RPINaptaBASE: a Database for RNA-Protein Interaction Network Analysis and Aptamer Design

Mahsa Torkamanian-Afshar

University of Tehran

Hossein Lanjanian

University of Tehran

Sajjad Nematzadeh

Yildiz Teknik Universitesi

Maryam Tabar zad

Shaheed Beheshti University of Medical Sciences

Ali Najafi

Baqiyatallah University of Medical Sciences

Farzad Kiani

Istanbul Sabahattin Zaim Universitesi

Ali Masoudi-Nejad (✉ amasoudin@ibb.ut.ac.ir)

University of Tehran <https://orcid.org/0000-0003-0659-5183>

Database

Keywords: RNA-protein interactions, Nested query, Macromolecule features, Negative sample, RNA-protein database

Posted Date: May 14th, 2019

DOI: <https://doi.org/10.21203/rs.2.9610/v1>

License: This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Abstract

Abstract Background The RNA-protein interactions play crucial roles in the biological processes. Recent developments to clarify RNA and protein structural features have the urgent need for designing various databases, related to the specificity and the mechanism of the underlying interactions between a protein and an RNA molecule. The majority of these databases have focused on RNAs or proteins macromolecules independently, and they do not have the capability to run integrated queries on the RNA-protein complex. These existing databases have a linear query structure. Furthermore, they only focus on interacting (positive) samples and they do not contain non-interacting (negative) samples. **Results** We developed a Database for RNA-Protein Interaction Network Analysis and Aptamer Design (RPinaptaBASE). RPinaptaBASE has a nested query approach that enables users to apply nonlinear query analysis. The query engine module contains a wide range of features related to RNA and protein sequences and secondary structure elements of these macromolecules, which are helpful to generate custom datasets, especially for machine learning approaches. In this version, more than 175 features were calculated and available to users. It provides a web interface with download management services allowing users to generate desired datasets of unique RNA or protein sequences in independent lists. Furthermore; the web service empowers users to create artificial datasets of positive and negative samples from RNA-protein complexes. In order to present negative samples, the idea of distinguishing protein sequences by their clans and families was employed to efficiently generate non-interacting pairs. **Conclusion** This database prepares a user-friendly platform to study RNA-protein interactions. It also provides an important simplified contribution to the oligonucleotide-aptamer design process using machine learning algorithms. RPinaptaBASE is freely available at <http://rpinbase.com>

Background

RNA-protein interaction is one of the most important concepts in the biological processes and receives considerable critical attention. It carries out an essential role in various fields, such as regulating gene expression [1-3], protein synthesis [4, 5], viral replication and cellular defense mechanism [6-8]. Recent developments to clarify RNA and protein structural features have heightened the need for designing various databases, to investigate the RNA-protein interactions. Several studies, yielding various databases, have been carried out in this regard. The most important databases are: The Nucleic Acid Database (NDB) [9], The RNA secondary structure and statistical analysis database (RNA Strand) [10], The RNA Characterization of Secondary Structure Motifs (RNA CoSSMos) [11], The RNA FRAGMENT search engine & dataBASE (RNA FRABASE) [12], The Universe of RNA Structures DataBase (URSDB) [13], The Nucleic Acids-Protein Interaction DataBase (NPIDB) [14], The RNA-Binding Protein DataBase (RBPDB) [15], The noncoding RNAs and protein related biomacromolecules interaction database (NPInter) [16], RNA-protein Association and Interaction Networks (RIAN) [17], Protein-RNA interaction predictions for model organisms with supporting experimental data (RNAct) [18], and The Protein-RNA Interface Database (PRIDB) [19], which will be discussed in the following text.

NDB [9], contains structural information of oligonucleotides such as nucleic acid structural conformation, RNA 3D motifs and RNA 3D interactions. Furthermore, in this database, searching with predefined parameters and logical operators has been provided. RNA Strand [10], includes general and structural information of RNAs, such as the length of structural motifs, type of structural elements and number of non-canonical base pairs per molecule. RNA CoSSMos [11], consists of structures derived from The Protein Data Bank (PDB) [20] and a collection of motifs extracted from the structural elements. The database provides the option of searching for the desired data according to the general parameters, experimental parameters, structural motifs, and nucleic acid sequences. RNA FRABASE [12], includes RNA structures (derived from the PDB and NDB databases), the dot-bracket format of the RNA sequences, and the comprehensive series of parameters related to structural elements. URSDB [13], contains structural elements of RNAs, which have been derived from the PDB database. In this database, pseudoknots in secondary structures have also been presented; moreover, executing queries with respect to the structural features is available. NPIDB [14], includes information derived from PDB complexes. It provides a user interface and a set of tools for extracting features, such as the calculations of intermolecular interactions, the classification of proteins based on the Structural Classification of Proteins (SCOP) [21] and the Protein Families database (Pfam) databases [22], and some information on DNA-protein interactions. RBPDB [15], as a database of RNA-binding proteins, consists of the experimental data of four species (human, mouse, fly and worm). To recognize the similarity of proteins in various species, RBPDB also provides their orthology relationships. NPInter [16], includes functional interactions between noncoding RNAs and other biomolecules. It covers the largest number of interaction data on tissues or cell lines, binding sites, conservation, and other factors. RIAN [17], contains noncoding RNA associations in four organisms (human, mouse, rat, and yeast). These are generated from curated samples, experiments, interaction predictions and text mining. Also, it provides an integrative scoring scheme to determine the reliability of each interaction. RNAct [18], covers predicted and experimentally identified information on RNA-protein interactions in three organisms (human, mouse, and yeast). Moreover, it contains pairwise search for sets of proteins and RNAs. PRIDB [19], contains information on RNA-protein complexes; moreover, the sequential motifs of complexes were identified. Additionally, three benchmarks datasets are available for evaluating machine learning methods. In the latest version of this database, RNA-protein interactions using only sequence information (PRIsq) [23, 24], the analysis of RNA-protein interactions are provided with using Support Vector Machine (SVM) [25] and Random Forests (RF) [26] algorithms.

The majority of these databases have focused on RNAs or proteins macromolecules independently, moreover, these RNA-protein complex databases do not have the capability to run integrated queries on the RNA-protein complex. On the other hand, existing databases have a linear query structure. All conditional statements in this query structure are in the equivalent degree of priority. Consequently, the linear query structure is inconvenient for complex requests. However, the nested query structure enables one to achieve this goal. In other words, the precedence of query evaluation of nested queries is analogous to that of parentheses in mathematical operations where subqueries are evaluated first. 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 structure of macromolecules, then

these subqueries are aggregated to form one nested query and this query is sent to the database. Another problem, facing most databases in the field of nucleic acid-protein interaction, is that they only focus on interacting (positive) samples and they do not contain validated non-interacting (negative) samples. Therefore, the majority of these studies addressed this problem by using atomic-distance with an arbitrary threshold or by random pairing [24, 27-32]. In another study, Cheng et al. [33] applied the scoring method using Gene Ontology (GO) [34], Pfam and The Universal Protein Resource (UniProt) [35] databases to generate the negative datasets.

RPINaptaBASE is a repository of all RNA-protein complexes stored in the PDB database. The query executor module of RPINaptaBASE contains a wide range of features related to 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 to search various substrings between existing sequences. Moreover, the phylogeny data about the family and clan of protein sequences are available. At the secondary structure level, it allows queries containing different related secondary structural features. Furthermore, the query structure of RPINaptaBASE is designed based on the nested object concepts in object-oriented programming[36-38] to respond to diverse demands related to the study of different aspects of RNA-protein binding. However, the importance of machine learning as a fast-growing approach in this field is undeniable. Therefore, RPINaptaBASE presents an option to create downloadable positive and negative datasets with non-redundant sequences for training and testing classifiers. Non-interacting (negative) samples in this database were generated using protein clans whereas interacting pairs (positive) samples were generated using atomic-distances.

Results

RPINaptaBASE provides a 'Help' section with comprehensive information and a practical example of RPINaptaBASE's web interface usage (Additional file 1: Help).

There is a menu bar on the homepage. In this bar four stages are represented; 1-'Start', 2-'Select mode', 3-'Make a query' and 4- 'Check the result & download'. The 'Start', is the entry point of the database. There is also a start button. By clicking this button user can go to the second stage. In the 'select mode' page, users can select three types of datasets: 'Complex', 'Protein' and 'RNA'. Figure 1 shows the 'select mode' page. By choosing one of these modes, the 'make a query' page appears.

In the 'Make a query' stage users create a query to request a dataset from the database using specific features. These pages contain logical blocks to create specific queries. These blocks are executed according to the rules and precedence of parentheses. This means all the items in parentheses are evaluated separately. Items with nested parentheses are evaluated, from inside to outside.

- By selecting the 'Complex' mode on the 'select mode' page, users can create positive and negative datasets. Each of them divided into four parts: 'Query setting', 'General', 'RNA' and 'Protein'. In the Query setting, users can specify the number of desired samples and select the desired threshold. In 'General', 'RNA' and 'Protein' parts users can create their requests based on the general information,

primary and secondary structure of RNA and protein macromolecules. Figure 2 shows the page of the complex mode.

- By selecting the 'RNA' mode on the 'select mode' page, users can create RNA datasets. This page is divided into two parts: 'Query setting' and 'RNA'. In Query setting, users can specify the number of output sequences in their RNA samples. In the RNA part, users can create their queries based on the primary and secondary structure of RNA sequences. Figure 3 shows the RNA mode.
- By selecting the 'Protein' mode, on the 'select mode' page users can create protein datasets. This page is divided into two parts: 'Query setting' and 'Protein'. In Query setting, users can specify the number of output sequences in their protein samples. In the protein part, users can create their queries based on the protein family, primary and secondary structure of protein sequences. Figure 4 shows the protein mode.

By selecting the 'Check result' button in all of the three modes, users have access to the 'Check the result & download' page. In this page, users can view the statistics of the created set and are able to download and save the results in one of two forms (CSV and Text). Figure 5 shows the result page.

Discussion

With regarding RNA-protein interactions, RPiNaptaBASE provides the nested query mechanism. It allows users to easily search sequential motifs, collect phylogeny data about protein sequences and also collect various types of statistical data related to RNA and protein macromolecules. Furthermore, RPiNaptaBASE can be used to create three types of data sets; 'RNA', 'Protein' and 'RNA-protein Complex'. From the other hand, the negative datasets of RPiNaptaBASE create based on the view of family and clan of protein sequences.

In most of the RNA-protein databases [14, 15, 19] users cannot simultaneously query primary and secondary structural features of both RNA and protein macromolecules. It is essential for the RNA-protein database to have the capability to run complex queries, which consists of many conditional statements and subqueries. In addition, there are databases [10-13] that, perform primary and secondary feature queries on RNAs and have a linear query mechanism. It means that all required conditions in the database are evaluated at a single level. Therefore, it is impossible to design a query that utilizes results from subqueries. Consequently, the linear query structure is inconvenient for complex requests. In the linear query structure, users have to use multiple queries to achieve what a single nested query achieves.

As for the RPiNaptaBASE, its web interface allows one to create a nested query structure on different features of both RNA and protein macromolecules.

In the point of negative samples, PRiSeq [23, 24] provides benchmark datasets, that they only focus on positive samples and they do not contain negative samples. Many studies used these positive datasets for machine learning applications. On the other hand, to create a balanced dataset of negative samples, some of them used atomic distance and the others, randomly paired the RNAs and proteins. These solutions can lead to the generation of redundant and irrelevant information which increase the number of false negative samples. RPiNaptaBASE provides negative dataset based on family and clan of the protein sequence.

This database application provides an important simplified contribution to the oligonucleotide-aptamer design process using machine learning algorithms. Aptamers are a short sequence of oligonucleotides with high specificity and affinity, which can bind to dedicated small to large targets [39-42]. In the Future version of RPiNaptaBASE, we intend to insert sequence and structural information of DNA-Protein complexes. In addition, the feature vector of these macromolecules will be accessible.

Conclusions

RPiNaptaBASE provides a user-friendly effective tool for researchers who need to create the accurate and minimal datasets of protein, RNA and RNA-protein complexes easily and quickly by investigating in 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 and available in this database. Users can properly prepare datasets in the form of 'complex', 'protein' and 'RNA' targets that typically contain specific features for machine learning purposes. In this database, users can also select the negative dataset typically generated according to the family of protein sequences based on their specific characteristics. RPiNaptaBASE updates regularly from PDB, which is considered as an acceptable source for users requiring information on these complexes.

Methods

The data stored on RPiNaptaBASE was prepared as follows; at first, the contents of PDB structures were extracted, cleaned and converted to a suitable format for database design. Then, the values of macromolecules' structural features were calculated and appended to the database. Finally, positive and negative datasets were constructed.

Data gathering and preprocessing

RPiNaptaBASE is a repository of RNA-protein complexes. In the beginning, all structures were extracted from the PDB, then, the PDB files that contain at least one protein and one RNA sequence were stored as the target sample. Our database performed the analysis of 2258 complexes (as of June 2018). In the preprocessing step, duplicated chains and sequences which contain 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). The A and B in protein chains also C and D in RNA chains were identical. Consequently, we removed one protein and one RNA chain, then obtained the B-C as a positive and non-redundant sample. Table 1 shows the number of raw and preprocessed data for developing RPiNaptaBASE.

To investigate which chains of proteins directly interact with the RNA chains, the analysis of 3D structures of macromolecules in the complexes were done. Different thresholds were used to differentiate the interaction between chains of macromolecules [24, 28-32]. Therefore, if the atomic distance between RNA and a chain of protein in the PDB files was less than the chosen threshold, then these two chains were detected as interacting pairs. One of 3.4 Å, 3.7 Å, 5 Å, 7 Å, and 10 Å was chosen as a threshold to distinguish strongly interacting pairs. Users can also select one of these distances to select highly interacting pairs. Afterward, the secondary structure of the protein sequences assigned by Define Secondary Structure of Proteins (DSSP) [43] was extracted from PDB. The Protein Secondary Structure Prediction server (JPred) [44] was used to predict the secondary structure of sequences that didn't include any DSSP assignment. Besides, the RNAfold from the Vienna package [45] was used 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 the information of sequences is shown in figure 6.

Structural processing of macromolecules

The main aim of RPiNaptaBASE is to provide a suitable resource to construct the desired datasets based on different features of RNA macromolecule or protein macromolecule in a complex. Therefore, the final stored data in the last step are processed to extract all of the possible features of first and secondary structures. Furthermore, the phylogeny information of macromolecules was extracted from the Pfam database, hence, it is possible to create the desired dataset based on a specific family or clan. In the case of primary structure, RPiNaptaBASE supports substrings and length of RNA and protein sequences. On the secondary structure level, this database contains the information of RNA secondary structure elements (Stem, Hairpin, Bulge, Internal loop, and Multi-Loop), and protein secondary structure (Alpha Helix, Beta Strand, and Coil). According to the DSSP algorithm, the specified protein secondary structure assignment has eight elements. These eight letters were translated into three letters for ease of interpretation [46]. Here, the results of the study [47] were used to calculate the number of parallel and antiparallel beta sheets. In addition, general information related to the structural elements was processed using different algorithms [48-50] and their results were stored as searchable data in the database. In the current version of RPiNaptaBASE, more than 175 features (Additional file 2: Macromolecules features) of protein and RNA macromolecules were calculated and presented to be used in queries.

Generating Negative and Positive datasets

In this study, the family and clan of protein sequences were used to generate non-interacting pairs. To this aim, an infrastructure was constructed to interact with RNA-protein pairs by observing experimental reports and construct non-interacting RNA-protein pairs based on family and clan of protein sequences. There were 9,367 protein families, and these families were classified into 604 clans in the Pfam

database. Interestingly, RNA-protein complexes only cover 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 take into account only clans of RNA binding proteins to ensure that only related features are considered. Moreover, discrimination between positive and negative samples is done on the base of specific features of macromolecules capable of composing the RNA-protein complex. Positive samples were composed by combining protein chains with RNA sequences of complexes whose distance was less than the selected thresholds of 3.4 Å, 3.7 Å, 5 Å, 7 Å, and 10 Å. Negative samples were generated by choosing RNA sequences in a complex and combining with the protein sequences that had not been observed in the same clan of that complex. Consequently, users are enabled to construct positive and negative datasets based on the desired family and clan of protein sequences. Figure 7 shows an overview of the formation of positive and negative samples.

Nested query

In this database, recursive functions were used to obtain the evaluation of subqueries in nested structures. This means that various subqueries can be created according to the characteristics of the primary and secondary structure of macromolecules, then these subqueries are aggregated and sent to the database. Nested queries are defined in the form of nested objects. Each query object is the parent of their 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 [51, 52]. This nonlinear query structure is helpful when it comes to complex queries and mixed conditional statements on different types of features.

Implementation

The database first approach was chosen and the data models were created from tables. The RPiNaptaBASE application was developed in a standard three layers of architecture. First, the data access layer, which was implemented using Microsoft SQL Server, has the data storing and retrieving task. The second layer is the business layer. This layer was developed using Active Server Pages (ASP.Net) and is run 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, validated and forwarded to the data access layers. The results which come from the database are converted to the JavaScript Object Notation (JSON) format and are sent to the presentation layer. The presentation layer (web interface) was created with HyperText Markup Language (HTML) and AngularJS which provide the dynamic script execution on the clients' browsers. Figure 8 shows the three layers making up the RPiNaptaBASE infrastructure.

Abbreviations

NDB: The Nucleic Acid Database

RNA Strand: The RNA secondary structure and statistical analysis database

RNA CoSSMos: The RNA Characterization of Secondary Structure Motifs

RNA FRABASE: RNA FRAGment search engine & dataBASE

URSDB: The Universe of RNA Structures DataBase

NPIDB: Nucleic Acids-Protein Interaction DataBase

RBPDB: The RNA-Binding Protein DataBase

NPInter: The noncoding RNAs and protein related bio macromolecules interaction database

RIAN: RNA-protein Association and Interaction Networks

RNAct: Protein-RNA interaction predictions for model organisms with supporting experimental data

PRIDB: The Protein-RNA Interface Database

PDB: The Protein Data Bank

SCOP: Structural Classification of Proteins

Pfam: The Protein Families database

PRlseq: RNA-protein interactions using only sequence information

SVM: Support Vector Machine

RF: Random Forests

GO: Gene Ontology

Uniprot: The Universal Protein Resource

DSSP: Define Secondary Structure of Proteins

JPred: The Protein Secondary Structure Prediction server

ASP: Active Server Pages

IIS: Internet Information Services

MVC: The Model-View-Controller

JSON: JavaScript Object Notation

HTML: Hyper Text Markup Language

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data and materials

RPINaptaBASE is freely available for public use at <http://rpinbase.com>. It is accessed from a web browser; therefore, it is platform-independent. Also it has been extensively tested with Google Chrome, Safari, Mozilla Firefox, and Microsoft Edge. RPINaptaBASE is free to use by everyone; there are no restrictions.

Competing interests

The authors have no competing interests.

Funding

No funding

Authors' contribution

MTA, SN, HL and MT: Problem definition and analyze, MTA and SN: Design and implementation of the data access layer, data miners, sequence processors, structural feature of macromolecules extractors, web services, business and control layer, user interface, query making tools and download job manager. HL, MT and AN: Testing, validation and visualization. FK: Machine learning, network infrastructure and web-server configuration consulter. MTA, SN, HL and MT: Wrote the manuscript. AMN: Problem definition, supervision, project administration, and editing the manuscript. All authors read and approved the final manuscript.

Acknowledgements

Not applicable.

References

1. Siomi H, Dreyfuss GJCoig, development: **RNA-binding proteins as regulators of gene expression**. 1997, **7**(3):345-353.
2. Derrien T, Johnson R, Bussotti G, Tanzer A, Djebali S, Tilgner H, Guernec G, Martin D, Merkel A, Knowles DGJGr: **The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression**. 2012, **22**(9):1775-1789.
3. Mata J, Marguerat S, Bähler JJTibs: **Post-transcriptional control of gene expression: a genome-wide perspective**. 2005, **30**(9):506-514.
4. Lunde BM, Moore C, Varani GJNrMcb: **RNA-binding proteins: modular design for efficient function**. 2007, **8**(6):479.
5. Ramakrishnan V, White SWJTibs: **Ribosomal protein structures: insights into the architecture, machinery and evolution of the ribosome**. 1998, **23**(6):208-212.
6. Li Z, Nagy PDJRb: **Diverse roles of host RNA binding proteins in RNA virus replication**. 2011, **8**(2):305-315.
7. Hall KBJCoisb: **RNA–protein interactions**. 2002, **12**(3):283-288.
8. Sola I, Mateos-Gomez PA, Almazan F, Zuniga S, Enjuanes LJRB: **RNA-RNA and RNA-protein interactions in coronavirus replication and transcription**. 2011, **8**(2):237-248.
9. Coimbatore Narayanan B, Westbrook J, Ghosh S, Petrov AI, Sweeney B, Zirbel CL, Leontis NB, Berman HMJNar: **The Nucleic Acid Database: new features and capabilities**. 2013, **42**(D1):D114-D122.
10. Andronescu M, Bereg V, Hoos HH, Condon AJBb: **RNA STRAND: the RNA secondary structure and statistical analysis database**. 2008, **9**(1):340.
11. Vanegas PL, Hudson GA, Davis AR, Kelly SC, Kirkpatrick CC, Znosko BMJNar: **RNA CoSSMos: characterization of secondary structure motifs—a searchable database of secondary structure motifs in RNA three-dimensional structures**. 2011, **40**(D1):D439-D444.
12. Popena M, Szachniuk M, Blazewicz M, Wasik S, Burke EK, Blazewicz J, Adamiak RWJBB: **RNA FRABASE 2.0: an advanced web-accessible database with the capacity to search the three-dimensional fragments within RNA structures**. 2010, **11**(1):231.
13. Baulin E, Yacovlev V, Khachko D, Spirin S, Roytberg MJD: **URS DataBase: universe of RNA structures and their motifs**. 2016, **2016**.
14. Zanagina O, Kirsanov D, Baulin E, Karyagina A, Alexeevski A, Spirin SJNar: **An updated version of NPIDB includes new classifications of DNA–protein complexes and their families**. 2015, **44**(D1):D144-D153.
15. Cook KB, Kazan H, Zuberi K, Morris Q, Hughes TRJNar: **RBPDB: a database of RNA-binding specificities**. 2010, **39**(suppl_1):D301-D308.
16. Wu T, Wang J, Liu C, Zhang Y, Shi B, Zhu X, Zhang Z, Skogerbø G, Chen L, Lu HJNar: **NPInter: the noncoding RNAs and protein related biomacromolecules interaction database**. 2006, **34**(suppl_1):D150-D152.

17. Junge A, Refsgaard JC, Garde C, Pan X, Santos A, Alkan F, Anthon C, von Mering C, Workman CT, Jensen LJJD: **RAIN: RNA–protein association and interaction networks**. 2017, **2017**.
18. Lang B, Armaos A, Tartaglia GGJNar: **RNAct: Protein–RNA interaction predictions for model organisms with supporting experimental data**. 2018, **47**(D1):D601-D606.
19. Lewis BA, Walia RR, Terribilini M, Ferguson J, Zheng C, Honavar V, Dobbs DJNar: **PRIDB: a protein–RNA interface database**. 2010, **39**(suppl_1):D277-D282.
20. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PEJNar: **The protein data bank**. 2000, **28**(1):235-242.
21. Andreeva A, Howorth D, Chandonia J-M, Brenner SE, Hubbard TJ, Chothia C, Murzin AGJNar: **Data growth and its impact on the SCOP database: new developments**. 2007, **36**(suppl_1):D419-D425.
22. El-Gebali S, Mistry J, Bateman A, Eddy SR, Luciani A, Potter SC, Qureshi M, Richardson LJ, Salazar GA, Smart AJNar: **The Pfam protein families database in 2019**. 2018, **47**(D1):D427-D432.
23. Muppirlala UK, Lewis BA, Dobbs DLJJoCS, Biology C: **Computational tools for investigating RNA-protein interaction partners**. 2013, **6**(4):182.
24. Muppirlala UK, Honavar VG, Dobbs DJBb: **Predicting RNA-protein interactions using only sequence information**. 2011, **12**(1):489.
25. Cortes C, Vapnik VJMI: **Support-vector networks**. 1995, **20**(3):273-297.
26. Breiman LJMI: **Random forests**. 2001, **45**(1):5-32.
27. Cheng Z, Zhou S, Guan JJJob, biology c: **Computationally predicting protein-RNA interactions using only positive and unlabeled examples**. 2015, **13**(03):1541005.
28. Wang Y, Chen X, Liu Z-P, Huang Q, Wang Y, Xu D, Zhang X-S, Chen R, Chen LJMB: **De novo prediction of RNA–protein interactions from sequence information**. 2013, **9**(1):133-142.
29. Suresh V, Liu L, Adjeroh D, Zhou XJNar: **RPI-Pred: predicting ncRNA-protein interaction using sequence and structural information**. 2015, **43**(3):1370-1379.
30. Pancaldi V, Bähler JJNar: **In silico characterization and prediction of global protein–mRNA interactions in yeast**. 2011, **39**(14):5826-5836.
31. Lu Q, Ren S, Lu M, Zhang Y, Zhu D, Zhang X, Li TJBg: **Computational prediction of associations between long non-coding RNAs and proteins**. 2013, **14**(1):651.
32. Bellucci M, Agostini F, Masin M, Tartaglia GGJNm: **Predicting protein associations with long noncoding RNAs**. 2011, **8**(6):444.
33. Cheng Z, Huang K, Wang Y, Liu H, Guan J, Zhou SJBsb: **Selecting high-quality negative samples for effectively predicting protein-RNA interactions**. 2017, **11**(2):9.
34. research GOCJNa: **Gene Ontology annotations and resources**. 2012, **41**(D1):D530-D535.
35. research UCJNa: **Update on activities at the Universal Protein Resource (UniProt) in 2013**. 2012, **41**(D1):D43-D47.
36. Kim WJAToDS: **On optimizing an SQL-like nested query**. 1982, **7**(3):443-469.

37. Han JLJTVJTIJoVLDB: **Optimizing relational queries in connection hypergraphs: nested queries, views, and binding propagations.** 1998, **7**(1):1-11.
38. Chen JT, Yang WP, Hsu MJJotCloE: **Optimization on a case of type-d nested query.** 1988, **11**(3):309-316.
39. Sun H, Zhu X, Lu PY, Rosato RR, Tan W, Zu YJMT-NA: **Oligonucleotide aptamers: new tools for targeted cancer therapy.** 2014, **3**:e182.
40. Hoinka J, Zotenko E, Friedman A, Sauna ZE, Przytycka TMJB: **Identification of sequence–structure RNA binding motifs for SELEX-derived aptamers.** 2012, **28**(12):i215-i223.
41. Kedzierski S, Khoshnejad M, Caltagirone GTBBJ: **Synthetic antibodies: the emerging field of aptamers.** 2012, **11**:46-49.
42. Sun H, Zu YJM: **A highlight of recent advances in aptamer technology and its application.** 2015, **20**(7):11959-11980.
43. Kabsch W, Sander CJBORoB: **Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features.** 1983, **22**(12):2577-2637.
44. Drozdetskiy A, Cole C, Procter J, Barton GJJNar: **JPred4: a protein secondary structure prediction server.** 2015, **43**(W1):W389-W394.
45. Hofacker ILJNar: **Vienna RNA secondary structure server.** 2003, **31**(13):3429-3431.
46. Kloczkowski A, Ting KL, Jernigan R, Garnier JJPS, Function,, Bioinformatics: **Combining the GOR V algorithm with evolutionary information for protein secondary structure prediction from amino acid sequence.** 2002, **49**(2):154-166.
47. Liu T, Jia CJJotb: **A high-accuracy protein structural class prediction algorithm using predicted secondary structural information.** 2010, **267**(3):272-275.
48. Zhang S, Ding S, Wang TJB: **High-accuracy prediction of protein structural class for low-similarity sequences based on predicted secondary structure.** 2011, **93**(4):710-714.
49. Ding S, Zhang S, Li Y, Wang TJB: **A novel protein structural classes prediction method based on predicted secondary structure.** 2012, **94**(5):1166-1171.
50. Aldwairi M, Al-Hajasad B, Khamayseh YJljobr, applications: **A classifier system for predicting RNA secondary structure.** 2014, **10**(3):307-320.
51. Hurkens AJC, McArthur M, Moschovakis YN, Moss LS, Whitney GTJTJoSL: **The logic of recursive equations.** 1998, **63**(2):451-478.
52. Krauss AJJoAR: **Partial and nested recursive function definitions in higher-order logic.** 2010, **44**(4):303-336.

Tables

Table 1: RPiNaptaBASE statistics

Type		Total number	Unique
RNA-protein complexes		2258	2258
RNA chains		7746	1892
Protein chains		52397	5101

Additional File Legend

Additional file 1: Help. Word file containing RPiNaptaBASE’s web interface usage.

Additional file 2: Macromolecules features. Excel file containing features of RNA and protein macromolecules.

Figures

RPIN dataset

Complex dataset

Generating positive and negative datasets of RNA-Protein complexes with desired features. This mode includes: Query setting, General, RNA and Protein.

Complex

Protein dataset

Generating protein datasets of RNA-protein complexes with desired features. This mode contains two parts: Query setting and protein.

Protein

RNA dataset

Generating RNA datasets of RNA-protein complexes with desired features. This mode contains two parts: Query setting and RNA.

RNA

© 2019 - RPINaptaBASE

Figure 1

A screenshot of the Select Mode

RPINaptaBASE

Help

Contact us

About us

1

START

2

SELECT MODE

3

MAKE A QUERY

4

CHECK THE RESULT &

Result

Check result >>

Positive

Setting

General

RNA

Protein

Negative

Setting

General

RNA

Protein

+

Positive set

Query setting

200

General

RNA

Protein

-

Negative set

Query setting

200

Items limit

200

General

RNA

Protein

© 2019 - RPINaptaBASE

Figure 2

A screenshot of complex page

Page 16/22

Figure 3

A screenshot of RNA page

Figure 4

A screenshot of protein page

[RPINaptaBASE](#)
[Help](#)
[Contact us](#)
[About us](#)

1

START

2

SELECT MODE

3

MAKE A QUERY

4

CHECK THE RESULT & DOWNLOAD

	Positive Set	Negative Set	Total
Count	0	0	0
Query result %	%	%	100 %
Max Sequences	<input type="text" value="0"/>	<input type="text" value="0"/>	0
Dataset composition %	%	%	100 %
Includes in csv	☒ Headers	☐ PDB names and chains	☒ Sequences

Download

raw text

csv dataset

The preview can be different from downloaded dataset due to generating random records.

Refresh Preview

Positive preview

Row	RNA	PDBInfoRNA	Protein	PDBInfoProtein
-----	-----	------------	---------	----------------

Negative preview

Row	RNA	PDBInfoRNA	Protein	PDBInfoProtein
-----	-----	------------	---------	----------------

Back to query

© 2019 - RPINaptaBASE

Figure 5

A screenshot of result page

Figure 6

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

Figure 7

Generation of positive and negative samples

Figure 8

The RPINaptaBASE architecture

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [supplement1.xlsx](#)
- [supplement2.docx](#)