Genomics 112 (2020) 2623–2632

Contents lists available at ScienceDirect

Genomics

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

RPINBASE: An online toolbox to extract features for predicting RNA-protein interactions

Mahsa Torkamanian-Afshar[a,b], Hossein Lanjanian[b], Sajjad Nematzadeh[c], Maryam Tabarzad[d], Ali Najafi[e], Farzad Kiani[f], Ali Masoudi-Nejad[a,b,*]

[a] Laboratory of Systems Biology and Bioinformatics (LBB), Department of Bioinformatics, Kish International Campus, University of Tehran, Kish Island, Iran
[b] Laboratory of Systems Biology and Bioinformatics (LBB), Institute of Biochemistry and Biophysics, University of Tehran, Tehran, Iran
[c] Department of Computer Technologies, Beykent University, Istanbul, Turkey
[d] Protein Technology Research Center, Shahid Beheshti University of Medical Sciences, Tehran, Iran
[e] Molecular Biology Research Center, Systems Biology and Poisonings Institute, Tehran, Iran
[f] Department of Computer Engineering, Engineering and Architecture Faculty, Istanbul Arel University, Istanbul, Turkey

ARTICLE INFO

Keywords:
Biological applications
Feature extraction
Machine learning
RNA-protein interactions
RPINBASE

ABSTRACT

Feature extraction is one of the most important preprocessing steps in predicting the interactions between RNAs and proteins by applying machine learning approaches. Despite many efforts in this area, still, no suitable structural feature extraction tool has been designed. Therefore, an online toolbox, named RPINBASE which can be applied to different scopes of biological applications, is introduced in this paper. This toolbox employs efficient nested queries that enhance the speed of the requests and produces desired features in the form of positive and negative samples. To show the capabilities of the proposed toolbox, the developed toolbox was investigated in the aptamer design problem, and the obtained results are discussed. RPINBASE is an online toolbox and is accessible at http://rpinbase.com.

1. Introduction

RNA and protein are two major biological macromolecules and their interaction can have profound effects in different fields including the regulation of gene expression [1–4], protein synthesis [5,6], viral replication, and cellular defense mechanism [7–10]. Despite the importance of these structures, it is difficult to identify their interaction using experimental methods, as they are expensive and time-consuming. Thus, there is an increasing need to have machine learning approaches to accurately predict these interactions [11–16]. Structural feature extraction is one of the most important preprocessing steps in this area. Recent developments in clarifying RNA and protein structural features have increased the need to design different tools, aiming at the investigation of interactions between RNA and protein. Despite several studies on the issue, no suitable structural feature extraction tool has been designed, yet.

RPINBASE is a repository of all RNA-protein complexes stored in the Protein Data Bank (PDB) [17] with a toolbox to quickly execute the queries, generate and download ready- to- learn datasets. It is available at http://rpinbase.com for free. The query executor module of RPINBASE contains a wide range of features (Supplementary File 1: Macromolecules Features) related to the primary and secondary structural elements of RNA and protein macromolecules. At the level of the primary structure, it allows queries to request protein and RNA sequences with different lengths and searches various substrings between the existing sequences. Moreover, the phylogeny data on the family and clan of protein sequences are available. At the secondary structural level, it allows queries that contain different relevant secondary structure features of RNA and protein macromolecules. The query structure of RPINBASE is designed based on the nested object concepts in object-oriented programming [18–20] to respond to diverse demands related to the study of different aspects of RNA-protein binding. Therefore, one complex query can be broken down into a series of logical subqueries. These subqueries are created based on the characteristics of the primary and secondary structures of macromolecules. Thus, these subqueries are aggregated to form one nested query and then, this query is sent to the database. There is no dispute over the importance of machine learning as a fast-growing approach in this field. Accordingly, RPINBASE can be

* Corresponding author at: Laboratory of Systems Biology and Bioinformatics (LBB), Institute of Biochemistry and Biophysics, University of Tehran, Tehran, Iran.
E-mail addresses: torkamanianafshar@ut.ac.ir (M. Torkamanian-Afshar), h.lanjanian@ut.ac.ir (H. Lanjanian), sajjadnematzadeh@beykent.edu.tr (S. Nematzadeh), m_tabarzad@sbmu.ac.ir (M. Tabarzad), najafi74@bmsu.ac.ir (A. Najafi), amasoudin@ut.ac.ir (A. Masoudi-Nejad).

https://doi.org/10.1016/j.ygeno.2020.02.013
Received 5 November 2019; Received in revised form 4 January 2020; Accepted 13 February 2020
Available online 21 February 2020
0888-7543/

an option to create downloadable positive and negative datasets with non-redundant sequences to train and test classifiers. The type of content of these datasets can be chosen by users as raw sequences or pre-extracted feature vectors. Here, non-interacting (negative) samples are generated using protein clans while interacting (positive) samples are generated utilizing atomic-distances.

2. Related work

In this section, we have investigated the resources used in the scope of RNA-protein interaction. We divided the databases into two categories as following:

(i) Protein annotation databases: several databases store RNA–protein interactions along with annotations on protein structures such as the Nucleic Acids-Protein Interaction Database (NPIDB) [21], the RNA-Binding Protein Database (RBPDB) [22], the noncoding RNAs and protein related biomacromolecules interaction database (NPInter) [23], Protein-RNA interaction predictions for model organisms with supporting experimental data (RNAct) [24], and the Protein-RNA Interface Database (PRIDB) [25]. By integration different data sources and their unification, some of these databases have tried to be comprehensive and broader, while there are some others have only focused on one category. The majority of these databases are not capable of running desired queries of RNA and protein structural features on the RNA-protein complexes.

(ii) RNA annotation databases: some databases such as the Nucleic Acid Database (NDB) [26], the RNA secondary structure and statistical analysis database (RNA Strand) [27], the RNA Characterization of Secondary Structure Motifs (RNA CoSSMos) [28], the RNA FRAgment search engine & dataBASE (RNA FRABASE) [29], and the Universe of RNA Structures DataBase (URSDB) [30], have provided beneficial data on RNA structures and structural motifs. They have been used to analyze RNA secondary structures in different studies. The advanced search of existing databases has a linear query structure which is inconvenient for complex requests.

In addition, there are other types of tools such as RPISeq [31,32] and RNA–Protein interaction predictor (RPI-Pred) [33] being developed to study the interaction between RNA and protein macromolecules with pre-defined features. Generally, these predictive tools focus on some of the pre-defined features of macromolecules.

Despite the need to execute queries for RNA and protein features, simultaneously in the case of investigating complexes, these databases have just focused on RNAs or protein macromolecules separately. On the other hand, existing databases have linear query structures. All conditional statements in this query structure have the same priority. Consequently, the linear query structure is inconvenient for complex requests, while the nested query structure enables one to achieve this goal. Another challenge about most databases in the field of nucleic acid-protein interaction is that they only focus on interacting (positive) samples and do not contain validated non-interacting (negative) samples. Hence, the majority of these studies have addressed the challenge by using atomic-distance with an arbitrary threshold or by random pairing [31,33–40]. In a study by Cheng et al. [41] a scoring method was applied using Gene Ontology (GO) [42], the Protein Families database (Pfam) [43], and the Universal Protein Resource (UniProt) [44] databases to generate negative datasets.

Given the pros and cons of the mentioned tools and databases, we aimed to develop the RPINBASE accordingly. Utilizing RPINBASE has several merits including (i) generating queryable positive samples from PDB complexes (based on multilevel atomic distances: 3.4 Å, 3.7 Å, 5 Å, 7 Å, and 10 Å), (ii) generating queryable artificial proposed negative samples (based on family and clan), (iii) providing a wide range of extracted features for primary and secondary structures of RNA and protein macromolecules, (iv) selecting efficient and high-performance dataset powered by preprocessed and stored data to download feature vectors quickly, (v) providing powerful query engine to eliminate duplicated queries by nested structures, and finally (vi) the possibility of sample filtering by PDB information and structural features of macromolecules.

3. Materials and methods

Stored data on RPINBASE were prepared as the following: initially, the contents of PDB structures were extracted, cleaned, and converted to a suitable format for database design. Then, the values of structural features of the macromolecules were calculated and appended to the database and finally, the positive and negative datasets were constructed.

3.1. Data gathering and preprocessing

RPINBASE is a repository of RNA-protein complexes. Initially, all structures were extracted from the PDB and then, the PDB files, with at least one protein and one RNA sequence, were stored as the target samples. The database of our toolbox analyzed 2258 complexes (since June 2018). In the preprocessing step, duplicated chains and sequences which contained unknown alphabets were recognized and ignored. For example, the complex with the PDB id '3oij' has two protein chains (A, B) and two RNA chains (C, D). Moreover, A, B in protein chains and C, D in RNA chains were identical. Consequently, we removed one protein and one RNA chain and then, obtained the B—C as a positive and non-redundant sample. To investigate which chains of proteins directly interacted with the RNA chains, the analysis of 3D structures of macromolecules in the complexes was carried out. Different thresholds were used to differentiate the interaction between chains of macromolecules and have subtle effects on the qualities of various methods [45]. Suresh et al. [33] and Adjeroh et al. [46] used the threshold of 3.4 Å. In another studies, BindN+ [47], RISP [48] and PRBR [49] used the 3.5 Å cutoff. ProteRNA [50] and RNABindR [51] used the threshold of 5 Å. Few studies have used larger thresholds such as 7 Å [35] and 8 Å [31]. It should be noted that smaller cutoffs are used most often in prediction methods. In this study, we applied five frequently-used thresholds between 3.4 Å and 10 Å. Therefore, if the atomic distance between RNA and a chain of protein in the PDB files was less than the selected threshold, these two chains were identified as interacting pairs. One out of 3.4 Å, 3.7 Å, 5 Å, 7 Å, and 10 Å was selected as a threshold to distinguish strongly interacting pairs. Users can also choose one of these distances to select highly interacting pairs. Afterward, the secondary structures of the protein sequences assigned by Define Secondary Structure of Proteins (DSSP) [52] were extracted from PDB. The Protein Secondary Structure Prediction server (JPred) [53] was used to predict the secondary structure of sequences that did not include any DSSP assignment. Besides, the RNA fold from the Vienna package [54] was applied to predict the secondary structure of RNA sequences. Finally, the family and clan of protein sequences were extracted from the Pfam. The process of storing data on sequences is illustrated in Fig. 1.

3.2. Structural processing of macromolecules

The main aim of RPINBASE is to provide a suitable resource to construct the desired datasets based on diverse features of RNA macromolecule or protein macromolecule in a complex. Therefore, the final stored data in the last step were processed to extract all possible features of the first and secondary structures. Furthermore, phylogenic information of the macromolecules was extracted from the Pfam database. Hence, it was possible to create the desired dataset based on a specific family or clan. In the case of a primary structure, RPINBASE supported substrings and lengths of RNA and protein sequences. On the secondary structure level, this toolbox contained the information of RNA secondary structure elements (Stem, Hairpin, Bulge, Internal loop,

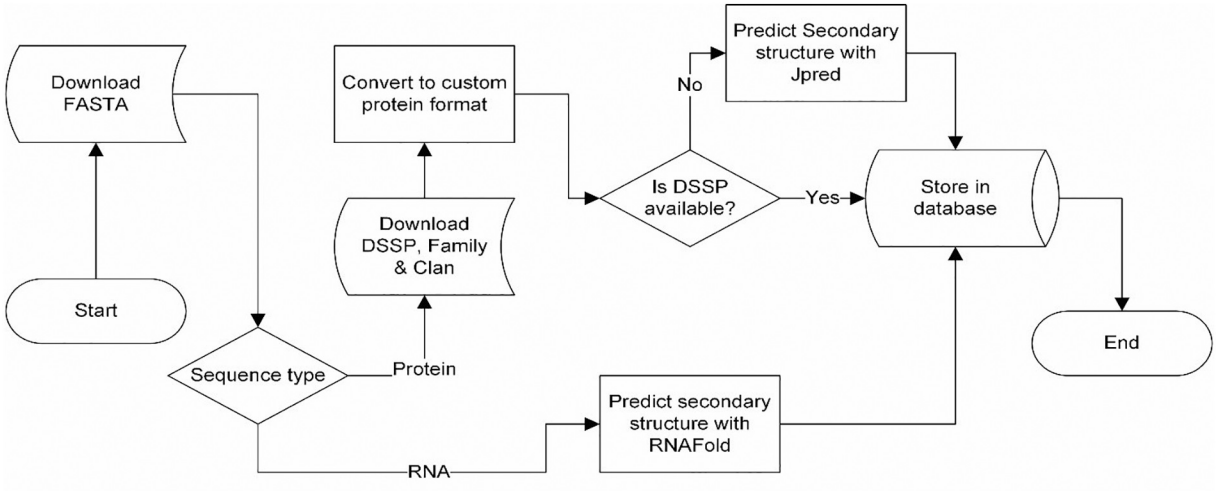

Fig. 1. The process of storing macromolecules; contains external data, processes and conditions.

and Multi-Loop), and protein secondary structure (Alpha Helix, Beta Sheet, and Coil). According to the DSSP algorithm, the assignment of the specified protein secondary structure has eight elements. These eight letters were translated into three letters to ease their interpretation [55]. Here, the results of a study by Liu et al. [56] were used to calculate the number of parallel and antiparallel beta sheets. In addition, general information concerning structural elements was processed using different algorithms [57–59] and their results were presented to be used in queries. Moreover, the values of some features of protein chains were calculated using "protein encoding toolbox" [60] and appended to the database. Here, those proteins with a sequence length of fewer than 20 amino acids were excluded. Supplementary File 1: Macromolecules Features, describes these features.

3.3. Generating negative and positive datasets

In the present study, the family and clan of protein sequences were applied to generate non-interacting pairs. To this aim, an infrastructure was constructed to interact with RNA-protein pairs by observing experimental reports and constructing non-interacting RNA-protein pairs based on family and clan of protein sequences. There were 9367 protein families, and these families were classified into 604 clans in the Pfam database. Interestingly, RNA-protein complexes only covered 111 clans and 620 families in the Pfam. Owing to this fact, we introduced the idea of using family and clan of protein sequences to generate non-interacting pairs (negative samples). We took into account only the clans of RNA binding proteins to ensure that only the relevant features are considered. Moreover, the distinction between positive and negative samples was made based on specific features of macromolecules being capable of composing the RNA-protein complex. Positive samples were composed by combining protein chains with RNA sequences of complexes whose distances were less than those in the selected thresholds of 3.4 Å, 3.7 Å, 5 Å, 7 Å, and 10 Å. Negative samples were generated by selecting RNA sequences in a complex and combining them with the protein sequences that had not been spotted in the same clan of the given complex. Consequently, users could construct positive and negative datasets based on the desired family and clan of protein sequences. Fig. 2 provides an overview of the formation of positive and negative samples.

3.4. Nested query

In the present toolbox, recursive functions were applied to evaluate subqueries in the nested structures. It means that various subqueries can be created according to the features of the primary and secondary structures of macromolecules. Then, these subqueries were aggregated and sent to the database. Nested queries were defined in the form of nested objects. Each query object is the parent of its own child objects. In other words, with the assumption of the parent's conditions, each child node appends more details in a total query and redefines its parent more accurately. This type of query was implemented using recursive logic [61,62]. This nonlinear query structure is efficient when it comes to complex queries and mixed conditional statements on different types of features.

3.5. Implementation

The 'database first' approach was chosen, and the data models were created from tables. The RPINBASE application was developed drawing upon a standard three layers of architecture. First, the data access layer was implemented using Microsoft SQL Server which has the data storing and retrieving tasks. The second layer is the business layer. This layer was developed using Active Server Pages (ASP.Net) and is running on Internet Information Services (IIS). This layer was based on the Model-View-Controller (MVC) and Entity Framework. All queries are sent from the business layer to the data access layer and the data access layer responds appropriately to the query. The third layer is the presentation layer. Every query is received from the presentation layer via web services and is validated and forwarded to the data access layers. The results come from the database and temporary datasets are generated by related services and functions. Then these datasets are converted to the JavaScript Object Notation (JSON) format and are transmitted to the presentation layer. The presentation layer (web interface) is created with HyperText Markup Language (HTML) and AngularJS which provides a dynamic script execution on the clients' browsers.

4. Results and discussion

RPINBASE is a novel source of RNA-protein interactions that integrates the annotation of protein and RNA structural elements. Furthermore, this toolbox can be used to generate three types of data sets: 'RNA', 'protein' and 'RNA-protein interaction'. In this section, first, the potentials of using RPINBASE are discussed. Second, the performance comparison between using our method and a random one is indicated. Third, using different thresholds for generating positive samples are evaluated. Finally, two cases of using RPINBASE are demonstrated.

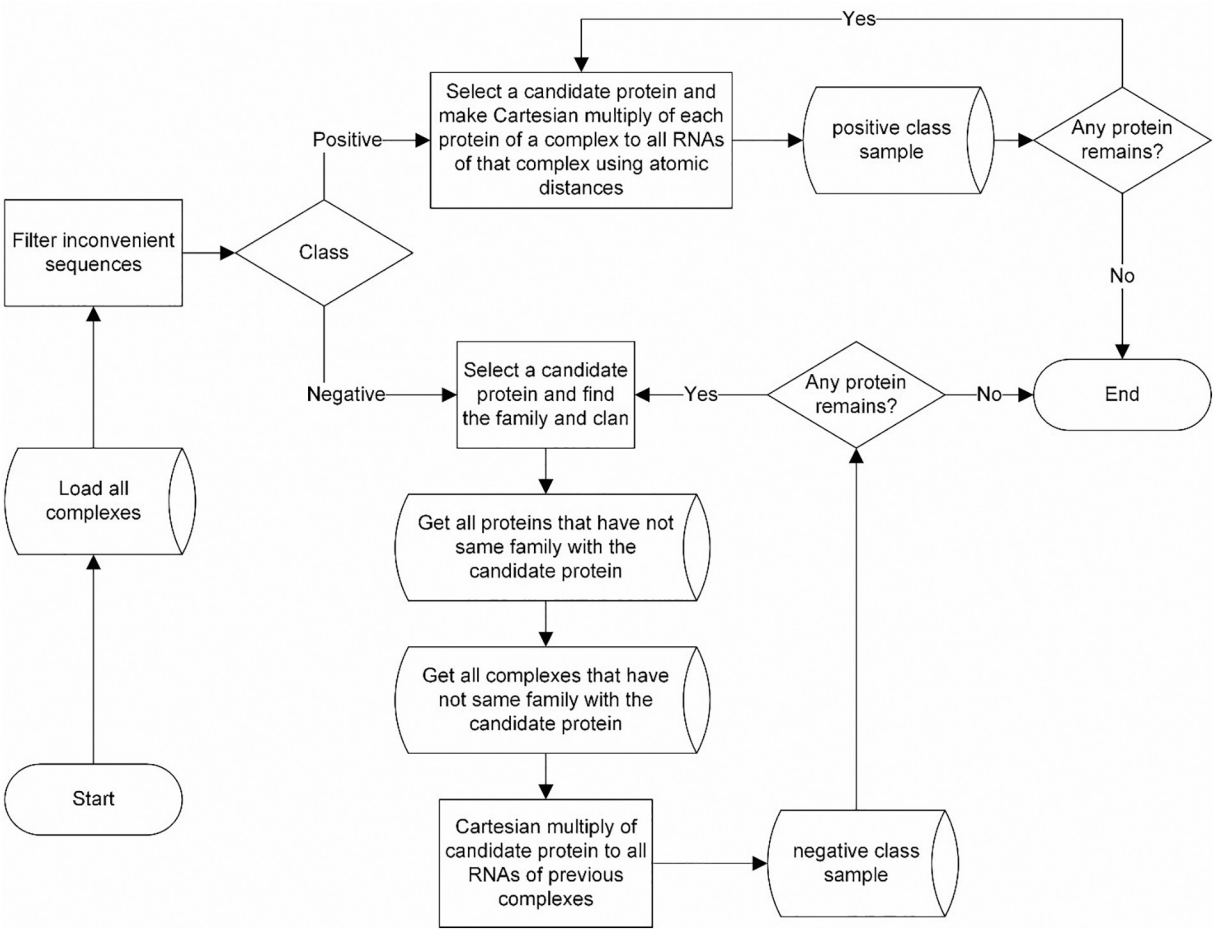

Fig. 2. Generation of positive and negative samples.

4.1. Web interface

RPINBASE has been developed in a user-friendly manner consisting of three main parts of 1-'Select mode', 2-'Make a query' and 3- 'Result & download'.

The 'select mode' is the entry point of the toolbox and users can create three types of datasets: 'RNA-protein pairs', 'protein' and 'RNA'. In the 'make a query' stage, users create a query based on a selected mode to request a dataset from the toolbox utilizing specific features. These pages contain logical blocks to create specific queries. These blocks are executed according to the rules and precedence of parentheses. It means that all items in parentheses are evaluated independently. Items with nested parentheses are evaluated from inside to outside. In the 'Result & download' stage, users have access to statistics of the created set and can download and save the results in three forms (Sequence and PDB, Dataset for Machine Learning and Feature

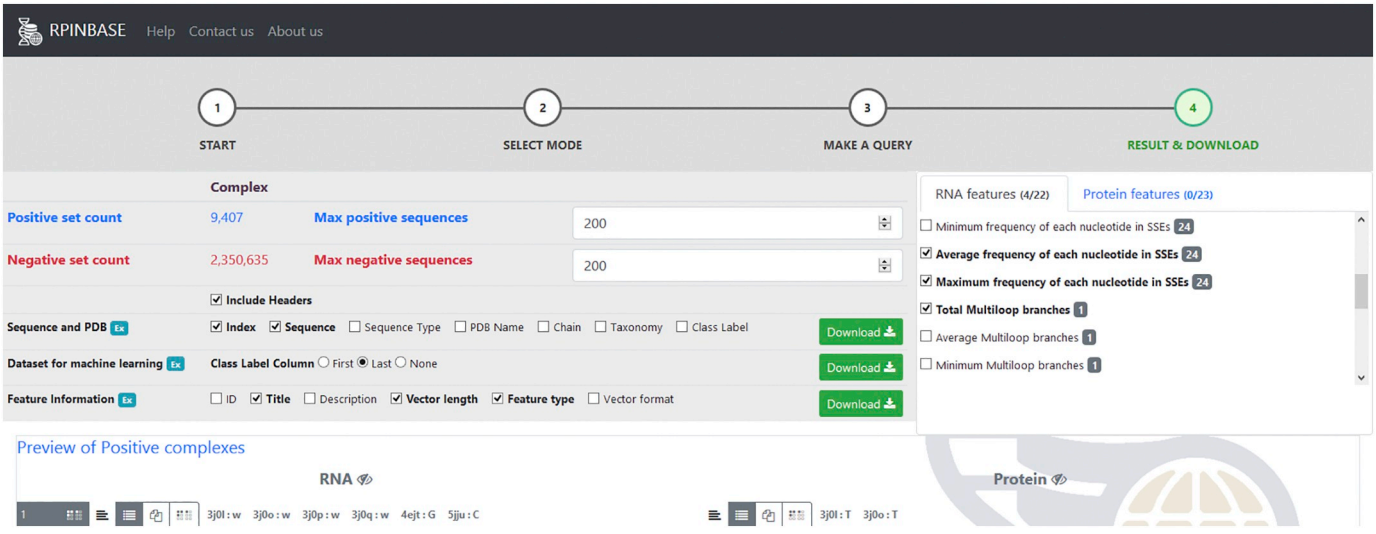

Fig. 3. A screenshot of the result page.

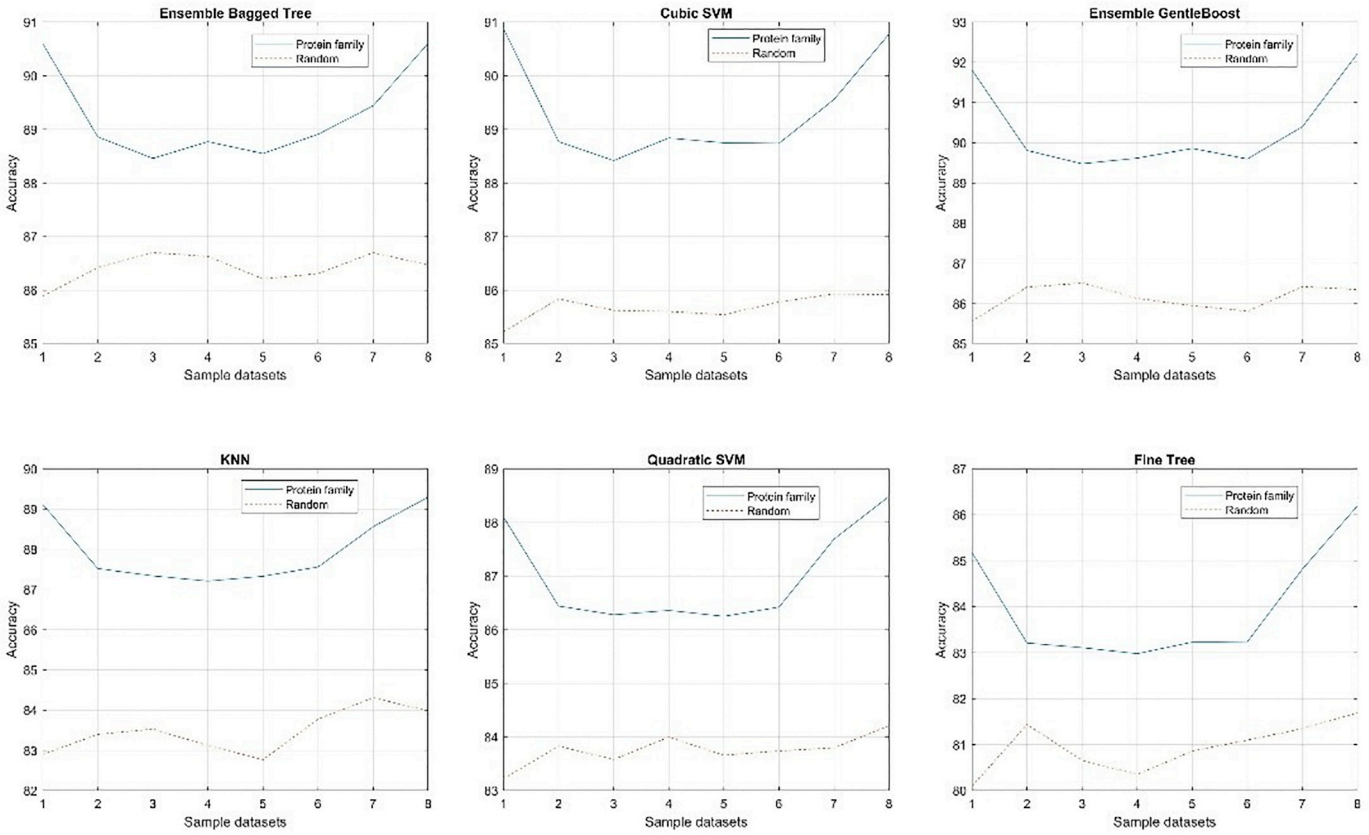

Fig. 4. The accuracy of machine learning algorithms on generated datasets.

Information). Fig. 3 shows the result page. First, in "Sequence and PDB" section, users can download the sequences of macromolecules. Second, in "Dataset for Machine Learning" section, users can select desired features and download a dataset for machine learning algorithms. Third, in "Feature Information" section, users can download additional information concerning the selected features. The format of output files is comma-separated CSV with selected columns which can be chosen by users. Moreover, this toolbox provides a 'Help' section with comprehensive information and a practical example of RPINBASE's web interface usage (Supplementary File 2: Help).

4.2. Positive and negative samples and their feature vectors

The data storage section of this toolbox contains complex information that discriminates RNA-protein pairs as positive samples by calculating atomic distance, negative samples by protein clans, and extracted feature vectors for each sequence. To generate complex datasets that contain positive and/or negative samples, RPINBASE concatenates user-preferred feature vectors of sequences in filtered sample pairs. Currently, this tool provides hand-designed features. In the other versions of this toolbox, deep learning-based features are applied to identify other important characteristics of RNA-protein interactions. These features are extracted using the deep layers from the various deep learning models such as Recurrent Neural Network (RNN), Convolutional neural network (CNN(, Autoencoder (AE), and Deep Belief Network (DBN) [63–65]. They have a broad range of applications in biological problems like aptamer-protein interaction [66], protein-protein interaction [67], and drug-target interaction prediction [68]. Contrary to hand-designed features, deep learning-based features generalize well to learn abstract feature representation from the raw data.

4.2.1. Performance comparison of proposed datasets

Generally, in supervised machine learning problems with a proper training process, acquiring desired true results against false ones proves that the machine succeeds to distinguish between samples by given features. On the other hand, it indicates that labels of samples are marked correctly. We created a series of datasets from RPINBASE that contain all positive samples, without any filtering except for distance, along with balanced negative samples and give them various binary classifier algorithms. The acceptable and logical relation of features and classes can be demonstrated by the performance comparison of training and testing results [69].

Precision (PRE), recall (REC), accuracy (ACC), and F-Score (FSC) metrics are used to measure the performance. These metrics are evaluated as follows: PRE = TP / (TP + FP), REC = TP / (TP + FN), ACC = (TP + TN) / (TP + TN + FP + FN), FSC = 2 * (PRE *REC) / (PRE + REC), where TP is the number of True Positives, TN is the number of True Negatives, FP is the number of False Positives, and FN is the number of False Negatives. The performance of various machine learning algorithms on the protein family dataset and random one was evaluated. At first, a positive set was generated. This set included the combination of protein chains with RNA sequences of complexes whose distances were less than or equal to 3.4 Å. In addition, we generated sixteen negative sets, eight sets of which were composed of family and clan method, while the other eight sets were composed utilizing a random method. The protein family method is mentioned in Section 3.3. "Generating Negative and Positive datasets". In the random method, negative samples included a combination of random protein chains with RNA sequences that were absent in the positive set. The positive set was combined with all negative sets, one by one, to create datasets for evaluation in machine learning algorithms. Further, protein and RNA with sequence lengths of < 20 were excluded from these positive and negative datasets. Then, the sequential and structural features of RNA and protein macromolecules were extracted (Supplementary File 1: Macromolecules features). Finally, for each dataset, a five-cross validation was performed on Fine Tree, Quadratic

Table 1
performance evaluation of family method and random method.

	Family & clan method				Random method			
	REC	PRE	FSC	ACC	REC	PRE	FSC	ACC
Fine tree	83.80%	84.14%	83.97%	84%	82.87%	79.80%	81.31%	80.94%
Quadratic SVM	86.10%	87.69%	86.89%	87%	84.74%	83.10%	83.92%	83.75%
Cubic SVM	90.59%	88.39%	89.48%	89.34%	88.87%	83.54%	86.13%	85.68%
Fine KNN	90.96%	85.88%	88.35%	87.99%	89.18%	80.05%	84.35%	83.47%
Ensemble bagged tree	87.82%	90.45%	89.12%	89.28%	86.26%	86.53%	86.40%	86.42%
Ensemble gentle	89.71%	90.88%	90.35%	90.35%	87.09%	85.47%	86.28%	86.14%

Table 2
Results of the executed query.

Index	Sequence	PDB name
1	AGGUGCUGCAUGGCCGUCGUCAACGACGUCUGGUCAGCAUGGCCC	1MVR:A
2	AUCGAAUCGCCACCUACAAGACUGGAGCUUGCUCCCUCGAAGGCGCCAAGUAUAUUCAUGAUCACAAGACA	6AZ3:6
3	GCGGAUUUAACUCAGUUGGGAGAGCGCCUUCGGGAGGUCCUGUGUUCGAUCCACAGAAUUCGCACCA	3ICQ:D&E
4	GGAGGUAGUAGGUUGUAUAGUAGUAAGACCAGACCCUAGACCAAUUCAUGCC	6BU9:B
5	GGCAUGAAUUGGUCUAGGGUCUGGUCUUACUACUAUACAACCUACUACCUCC	6BU9:C
6	GGGAGUAUAUGGGCGCACUUCGGUGACGGUACAGGCUCCU	4PMI:A
7	GGGAUGCGUAGGAUAGGUGGGAGCCGCAAGGCGCCGGUGAAAUACCACCCUUCCC	1MZP:B
8	GGGGGCGGAAAGGAUUCGACGGGGACUUCGGUCCUCGGACGCGGGUUCGAUUCCCGCCGCCUCCACCA	1P6V:B&D
9	GUCACACCAUGGGAGUGGGUAUUCAUGACUGGGGUGAAGU	1ZN0:C;1ZN1:C
10	GUGCCGGAAGGUCAAGGGGAGGGGUGCAAGCCCCGAACCGAAGCCCCGGUGAAC	2OM7:G

SVM, Cubic SVM, Fine KNN, Ensemble Bagged Tree, and Ensemble Gentle classifiers. The comparative accuracies of machine learning algorithms between our method and the random one on the generated datasets are depicted in Fig. 4.

Also, the evaluation of average performances between our method and the random one is provided in Table 1. The findings revealed that the performances of all machine learning algorithms for protein family negative sets were better than those found in random negative sets.

It was also evident that the Ensemble Gentle classifier performed better than the other classifiers on these datasets. These findings demonstrate that the present method can improve the predictive performance of RNA-protein interaction.

In addition, further analyses were performed to illustrate the advantage of generated negative samples using the family & clan method. The Ensemble Gentle classifier was used to perform these analyses. We divided them into the following cases:

Case 1: the model was trained using negative samples generated by the family & clan method and performed on a test set using negative samples generated by random method.

Case 2: the model was trained using negative samples generated by the random method and performed on a test set using negative samples generated by the family & clan method.

By comparing the training results of these two methods, the family & clan method showed better performance. Along with this, in the test results of these two cases, the advantages of the family & clan method were identified. Also, due to the True Negative Rate (TNR) results of training models and these two cases, we observed that the random negative datasets are noisier than family & clan datasets. Because the difference between the TNR value of case1 and TNR value of trained model using the clan method is higher than the TNR value of case2 and TNR value of trained model using the random method (Supplementary File 3: Analyses. Table S1).

The suggested method for generating negative samples implies that any RNA is less likely to interact with two proteins from different clans. When the constructed samples in the negative and positive sets were compared, no common pairs were found in both positive and negative sets. Further, to support this claim, first, we considered the set of all protein-binding RNAs. The similarity between any two RNAs was computed using the normalized Levenshtein distance metric. If they shared more than 80% sequence identity are kept in the set, otherwise, they would be discarded. Then, we followed whether these two sequences bind to proteins from the same clan or not. In order to evaluate this, the normalized penalty value was computed according to this equation as follows: Normalized penalty = 1- abs (number of Clan1ExceptClan2 in RNA1 – number of Clan2ExceptClan1 in RNA2) / (number of Clan1ExceptClan2 in RNA1 + number of Clan2ExceptClan1 in RNA2).

A normalized penalty is a metric in similar RNAs, which shows the proportion of non-shared clans relative to the union of their clans. If the Normalized penalty value is zero, it will indicate that clans of RNA are the subset of its mutual pair; otherwise, they bind to different protein clans (Supplementary File 3: Analyses). We repeated this metric to all pairs of RNAs. The results were reasonable and showed that only about 1% of this set could not bind to the same clan. Also, by investigating this error rate in some of the pairs, we observed that although the primary structure of the RNA sequences was similar, the secondary structures were different from each other.

4.3. Performance evaluation of different cutoffs

We generated five positive datasets. These sets included the combination of protein chains with RNA sequences of complexes whose distances were less than or equal to 3.4 Å, 3.7 Å, 5 Å, 7 Å, and 10 Å. Also, a negative set was generated using the family & clan method. The negative set was combined with all positive sets, one by one, to create datasets for evaluation in the Fine Decision Tree classifier. Table S2 in Supplementary File 3 shows these results. According to Table S2, smaller thresholds are recommended as an important criterion for constructing positive interacting pairs. The prediction qualities on the larger thresholds are characterized by a decrease in sensitivity compared with the smaller thresholds.

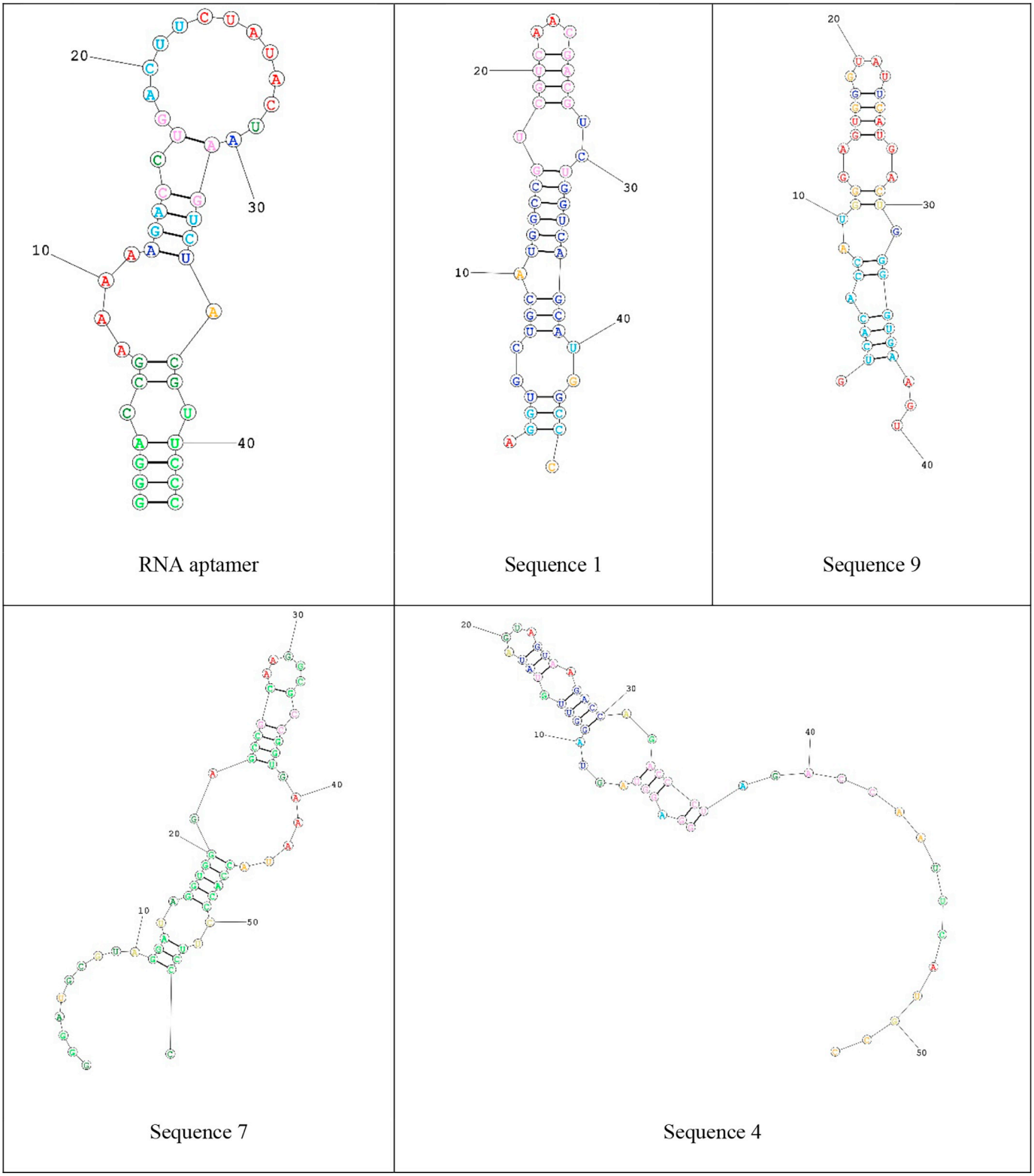

Fig. 5. The RNA structure-predicted secondary structures of RNA aptamer to PSMA and RNA sequences of RPINBASE.

4.4. Case studies

Two cases of using RPINBASE and its web interface are indicated below. The first case demonstrates the advantages of the RPINBASE interface, and the second case indicates the use-case examples of this toolbox.

4.4.1. Case I: utility of the user interface

In this case, a comparison of the search criteria of three different databases (RCSB PDB [17], URSDB [30], and RPINBASE) is shown as the following:

Evaluate the following example:

The selection of all RNA sequences and protein sequences comprises a combination of the following conditions:

Table 3
interaction probabilities of result sequences.

Index	Sequence	SVM (ACC)	RF (ACC)
1	UCUGGUGACUAUAGCAAGGAGGUCACACCUGUUCCCAUGCCGAACACAGAAGUUAAGGUCUUUAGCGACGAUGGUAGCCA ACUUACGUUCCGCUAGAGUAGAACGUUGCCAGGC	0.79	0.8
2	GUUCGCGAAGUAACCCUUCGUGGACAUUUGGUCAAUUUGAAACAAUACAGAGAUGAUCAGCAGUUCCCCUGCAUAAGGAU GAACCGUUUUACAAAGAGAUUUAUUUCGUUUU	0.93	0.85
3	CGACUCUUAGCGGUGGAUCACUCGGCUCGUGCGUCGAUGAAGAACGCAGCUAGCUGCGAGAAUUAAUGUGAAUUGCAGGA CACAUUGAUCAUCGACACUUCGAACGCACUUGCGGCCCCGGGUUCCUCCCGGGGCUACGCCUGUCUGAGCGUCGCUU	0.59	0.65
4	CGACUCUUAGCGGUGGAUCACUCGGCUCGUGCGUCGAUGAAGAACGCAGCUAGCUGCGAGAAUUAAUGUGAAUUGCAGGU UGAUCAUCGACACUUCGAACGCACUUGCGGCCCCGGGUUCCUCCCGGGGCUACGCCUGUCUGAGCGUCGCU	0.72	0.65
5	GGUUGCGGCCAUAUCUACCAGAAAGCACCGUUUCCCGUCCGAUCAACUGUAGUUAAGCUGGUAAGAGCCUGACCGAGUAG UGUAGUGGGUGACCAUACGCGAAACUCAGGUGCUGCAAUCU	0.93	0.45

Note: the PRIseq server sets the threshold range from 0 to 1 to distinguish the positive (bindable) and negative (non-bindable) pairs. The predictions with probabilities > 0.5 were considered as positive pairs.

4.4.1.1. RNA

- The frequency of Multi Loop and Bulge is greater than one.

OR

- The length of the RNA sequences is greater than 100 ***AND*** including "CGCG" substring.

4.4.1.2. Protein

- The length of the protein sequences is less than 200 ***AND*** containing one parallel beta sheet ***OR*** one antiparallel beta sheet.

OR

- The length of the protein sequences is greater than 200, ***AND*** the average coil percentage is equal to 50.

(i) **RCSB PDB (http://www.rcsb.org/pdb/search/advSearch.do?search=new)**

Despite the diversity of parameters related to the analysis of protein structures in this database, there exist some limitations in this regard. PDB has a linear query structure with only an AND-junction or an OR-junction. Hence, it does not allow the combination of these clauses. Further, it does not contain structural features of RNA.

(ii) **URS (http://server3.lpm.org.ru/urs/struct.py)**

The search option for URS contains different parameters related to RNA structures and pseudoknots. However, the query engine module enables us to combine "AND" and "OR" clauses, but they do not have any nested query structures. In addition, the structural features of proteins are not supported.

(iii) **RPINBASE (http://rpinbase.com)**

The search option of RPINBASE has a nested query approach. It allows us to respond to the diverse demands related to different features of RNA and protein macromolecules, concurrently. Appending a query block with its specific criteria inside another query block can be separately performed for RNAs and proteins. Figs. 3 and 4 in supplementary File 2 show how the problem of the example can be solved using nested query tools as they are provided by the RPINBASE. These properties make this toolbox to be more flexible than the others.

4.4.2. Case II: use-case examples

For researchers who need statistics about RNA-protein interactions or datasets for machine learning algorithms, this toolbox is more efficient. Also, the design of oligonucleotide aptamer and regulatory RNAs such as ribozyme and riboswitch are the fields that require accurate datasets for computational processes to predict the interactions. For example, an aptamer is a short sequence of oligonucleotides with high specificity and affinity, which can bind itself to dedicated small to large targets [70–75]. Users can order their custom dataset to train a model, predict and find appropriate sequences of RNA as the initial population of the aptamer pool.

Consider the following examples:

(i) In Xu et al. [76] study, they determined the structural motifs of an RNA aptamer to Prostate-Specific Membrane Antigen (PSMA). This aptamer has a hairpin loop, a bulge loop, two internal loops, and four stems. Consider an example scenario that an aptamer pool should be designed to choose high-affinity novel sequences with the genetic algorithm. So, an initializing population including some RNA sequences is required to execute the algorithm. Filtered RNA sequences with the above conditions from RPINBASE can be used instead of any random RNA set. Through these structural motifs, RNA sequences from the RPINBASE can be filtered to create an initial aptamer pool. Also, the lengths of aptamers are recommended between 20 and 80 nucleotides; these criteria are applied to query. The output results are shown in Table 2.

To gain additional insights into the extracted RNA sequences, RNA structure was utilized [77] to model the secondary structure prediction of the RNA sequences. Fig. 5 shows some of these RNA sequences. The extracted RNA sequences from RPINBASE have similar structure motifs of RNA aptamer which is indicated in the example. Using these motifs, novel RNA sequences can be generated.

(ii) Regarding RNA-protein interactions, the RPINBASE creates a dataset that can be applied to predict the interaction between desired protein and existing RNAs. To examine the capabilities of the RPINBASE, the CD44 biomarker was selected based on its applications. The results are described as:
CD44 is one of the most commonly used surface markers to identify cancer stem cells. It plays a key role in the invasion of a wide range of tumor cells [78,79]. Therefore, finding RNA sequences capable of binding to CD44 may be promising in designing novel RNA sequences. In this regard, the desired dataset was extracted from RPINBASE and the model was created to train and test the classifiers as mentioned in "section 3.3". This model was utilized to predict the RNA-protein interaction. The input of this model's predictor function is a vector that contains RNA and protein features values. So, the input dataset was generated from vectors that were combined from CD44 and all RNAs' features. The output result shows the RNA bindable CD44 sequences. Output sequences were

examined by the PRIseq web-based server [31,32] that demonstrated the efficiency of this approach (Table 3).

5. Conclusions

RPINBASE provides a user-friendly and effective tool for researchers to easily and quickly establish accurate and minimal datasets of proteins, RNAs, and RNA-protein complexes by investigating the structures of RNA-protein complexes. The extracted features are completely classified in the specific form of primary and secondary structures of protein and RNA sequences available in this toolbox. Users can properly prepare datasets (raw sequences or feature vectors) in the form of 'complex', 'protein' and 'RNA' targets which typically contain specific features for machine learning purposes. In this toolbox, users can also select the negative dataset typically generated according to the family of protein sequences based on their specific characteristics. RPINBASE regularly updates itself from PDB, which is considered as an acceptable source for users who need information on these complexes.

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ygeno.2020.02.013.

Declaration of Competing Interest

All the authors have read and confirmed the paper.

References

[1] H. Siomi, G. Dreyfuss, RNA-binding proteins as regulators of gene expression, Curr. Opin. Genet. Dev. 7 (1997) 345–353.
[2] T. Derrien, R. Johnson, G. Bussotti, A. Tanzer, S. Djebali, H. Tilgner, G. Guernec, D. Martin, A. Merkel, D.G. Knowles, The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression, Genome Res. 22 (2012) 1775–1789.
[3] W. Prall, B. Sharma, B.D. Gregory, Transcription is just the beginning of gene expression regulation: the functional significance of RNA-binding proteins to post-transcriptional processes in plants, Plant Cell Physiol. 60 (2019) 1939–1952.
[4] V.P. Belancio, A.M. Roy-Engel, Deininger PL: All y'all need to know 'bout retroelements in cancer, Seminars in Cancer Biology, Elsevier, 2010, pp. 200–210.
[5] B.M. Lunde, C. Moore, G. Varani, RNA-binding proteins: modular design for efficient function, Nat. Rev. Mol. Cell Biol. 8 (2007) 479.
[6] V. Ramakrishnan, S.W. White, Ribosomal protein structures: insights into the architecture, machinery and evolution of the ribosome, Trends Biochem. Sci. 23 (1998) 208–212.
[7] Z. Li, P.D. Nagy, Diverse roles of host RNA binding proteins in RNA virus replication, RNA Biol. 8 (2011) 305–315.
[8] K.B. Hall, RNA–protein interactions, Curr. Opin. Struct. Biol. 12 (2002) 283–288.
[9] I. Sola, P.A. Mateos-Gomez, F. Almazan, S. Zuniga, L. Enjuanes, RNA-RNA and RNA-protein interactions in coronavirus replication and transcription, RNA Biol. 8 (2011) 237–248.
[10] H. Cao, K. Zhao, Y. Yao, J. Guo, X. Gao, Q. Yang, M. Guo, W. Zhu, Y. Wang, C. Wu, RNA binding protein 24 regulates the translation and replication of hepatitis C virus, Protein Cell 9 (2018) 930–944.
[11] K.S. Moore, P.A.C. Hoen, Computational approaches for the analysis of RNA–protein interactions: a primer for biologists, J. Biol. Chem. 294 (2019) 1–9.
[12] B. Niu, C. Liang, Y. Lu, M. Zhao, Q. Chen, Y. Zhang, L. Zheng, K.-C. Chou, Glioma stages prediction based on machine learning algorithm combined with protein-protein interaction networks, Genomics 112 (2020) 837–847.
[13] X. Pan, Y. Yang, C.Q. Xia, A.H. Mirza, H.B. Shen, Recent methodology progress of deep learning for RNA–protein interaction prediction, Wiley Interdisciplinary Reviews: RNA, 2019 e1544.
[14] H. Nematzadeh, R. Enayatifar, M. Mahmud, E. Akbari, Frequency based feature selection method using whale algorithm, Genomics 111 (2019) 1946–1955.
[15] Y. Masoudi-Sobhanzadeh, Y. Omidi, M. Amanlou, A. Masoudi-Nejad, Trader as a new optimization algorithm predicts drug-target interactions efficiently, Sci. Rep. 9 (2019) 9348.
[16] A. Meshkin, A. Shakery, A. Masoudi-Nejad, GPS: identification of disease genes by rank aggregation of multi-genomic scoring schemes, Genomics 111 (2019) 612–618.
[17] H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I.N. Shindyalov, P.E. Bourne, The protein data bank, Nucleic Acids Res. 28 (2000) 235–242.
[18] J.T. Chen, W.P. Yang, M. Hsu, Optimization on a case of type-d nested query, J. Chin. Inst. Eng. 11 (1988) 309–316.
[19] J.L. Han, Optimizing relational queries in connection hypergraphs: nested queries, views, and binding propagations, VLDB J. 7 (1998) 1–11.
[20] W. Kim, On optimizing an SQL-like nested query, ACM Trans. Database Syst. 7 (1982) 443–469.
[21] O. Zanegina, D. Kirsanov, E. Baulin, A. Karyagina, A. Alexeevski, S. Spirin, An updated version of NPIDB includes new classifications of DNA–protein complexes and their families, Nucleic Acids Res. 44 (2015) D144–D153.
[22] K.B. Cook, H. Kazan, K. Zuberi, Q. Morris, T.R. Hughes, RBPDB: a database of RNA-binding specificities, Nucleic Acids Res. 39 (2010) D301–D308.
[23] T. Wu, J. Wang, C. Liu, Y. Zhang, B. Shi, X. Zhu, Z. Zhang, G. Skogerbø, L. Chen, H. Lu, NPInter: the noncoding RNAs and protein related biomacromolecules interaction database, Nucleic Acids Res. 34 (2006) D150–D152.
[24] B. Lang, A. Armaos, G.G. Tartaglia, RNAct: protein–RNA interaction predictions for model organisms with supporting experimental data, Nucleic Acids Res. 47 (2018) D601–D606.
[25] B.A. Lewis, R.R. Walia, M. Terribilini, J. Ferguson, C. Zheng, V. Honavar, D. Dobbs, PRIDB: a protein–RNA interface database, Nucleic Acids Res. 39 (2010) D277–D282.
[26] B. Coimbatore Narayanan, J. Westbrook, S. Ghosh, A.I. Petrov, B. Sweeney, C.L. Zirbel, N.B. Leontis, H.M. Berman, The nucleic acid database: new features and capabilities, Nucleic Acids Res. 42 (2013) D114–D122.
[27] M. Andronescu, V. Bereg, H.H. Hoos, A. Condon, RNA STRAND: the RNA secondary structure and statistical analysis database, BMC Bioinformatics 9 (2008) 340.
[28] P.L. Vanegas, G.A. Hudson, A.R. Davis, S.C. Kelly, C.C. Kirkpatrick, B.M. Znosko, RNA CoSSMos: characterization of secondary structure motifs—a searchable database of secondary structure motifs in RNA three-dimensional structures, Nucleic Acids Res. 40 (2011) D439–D444.
[29] M. Popenda, M. Szachniuk, M. Blazewicz, S. Wasik, E.K. Burke, J. Blazewicz, R.W. Adamiak, RNA FRABASE 2.0: an advanced web-accessible database with the capacity to search the three-dimensional fragments within RNA structures, Bmc Bioinformatics 11 (2010) 231.
[30] E. Baulin, V. Yacovlev, D. Khachko, S. Spirin, M. Roytberg, URS DataBase: universe of RNA structures and their motifs, Database 2016 (2016).
[31] U.K. Muppirala, V.G. Honavar, D. Dobbs, Predicting RNA-protein interactions using only sequence information, BMC Bioinformatics 12 (2011) 489.
[32] U.K. Muppirala, B.A. Lewis, D.L. Dobbs, C. Biology, Computational tools for investigating RNA-protein interaction partners, J. Comput. Sci. 6 (2013) 182.
[33] V. Suresh, L. Liu, D. Adjeroh, X. Zhou, RPI-Pred: predicting ncRNA-protein interaction using sequence and structural information, Nucleic Acids Res. 43 (2015) 1370–1379.
[34] Z. Cheng, S. Zhou, J. Guan, Computationally predicting protein-RNA interactions using only positive and unlabeled examples, J. Bioinforma. Comput. Biol. 13 (2015) 1541005.
[35] M. Bellucci, F. Agostini, M. Masin, G.G. Tartaglia, Predicting protein associations with long noncoding RNAs, Nat. Methods 8 (2011) 444.
[36] Q. Lu, S. Ren, M. Lu, Y. Zhang, D. Zhu, X. Zhang, T. Li, Computational prediction of associations between long non-coding RNAs and proteins, BMC Genomics 14 (2013) 651.
[37] V. Pancaldi, J. Bähler, In silico characterization and prediction of global protein–mRNA interactions in yeast, Nucleic Acids Res. 39 (2011) 5826–5836.
[38] Y. Wang, X. Chen, Z.-P. Liu, Q. Huang, Y. Wang, D. Xu, X.-S. Zhang, R. Chen, L. Chen, De novo prediction of RNA–protein interactions from sequence information, Mol. BioSyst. 9 (2013) 133–142.
[39] X. Pan, P. Rijnbeek, J. Yan, H.-B. Shen, Prediction of RNA-protein sequence and structure binding preferences using deep convolutional and recurrent neural networks, BMC Genomics 19 (2018) 511.
[40] L. Wang, X. Yan, M.-L. Liu, K.-J. Song, X.-F. Sun, W.-W. Pan, Prediction of RNA-protein interactions by combining deep convolutional neural network with feature selection ensemble method, J. Theor. Biol. 461 (2019) 230–238.
[41] Z. Cheng, K. Huang, Y. Wang, H. Liu, J. Guan, S. Zhou, Selecting high-quality negative samples for effectively predicting protein-RNA interactions, BMC Syst. Biol. 11 (2017) 9.
[42] G.O. Consortium, Gene ontology annotations and resources, Nucleic Acids Res. 41 (2012) D530–D535.
[43] S. El-Gebali, J. Mistry, A. Bateman, S.R. Eddy, A. Luciani, S.C. Potter, M. Qureshi, L.J. Richardson, G.A. Salazar, A. Smart, The Pfam protein families database in 2019, Nucleic Acids Res. 47 (2018) D427–D432.
[44] U. Consortium, Update on activities at the universal protein resource (UniProt) in 2013, Nucleic Acids Res. 41 (2012) D43–D47.
[45] R. Nagarajan, M.M. Gromiha, Prediction of RNA binding residues: an extensive analysis based on structure and function to select the best predictor, PLoS One 9 (2014) e91140.
[46] D. Adjeroh, M. Allaga, J. Tan, J. Lin, Y. Jiang, A. Abbasi, X. Zhou, Feature-based and string-based models for predicting RNA-protein interaction, Molecules 23 (2018) 697.
[47] L. Wang, C. Huang, M.Q. Yang, J.Y. Yang, BindN+ for accurate prediction of DNA and RNA-binding residues from protein sequence features, BMC Syst. Biol. 4 (2010) S3.
[48] J. Tong, P. Jiang, Lu Z-h: RISP: a web-based server for prediction of RNA-binding sites in proteins, Comp. Methods Progr. Biomed. 90 (2008) 148–153.
[49] X. Ma, J. Guo, J. Wu, H. Liu, J. Yu, J. Xie, X. Sun, Prediction of RNA-binding residues in proteins from primary sequence using an enriched random forest model with a novel hybrid feature, Protein. Struct. Funct. Bioinform. 79 (2011) 1230–1239.
[50] Y.-F. Huang, L.-Y. Chiu, C.-C. Huang, C.-K. Huang, Predicting RNA-binding residues from evolutionary information and sequence conservation, *BMC Genomics*. BioMed Central, S2 2010.
[51] M. Terribilini, J.D. Sander, J.-H. Lee, P. Zaback, R.L. Jernigan, V. Honavar, D. Dobbs, RNABindR: a server for analyzing and predicting RNA-binding sites in proteins, Nucleic Acids Res. 35 (2007) W578–W584.

[52] W. Kabsch, C. Sander, Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features, Biopolymers 22 (1983) 2577–2637.
[53] A. Drozdetskiy, C. Cole, J. Procter, G.J. Barton, JPred4: a protein secondary structure prediction server, Nucleic Acids Res. 43 (2015) W389–W394.
[54] I.L. Hofacker, Vienna RNA secondary structure server, Nucleic Acids Res. 31 (2003) 3429–3431.
[55] A. Kloczkowski, K.L. Ting, R. Jernigan, J. Garnier, Combining the GOR V algorithm with evolutionary information for protein secondary structure prediction from amino acid sequence, Proteins 49 (2002) 154–166.
[56] T. Liu, C. Jia, A high-accuracy protein structural class prediction algorithm using predicted secondary structural information, J. Theor. Biol. 267 (2010) 272–275.
[57] S. Zhang, S. Ding, T. Wang, High-accuracy prediction of protein structural class for low-similarity sequences based on predicted secondary structure, Biochimie 93 (2011) 710–714.
[58] S. Ding, S. Zhang, Y. Li, T. Wang, A novel protein structural classes prediction method based on predicted secondary structure, Biochimie 94 (2012) 1166–1171.
[59] M. Aldwairi, B. Al-Hajasad, Y. Khamayseh, A classifier system for predicting RNA secondary structure, Int. J. Bioinforma. Res. Appl. 10 (2014) 307–320.
[60] W. Zhang, M. Ke, Protein encoding: a Matlab toolbox of representing or encoding protein sequences as numerical vectors for bioinformatics, J. Chem. Pharm. Res. 6 (2014) 8.
[61] A.J.C. Hurkens, M. McArthur, Y.N. Moschovakis, L.S. Moss, G.T. Whitney, The logic of recursive equations, J. Symb. Log. 63 (1998) 451–478.
[62] A. Krauss, Partial and nested recursive function definitions in higher-order logic, J. Autom. Reason. 44 (2010) 303–336.
[63] X. Pan, Y.-X. Fan, J. Yan, H.-B. Shen, IPMiner: hidden ncRNA-protein interaction sequential pattern mining with stacked autoencoder for accurate computational prediction, BMC Genomics 17 (2016) 582.
[64] C. Angermueller, T. Pärnamaa, L. Parts, O. Stegle, Deep learning for computational biology, Mol. Syst. Biol. 12 (2016).
[65] B. Tang, Z. Pan, K. Yin, A. Khateeb, Recent advances of deep learning in bioinformatics and computational Biology, Front. Genet. 10 (2019).
[66] Q. Yang, C. Jia, T. Li, Prediction of aptamer–protein interacting pairs based on sparse autoencoder feature extraction and an ensemble classifier, Math. Biosci. 311 (2019) 103–108.
[67] Y.-B. Wang, Z.-H. You, X. Li, T.-H. Jiang, X. Chen, X. Zhou, L. Wang, Predicting protein–protein interactions from protein sequences by a stacked sparse autoencoder deep neural network, Mol. BioSyst. 13 (2017) 1336–1344.
[68] L. Wang, Z.-H. You, X. Chen, S.-X. Xia, F. Liu, X. Yan, Y. Zhou, K.-J. Song, A computational-based method for predicting drug–target interactions by using stacked autoencoder deep neural network, J. Comput. Biol. 25 (2018) 361–373.
[69] M.S. Rahman, U. Aktar, M.R. Jani, S. Shatabda, iPromoter-FSEn: identification of bacterial σ70 promoter sequences using feature subspace based ensemble classifier, Genomics 111 (2019) 1160–1166.
[70] H. Sun, X. Zhu, P.Y. Lu, R.R. Rosato, W. Tan, Y. Zu, Oligonucleotide aptamers: new tools for targeted cancer therapy, Mol. Ther. Nucleic Acid. 3 (2014) e182.
[71] J. Hoinka, E. Zotenko, A. Friedman, Z.E. Sauna, T.M. Przytycka, Identification of sequence–structure RNA binding motifs for SELEX-derived aptamers, Bioinformatics 28 (2012) i215–i223.
[72] S. Kedzierski, M. Khoshnejad, G.T. Caltagirone, Synthetic antibodies: the emerging field of aptamers, Bioprocess. J. 11 (2012) 46–49.
[73] H. Sun, Y. Zu, A highlight of recent advances in aptamer technology and its application, Molecules 20 (2015) 11959–11980.
[74] G.-Q. Zhang, L.-P. Zhong, N. Yang, Y.-X. Zhao, Screening of aptamers and their potential application in targeted diagnosis and therapy of liver cancer, World J. Gastroenterol. 25 (2019) 3359.
[75] A.T. Ponce, K.L. Hong, A mini-review: clinical development and potential of Aptamers for thrombotic events treatment and monitoring, Biomedicines 7 (2019) 55.
[76] X. Xu, D.D. Dickey, S.-J. Chen, P.H.J.M. Giangrande, Structural computational modeling of RNA aptamers, Methods 103 (2016) 175–179.
[77] J.S. Reuter, D.H. Mathews, RNAstructure: software for RNA secondary structure prediction and analysis, BMC Bioinformatics 11 (2010) 129.
[78] C. Chandola, M.G. Casteleijn, U.M. Chandola, L.N. Gopalan, A. Urtti, M. Neerathilingam, CD44 aptamer mediated cargo delivery to lysosomes of retinal pigment epithelial cells to prevent age-related macular degeneration, Biochem. Biophys. Rep. 18 (2019) 100642.
[79] N. Ababneh, W. Alshaer, O. Allozi, A. Mahafzah, M. El-Khateeb, H. Hillaireau, M. Noiray, E. Fattal, S. Ismail, In vitro selection of modified RNA aptamers against CD44 cancer stem cell marker, Nucleic Acid Ther. 23 (2013) 401–407.