International Conference on

Frontiers of Interface Between Statistics and Sciences

CONFERENCE PROCEEDINGS

December 30, 2009 to January 2, 2010

Organised in honour of Padma Vibhushan Prof. C.R. Rao in his 90th year
by CR Rao Advanced Institute of Mathematics, Statistics & Computer Science
and University of Hyderabad

Papers/Extended Abstracts

CR RAO Advanced Institute of Mathematics, Statistics and Computer Science
Prof. C.R. Rao Road, University of Hyderabad Campus
Gachibowli, Hyderabad - 500 046. Website: www.crraoaimscs.org

University of Hyderabad
Prof. C.R. Rao Road,
Gachibowli, Hyderabad - 500 046.
Website: www.uohyd.ernet.in

HD-RNAS: Hierarchical Database of RNA Structures

Shubhra Shankar Ray[1], Sukanya Halder, Dhananjay Bhattacharyya[*]

Biophysics Division, Saha Institute of Nuclear Physics

1/AF, Bidhannagar, Kolkata – 700 064, India

Abstract:

The number of RNA crystal structures, available in PDB, is growing exponentially but little advances have been made regarding management of this huge amount of information and its proper representation. There are redundant files, ambiguous synthetic sequences etc, which misguide proper analysis of these structures in a good statistical manner. Moreover, a hierarchical organization of these functional RNAs is missing in PDB or any other server. In this investigation we propose a web-based server, called *Hierarchical Database of RNA Structures* (**HD-RNAS**), for handling these issues. In this approach we have programmatically classified all available RNA crystal structures solved by X-Ray crystallography with chain length greater than nine into (i) firstly their functional type and (ii) secondly the source of the molecule. As the classification is done automatically by running an octave program on all the downloaded PDB files, it can be repeated frequently to keep pace with structure determination. We have classified the structures into one of eight classes, namely, mRNA, rRNA, tRNA, SRP RNA, Ribozyme, Riboswitch, Ribonuclease and Unclassified function at the topmost level. All the structures are finally classified according to their sources. Thus, we can determine (i) a non-redundant set of RNA structures and (ii) if available, a set of structures of similar sequences. The classification of RNA structures is available online at http://www.saha.ac.in/biop/HD-RNAS.html.

[1] Present address: Machine Intelligence Unit, Indian Statistical Institute, 203 B.T. Road, Kolkata, India

*** Address for all correspondence: E-mail: dhananjay.bhattacharyya@saha.ac.in**

The web interface also allows the user to search the database with specific criteria, *e.g.* source organism, functional type, resolution, length of the RNA chain, etc.As a part of our analysis, we have also determined the source organisms of a large number of sequences that have been otherwise designated as 'Synthetic' in PDB, by using BLAST search against the binary nucleotide sequence database available at NCBI.

Introduction:

The number of RNA structures solved by crystallography as well as NMR studies is increasing exponentially with advancement in different experimental techniques. The total number of crystal structures of RNA with oligo or polymeric length, as available in Protein Data Bank (PDB) [1] in Aug. 2009, is 1095 and the number is increasing at a pace of about 100 per year. The determination of various RNA structures, such as the hammerhead ribozyme, SRP RNA and the 5S, 16S and 23S ribosomal RNAs has greatly increased our knowledge of RNA folds and the three-dimensional organization of RNA chains. Collectively, these structures provide a large amount of information about RNA structural motifs. Similar exponential growth of number of crystal structures of proteins is also taking place in the PDB. Considering the need of classification of these proteins, there are a number of methods available, such as SCOP, FSSP, Pisces etc [2-5]. These methods can classify a protein structure based on its structural class, source organism, secondary structure content, resolution, etc. One can further determine a set of non-redundant structures of proteins, which are not evolutionarily related, for a statistical analysis in an unbiased method. In a similar manner, RNABase and SCOR [6, 7] attempted to classify the available RNA crystal structures in 2001 but failed to maintain pace of RNA structure determination as the number of structures is increasing quite fast. Furthermore, it is often important to compare several structures of RNA, which have identical sequence, same function and from same source, to understand the effect of ligand binding, crystallization environments etc. on the three-dimensional folding. As the earlier attempts to classify RNA structures failed to keep pace with the growing number of determined structures in high speed, we have adapted an automated programmatic classification scheme with minimal or no manual feedback. In order to organize and classify the information of RNA structures in PDB-files and make it available to the non-specialists, based on a user-defined criterion, we developed a web-server, called Hierarchical Database of RNA Structures (HD-RNAS).

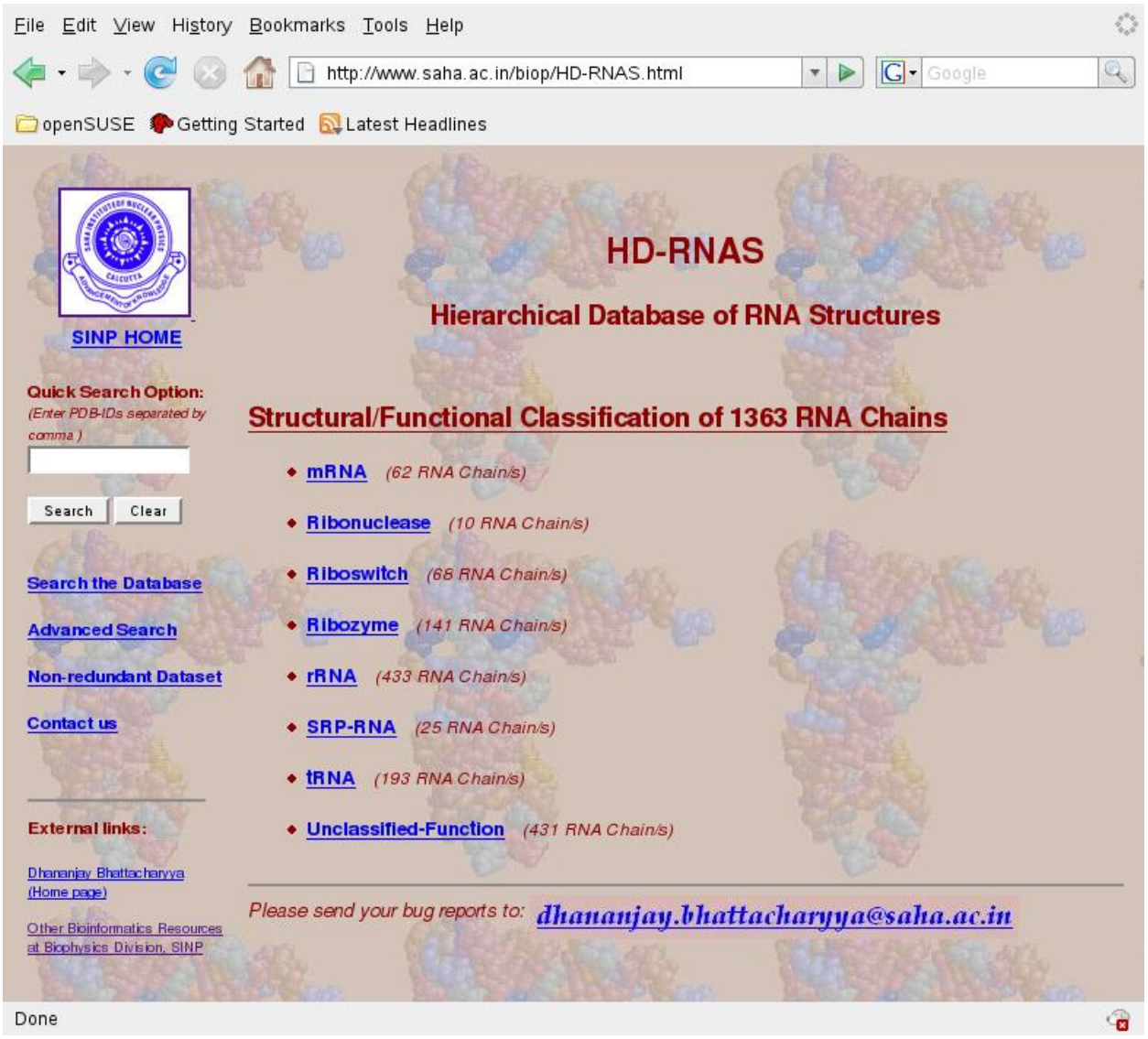

Figure 1: Homepage of HD-RNAS web-server

Materials and Methods:

We mainly focus on PDB-files containing RNA chains with chain length greater than nine, as one is more interested about biologically significant molecules and they are generally of large length, and solved by X-Ray crystallography with resolution between better than 3.5Å. We have not looked at the NMR derived structures, as there is no way to determine quality of the structure from indices like resolution, refinement quality (R-factor) etc. We have developed a software in GNU-Octave, which is similar to MATLAB scripting language, that examines and

reads the information of all the RNA structures, classifies them and creates the necessary database files and web-layout of HTML-pages displayed in the server, containing major information of each RNA chain. These HTML files then can be published in the web. As the classification and webpage database creation is done automatically by the Octave program, our automated tool is capable of frequently classifying the newly released structures with minimal manual intervention.

At the first stage, the RNA structures are classified according to their functional classes, e.g., tRNA, rRNA, mRNA etc. Along with the most common ones, we have also included some other specific types namely ribozymes, riboswitches, ribonucleases, signal recognition particle RNA, keeping in mind their growing significance in maintaining cellular machinery. A second level of classification is also needed for tRNA and rRNA based on their coding amino acids and sedimentation coefficients of the RNA chains, respectively. At the next stage, the RNA chains are divided into sub-classes according to their source organism. A number of PDB file correspond to multi-molecular complexes of several RNA as well as protein chains. We have classified all the RNA chains present in all the available PDB files. We have made no attempt to classify the structures of DNA or protein chains. Our classified database is maintained in a flat-file format, without any database management system. This has been possible as we do not keep the large PDB files at the web server and our complete database is quite small. The web-server also provides different search options with user-specified criteria. The database can be searched for specific organism or functional type of RNA. Sequence of the RNA chains in plain text formats can be obtained from the search result pages. Similarly, one can search for PDB-files in the database containing a given sequence motif. Advanced options for search allow the browser to retrieve structural entries with specified length and resolution. One can also search the database for a PDB entry to find out detailed information about the RNA chains present in the structure as well as the other members of its class. As the web-server gives a thorough classification based on functionality and sequence variation, we can determine (i) a non-redundant set of RNA structures, taking representatives from all the classes and (ii) if available, a set of structures of identical sequences.

Results and Discussions:

Out of 1601 entries obtained from PDB, as on August 2009, containing at least one RNA chain in each entry, we found that 1095 structures have been solved by X-ray crystallography. We further rejected 158 X-ray crystal structures as these either do not contain significant length of RNA chains or contain protein chains only, such as RNA polymerase. At present, the database contains 937 PDB files having structures of 1363 RNA chains with significant length. They are classified into seven major classes – mRNA, tRNA, rRNA, Signal recognition particle RNA, Ribozyme, Ribonuclease and Riboswitch. There are also RNA sequences for which annotation of any particular function is not available. We have clustered these sequences under the class named 'Unclassified-Function'. A look at this class tells us that most of the unannotated RNA chains are shorter than 30-nucleotides, whereas functional RNAs are generally longer than that. Around 400 structures among the 422 entries in the unclassified group are shorter than 30-residues.

As the classification shows, there are many RNA classes where the numbers of PDB-files are quite large. As examples, there are 66 and 60 entries of 23S and 5S rRNA of *Haloarcula marismortui*, respectively, there are 31 structures of 23S rRNA of *Deinococcus radiodurans*, 66 structures of 16S rRNA of *Thermus thermophilus*, 17 structures of $tRNA^{Phe}$ of E. *col*i, etc. Each of these classes holds a set of structures of the same molecule, but crystallized with diverse ligands or under varying physicochemical environments. Thus they carry signatures that may indicate variations introduced in the molecular structure due to ligand binding or alteration of crystallization conditions. Eventually, they can be referred to as crystallographic ensembles in analogy with statistical ensembles obtained from molecular dynamics or Monte Carlo simulations, from which ligand binding thermodynamic quantities can be calculated [8].

There are many sequences/structures where the source organism is mentioned as synthetic by the depositors. While, many of these RNA structures are truly of synthetic sequences, there are also some sequences actually taken from a natural source, but the information is not provided in machine-interpretable format. In order to determine the actual source of these RNA chains, we have used BLAST at our local site. We have used nucleotide sequence database in binary format from NCBI as available on Aug 2009 and compared our synthetic RNA sequence with all of them. We have picked up the hits having e-value less than

1.0e-5, number of aligned bases greater than 99% of the complete chain length of synthetic sequence, and sequence identity 99% or greater. The proper source organisms of 63 RNA sequences have been revealed by BLAST search that were wrongly designated as synthetic.

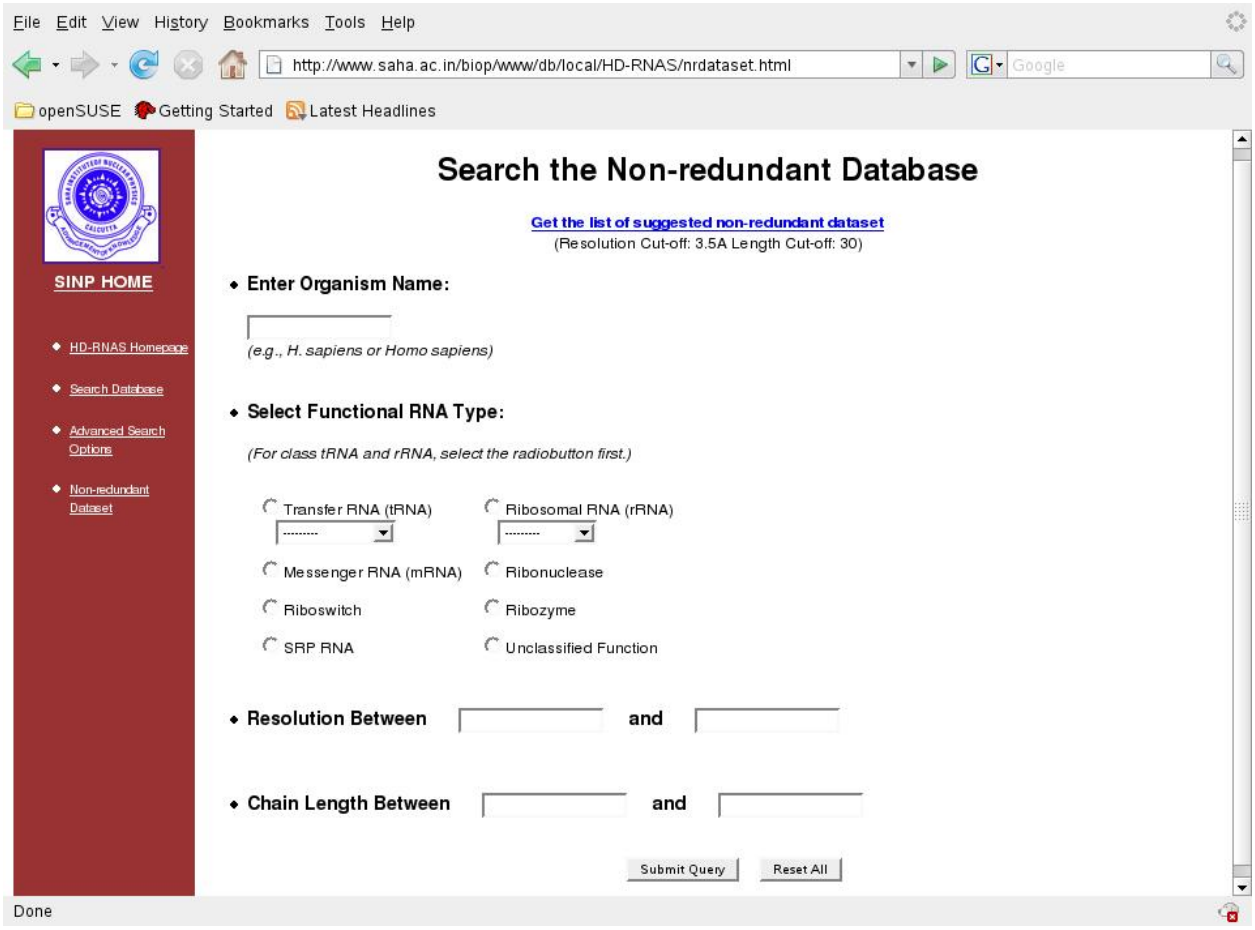

Figure 2: Searchpage of non-redundant dataset

In our attempt to obtain an unbiased set of RNA structures, we have derived a non-redundant dataset consisting of the best representative structures from each of the classes. The representative structure is the one with best resolution or, in case there are many structures having same resolution, with smallest R-factor, and among one of the larger length. As mentioned earlier there are about 400 structures of Unclassified-function and synthetic sequence and most of them are of short length. Taking a single representative of this huge number of structures of various length and probable function is also unwise. We have therefore, considered most of the structures from this sub-set as non-redundant. In order to remove several structures of same sequence being considered, we have calculated sequence identity among the structures

of this group. In cases where two sequences are 100% identical, we have considered that of best resolution and R-factor as the representative one. We have also given priority to a molecule with its full structure described rather than any fragmented part, even if the fragmented part is the one with better resolution. The non-redundant dataset thus obtained contains 353 structures. As the functional RNA molecules are generally of length larger than 30-residues, we have also tabulated a suggested non-redundant set containing representative structures from each of the functional classes as well as representatives from the unannotated groups with larger length. The members are selected with resolution better than 3.5A to make it a meaningful set of structures for real applications.

Such collection of RNA structures were used earlier to find out frequencies and structural features of different canonical and non-canonical base pairs in RNA structures [9], however, a redundant dataset was used. The nucleotide bases can form multiple hydrogen bonds using three different edges of each base in two relative orientations (cis or trans) to adopt a planar base pair, which can stack on top of each other to form a double helix. As the double helices are the only secondary structural elements of RNA, finding out such structural motifs from crystal structures is crucial in RNA structural bioinformatics. These base pairs were detected programmatically in crystal structures of all available functional RNA molecules [9]. Earlier attempt to find frequencies of these base pairs from a redundant dataset gave rise to exaggerated importance of several non-canonical base pairs. Usage of a non-redundant dataset of PDB structures can give proper unbiased estimate of relative importance of these non-canonical base pairs as shown in Table 1. It is clearly indicated in the Table that few base pairs, such as A:C W:S T, C:U W:W C, A:U S:W C, etc had shown large frequency of occurrence in the previous study [9], which reduces significantly indicating their less importance in RNA folding.

Conclusion:

Thus HD-RNAS has every potential to become a necessary tool of RNA research in comparing similar structures of the same molecule as well as among different functional classes, determining free energy of ligand-binding, protein-induced conformational alterations etc. Of

course there are some limitations of this study, mainly due to inadequate information provided by the PDB files. The database may need additional manual intervention when structures of new class of RNA become available. For example, one can expect structures of viral RNA to be classified into a separate group.

Acknowledgement:

We are thankful to Rahul Pal for technical support. We are grateful to Department of Biotechnology, Government of India for partial financial support.

References:

1. Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N. and Bourne, P.E. (2000) The Protein Data Bank. *Nucl. Acids Res.*, **28**, 235-242.
2. Hubbard, T.J.P., Murzin, A.G., Brenner, S.E. and Chothia, C. (1997) SCOP: A structural classification of proteins database. *Nucl. Acids Res.*, **25**, 236-239.
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. Holm, L. and Sander, C. (1997) Dali/FSSP classification of three-dimensional protein folds. *Nucl. Acids Res.*, **25**, 231-234.
5. Wang, G.L. and Dunbrack, R.L. (2005) PISCES: recent improvements to a PDB sequence culling server. *Nucl. Acids Res.*, **33**, W94-W98.
6. Murthy, V.L. and Rose, G.D. (2003) RNABase: an annotated database of RNA structures. *Nucl. Acids Res.*, **31**, 502-504.
7. Klosterman, P.S., Tamura, M., Holbrook, S.R. and Brenner, S.E. (2002) SCOR: A Structural Classification of RNA database. *Nucl. Acids Res.*, **30**, 392-394.
8. Samanta S., Chakrabarti J., Bhattacharyya D. (2010) Changes in thermodynamic properties of DNA base pairs in protein-DNA recognition. *J. Biomol. Struct. Dynam.*, **27**, 429-442.
9. Das J, Mukherjee S, Mitra A and Bhattacharyya D (2006) *J. Biomol. Struct. Dynam.* **24**, 149-161.

Table 1: Frequencies of different non-canonical base pairs in cis as well as trans orientation, those occur in RNA structures. The frequencies calculated from earlier published [9] redundant dataset of RNA crystal structures are shown in parenthesis. Whenever the frequencies calculated from the redundant and the non-redundant datasets differ significantly, they are highlighted.

	Cis Basepair											
	Ade W	**Ade H**	**Ade S**	**Gua W**	**Gua H**	**Gua S**	**Cyt W**	**Cyt H**	**Cyt S**	**Ura W**	**Ura H**	**Ura S**
				(404)		(197)				(6861)		
Ade W	***(62) 12***			150	(32) 4	48	(184) 51		(143) 24	1816		(41) 10
Ade H				(28) 9						(201) 79		***(52) 10***
Ade S				(13) 4						***(73) 19***		
					(158)		(21614)			(2769)		
Gua W					51		6041			846		
Gua H												
Gua S										(39) 11	(12) 57	
Cyt W							(48) 13	(18) 5		***(48) 3***		
Cyt H									(13) 9			
Cyt S												
Ura W										(360) 84	(15) 7	
Ura H												
Ura S												

Trans Basepairs

	Ade W	Ade H	Ade S	Gua W	Gua H	Gua S	Cyt W	Cyt H	Cyt S	Ura W	Ura H	Ura S
	(266)	(204)				(350)						
Ade W	61	84	(42) 7		(16) 9	88	(37) 10		***(180) 3***	(210) 62		(97) 69
		(437)	(176)			(2323)				(1193)		
Ade H		109	51			558	(336) 69			415	(23) 8	
						(1517)						
Ade S			***(45) 6***			289				***(57) 9***		
					(65)							
Gua W				(1) 7	41		(132) 63			***(44) 4***	(18) 8	
Gua H							(22) 5	(33) 12				
						(130)						
Gua S						32	***(67) 13***			***(44) 5***		
Cyt W							(8) 3	(22) 8				
Cyt H									***(45) 9***			***(38) 6***
Cyt S												
Ura W										(50) 21	(53) 17	
Ura H												
Ura S												