-344	0.87	0.90	0.81	0.81	0.82	0.71	0.94
RNA-30	0.83	0.89	0.70	0.68	0.69	0.62	0.89

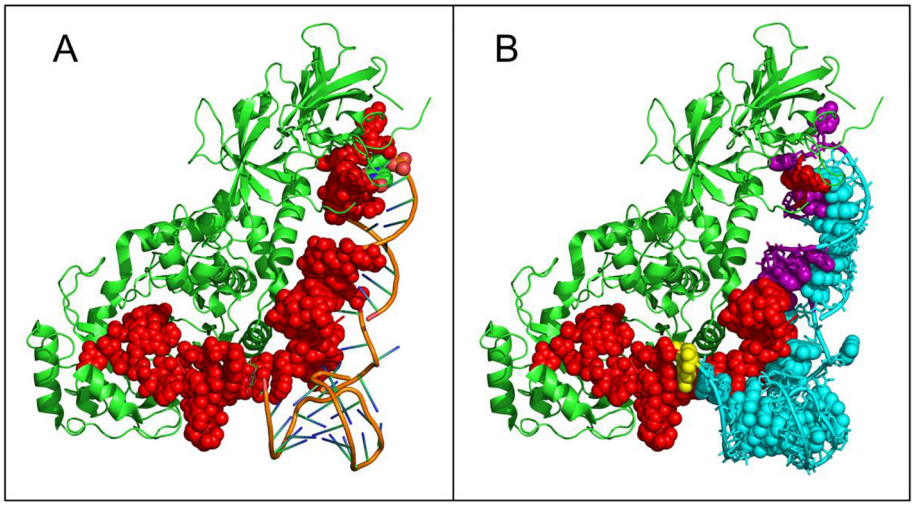

Fig. 9. Three-dimensional structure of tRNA (4JXZ_B, 71 nt) in complex with glutaminyl-tRNA synthetase (4JXZ_A). (A) The 34 protein-binding nucleotides, calculated based on SASA from the PDB structure, are shown as red spheres. The rest of the non-binding nucleotides are shown in orange cartoon. (B) The 24 protein-binding (TP) and the 36 non-binding (TN) nucleotides along with one false positive and 10 false negative nucleotides predicted by the developed classifier are represented in red, cyan, yellow and purple spheres, respectively. Protein is represented in green ribbons in both the structures. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

4.1. Sliding window optimization and feature selection

Binding sites are highly influenced by the surrounding adjacent sites as evident from the occurrence of contiguous stretches of nucleotides at the interface (Fig. 3). Hence, to incorporate the features of target and surrounding nucleotides in the sequence, we have employed a sliding-window strategy to represent each nucleotide of the non-redundant 194 RNA sequences comprising of structurally different types of RNA (tRNA, ssRNA, dsRNA and rRNA) in the training dataset. We have used different WS and evaluated performance for each sliding window to get optimum WS. Optimum performance for all the metrics is obtained at WS of 23

nt as evident from Table 4 and Fig. 6. An increase in all the performance metrics is observed with each incremental window till optimum WS is reached. Thus, it can be inferred that selection of window size plays a crucial role in improving the prediction performance while training.

Feature selection is an efficient way to discard information irrelevant for classification, and to reduce the search dimension. For this purpose, different global and local sequence features of RNA are calculated. These features are evalutaed to select the best features with more distinguishing capability. For performance assessment in RF, we have compared AUC and MCC obtained on 10-fold CV for all the individual feature and feature combinations (Table 3 and Fig. 5). NC-triplet and NC-quartet features improved the performance and increased the classification accuracy compared to other individual and combined features. This demonstrates the importance of triplet and quartet nucleotide compositions at the RNA binding sites.

4.2. Comparative analysis of predictive performance of ML classifiers

Ten different ML algorithms trained on PB-RNA194 dataset with selected features are compared based on 10-fold CV scores (Table 7). Among all the classifiers, RF provides the best results on 10-fold stratified CV. XGBoost with 0.05 learning rate and 500 estimators performed closer to RF. It also provides 0.93 AUC on 10-fold stratified CV. We have tested an ensemble of five best performing classifiers, i.e., Random Forest, XGBoost, SVM, KNN and MLP and achieved maximum 0.92 AUC and 0.67 MCC. The voting ensemble classifier performed well but does not increase the performance compared to RF classifier, which stand out as the most powerful method in discriminating binding and non-binding nucleotides.

4.3. Comparisions with existing predictive methods

We have compared the performance of the method developed in this study with a few existing methods for the prediction of protein binding sites in RNA. RNApin [53] utilizes composition profile of tri-nucleotides to discriminate the protein-interacting and non-interacting nucleotides. It achieved a maximum AUC of 0.88 and MCC of 0.62 with SVM model and MCC of 0.47 with RF model. The RF model developed in this study when tested on RNA-208 dataset shows improvement in performance measures with increase in MCC from 0.47 to 0.53 and increase in accuracy from 0.76 to 0.83 (Table 8). RPI-Bind [55] is a structure based method for identification of protein-RNA interactions. It achieved maximum accuracy and MCC of 0.63 and 0.27, respectively using nucleotide composition profile and RF algorithm. Including RNA local structural features, it further achieved a maximum accuracy and MCC of 0.71 and 0.4, respectively. On RNA-208 dataset, it achieved an accuracy of 0.81 and AUC of 0.88. On the other hand, our method achieved a higher accuracy of 0.83 and similar AUC of 0.88 on RNA-208 dataset (Table 8). RBPbinding [56] predicts protein binding regions in mRNA sequences using SVM model. It achieved high accuracy (0.87) and MCC (0.75) in cross-validation on a balanced dataset. However, it was designed to predict binding regions in mRNA sequences only. PRIdictor [57] combined both nucleotide and residue level information to predict binding sites. With SVM model, it achieved MCC of 0.69 in 10-fold cross validation, similar to that of our method. Few frameworks are developed recently to predict protein-binding nucleotide motifs in RNA sequences. iDeepS [58] is one such method. It applies one-hot encoding for the sequences and predicted secondary structures, and used these features in Convolutional Neural Networks (CNNs) to predict the protein-binding motifs in RNA. The method achieved high AUC of 0.87; however, it is a RBP-specific model and thus can predict binding targets only for the specific RBPs trained in their study. Moreover, it utilizes a RNA structure prediction tool for feature calculation, and thus depends on another external tool for the prediction, which is time consuming. These models were designed to predict binding motifs in the sequences rather than prediction of binding nucleotides, which is the focus of the present study. Hence, we do not perform direct comparision with these methods. All these existing methods developed using different datasets, binding site definition and features have their own advantages and disadvantages. Thus, fine-tuning of the features, refinement of input data and parameter optimization of ML algorithm provides considerable realm of scope for improvement in the overall prediction performance. In this study, we have performed parameter optimization of ten ML classifiers, and compared their performance. With RF model, we have achieved the best performance, which is comparable to that of few state-of-the-art methods developed to predict protein-binding sites in a given nucleotide sequence.

4.4. Predictive performance on validation sets and future direction

We have evaluated the model developed on PB-RNA194 training dataset on different validation and test datasets to check how well the model generalizes. We have extracted all RNA chains with sequence length below 500 nt from RNA-208 dataset [53], and discarded long ribosomal RNAs as they largely contain non-binding nucleotides (~95%). On this length-restricted dataset (RNA-150), we have achieved significantly better accuracy of 0.91 with high AUC and MCC of 0.95 and 0.75, respectively (Table 8). Further, we have merged the PB-RNA194 and RNA-150 datasets. On this combined dataset (RNA-344), we have performed 10-fold cross validation and achieved reasonably better performance with accuracy of 0.87, AUC and MCC of 0.94 and 0.71, respectively. On an independent test set, we have achieved accuracy of 0.83, specificity of 0.89, sensitivity of 0.70, MCC of 0.62 and AUC of 0.89. Although, we get reasonably good performance on our training dataset, test dataset shows average performance. The limitation of current method is the relatively smaller dataset used to train the classifier. This can be overcome in future with availability of more experimentally solved structures of RNA-protein complexes. Training and evaluation on larger datasets, capable of capturing more information at the sequence level, are required to improve the predictive performance on unknown data and to develop a more robust model. Further, different RNA types interact with proteins differently and hence possess distinct binding sites. Thus, with increase in structural data for each type of RNA, which is currently very limited, specific training models for each RNA type can be developed.

5. Conclusion

RNA interacts with proteins to drive many cellular processes. Knowledge of interaction sites is thus crucial to decipher the functional implications of binding. In this study, we have compared ten different ML approaches to predict the protein binding nucleotides in a RNA sequence. We have used window size optimization and feature selection-based approach to distinguish the interacting and non-interacting patterns. Triplet nucleotide compositions are used as primary feature for the prediction. RF classifier with properly tuned classifier parameters, such as bootstrap number, depth of trees and number of estimators provides the best prediction results compared to other classifiers. All the models are evaluated using repeated stratified 10-fold cross validation technique. Reasonably good predictive performance obtained on an independent test case shows the generalized predictive capability of the trained classifier on unknown sequences. The final model built with the best performing nucleotide-triplet feature based RF algorithm at

optimized window-size of 23 achieved an AUC of 0.93, accuracy of 0.85 and MCC of 0.70. The performance of the model can be enhanced in future with employment of more distinctive nucleotide features of RNA obtained from larger datasets.

Author contributions

R.P.B and A.A conceived the idea. A.A performed the experiments and analyses. K.S performed the analyses. A.A and S.K designed and implemented the webserver. R.P.B and A.A contributed to the manuscript writing.

CRediT authorship contribution statement

Ankita Agarwal: Conceptualization, Methodology, Investigation, Data curation, Software, Validation, Formal analysis, Writing – original draft. **Kunal Singh:** Investigation, Data curation, Formal analysis. **Shri Kant:** Software, Writing – original draft. **Ranjit Prasad Bahadur:** Conceptualization, Resources, Supervision, Project administration, Writing – review & editing.

Declaration of Competing Interest

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.

Acknowledgements

A.A acknowledges the fellowship and computing facilities received from IIT Kharagpur. K.S and S.K acknowledges the fellowship from IIT Kharagpur. R.P.B acknowledges the support from CSIR, India.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.csbj.2022.06.036.

References

[1] Anantharaman V, Koonin EV, Aravind L. Comparative genomics and evolution of proteins involved in RNA metabolism. Nucl Acids Res 2002;30:1427. https://doi.org/10.1093/NAR/30.7.1427.
[2] Kishore S, Luber S, Zavolan M. Deciphering the role of RNA-binding proteins in the post-transcriptional control of gene expression. Brief Funct Genomics 2010;9:391–404. https://doi.org/10.1093/BFGP/ELQ028.
[3] Kloetgen A, Münch PC, Borkhardt A, Hoell JI, McHardy AC. Biochemical and bioinformatic methods for elucidating the role of RNA-protein interactions in posttranscriptional regulation. Brief Funct Genomics 2015;14:102–14. https://doi.org/10.1093/bfgp/elu020.
[4] Armaos A, Zacco E, de Groot NS, Tartaglia GG. RNA-protein interactions: Central players in coordination of regulatory networks. BioEssays 2021;43:2000118. https://doi.org/10.1002/BIES.202000118.
[5] Licatalosi DD, Darnell RB. RNA processing and its regulation: global insights into biological networks. Nat Rev Genet 2010;11:75–87. https://doi.org/10.1038/nrg2673.
[6] Nahalka J. Protein–RNA recognition: cracking the code. J Theor Biol 2014;343:9–15. https://doi.org/10.1016/j.jtbi.2013.11.006.
[7] Lewis CJT, Pan T, Kalsotra A. RNA modifications and structures cooperate to guide RNA-protein interactions. Nat Rev Mol Cell Biol 2017;18:202–10. https://doi.org/10.1038/nrm.2016.163.
[8] Jolma A, Zhang J, Mondragón E, Morgunova E, Kivioja T, Laverty KU, et al. Binding specificities of human RNA-binding proteins toward structured and linear RNA sequences. Genome Res 2020;30:962–73. https://doi.org/10.1101/gr.258848.119.
[9] Mittal N, Scherrer T, Gerber AP, Janga SC. Interplay between posttranscriptional and posttranslational interactions of RNA-binding proteins. J Mol Biol 2011;409:466–79. https://doi.org/10.1016/j.jmb.2011.03.064.
[10] Mihailovic MK, Chen A, Gonzalez-Rivera JC, Contreras LM. Defective ribonucleoproteins, mistakes in RNA processing, and diseases. Biochemistry 2017;56:1367–82. https://doi.org/10.1021/acs.biochem.6b01134.
[11] Barta A, Jantsch MF. RNA in Disease and development. RNA Biol 2017;14:457–9. https://doi.org/10.1080/15476286.2017.1316929.
[12] Carey KT, Wickramasinghe VO. Regulatory potential of the RNA processing machinery: implications for human disease. Trends Genet 2018;34:279–90. https://doi.org/10.1016/j.tig.2017.12.012.
[13] Khalil AM, Rinn JL. RNA-protein interactions in human health and disease. Semin Cell Dev Biol 2011;22:359–65. https://doi.org/10.1016/j.semcdb.2011.02.016.
[14] Gebhart NN, Hardy RW, Sokoloski KJ. Comparative analyses of alphaviral RNA: protein complexes reveals conserved host-pathogen interactions. PLoS ONE 2020;15:e0238254–e. https://doi.org/10.1371/journal.pone.0238254.
[15] Li Z, Nagy PD. Diverse roles of host RNA binding proteins in RNA virus replication. RNA Biol 2011;8:305–15. https://doi.org/10.4161/rna.8.2.15391.
[16] Li D, Zhang J, Li X, Chen Y, Yu F, Liu Q. Insights into lncRNAs in Alzheimer's disease mechanisms. RNA Biol 2021;18:1037–47. https://doi.org/10.1080/15476286.2020.1788848.
[17] Butti Z, Patten SA. RNA Dysregulation in amyotrophic lateral sclerosis. Front Genet 2019;9:712. https://doi.org/10.3389/fgene.2018.00712.
[18] Neueder A. RNA-mediated disease mechanisms in neurodegenerative disorders. J Mol Biol 2019;431:1780–91. https://doi.org/10.1016/j.jmb.2018.12.012.
[19] Gebauer F, Schwarzl T, Valcárcel J, Hentze MW. RNA-binding proteins in human genetic disease. Nat Rev Genet 2021;22:185–98. https://doi.org/10.1038/s41576-020-00302-y.
[20] Zhang B, Babu KR, Lim CY, Kwok ZH, Li J, Zhou S, et al. A comprehensive expression landscape of RNA-binding proteins (RBPs) across 16 human cancer types. RNA Biol 2020;17:211–26. https://doi.org/10.1080/15476286.2019.1673657.
[21] Jonas K, Calin GA, Pichler M. RNA-binding proteins as important regulators of long non-coding RNAs in cancer. Int J Mol Sci 2020;21:2969. https://doi.org/10.3390/ijms21082969.
[22] Zhang Q, Wei Y, Yan Z, Wu C, Chang Z, Zhu Y, et al. The characteristic landscape of lncRNAs classified by RBP–lncRNA interactions across 10 cancers. Mol Biosyst 2017;13:1142–51. https://doi.org/10.1039/C7MB00144D.
[23] Wang J, Liu Q, Shyr Y. Dysregulated transcription across diverse cancer types reveals the importance of RNA-binding protein in carcinogenesis. BMC Genomics 2015;16(Suppl 7):S5–S. https://doi.org/10.1186/1471-2164-16-S7-S5.
[24] Gupta A, Gribskov M. The role of RNA sequence and structure in RNA–protein interactions. J Mol Biol 2011;409:574–87. https://doi.org/10.1016/j.jmb.2011.04.007.
[25] Parmar C, Grossmann P, Bussink J, Lambin P, Aerts HJWL. Machine learning methods for quantitative radiomic biomarkers. Sci Rep 2015;5:13087. https://doi.org/10.1038/srep13087.
[26] Field M, Hardcastle N, Jameson M, Aherne N, Holloway L. Machine learning applications in radiation oncology. Phys Imaging Radiat Oncol 2021;19:13–24. https://doi.org/10.1016/j.phro.2021.05.007.
[27] Deist TM, Dankers FJWM, Valdes G, Wijsman R, Hsu I-C, Oberije C, et al. Machine learning algorithms for outcome prediction in chemoradiotherapy: an empirical comparison of classifiers. Med Phys 2018;45:3449–59. https://doi.org/10.1002/mp.12967.
[28] Beunza JJ, Puertas E, García-Ovejero E, Villalba G, Condes E, Koleva G, et al. Comparison of machine learning algorithms for clinical event prediction (risk of coronary heart disease). J Biomed Inform 2019;97:. https://doi.org/10.1016/j.jbi.2019.103257103257.
[29] Inza I, Calvo B, Armañanzas R, Bengoetxea E, Larrañaga P, Lozano JA. Machine learning: an indispensable tool in bioinformatics. Bioinformatics Methods in Clinical Research. In: Matthiesen R, editor., Totowa, NJ: Humana Press; 2010, p. 25–48. doi: 10.1007/978-1-60327-194-3_2.
[30] Olson RS, La CW, Mustahsan Z, Varik A, Moore JH. Data-driven advice for applying machine learning to bioinformatics problems. Pac Symp Biocomput 2018;23:192–203. https://doi.org/10.1142/9789813235533_0018.
[31] Ward JJ, McGuffin LJ, Buxton BF, Jones DT. Secondary structure prediction with support vector machines. Bioinformatics 2003;19:1650–5. https://doi.org/10.1093/bioinformatics/btg223.
[32] Yang Z, Wang L, Young N, Chou KC. Pattern recognition methods for protein functional site prediction. Curr Protein Pept Sci 2005;6:479–91. https://doi.org/10.2174/138920305774329322.
[33] Patel N, Wang JTL. Semi-supervised prediction of gene regulatory networks using machine learning algorithms. J Biosci 2015;40:731–40. https://doi.org/10.1007/s12038-015-9558-9.
[34] Cai Y, Lin SL. Support vector machines for predicting rRNA-, RNA-, and DNA-binding proteins from amino acid sequence. Biochim Biophys Acta – Proteins Proteomics 2003;1648:127–33. https://doi.org/10.1016/S1570-9639(03)00112-2.
[35] Shao X, Tian Y, Wu L, Wang Y, Jing L, Deng N. Predicting DNA- and RNA-binding proteins from sequences with kernel methods. J Theor Biol 2009;258:289–93. https://doi.org/10.1016/j.jtbi.2009.01.024.
[36] Liu ZP, Wu L-Y, Wang Y, Zhang XS, Chen L. Prediction of protein–RNA binding sites by a random forest method with combined features. Bioinformatics 2010;26:1616–22. https://doi.org/10.1093/bioinformatics/btq253.
[37] Miao Z, Westhof E. A large-scale assessment of nucleic acids binding site prediction programs. PLOS Comput Biol 2015;11:e1004639.
[38] Nguyen TTD, Le NQK, Kusuma RMI, Ou YY. Prediction of ATP-binding sites in membrane proteins using a two-dimensional convolutional neural network. J

Mol Graph Model 2019;92:86–93. https://doi.org/10.1016/j.jmgm.2019.07.003.
[39] Breiman L. Random Forests. Mach Learn 2001;45:5–32. https://doi.org/10.1023/A:1010933404324.
[40] Chen T, Guestrin C. XGBoost: A scalable tree boosting system. Proc ACM SIGKDD Int Conf Knowl Discov Data Min 2016;13-17-Augu:785–94. doi: 10.1145/2939672.2939785.
[41] Friedman JH. Greedy function approximation: A gradient boosting machine. Ann Stat 2001;29:1189–232. https://doi.org/10.1214/aos/1013203451.
[42] Vapnik VN. The Nature of Statistical Learning Theory. Springer New York; 1995. doi: 10.1007/978-1-4757-2440-0.
[43] Nithin C, Mukherjee S, Bahadur RP. A non-redundant protein–RNA docking benchmark version 2.0. Proteins Struct Funct Bioinforma 2017;85:256–67. https://doi.org/10.1002/prot.25211.
[44] Berman HM. The protein data bank. Nucl Acids Res 2000;28:235–42. https://doi.org/10.1093/nar/28.1.235.
[45] Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol 1990;215:403–10. https://doi.org/10.1016/S0022-2836(05)80360-2.
[46] Wang L, Huang C, Yang MQ, Yang JY. BindN+ for accurate prediction of DNA and RNA-binding residues from protein sequence features. BMC Syst Biol 2010;4:S3. https://doi.org/10.1186/1752-0509-4-S1-S3.
[47] Walia RR, Caragea C, Lewis BA, Towfic F, Terribilini M, El-Manzalawy Y, et al. Protein-RNA interface residue prediction using machine learning: an assessment of the state of the art. BMC Bioinf 2012;13:89. https://doi.org/10.1186/1471-2105-13-89.
[48] Bahadur RP, Zacharias M, Janin J. Dissecting protein-RNA recognition sites. Nucleic Acids Res 2008;36:2705–16. https://doi.org/10.1093/nar/gkn102.
[49] Hubbard SJ, Thornton JM. NACCESS. Comput Progr 1993.
[50] Lee B, Richards FM. The interpretation of protein structures: Estimation of static accessibility. J Mol Biol 1971;55. https://doi.org/10.1016/0022-2836(71)90324-X.
[51] Pedregosa F, Grisel O, Weiss R, Passos A, Brucher M, Varoquax G, et al. Scikit-learn: machine learning in python. J Mach Learn Res 2011;12:2825–30.
[52] Baldi P, Brunak S, Chauvin Y, Andersen CAF, Nielsen H. Assessing the accuracy of prediction algorithms for classification: an overview. Bioinformatics 2000;16:412–24. https://doi.org/10.1093/bioinformatics/16.5.412.
[53] Panwar B, Raghava GPS. Identification of protein-interacting nucleotides in a RNA sequence using composition profile of tri-nucleotides. Genomics 2015;105:197–203. https://doi.org/10.1016/j.ygeno.2015.01.005.
[54] Loughlin FE, Lukavsky PJ, Kazeeva T, Reber S, Hock EM, Colombo M, et al. The solution structure of FUS bound to RNA reveals a bipartite mode of RNA recognition with both sequence and shape specificity. Mol Cell 2019;73:490–504.e6. https://doi.org/10.1016/j.molcel.2018.11.012.
[55] Luo J, Liu L, Venkateswaran S, Song Q, Zhou X. RPI-Bind: a structure-based method for accurate identification of RNA-protein binding sites. Sci Rep 2017;7:614. https://doi.org/10.1038/s41598-017-00795-4.
[56] Choi D, Park B, Chae H, Lee W, Han K. Predicting protein-binding regions in RNA using nucleotide profiles and compositions. BMC Syst Biol 2017;11:16. https://doi.org/10.1186/s12918-017-0386-4.
[57] Tuvshinjargal N, Lee W, Park B, Han K. PRIdictor: protein–RNA interaction predictor. Biosystems 2016;139:17–22. https://doi.org/10.1016/j.biosystems.2015.10.004.
[58] Pan X, Rijnbeek P, Yan J, Shen HB. Prediction of RNA-protein sequence and structure binding preferences using deep convolutional and recurrent neural networks. BMC Genomics 2018;19:511. https://doi.org/10.1186/s12864-018-4889-1.