D202–D205 *Nucleic Acids Research, 2018, Vol. 46, Database issue* *Published online 23 October 2017*
doi: 10.1093/nar/gkx966

RNArchitecture: a database and a classification system of RNA families, with a focus on structural information

Pietro Boccaletto[1], Marcin Magnus[1], Catarina Almeida[1], Adriana Żyła[1], Astha Astha[1], Radosław Pluta[1], Błażej Bagiński[1], Elżbieta Jankowska[1], Stanisław Dunin-Horkawicz[1], Tomasz K. Wirecki[1], Michał J. Boniecki[1], Filip Stefaniak[1] and Janusz M. Bujnicki[1,2,*]

[1]Laboratory of Bioinformatics and Protein Engineering, International Institute of Molecular and Cell Biology in Warsaw, ul. Ks. Trojdena 4, PL-02-109 Warsaw, Poland and [2]Faculty of Biology, Adam Mickiewicz University, ul. Umultowska 89, PL-61-614 Poznan, Poland

Received August 18, 2017; Revised September 28, 2017; Editorial Decision October 06, 2017; Accepted October 16, 2017

ABSTRACT

RNArchitecture is a database that provides a comprehensive description of relationships between known families of structured non-coding RNAs, with a focus on structural similarities. The classification is hierarchical and similar to the system used in the SCOP and CATH databases of protein structures. Its central level is Family, which builds on the Rfam catalog and gathers closely related RNAs. Consensus structures of Families are described with a reduced secondary structure representation. Evolutionarily related Families are grouped into Superfamilies. Similar structures are further grouped into Architectures. The highest level, Class, organizes families into very broad structural categories, such as simple or complex structured RNAs. Some groups at different levels of the hierarchy are currently labeled as 'unclassified'. The classification is expected to evolve as new data become available. For each Family with an experimentally determined three-diemsional (3D) structure(s), a representative one is provided. RNArchitecture also presents theoretical models of RNA 3D structure and is open for submission of structural models by users. Compared to other databases, RNArchitecture is unique in its focus on structure-based RNA classification, and in providing a platform for storing RNA 3D structure predictions. RNArchitecture can be accessed at http://iimcb.genesilico.pl/RNArchitecture/.

INTRODUCTION

RNA molecules play fundamental roles in cellular processes. They have been long known to carry genetic information and to synthesize proteins. They may detect the presence of ions or small molecules in the environment, regulate gene expression at various levels (from DNA to RNA, to proteins) and catalyze chemical reactions (reviewed comprehensively in (1)). Many RNAs that have been structurally characterized form compact, functional, three-dimensional (3D) structures that determine their function and interactions with other molecules, in a similar manner to sequence-structure-function relationships that have been well described for proteins.

Knowledge of similarity between biological macromolecules enables us to cluster them, to group them into families, superfamilies and higher-level organizations, to infer their evolutionary history, to detect functional motifs and thus to predict the mechanism of their action (2). The need to compare and classify protein structures has led to the development of commonly used databases and hierarchical structural classifications, such as SCOP (3) and CATH (4). Thanks to these and other computational resources, comparing and classifying proteins was demonstrated to be of crucial importance for function inference and it continues to be used in many applications.

For RNA structures several databases have been developed. However, some of them, including RNABase (5) and SCOR (6) are not updated any more, while others, that implement automated clustering, such as HD-RNAS (7) or RNA Structure Atlas (8), provide information only about very close similarities. Currently, no comprehensive database of RNA families exists that provides information about structural similarities and dissimilarities at a level analogous to superfamily or fold in SCOP, or to topology or architecture in CATH. This unmet need has prompted us

*To whom correspondence should be addressed. Tel: +48 22 597 0750; Fax: +48 22 597 0715; Email: iamb@genesilico.pl

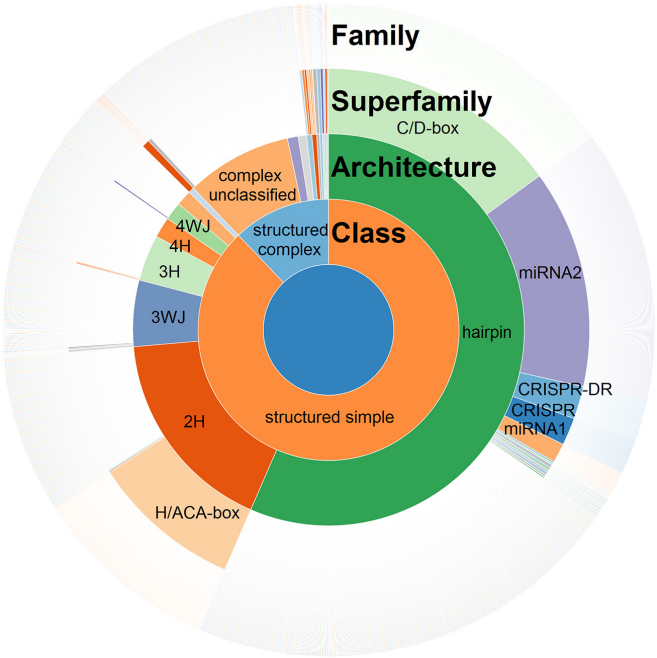

Figure 1. A sunburst plot illustrating the hierarchy of RNArchitecture and the content of the 1.0 release. The outermost layer indicates 2688 RNA Families. The successive layers combine these Families into 1721 Superfamilies, 22 Architectures and finally into two Classes. Names are shown for Classes and largest Architectures, and Superfamilies.

to develop the RNArchitecture database and its new hierarchical classification system. RNArchitecture is based on an established catalogue of Rfam families and extends this classification toward higher levels of organization built on structural considerations.

DATABASE CONTENT

The RNArchitecture database has been developed to provide a comprehensive classification system that describes relationships between RNA families, with a focus on structural similarities (Figure 1). RNArchitecture uses and organizes information from several databases, including Rfam (9) and Protein Data Bank (10), and introduces a SCOP/CATH-like hierarchical classification. The central level of classification is Family, which has been largely taken from the Rfam database. RNArchitecture includes 2688 Families of which only 2.54% (74 Families) have a structural model solved experimentally. Families group together evolutionarily related RNAs with conserved structure and detectable sequence similarity. Families whose members exhibit structural variation are further subdivided into Subfamilies. On the other hand, Families with similar structures and functions, and likely to be evolutionarily related (or at least converged to fulfill the same role in essentially the same way) are grouped into Superfamilies (which are more extensive than clans currently defined by Rfam). Superfamilies that share a similar core structure, but which are not clear homologs, are grouped into Architectures. The highest level, Class, organizes the data into very broad structural and functional categories. The coarse-graining of secondary structure in RNArchitecture is not much used at the 'evolutionary' level of classification into Superfamilies, but mostly at higher levels of Architecture and Class, which group RNAs according to structural similarities, in analogy to fold and class in SCOP. RNArchitecture also serves as a repository of theoretical models of RNA 3D structures, open for submission from users.

Dataset acquisition and processing

The dataset of Families (in each case including the multiple sequence alignment, the consensus secondary structure, and the consensus sequence) was constructed based on Rfam (release 12.3), and expanded to include known RNA families that are currently not covered by Rfam. In particular, we included group I and group II intron Families which are not included in Rfam as complete full-length sequences. Group I intron sequences and alignments were obtained from GISSD: Group I Intron Sequence and Structure Database (11), and group II intron sequences were obtained from The Database for Bacterial Group II Introns (12) and the alignments were generated by us. These Families were further subdivided into Subfamilies, as proposed in these databases.

The consensus structures were used to calculate reduced shape representations, similarly to the RNA shapes approach (13,14). We used an in-house program to convert full secondary structure representation to Level1 reduced representation, in which a single pair of brackets corresponds to two uninterrupted segments of paired residues, and then to Level2 reduced representation, in which a single pair of brackets corresponds to two series of mutually paired segments that can be interrupted by bulges and loops, but are not interrupted by residues paired with other segments (Figure 2). Level2 reduced representations were compared and the most common shapes were used, along with the 3D structural information, wherever available, to define the Architectures.

For each Family, a representative member was identified to illustrate exemplary structural information. First, for each Family with a member that has an experimentally determined 3D structure (according to Rfam annotation), the structural coordinates were obtained from the Protein Data Bank (10). For these structurally characterized RNAs, sequences were extracted from the PDB file using do_x3dna (15), and secondary structure was annotated using ClaRNA (16). Additional data, such as header, title, molecule, and information about ligands, cofactors and hetero atoms, were derived from the associated PDB file using pypdb (17). Second, for each Family without an experimentally determined 3D structure, we selected the closest sequence to the consensus sequence, in terms of sequence identity. Its secondary structure was assigned based on the Rfam consensus secondary structure. In a few selected cases (e.g., in the case of RNAs with additional structural and functional information), a different family representative was identified, and secondary structure was assigned manually, based on literature and database information.

For a number of Families without experimentally determined 3D structure, we generated 3D structural models, aiming to provide at least one 3D structural model for

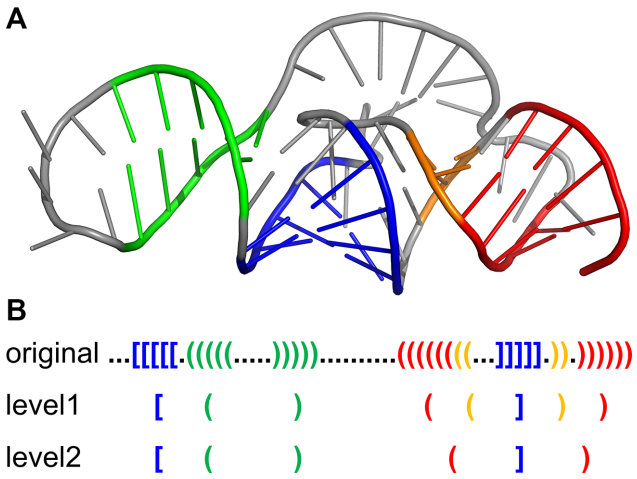

Figure 2. Example of key new information offered by the RNArchitecture database as a basis of the classification system for RNA Families. (**A**) 3D structural model of the DP (downstream peptide) Family representative. (**B**) Reduction of the DP Family representative secondary structure assignment to simplified representations used for classifications into Architectures and Classes—in this case a pseudoknot. Colors indicate conserved elements of secondary structure.

each Architecture and for the largest Superfamilies. Exceptions, for which we decided not to generate 3D structure models, include Families considered as largely unstructured, and hence unlikely to possess a stable unique 3D structure, or those without reliable structural information. Briefly, depending on the availability of a tentative structural template, a preliminary model for a family representative was generated with ModeRNA (18) or RNA Composer (19). This starting model was then refolded with SimRNA (20), with default parameters, using secondary structure restraints. Unless otherwise noted, the modeling was largely automated, and the resulting models have not been inspected for agreement with published literature, therefore they must be considered as tentative and still to be improved by more exhaustive studies.

Search

Options for database searching and querying have been implemented, including search using PDB IDs, names of Class, Architecture, Superfamily, Family, Subfamily, Rfam accession number and RNA type. The database also includes a powerful and dedicated RNA shape search tool for the exploration of the Families that contain a particular architectural motif, e.g., a simple pseudoknot, by querying the database with '([])'. Currently such searches are either not possible or very difficult to make with other databases of RNA families or RNA structures.

Database implementation

An object-relational PostgreSQL database management system is used to store all the information. The web server is implemented for the Linux operating system under Django. Programs to export, to import and to visualize have been implemented in Python 3, including our in-house package rna-pdb-tools (freely available https://github.com/mmagnus/rna-pdb-tools/). The secondary structure visualizations are generated with VARNA (21), the tertiary structure visualizations are rendered with JSmol (22).

Future prospects

The number of experimentally determined RNA molecules is increasing rapidly, in line with recent discoveries and growing interest in RNA functions. The number of experimentally determined RNA structures is also growing, albeit at a much slower rate. Therefore, RNArchitecture is expected to be updated systematically with new information. In particular, we envisage updating it, following all major updates of the Rfam database. The catalog of Families is intended to be systematically expanded, to include additional sequences e.g., from the RNAcentral database (23). Structural information will be updated with new experimentally determined RNA structures, as well as with improved theoretical models. The generation of 3D structure representatives for all Families that are predicted to form stable 3D structures is an ongoing process and we intend to expand the current dataset of models to maximize the coverage. We encourage the users of RNArchitecture to submit models of RNA 3D structures, preferably ones with experimental support, to be included in the database, as well as suggestions for improvement of the existing classification.

For the next release, we plan to expand the structural repertoire to include multiple structures (e.g., for different members of the same Family, for different functional/structural states of the same RNA, or for alternative theoretical models) and structural superpositions. New structural data will be used to update and potentially revise the classification system, in particular, the assignment to Superfamilies and Architectures, and the diversification of Classes. Another envisaged next step of the RNArchitecture database development is to link it with databases on other aspects of RNA structure, such as RMDB (24). In particular, we envisage that RNArchitecture will evolve in concert with the developments of the RNA-Puzzles experiment (25), and will serve the community of RNA structure predictors. We hope that the RNArchitecture RNA structure classification project will prompt new advances in the field, for instance facilitating and stimulating the choice of targets for theoretical prediction and experimental determination of RNA 3D structures.

AVAILABILITY

The data are accessible freely for research purposes at http://iimcb.genesilico.pl/RNArchitecture/. All RNA structures in the PDB format, images and alignments in the Stockholm format are available for download. The scripts used to process the data are part of our in-house package rna-pdb-tools (freely available at https://github.com/mmagnus/rna-pdb-tools/).

ACKNOWLEDGEMENTS

We would like to thank all members of the Bujnicki laboratory and all participants of the RNA Puzzles experiment for

fruitful discussions. We are indebted to the authors of primary databases and services, whose content could be reused or linked to by RNArchitecture, in particular, Rfam, PDB and RNAcentral.

FUNDING

Polish National Science Centre [2012/04/A/NZ2/00455 to J.M.B]. Funding for open access charge: Polish National Science Centre [2012/04/A/NZ2/00455 to J.M.B.].
Conflict of interest statement. Janusz M. Bujnicki is an Executive Editor of *Nucleic Acids Research*.

REFERENCES

1. Gesteland,R.F., Cech,T.R. and Atkins,J.F. (2005) *The RNA World*. Cold Spring Harbor Laboratory Press, NY.
2. Atkinson,H.J., Morris,J.H., Ferrin,T.E. and Babbitt,P.C. (2009) Using sequence similarity networks for visualization of relationships across diverse protein superfamilies. *PLoS One*, **4**, e4345.
3. Murzin,A.G., Brenner,S.E., Hubbard,T. and Chothia,C. (1995) SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.*, **247**, 536–540.
4. Orengo,C.A., Michie,A.D., Jones,S., Jones,D.T., Swindells,M.B. and Thornton,J.M. (1997) CATH–a hierarchic classification of protein domain structures. *Structure*, **5**, 1093–1108.
5. Murthy,V.L. and Rose,G.D. (2003) RNABase: an annotated database of RNA structures. *Nucleic Acids Res.*, **31**, 502–504.
6. Klosterman,P.S., Tamura,M., Holbrook,S.R. and Brenner,S.E. (2002) SCOR: a structural classification of RNA database. *Nucleic Acids Res.*, **30**, 392–394.
7. Ray,S.S., Halder,S., Kaypee,S. and Bhattacharyya,D. (2012) HD-RNAS: an automated hierarchical database of RNA structures. *Front. Genet.*, **3**, 59.
8. Leontis,N.B. and Zirbel,C.L. (2012) Nonredundant 3D structure datasets for RNA knowledge extraction and benchmarking. In: Leontis,NB and Westhof,E (eds). *RNA 3D Structure Analysis and Prediction*. Springer, Berlin, Heidelberg, Vol. **27**, pp. 281–298.
9. Nawrocki,E.P., Burge,S.W., Bateman,A., Daub,J., Eberhardt,R.Y., Eddy,S.R., Floden,E.W., Gardner,P.P., Jones,T.A., Tate,J. *et al.* (2015) Rfam 12.0: updates to the RNA families database. *Nucleic Acids Res.*, **43**, D130–D137.
10. Rose,P.W., Prlic,A., Altunkaya,A., Bi,C., Bradley,A.R., Christie,C.H., Costanzo,L.D., Duarte,J.M., Dutta,S., Feng,Z. *et al.* (2017) The RCSB protein data bank: integrative view of protein, gene and 3D structural information. *Nucleic Acids Res.*, **45**, D271–D281.
11. Zhou,Y., Lu,C., Wu,Q.J., Wang,Y., Sun,Z.T., Deng,J.C. and Zhang,Y. (2008) GISSD: Group I intron sequence and structure database. *Nucleic Acids Res.*, **36**, D31–D37.
12. Candales,M.A., Duong,A., Hood,K.S., Li,T., Neufeld,R.A., Sun,R., McNeil,B.A., Wu,L., Jarding,A.M. and Zimmerly,S. (2012) Database for bacterial group II introns. *Nucleic Acids Res.*, **40**, D187–D190.
13. Giegerich,R., Voss,B. and Rehmsmeier,M. (2004) Abstract shapes of RNA. *Nucleic Acids Res.*, **32**, 4843–4851.
14. Voss,B., Giegerich,R. and Rehmsmeier,M. (2006) Complete probabilistic analysis of RNA shapes. *BMC Biol.*, **4**, 5.
15. Kumar,R. and Grubmuller,H. (2015) do_x3dna: a tool to analyze structural fluctuations of dsDNA or dsRNA from molecular dynamics simulations. *Bioinformatics*, **31**, 2583–2585.
16. Walen,T., Chojnowski,G., Gierski,P. and Bujnicki,J.M. (2014) ClaRNA: a classifier of contacts in RNA 3D structures based on a comparative analysis of various classification schemes. *Nucleic Acids Res.*, **42**, e151.
17. Gilpin,W. (2016) PyPDB: a Python API for the Protein Data Bank. *Bioinformatics*, **32**, 159–160.
18. Rother,M., Rother,K., Puton,T. and Bujnicki,J.M. (2011) ModeRNA: a tool for comparative modeling of RNA 3D structure. *Nucleic Acids Res.*, **39**, 4007–4022.
19. Popenda,M., Szachniuk,M., Antczak,M., Purzycka,K.J., Lukasiak,P., Bartol,N., Blazewicz,J. and Adamiak,R.W. (2012) Automated 3D structure composition for large RNAs. *Nucleic Acids Res.*, **40**, e112.
20. Boniecki,M.J., Lach,G., Dawson,W.K., Tomala,K., Lukasz,P., Soltysinski,T., Rother,K.M. and Bujnicki,J.M. (2016) SimRNA: a coarse-grained method for RNA folding simulations and 3D structure prediction. *Nucleic Acids Res.*, **44**, e63.
21. Darty,K., Denise,A. and Ponty,Y. (2009) VARNA: Interactive drawing and editing of the RNA secondary structure. *Bioinformatics*, **25**, 1974–1975.
22. Hanson,R.M., Prilusky,J., Renjian,Z., Nakane,T. and Sussman,J.L. (2013) JSmol and the next-generation web-based representation of 3D molecular structure as applied to proteopedia. *Isr. J. Chem.*, **53**, 207–216.
23. Petrov,A.I., Kay,S.J., Gibson,R., Kulesha,E., Staines,D., Bruford,E.A., Wright,M.W., Burge,S., Finn,R.D., Kersey,P.J. *et al.* (2015) RNAcentral: an international database of ncRNA sequences. *Nucleic Acids Res.*, **43**, D123–D129.
24. Cordero,P., Lucks,J.B. and Das,R. (2012) An RNA Mapping DataBase for curating RNA structure mapping experiments. *Bioinformatics*, **28**, 3006–3008.
25. Cruz,J.A., Blanchet,M.F., Boniecki,M., Bujnicki,J.M., Chen,S.J., Cao,S., Das,R., Ding,F., Dokholyan,N.V., Flores,S.C. *et al.* (2012) RNA-Puzzles: A CASP-like evaluation of RNA three-dimensional structure prediction. *RNA*, **14**, 610–625.