Nucleic Acids Research, 2002, Vol. 30, No. 1 395–397

NCIR: a database of non-canonical interactions in known RNA structures

Uma Nagaswamy, Maia Larios-Sanz, James Hury, Shakaala Collins, Zhengdong Zhang, Qin Zhao and George E. Fox*

Department of Biology and Biochemistry, University of Houston, 369 Science and Research Building 2, Houston, TX 77204-5001, USA

Received September 17, 2001; Accepted September 18, 2001

ABSTRACT

The secondary and tertiary structure of an RNA molecule typically includes a number of non-canonical base–base interactions. The known occurrences of these interactions are tabulated in the NCIR database, which can be accessed from http://prion.bchs.uh.edu/bp_type/. The number of examples is now over 1400, which is an increase of >700% since the database was first published. This dramatic increase reflects the addition of data from the recently published crystal structures of the 50S (2.4 Å) and 30S (3.0 Å) ribosomal subunits. In addition, non-canonical interactions observed in published crystal and NMR structures of tRNAs, group I introns, ribozymes, RNA aptamers and synthetic oligonucleotides are included. Properties associated with these interactions, such as sequence context, sugar pucker conformation, glycosidic angle conformation, melting temperature, chemical shift and free energy, are also reported when available. Out of the 29 anticipated pairs with at least two hydrogen bonds, 28 have been observed to date. In addition, several novel examples, not generally predicted, have also been encountered, bringing the total of such pairs to 36. Added to this list are a variety of single, bifurcated, triple and quadruple interactions. The most common non-canonical pairs are the sheared GA, GA imino, AU reverse Hoogsteen, and the GU and AC wobble pairs. The most frequent triple interaction connects N3 of an A with the amino of a G that is also involved in a standard Watson–Crick pair.

INTRODUCTION

More than half the nucleotides in typical RNAs participate in normal Watson–Crick base pairing, which constitutes the regular A-form helices. The two standard pairs are remarkably isosteric, each of the four combinations capable of substituting one another. As a consequence of their interchangeability, there is no distortion in the canonical A-form duplex. Besides the regular Watson–Crick pairs, a large collection of specific base–base interactions have been enumerated and frequently observed in several crystal and NMR structures of RNA molecules (1–3). It is now widely accepted that such non-standard interactions stabilize the secondary as well as the tertiary structures of RNA (4). Recent progress in RNA structural biology has revealed the presence of a multitude of motifs involving non-standard interactions (5,6). In fact, in the past year, the RNA structure database has expanded several-fold due to the release of atomic resolution structures of the 50S and 30S ribosomal subunits (7,8). These structures highlight the importance of non-standard interactions in shaping complex molecular machinery such as the ribosome. While the occurrences of these non-standard interactions are well recognized, general principles governing their associations are still emerging and will be facilitated by an extensive data set.

Several compilations of non-canonical base–base interactions have been presented in previous reviews (9,10). These compilations were based on the optimization of hydrogen bonding, symmetry and base type (purine and pyrimidine). The best-characterized examples of base–base interactions are those involving two bases connected by at least two hydrogen bonds. A recent tabulation by Burkard *et al*. (2) lists 29 possible pairs of this type including the standard pairs. There are two major types of pairs, normal and flipped. In what are considered to be normal pairs, the strands are antiparallel and the base is in the *anti* orientation relative to the ribose. In the flipped arrangements either the strands must be parallel or the base must be switched to the *syn* configuration. Within each category one can further distinguish purine–purine, purine–pyrimidine and pyrimidine–pyrimidine types. An alternative classification scheme was more recently proposed by Leontis and Westhof (10), based on the three interacting edges of the purines and pyrimidines that can potentially act as hydrogen bond donors and acceptors. These are the Watson–Crick, the Hoogsteen edge for purines and the CH edge for pyrimidines, and the sugar edge or the shallow groove edge. The bases can adopt either *cis* or *trans* orientation with respect to the glycosidic bond. The interacting edges along with the glycosidic bond orientation results in 12 distinctive edge-to-edge families. In constructing the database of non-canonical interactions in RNA, we attempted to classify each base pair in the two-hydrogen-bond category according to these two classification schemes. Together, these classifications help in organizing the

*To whom correspondence should be addressed. Tel: +1 713 743 8363; Fax: +1 713 743 8351; Email: fox@uh.edu

base–base interactions into isosteric families and hence aid in motif recognition in RNA sequences (11).

DATABASE DESCRIPTION

The primary objective of this database is to provide rapid access to all the RNA structures in which a particular, rare base–base interaction has been observed. A secondary objective is to summarize the important properties that have been associated with each non-canonical interaction. A literature survey of RNA structures was performed to obtain information on the non-canonical interactions. At the time of writing, approximately 1400 occurrences of non-canonical interactions have been identified. Each of these is tabulated in accordance with the classification scheme described above.

The primary page provides a description of the database, terminologies, conventions, contributors and links to several records. The database has two non-interactive records: base pair type and author type. The base pair type record provides a summary of all the entries grouped into the aforementioned interaction categories, author information and a link to individual pages. Individual entries provide access to citation information, the immediate sequence context, a brief common name for the structures in which a particular interaction was observed, and a visual representation of the geometry of the interaction. Additional structural information, such as melting point, chemical shift, comments, etc., is provided when this information is available.

ANALYSIS OF INTERACTIONS

One of the primary reasons for developing this database was the expectation that, as it grows, it will probably become clear that certain non-standard interactions will frequently be associated with particular local environments. Of the 29 anticipated two-hydrogen-bond base pairs (2), the CC N3–amino, symmetric pair is the only pair that has not been observed. In addition to the two-hydrogen-bond category, numerous pairs connected by a single hydrogen bond have been observed. Typically these can be classified as a variant of one or two of the primary types. In terms of total numbers, the vast majority of the non-canonical pairs encountered are normal in that the base is in the *anti* configuration relative to the sugar and the strands are antiparallel. A few cases of the unusual *syn* and C2′-*endo* combination with strands in the parallel orientation have also been encountered. Overall, GU wobble is the most prevalent, followed by GA sheared, AU reverse Hoogsteen and GA imino. Because of its geometric property, GU wobble pairs fit very smoothly in a regular A-form helix. GU wobble pairs frequently exchange with regular Watson–Crick pairs and, depending on the sequence environment, with AC pairs (12,13). When A is protonated at the N1 position, it can hydrogen bond with a C in a geometry that is isosteric with the GU wobble pair. Detailed explanation of the GU and the related pairs is provided elsewhere (14).

The GA sheared pair serves as a conserved building block in tertiary folding of many RNAs. In GNRA-type tetraloops and several hexaloops, GA sheared mismatch occurs as an integral part of the U-turn folding motif, a common structural motif observed in several RNA structures (5,15). The GA imino pair is frequently found at the ends of helices and only rarely in internal loops. One of the neighbors is always non-canonical, while the other is a standard Watson–Crick pair.

The AU reverse Hoogsteen base pair is the most abundant AU interaction observed in the ribosomal RNAs. It occurs as an integral part of commonly observed RNA motifs such as the bacterial loop E (internal loop) and the Sarcin/Ricin loop motifs (16). In these internal loops, this base pair is always flanked by non-canonical base pairs. The types of neighboring pairs include the AA N7-amino symmetric, a sheared GA base pair and the GU wobble base pair, but never a standard Watson–Crick pair. The AC reverse Hoogsteen base pair is perfectly isosteric with an AU reverse Hoogsteen pair and is always flanked by non-canonical base pairs.

The unusual, never before encountered, GU N3-imino, amino-4-carbonyl, AC reverse wobble and GU reverse wobble pairs have been observed in the 50S ribosomal subunit as a part of base triples. Several pairs involving the *trans* or the polarized CH edge have also been observed in both ribosomal subunits. The pH-dependent protonation of A, C and G provides additional hydrogen donors. Accordingly, a CA+ mismatch has been observed in the catalytic center of a hairpin ribozyme (17) and a GA+ mismatch has been observed in the crystal structure of an RNA duplex (18). A C+GC triple base interaction observed in mutant TAR RNA is the only known example of mismatches with protonated cytosine residues (19). Yet another variation of the base pair category is the bifurcated pairings. Bifurcated pairings were originally identified in tRNA (20). Since then, several examples of bifurcated pairings have been observed in RNAs including the loop E of bacterial 5S rRNA, UUCG tetraloops, the 50S and the 30S ribosomal subunits (3).

Base pairs of the standard and the non-canonical types can participate in hydrogen bonding with a third base, generally in a coplanar orientation. Triple base interactions were originally identified in the crystal structure of yeast $tRNA^{Phe}$ (21). Triple base interactions are of three general types. In type I, the core of the base triple is a Watson–Crick pair and the third base interacts with the pair on the major groove side. In type II, the third base interacts with a Watson–Crick pair on the minor groove side. In type III, all three bases form non-standard base pairs, connected by three to as many as six hydrogen bonds. To date, 174 different triple interactions have been observed. It is noteworthy that the most common base triple is an AGC (N3–amino; Watson–Crick) belonging to the minor groove category involving nucleotide A. Recently, interactions involving A residues in the minor groove of regular Watson–Crick pairs observed in the 50S ribosomal subunit have been designated as the A-minor motif (22).

The most spectacular base–base hydrogen bond interactions observed thus far in RNA are the base quadruples. A base quadruple interaction was first identified in the P4–P6 domain of group I introns (23). Since then, base quadruples have been observed in an RNA triplex of a ribosomal frameshifting viral pseudoknot structure (24), an RNA aptamer that binds the malachite green chromophore (25), and the 50S (7) and the 30S ribosomal subunits (8). Quadruple interactions occur at helical junctions, primarily providing stacking interface for cross-domain interactions. The information provided by this database will prove useful for developing predictive rules about non-canonical interactions in the structures of naturally occurring RNAs, and aid in the global understanding of the complex architecture of RNA.

ACKNOWLEDGEMENTS

This work was supported in part by grants from the Robert A. Welch foundation (E-1451) and the National Aeronautics and Space Administration (NAG5-1840) to G.E.F.

REFERENCES

1. Saenger,W. (1984) In Cantor,C.R. (ed.), *Principles of Nucleic Acid Structure*. Springer-Verlag, NY, pp. 116–158.
2. Burkard,M.E., Turner,D.H. and Tinoco,I.,Jr (1999) In Gesteland,R.F., Cech,T.R. and Atkins,J.F. (eds), *The RNA World*, 2nd edn. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY, pp. 675–680.
3. Nagaswamy,U., Voss,N., Zhang,Z. and Fox,G.E. (2000) Database of non-canonical base pairs found in known RNA structures. *Nucleic Acids Res.*, **28**, 375–376.
4. Hermann,T. and Patel,D.J. (1999) Stitching together RNA tertiary architectures. *J. Mol. Biol.*, **294**, 829–849.
5. Moore,P.B. (1999) Structural motifs in RNA. *Annu. Rev. Biochem.*, **68**, 287–300.
6. Batey,R.T., Rambo,R.P. and Doudna,J.A. (1999) Tertiary motifs in RNA structure and folding. *Angew. Chem. Int. Ed. Engl.*, **38**, 2326–2343.
7. Ban,N., Nissen,P., Hansen,J., Moore,P.B. and Steitz,T.A. (2000) The complete atomic structure of the large ribosomal subunit at 2.4 Å resolution. *Science*, **289**, 905–920.
8. Wimberly,B.T., Brodersen,D.E., Clemons,W.M.,Jr, Morgan-Warren,R.J., Carter,A.P., Vonrhein,C., Hartsch,T. and Ramakrishnan,V. (2000) Structure of the 30S ribosomal subunit. *Nature*, **407**, 327–339.
9. Leontis,N.B. and Westhof,E. (1998) Geometric nomenclature and classification of RNA base pairs. *Q. Rev. Biophys.*, **31**, 399–455.
10. Leontis,N.B. and Westhof,E. (2001) Geometric nomenclature and classification of RNA base pairs. *RNA*, **7**, 499–512.
11. Bourdeau,V., Ferbeyre,G., Pageau,M., Paquin,B. and Cedergren,R. (1999) The distribution of RNA motifs in natural sequences. *Nucleic Acids Res.*, **27**, 4457–4467.
12. Masquida,B. and Westhof,E. (2000) On the wobble GoU and related pairs. *RNA*, **6**, 9–15.
13. Rousset,F., Pelandakis,M. and Solignac,M. (1991) Evolution of compensatory substitutions through G.U intermediate state in *Drosophila* rRNA. *Proc. Natl Acad. Sci. USA*, **88**, 10032–10036.
14. Gautheret,D., Konings,D. and Gutell,R.R. (1995) G.U base pairing motifs in ribosomal RNA. *RNA*, **1**, 807–814.
15. Elgavish,T., Cannone,J.J., Lee,J.C., Harvey,S.C. and Gutell,R.R. (2001) AA.AG@helix.ends: A:A and A:G base-pairs at the ends of 16 S and 23 S rRNA helices. *J. Mol. Biol.*, **310**, 735–753.
16. Leontis,N.B. and Westhof,E. (1998) A common motif organizes the structure of multi-helix loops in 16 S and 23 S ribosomal RNAs. *J. Mol. Biol.*, **283**, 571–583.
17. Cai,Z. and Tinoco,I. (1996) Solution structure of loop A from the hairpin ribozyme from tobacco ringspot virus satellite. *Biochemistry*, **35**, 6026–6036.
18. Pan,B., Mitra,S.N. and Sundaralingam,M. (1999) Crystal structure of an RNA 16-mer duplex R(GCAGAGUUAAAUCUGC)2 with nonadjacent G(syn).A+(anti) mispairs. *Biochemistry*, **38**, 2826–2831.
19. Brodsky,A.S., Erlacher,H.A. and Williamson,J.R. (1998) NMR evidence for a base triple in the HIV-2 TAR C-G.C+ mutant–argininamide complex. *Nucleic Acids Res.*, **26**, 1991–1995.
20. Auffinger,P. and Westhof,E. (1999) Singly and bifurcated hydrogen-bonded base-pairs in tRNA anticodon hairpins and ribozymes. *J. Mol. Biol.*, **292**, 467–483.
21. Quigley,G.J. and Rich,A. (1976) Structural domains of transfer RNA molecules. *Science*, **194**, 796–806.
22. Nissen,P., Ippolito,J.A., Ban,N., Moore,P.B. and Steitz,T.A. (2001) RNA tertiary interactions in the large ribosomal subunit: the A-minor motif. *Proc. Natl Acad. Sci. USA*, **98**, 4899–4903.
23. Cate,J.H., Gooding,A.R., Podell,E., Zhou,K., Golden,B.L., Kundrot,C.E., Cech,T.R. and Doudna,J.A. (1996) Crystal structure of a group I ribozyme domain: principles of RNA packing. *Science*, **273**, 1678–1685.
24. Su,L., Chen,L., Egli,M., Berger,J.M. and Rich,A. (1999) Minor groove RNA triplex in the crystal structure of a ribosomal frameshifting viral pseudoknot. *Nature Struct. Biol.*, **6**, 285–292.
25. Baugh,C., Grate,D. and Wilson,C. (2000) 2.8 Å crystal structure of the malachite green aptamer. *J. Mol. Biol.*, **301**, 117–128.