Going beyond base-pairs: topology-based characterization of base-multiplets in RNA

SOHINI BHATTACHARYA,[1] AYUSH JHUNJHUNWALA,[1] ANTARIP HALDER,[1,3] DHANANJAY BHATTACHARYYA,[2] and ABHIJIT MITRA[1]

[1]Center for Computational Natural Sciences and Bioinformatics (CCNSB), International Institute of Information Technology (IIIT-H), Gachibowli, Hyderabad 500032, India

[2]Computational Science Division, Saha Institute of Nuclear Physics (SINP), 1/AF, Bidhannagar, Kolkata 700064, India

ABSTRACT

Identification and characterization of base-multiplets, which are essentially mediated by base-pairing interactions, can provide insights into the diversity in the structure and dynamics of complex functional RNAs, and thus facilitate hypothesis driven biological research. The necessary nomenclature scheme, an extension of the geometric classification scheme for base-pairs by Leontis and Westhof, is however available only for base-triplets. In the absence of information on topology, this scheme is not applicable to quartets and higher order multiplets. Here we propose a topology-based classification scheme which, in conjunction with a graph-based algorithm, can be used for the automated identification and characterization of higher order base-multiplets in RNA structures. Here, the RNA structure is represented as a graph, where nodes represent nucleotides and edges represent base-pairing connectivity. Sets of connected components (of n nodes) within these graphs constitute subgraphs representing multiplets of "n" nucleotides. The different topological variants of the RNA multiplets thus correspond to different nonisomorphic forms of these subgraphs. To annotate RNA base-multiplets unambiguously, we propose a set of topology-based nomenclature rules for quartets, which are extendable to higher multiplets. We also demonstrate the utility of our approach toward the identification and annotation of higher order RNA multiplets, by investigating the occurrence contexts of selected examples in order to gain insights regarding their probable functional roles.

Keywords: RNA structural bioinformatics; RNA structural elements; RNA as graph; graph mining; topology of RNA base-multiplets; nomenclature of RNA base-quartets

INTRODUCTION

Complex RNA structures are modular in nature (Grabow and Jaeger 2013) and are associated with multiple levels of structural hierarchy. At the bottommost level there is the stretch of nucleotide sequence and at the topmost level these sequences form complex folded RNA structures. Structural intricacies of RNA molecules determine their functions, including various structural roles (Schneemann 2006; Clemson et al. 2009), diverse regulatory roles (Sashital and Butcher 2006; Henkin 2008), and enzymatic roles (Cech et al. 1981; Guerrier-Takada et al. 1983; Doudna and Cech 2002) in different biochemical pathways. RNA structures are characterized by double helical stems and various types of loop motifs (Leontis and Westhof 2002, 2003) which constitute the RNA secondary structure or the stem–loop structure. The stem and loop elements fold to form more complex tertiary motifs via their mutual interactions. The large RNA structures can hence be described as a collection of noncovalently interacting 3D stem–loop motifs of different sizes, as interconnected modules, with different connectivity patterns and geometries. The compact folding of an RNA molecule brings together multiple nucleotides (usually not consecutive in sequence), which interact with each other via different noncovalent interactions to form higher order structural elements including base-multiplets. Base-pairs, composed of two interacting nucleobases connected by hydrogen bonds, which usually maintain a significantly planar geometry, may be considered as the fundamental building blocks of complex RNA structures. Apart from the canonical A:U

[3]**Present address:** Solid State and Structural Chemistry Unit, Indian Institute of Science, Bangalore 560012, India

Corresponding author: abi_chem@iiit.ac.in

Article is online at http://www.rnajournal.org/cgi/doi/10.1261/rna.068551.118.

and G:C pairs, which are the mainstay of the helical segments, a large variety of noncanonical base-pairs are observed in RNA structures. The geometries and stabilities of base-pairs (Šponer et al. 2005a,b,c; Sharma et al. 2008, 2010a,b; Mládek et al. 2009; Chawla et al. 2011; Halder et al. 2014), as well as the role of base-pairing interactions in dynamics and functionalities of different RNA molecules have been studied extensively (Réblová et al. 2011; Havrila et al. 2013; McPhee et al. 2014) in contemporary RNA research.

It is possible that the understanding of how base-pairs further interact with other bases and base-pairs to form multiplets, several of which recur in different contexts, would provide us with insights complementing the essentially hierarchical understanding of RNA structures. Some of the widely observed higher order multiplets are base-triplets (recurrent cluster of three bases connected by base-pairing interactions), and base-quartets (the cluster of four nucleotide bases, connected with each other by base-pairing interactions). The role of these multiplets in providing stability to different structural motifs in different structural and functional contexts has also been reported. For example, RNA triplets are reported to be involved in various functionalities in different structural contexts, such as, stabilization of L-shaped complex folded structures of t-RNAs (Kim et al. 1974a,b), stabilization of A-minor motifs (Nissen et al. 2001), formation of the sharp kink in the backbone and a typical hinge like motion of the k-turn motifs (Klein et al. 2001; Rázga et al. 2004; Réblová et al. 2011), etc. Similarly, base-quartets are also observed in different functional RNAs, where they show typical functionalities depending on their contexts of occurrences. The most widely studied RNA base-quartet is the G-quartet unit found in the RNA G-quadruplex structure, where more than one unit stack together, held by metal ions present in the center (Malgowska et al. 2016). RNA G-quadruplexes are involved in important functional roles (Agarwala et al. 2015; Malgowska et al. 2016), which include translation regulation, mRNA processing, mitochondrial transcription termination, post-transcriptional modification (such as 3′ end processing and alternative polyadenylation), mRNA localization, alternative splicing, RNA protein interaction and maintenance of telomeric chromosomes (Biffi et al. 2012; Montero et al. 2016). RNA base-quartets are also found in other functional RNA molecules. For example, in the aptamer domain of the purine riboswitch (Serganov et al. 2004; Sharma et al. 2009), and in the binding pocket of *T. maritima* Lysine riboswitch (Serganov et al. 2008). Quartets are also found to play important roles in RNA tetraloop-receptor interactions, which are most common and well-studied organizers of folded RNA structures (Wu et al. 2012). However, a comprehensive understanding about the role of the higher order multiplets in the structural dynamics of RNA is yet to be achieved, especially for multiplets beyond base-triplets.

Developing a reliable method for automated identification of base-multiplets present in large RNA structures is essential for addressing this requirement. This in turn requires a comprehensive characterization scheme for RNA base-multiplets, which can facilitate their algorithm driven automated detection. At the turn of the century, Nagaswamy et al. (2000, 2002) and then in 2009, Xin and Olson (Xin and Olson 2009) tried to classify and annotate RNA base-triplets based on the Sanger scheme (Hornby 1993) for nomenclature of noncanonical base-pairs. In this context, in 2001, Leontis and Westhof laid an excellent foundation by proposing an annotation scheme which characterizes base-pairs in terms of three interaction edges (Watson–Crick edge, Hoogsteen edge, and Sugar edge) and mutual orientation (cis and trans) of the two bases (Leontis and Westhof 2001). According to this scheme, popularly referred to as the LW scheme, RNA base-pairs are classified into 12 distinct geometric families (W:W, W:H, W:S, H:H, H:S, S:S base-pairs in both cis and trans geometries). The basis of the LW nomenclature scheme is explained in Figure 1 and is described in detail in the Materials and Methods section. Later, in 2012, Almakarem and coworkers extended the LW nomenclature and proposed a classification and annotation scheme for base-triplets based on combinatoric enumeration of different base-pair geometries (Abu Almakarem et al. 2012). For example, UAA *cis* Watson–Crick/Watson–Crick and trans Hoogsteen/Hoogsteen describes a base-triplet, where a central base adenine (A) pairs with Watson–Crick (WC) edge of a uracil (U) in cis orientation and Hoogsteen (H) edge of another adenine in trans orientation, via its WC and H edges, respectively. Based on their annotation scheme all possible triplets are classified into 108 potential geometric families,[4] of which 68 families are observed in known RNA structures. In this geometry-based approach, triplets are annotated by mentioning their base combination and their geometric family as defined by the constituent base-pairing geometries. Such a system for classification and annotation, in principle, opens up the possibility for developing automated tools for identification and characterization of base-pairs and base-multiplets present in complex RNA structures.

However, there are two complicating issues. Since higher order multiplets are composed of multiple base-pair units, one complication lies with the unambiguous detection of the constituent base-pairs. Available base-pair detection tools implement different algorithms to identify base-pairs and annotate them based on either the Sanger scheme or the LW scheme. There are two such widely used programs, namely "find_pair" (Lu et al. 2010) and

[4]Note that each triplet can be written in two different ways. For example, AUC cWW/cHW and CUA cWH/cWW are actually the same triplet but fall under two different families according to Almakarem's classification. Therefore, the actual number of unique triplet families is 108, which is exactly half of the total numerically possible varieties (i.e., 216) as explained in section "Topology is important for the nomenclature of RNA base-multiplets" under Results and Discussion section.

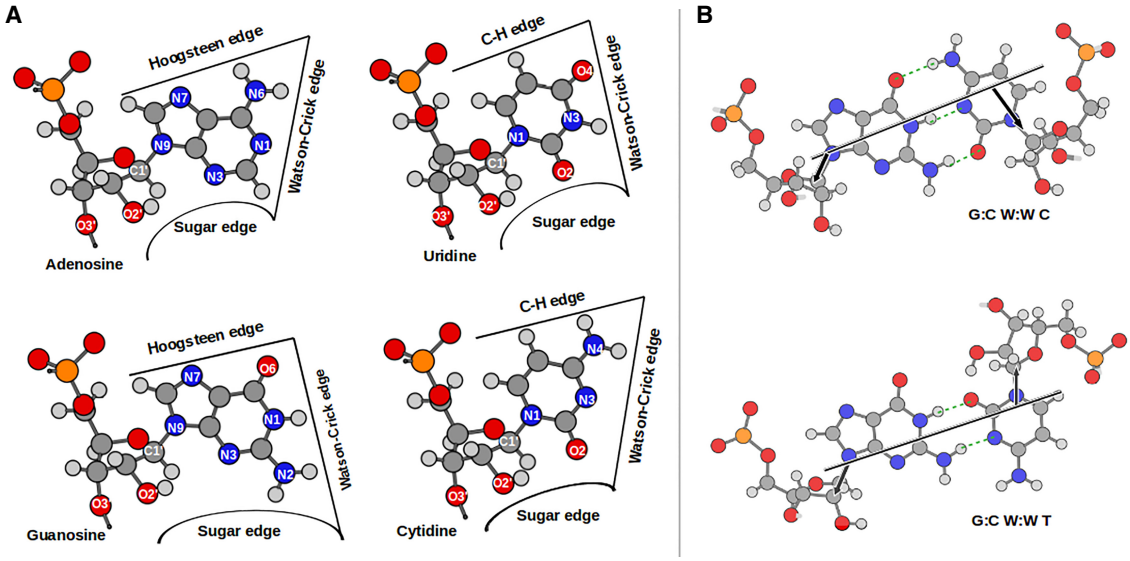


FIGURE 1. (*A*) Three distinct base-pairing edges of the RNA bases: the Watson–Crick edge (W), the Hoogsteen edge (H) (C-H edge in pyrimidines), and the Sugar edge (S). A given edge of one base can potentially pair up with any one of the three edges of a second base, which have compatible hydrogen bond donors and acceptors, to form a base-pair. (*B*) Two bases can approach in either cis or trans orientation of the glycosidic bonds around the axis defined by drawing a line parallel to and between the hydrogen bonds joining the edges. Broken lines represent interbase hydrogen bonding interactions.

"BPFind" (Das et al. 2006), which differ in the way base-pairs are defined. "Find_pair" is a major component of the 3DNA software suite (Lu and Olson 2008; Zheng et al. 2009), and is also used in the DSSR software tool (Lu et al. 2015). This program follows a purely geometry-based criteria for identification of base-pairs of all types—canonical, noncanonical, and base-pairs having modified bases, irrespective of their protonation and tautomeric state. BPFind, on the other hand implements a hypothesis driven approach (which is described in the Materials and Methods section) and checks for some predefined geometric criteria to detect and annotate base-pairs (as per the LW scheme). BPFind detects only those base-pairs that have at least two hydrogen bonds and significantly good geometry. In this study, we have chosen BPFind software for detection of base-pair units, as this software provides a comprehensively annotated list of base-pairs having significantly good geometry. It also explicitly identifies class-I protonated base-pairs (protonation in base-pairing edges) and base-pairs or base-multiplets having bifurcated geometry (one central base interacts with more than one other bases through a single interaction edge). The rationale for choosing BPFind software is more elaborately discussed in section S4 of the Supplemental Material.

A more problematic complicating issue, however, arises from the observation that constituent bases in the structural elements, having order higher than triplets, can interact with each other in distinctly different ways within the same order. In fact, this complication also arises in the cases of pairwise interconnected triplets when instances with bifurcated base-pairing are placed in a separate class, distinct from instances of normal cyclic triplets. Evidently, unambiguous characterization of base-multiplets must incorporate the information about topology, i.e., about the mutual interaction pattern of constituent bases.

One of the convenient ways to explain topologies is inbuilt in the graph theoretic approach. Graph representation is widely used to describe RNA secondary structures and motifs (Gan et al. 2004). One of the important examples is RNA tertiary structure cyclic motifs (Lemieux and Major 2006), which is represented by a graph, where vertices or nodes represent nucleotides and edges (termed as "interaction arc") represent all types of interactions, including base-pairing, stacking and phosphodiester linkage. Later, based on these cyclic motifs, a new more reliable approach for structure prediction has been evolved (Parisien and Major 2008). In addition to that, various graph-based algorithms like the graph partitioning approach (Kim et al. 2014) and subgraph isomorphism algorithms (Harrison et al. 2003; Djelloul and Denise 2008) have been used for finding modularity in RNA structures and mining recurrently observed substructural patterns and motifs present in large RNA molecules.

In order to develop a comprehensive characterization and annotation scheme for RNA base-multiplets, in conjunction with a BPFind based base-pair identification, here we present a graph-based representation scheme for various possible arrangements (topologies) of constituent nucleotides. We also propose a topology-based nomenclature convention of quartets which can be further extended to pentets, sextets, etc. We expect that the information about base-multiplets, and their proper characterization,

in conjunction with a comprehensive, self-explanatory nomenclature scheme will effectively facilitate research within the "RNA structure and function" paradigm.

RESULTS AND DISCUSSION

Identification of higher order base-multiplets

To facilitate automated searching of base-multiplets present in large RNA molecules, here RNA structures are preprocessed to generate graphs where nucleotides are considered as nodes and two nodes are connected by an edge, only if a base-pairing interaction exists between them. A wide variety of connectivity patterns may exist in a three-dimensional RNA structure. They include covalent phosphodiester linkage, and different types of noncovalent interactions (base-pairing, stacking, base-phosphate interaction, water, and ion mediated interaction, etc.). However, in contrast with RNA tertiary structure cyclic motifs (Lemieux and Major 2006), mentioned in the Introduction section, here we only consider sets of bases connected through base-pairing interactions as multiplets. Imposing this condition allows us to comprehensively list out a neatly defined set of RNA base-multiplets for further detailed investigation. Base-pairing interactions are identified by the BPFind program (Das et al. 2006). All the connected components (subgraphs, in which every pair of nodes are connected by a path) are identified by traversing the large RNA graph using Depth First Search (DFS) algorithm. These connected subgraph components (having "n" number of nodes) represent the higher (nth) order multiplets in RNA. Here the order of the multiplet is defined by the number of bases (n) in the set. Thus, base-pairs may be considered as multiplets of order two, and triplets, quartets, etc. are considered as higher order multiplets of corresponding orders. Therefore, in the context of RNA, the arrangement of constituent nucleobases in terms of base-pairing connectivity is captured by the topology of the connected subgraphs. In other words, topologies of multiplets correspond to the distinct nonisomorphic forms of the representative connected subgraphs. Essentially, it is the degree of the respective constituent nodes that differentiates one topology from the other. The possible topological variety for triplets, quartets and pentets are shown in Figure 2. The color codes of different nodes of each connected subgraph (base-multiplet) depict the degree of different nodes. In principle, topology of a base-multiplet can be defined in terms of the degrees of the constituent nodes of the connected subgraph. However, in some cases unambiguous topological classification is not possible simply by calculating the number of nodes with different degrees. For example, both the two pentet topologies, P3 and P7, consist of one node with degree = 3 (color green), one node with degree = 1 (color blue), and three nodes with degree = 2 (color red), respectively. However, P3 topology contains a three membered ring in it and P7 topology contains a four membered cyclic ring. In such situations, as discussed in this manuscript, a detailed study of the base-pairing interactions between constituent bases is the most reliable way to assign topology to a particular higher order base-multiplet.

As shown in Figure 2, only two topologies are possible for triplets. In quartets the number of possible topologies is six and in pentets it goes to 19. Naturally, the possible topological varieties of multiplets will increase exponentially, with the increase in the number of interacting nucleotides. Fortunately, several of these possibilities get ruled out due to constraints such as the requirement of multiple hydrogen bonds necessary for a good base-pair, or electrostatic repulsion between amino groups or hydrogen bond acceptors, etc. Each nucleobase has three interacting edges, respectively, with distinct characteristic hydrogen bond donor/acceptor patterns. Thus, assuming one partner nucleobase per edge, a nucleobase can ideally pair with at the most three other nucleobases. However, in each of the P9 through P19 pentet topologies shown in Figure 2, at least one nucleotide has degree = 4 (pink colored nodes). Such topologies are possible only if the nodes having degree = 4 form bifurcated pairing with two other bases involving a single interaction edge. Bifurcated geometries are often observed in multiplets and are detected and annotated by the BPFind program. However, steric constraints are likely to prevent any nucleotide to form proper base-pairing interaction with more than three other nucleotides simultaneously. Hence topologies, where one or more nodes have degree >3 are rarely observed in real examples.

In this study we have analyzed two nonredundant data sets of RNA structures maintained at the HD-RNAS database (Ray et al. 2012) and NDB BGSU server (Leontis and Zirbel 2012). To get a broader view, a large data set of RNA crystal structures (having resolution better than 3.5 Å) and available NMR structures obtained from the RCSB-PDB database (Westbrook et al. 2003), were also examined. All observed instances of RNA base-multiplets present in these data sets were identified using the BPFind program, in conjunction with our in-house developed computer programs (written in Python language) and have been curated for further analysis. Table 1 provides an overall view regarding occurrence frequencies of base-multiplets (quartets and higher) observed in all the four data sets studied. Note that, the highest multiplet order observed in existing RNA structures is as high as nine. The subsequent sections elaborately discuss the topologies and occurrence contexts of observed RNA multiplets.

Topology is important for the nomenclature of RNA base-multiplets

A characterization scheme for base-triplets has already been described in the Introduction section (Abu

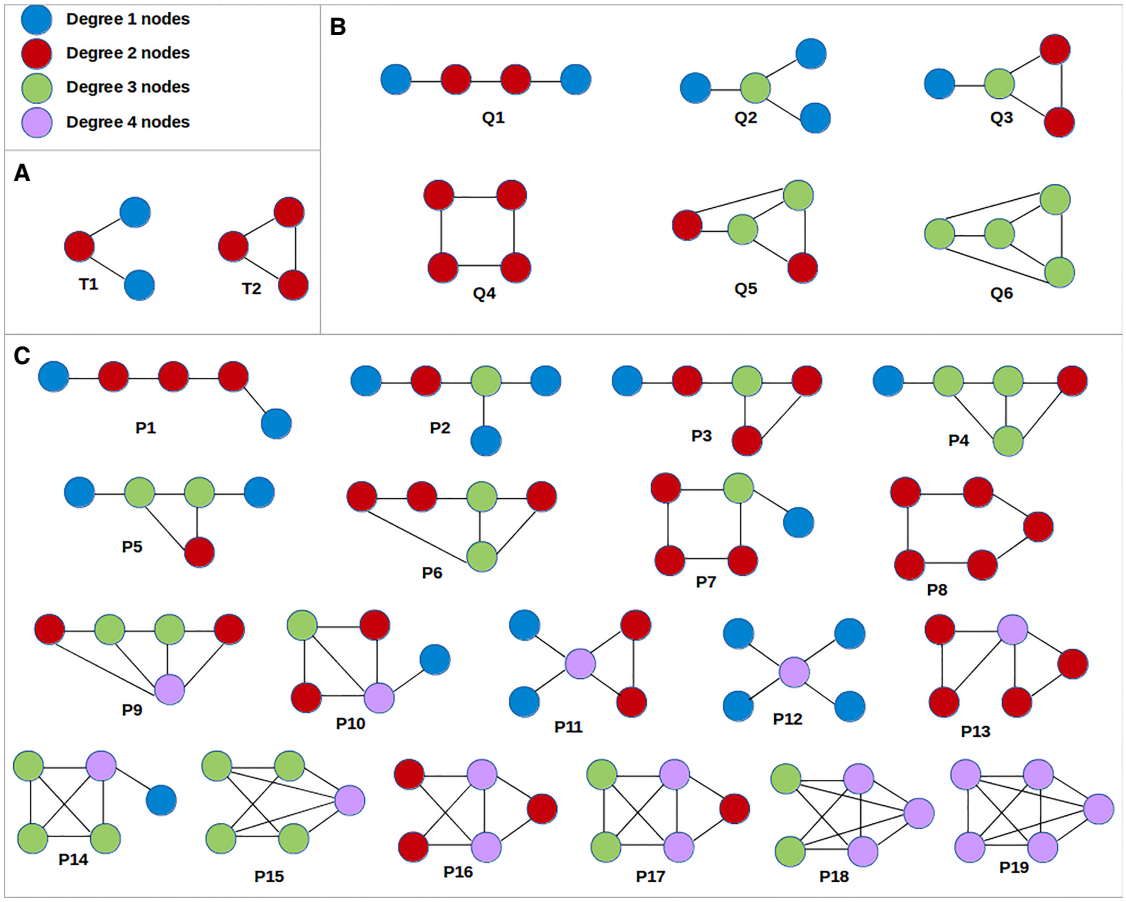


FIGURE 2. Possible topological varieties in (*A*) Triplets, (*B*) Quartets, and (*C*) Pentets. Among these, for quartets, only Q1, Q2, Q3, and Q4 are observed in nature; and for Pentets, P1, P2, P3, P8 (rare), and P12 (rare) are found in nature. Colored circles represent nodes with different degrees. An accurate nomenclature of higher order multiplets requires an unambiguous assignment of numbers to these nodes. This can be achieved by implementing the priority rules discussed in this work and is illustrated in Supplemental Figures S1 and S2.

Almakarem et al. 2012). However, though this scheme makes it possible to annotate the mutual arrangement of three bases for most triplets, it is not scalable for annotation of quartets or higher multiplets. This is mainly due to (i) exponential increase in the number of possible base combinations and (ii) absence of inbuilt information about topology (mutual arrangement of component bases). For example, for a triplet consisting of three bases B1, B2, and B3, base-pairing between B1 and B2 can involve $3 \times 3 = 9$ edge permutation (as each base has three edges), and that between B2 (with two free edges) and B3 can involve $2 \times 3 = 6$ permutations. Given that for each edge

TABLE 1. Occurrences of base-multiplets in the four data sets studied

Number of constituent nucleobases	Frequency in different data sets			
	HDRNAS nonredundant data set	BGSU NDB nonredundant data set	RCSB-PDB X-ray data set (<3.5 Å)	RCSB-PDB NMR data set
9 (9-mer)	0	0	11	0
8 (octet)	14	7	364	0
7 (septet)	26	6	397	4
6 (sextet)	100	65	2066	7
5 (pentet)	151	97	3981	9
4 (quartet)	678	536	19514	42
3 (triplet)	2331	2020	58616	136

permutation there are two possible orientations, cis and trans, the total number of possible geometric families in triplets comes to $9 \times 2 \times 6 \times 2 = 216$. If the same scheme is extended to quartets having a linear arrangement, the corresponding number for geometric families will increase $6 \times 2 = 12$ fold, taking into consideration the additional pairing of B3 with B4. In this way the total number of quartet geometries can be as large as 1592. In addition, the mutual arrangement of the four participating bases need not be linear and can adopt other possible topologies. For example, Figure 3A and B illustrate two different topologies of CGCG quartets. One of the main objectives of this work was to circumvent this complication related to family wise classification, and to develop an intuitive and self-explanatory characterization scheme for different topologies which can be scaled up to accommodate higher order multiplets beyond base-triplets. In order to explain the proposed characterization scheme for multiplets, we elaborate on a nomenclature convention of RNA base-quartets in the following section.

Topology-based nomenclature scheme for RNA base-quartets

RNA base-quartets display all types of topology related features that are present in higher multiplets. Keeping this in mind, here we present a scalable, comprehensive, unambiguous, and self-explanatory topology-based nomenclature convention, and demonstrate how it facilitates the formulation of topology specific rules for RNA base-quartets. In order to assign an unambiguous name to an RNA base-multiplet, the following four basic requirements should be addressed by the nomenclature convention:

i. Include the information about the topology.
ii. Follow an unambiguous numbering scheme for the nodes present in each topology.
iii. Clearly distinguish between nodes having different degrees.
iv. Follow an intuitive way of writing the interaction geometries between two interacting bases.

The basic requirements are addressed in a generic fashion which is applicable for all individual quartet topologies. For unambiguous numbering of the bases, a set of priority rules are applied iteratively, considering each base in a structural element as a node in a subgraph. The numbering rules are discussed below:

i. First priority is given to the highest order node which is labeled at N1 (exception: linear topology, where the name always starts from one of the terminal residues).

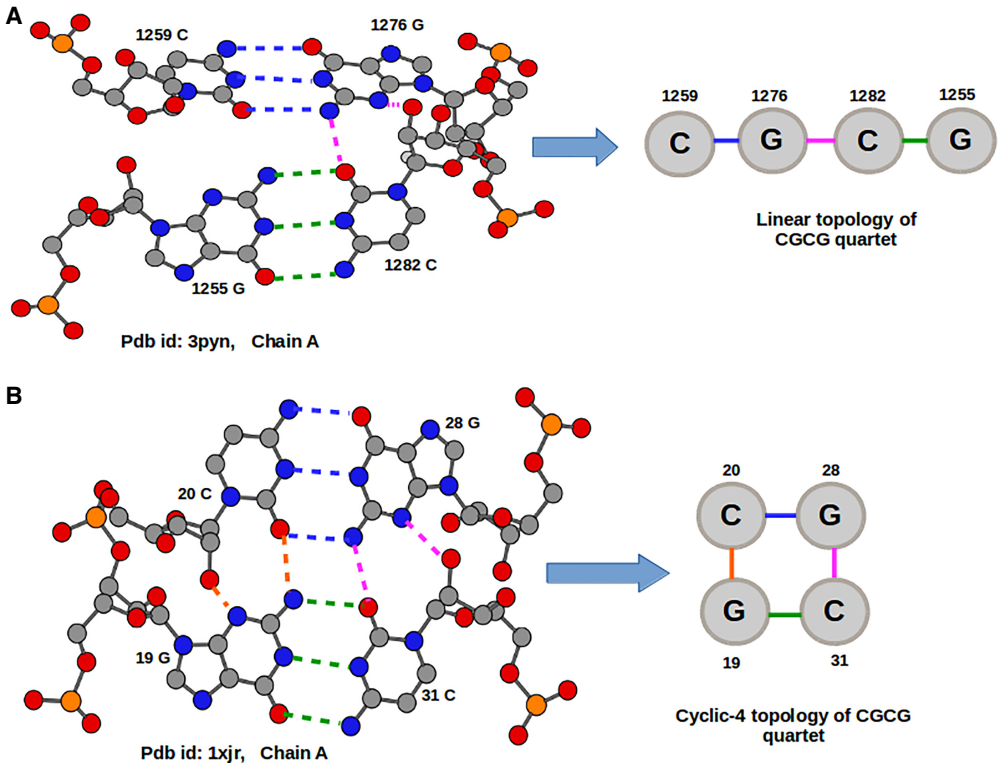


FIGURE 3. Geometry of CGCG quartets in (*A*) linear topology having base-pairs 1259C:1276G W:WC (blue), 1276G:1282C S:S C (magenta), and 1282C:1255G W:W C (green); and in (*B*) cyclic-4 topology having base-pairs 20C:28G W:W C (blue), 28G:31C S:SC (magenta), 31C:19G W:W C (green), and 19G:20 C S:S C (orange). Broken lines represent interbase hydrogen bonding interactions.

ii. For linear quartets, and in case of ambiguity in assignment with rule (i), priority is given to alphabetical order of the name of the bases (A>C>G>U). In case both of the above rules do not resolve the ambiguity in assignment, alphabetical order of the corresponding next neighboring residues (A>C>G>U), and then alphabetical order of base-pairing interaction (H:H C > H:H T > H:S C > H:S T > H:W C > H:W T > S:H C > S:H T > S:S C > S:S T > S:W C, S:W T > W:H C > W:HT > W:S C > W:S T > W:W C > W:W T) between starting residue and its neighboring residues are considered.

iii. If all of the above rules fail to resolve the ambiguity, lower nucleotide number in the RNA chain is given the preference.

The numbering scheme is explained in detail in Supplemental Figures S1 and S2. Note that the unambiguous numbering of the nodes depends on base identities and interaction geometries. As is explained in Supplemental Figure S2, ambiguities arise in the case of complex topologies for higher multiplet orders, if numbers are assigned to the nodes just based on degrees of the nodes and connectivities. For example, in P4 topology there are three nodes having degree = 3, which may be numbered 1–3, depending on base identity and interaction geometry. It is also important to note that priority is given to the alphabetical order of the base identities over the order of base-pair geometric families, since the structures of multiplets do not depend solely on the interaction family (base-pair family of constituent base-pairs). The main reason is that the isostericity pattern may vary in a specific interaction geometric family depending on base composition. Moreover, the isostericity of base-pairs loses a lot of its significance when they interact with some other molecules (detailed explanation is given in Section S1 of the Supplemental Material). Therefore, for unambiguous numbering, priority is given to the base identities. It may be mentioned here that this alphabetical order–based approach also addresses the issue of count duplication during auto detection of multiplets. Figure 2 shows all the six different quartet topologies possible, though Q5 and Q6 topologies have not been observed in existing RNA structures. The application of this numbering and nomenclature scheme is explained below, taking the specific examples of linear (Q1) and star (Q2) quartet topologies.

Q1: Linear quartet

Description

1. Graph of four nodes (N1, N2, N3, and N4).
2. Two internal nodes (N2 and N3) have degree = 2. Two terminal nodes (N1 and N4) have degree = 1.

Numbering convention

Numbering of the bases will start from one of the two terminals, selected on the basis of the following priority rules, and continued serially according to their connectivity.

1. Numbering starts from the alphabetically senior terminal base.
2. If the two terminal bases are identical, identities of the corresponding preterminal (adjacent) bases have to be considered, and numbering will start from the terminal base corresponding to the preferred preterminal base.
3. If both the preterminal bases are also identical, then the interacting edges of terminal bases and corresponding internal bases will be considered (N1:N2 and N4:N3 pairs) and priority will be given to the base-pairs as per alphabetical order mentioned above under generic priority rules (ii).
4. If both the terminal bases, and both the base-pair geometries, N1:N2 and N4:N3, are identical; then the name will start with the terminal base having a smaller nucleotide number in the RNA chain.

Nomenclature convention

1. In linear topology, connected nucleotides must be mentioned serially and therefore, nodes with different degrees need not be segregated here. They will be listed serially separated by "-" delimiters.
2. Interaction geometries of connected nucleotides will be written serially, separated by "-" delimiters.

Representation scheme

Name: Q1 or Linear N1-N2-N3-N4; Geometry: Bp12-Bp23-Bp34

Q2: Star quartet

Description

1. Graph of four nodes (N1, N2, N3, and N4).
2. One central node (N1) has degree = 3.
3. Three terminal nodes (N2, N3, and N4), each having degree = 1.

Numbering convention

1. Numbering of the bases will start from the central base (node which has the highest degree = 3).
2. Then three terminal bases (nodes having degree = 1) are numbered according to alphabetical order.

3. If two or more terminal bases are identical, then the base-pairing geometry between the central base and corresponding terminal bases (N1:N2, N1:N3 and N1:N4) will be considered. Here also priority is given to the base-pair geometry that comes first in alphabetical order (the priority order is the same as mentioned in generic priority rule [ii]).
4. If two or more terminal bases are identical and their corresponding interaction geometries with the central nucleotide are also the same, then the base having a lower nucleotide number in an RNA chain will be given priority.

Nomenclature convention

1. Here nodes with different degrees are segregated by using a square bracket. The central base is written outside the square bracket and all three terminal bases (having degree = 1) are written serially within the square bracket according to their assigned numbers.
2. Interaction geometries between the central base and the three terminal bases are written respectively separated by "/" delimiters.

Representation scheme

Name: Q2 or Star N1[N2 N3 N4]; Geometry: Bp12/Bp13/Bp14

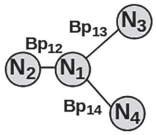
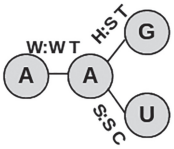
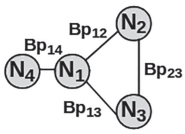
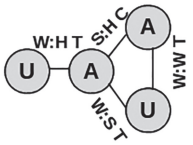
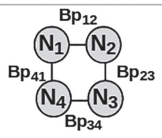
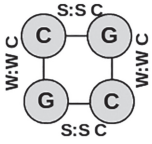
Figure 4A shows the schematic representation of all the four quartet topologies which are observed in existing RNA structures. In this figure the applications of topology specific nomenclature rules are illustrated with specific examples of observed quartets from each of the respective categories. The nomenclature convention of cyclic3 (Q3) and cyclic-4 (Q4) topologies are intuitive extensions of the convention as explained above for linear and star quartet topologies. These are also explained in Section S2 of the Supplemental Material. Extension of the quartet nomenclature rules to Q5 and Q6 topologies are described in Figure 4B, to demonstrate the applicability of the proposed topology-based nomenclature scheme for naming RNA base-pentets, sextets and other higher order multiplets. It may be noted that in multiplets having more than 4 nt bases (beyond quartets), several topologies may have branching at one or more points and they often contain both linear components as well as star type or cyclic components (for example P2, P3, P4, etc. in pentets). In order to avoid ambiguity in such cases a set of priority rules have been postulated for nomenclature of higher order base-multiplets. These are discussed in Section S3 of the Supplemental Material. Here, the essential additional points refer to complex topologies which are a combination of linear and cyclic components. The nomenclature of complex topologies which contain only open chains, with one or more branching points, follow the conventional representation scheme for trees in computer science. But, if the topology contains cyclic components, they have to be taken care of separately.

Topology wise distribution of RNA base-multiplets and analysis of their occurrence contexts

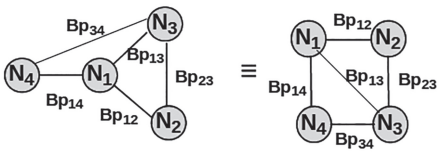
The nomenclature scheme outlined above was used to carry out a comprehensive analysis of the occurrence contexts of base-multiplets belonging to different topological classes. Table 2 shows the occurrence frequency of different topologies among base-quartets and pentets. In order to avoid duplication in count, care was taken such that the frequencies of lower order components of higher order multiplets did not get added with the statistics of the corresponding lower order elements. Accordingly, we have identified connected subgraphs extending up to its maximum order, and all reported occurrence frequencies are calculated considering those structural elements, which are not part of any higher order multiplet.

Table 2 gives an overview of the occurrence frequencies of different topologies, belonging to quartets and pentets, in RNA structures available in different databases. As expected, the linear quartet (Q1) and pentet (P1) topologies are most frequently observed due to the absence of any steric constraints. In contrast, topologies with a greater number of higher degree nodes are less frequent in nature. Thus, frequencies of occurrence of Q2, Q3, and Q4 quartets are significantly lower than that of Q1 quartets, and there are no reported examples of Q5 and Q6 quartets. Q1 quartets (Fig. 5A) are recurrently found in many structural contexts including tRNAs, ribozymes, 16S rRNAs, 23S rRNAs, tRNA-16S rRNA interface, etc. Q2 topology or star quartets (Fig. 5B) are also found in multiple structural contexts in tRNAs and ribosomal RNAs. Though, as is expected, the complex (larger number of higher degree nodes) Q3 and Q4 quartet topologies are significantly less frequently observed than the Q2 topology. It is interesting to note that, they are usually found in highly conserved structural contexts within specific functional RNA molecules. For example, A [<AU>U] Q3 quartets are found in both 16S and 23S rRNA of *T. thermophilus*. In 16S rRNA, the 547A [<397A 37U>498U] quartet shows cSH/tWS/tHW, tWW geometry, whereas in 23S rRNA 2077A [<2434A 2243U>2075U] shows cSS/cWW/cHW, tWS geometry. However, in both cases they are observed in conserved multihelix junction loops (five-way junction between helix 3, 4, 16, 17, and 18 in the 5′ domain of 16S rRNA and a three-way junction between helix 74, 75, and 82 in domain-V of 23S rRNA, respectively) and possibly help in providing a characteristic 3D structure of the respective regions. Similarly, instances of cyclic-4 topology (Q4) are rare in both the nonredundant data sets. Interestingly however, distinct geometric variants of this topology are present in conserved contexts. In the larger X-ray data

A

Topology type	Quartet nomenclature		Real example of each topology	
	Name	Geometry	Full name	Schematic representation
 Q1: Linear	N_1-N_2-N_3-N_4	Bp_{12}-Bp_{23}-Bp_{34}	Q1 CGAU cWW-cSS-tHW	
 Q2: Star	$N_1[N_2\ N_3\ N_4]$	$Bp_{12}/Bp_{13}/Bp_{14}$	Q2 A(AGU) tWW/tHS/cSS	
 Q3: Cyclic-3	$N_1[<N_2\ N_3>\ N_4]$	$Bp_{12}/Bp_{13}/Bp_{14}$, Bp_{23}	Q3 A[<AU>U] cSH/tWS/tHW, tWW	
 Q4: Cyclic-4	N_1-N_2-N_3-N_4->	Bp_{12}-Bp_{23}-Bp_{34}-Bp_{41}->	Q4 CGCG cSS-cWW-cSS-cWW->	

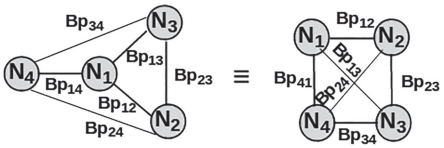
B



Q5 quartet topology

Nomenclature

N_1-N_2-N_3-N_4->, N_1-N_3
Bp_{12}-Bp_{23}-Bp_{34}-Bp_{41}, Bp_{13}



Q6 quartet topology

Nomenclature

N_1-N_2-N_3-N_4->, N_1-N_3, N_2-N_4
Bp_{12}-Bp_{23}-Bp_{34}-Bp_{41}, Bp_{13} , Bp_{24}

FIGURE 4. (*A*) Schematic representation and nomenclature rules for four different quartet topologies. An example from each topology has also been shown. (*B*) Nomenclature of Q5 and Q6 quartet topologies.

set, a frequently observed cyclic-4 quartet is 1280A-1149C-1124G-1125U-> **c**WS-cWW-cSH-cWH->, which is a conserved interaction found in 16S rRNA of *E. coli* and *T. thermoplilus* which mediates interaction between two internal loop regions present in helix 39 and helix 41, respectively, in the 3′M domain. Another cyclic-4 quartet G-G-G-G-> cHW-cHW-cHW-cHW-> is frequently observed as a part of G-quadruplex structures, in many synthetic RNA constructs. A thorough study (S Bhattacharya, A Jhunjhunwala, A Halder, et al., in prep.) on all the quartet topologies has been carried out and comprehensive information on all the quartet varieties belonging to different topologies, and their occurrence contexts are available in a database/web-server hosted at http://quarna.iiit.ac.in/.

Table 2 also shows that only five (P1, P2, P3, P8, and P12) out of 19 possible pentet topologies (as shown in Fig. 3), show any significant occurrence in the four data sets. As expected, the frequency of P1 or linear topology is significantly higher than that for all other topologies. P2 and P3 topologies are comparatively less frequent in available RNA structures and P8 and P12 topologies are rare in nature and are found only in the large X-ray structure data set. Figure 6A illustrates the structure of a P1 pentet present in *H. marismortui* 23S rRNA. This A-C-G-A-U cSS-cWW-cSS-tHW pentet mediates inter-domain interaction between domain-V and domain-VI of the 23SrRNA. There are other instances of P1 pentets present in *H. marismortui* 23S rRNA, which mediate inter-domain interactions

TABLE 2. Occurrences of distinct topologies of quartets and pentets in different data sets

		Frequency in different data sets			
	Topology	HDRNAS nonredundant data set	BGSU-NDB nonredundant data set	RCSB-PDB X-ray data set (<3.5 Å)	RCSB-PDB NMR data set
Quartets	Q1: N1-N2-N3-N4	653	486	18670	21
	Q2: N1[N2 N3 N4]	22	27	736	1
	Q3: N1[<N2 N3> N4]	2	2	27	0
	Q4: N1-N2-N3-N4->	1	21	81	20
Pentets	P1: N1-N2-N3-N4-N5	128	87	3489	5
	P2: N1[N2[N3] N4 N5]	18	7	362	2
	P3: N1[<N4 N5> N2[N3]]	5	2	122	2
	P8: N1[N2 N3 N4 N5]	0	0	5	0
	P12: N1-N2-N3-N4-N5->	0	0	3	0

Connectivities between nodes are described in Figure 4 (for quartet topologies) and Supplemental Figure S2 (for pentet topologies).

between domain-I and domain-II and between domain-II and domain-V, respectively. The P1 pentets are also observed in rRNAs from both large and small ribosomal subunits of *T. thermophilus*. However, their probable functional role may differ depending on their contexts of occurrences. For example, 397A-37U-547A-498U-404U tWW-tSW-tHW-cHW P1 pentet is present in a five-way junction (junction between helix 3, 4, 16, 17, and 18) of the 5′ domain of *T. thermophilus* 16S rRNA. Another conserved P1 pentet in *T. thermophilus* 16S rRNA is 769G-810C-899C-770C-809G cWW-cSH-cWS-cWW. This particular pentet is observed in the central domain, probably mediating a tertiary contact between a stem (helix 24) and a sequentially distantly placed hairpin loop (terminal hairpin loop of helix 27). P1 pentets are also found in multiple locations in the 23S rRNA of the same species. They are mostly found in large multihelix junctions. For example, 219G-234C-430G-262A-235U cSS-tWW-cSS-cWW pentet is

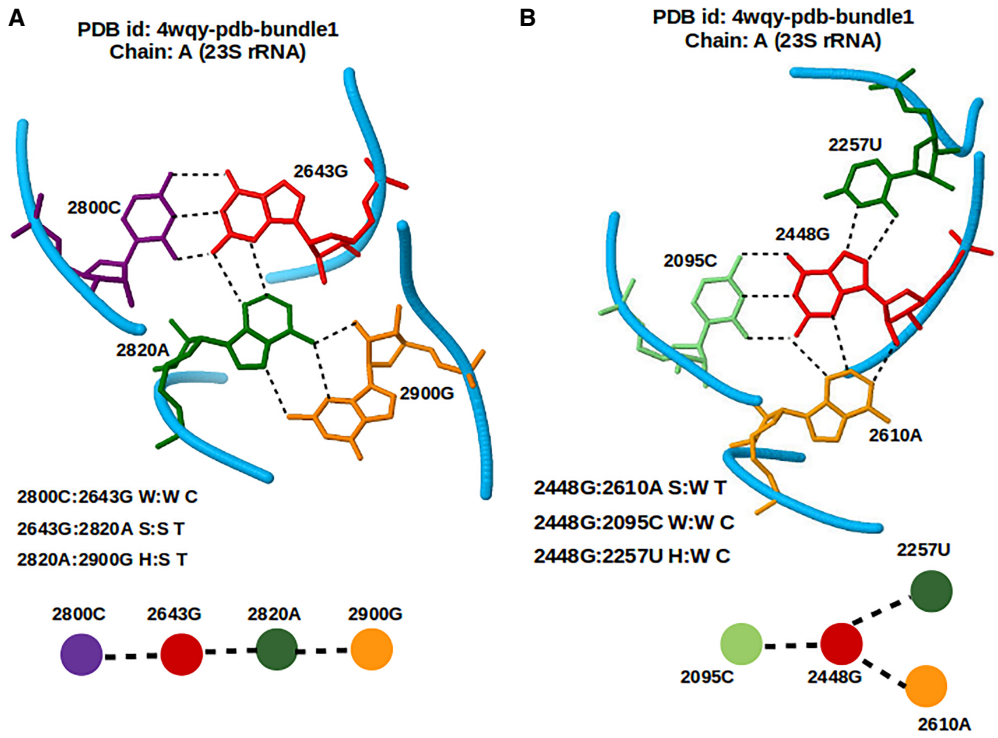


FIGURE 5. (*A*) Example of a quartet having Q1 or linear topology. The 2800C-G2643-2820A-2900G cWW-tSS-tHS quartet observed in *T. thermophilus* 23S rRNA. (*B*) Example of a quartet having Q2 or star topology. 2448G[2610A 2095C 2257U] tSW/cWW/cHW quartet present in *T. thermophilus* 23S rRNA.

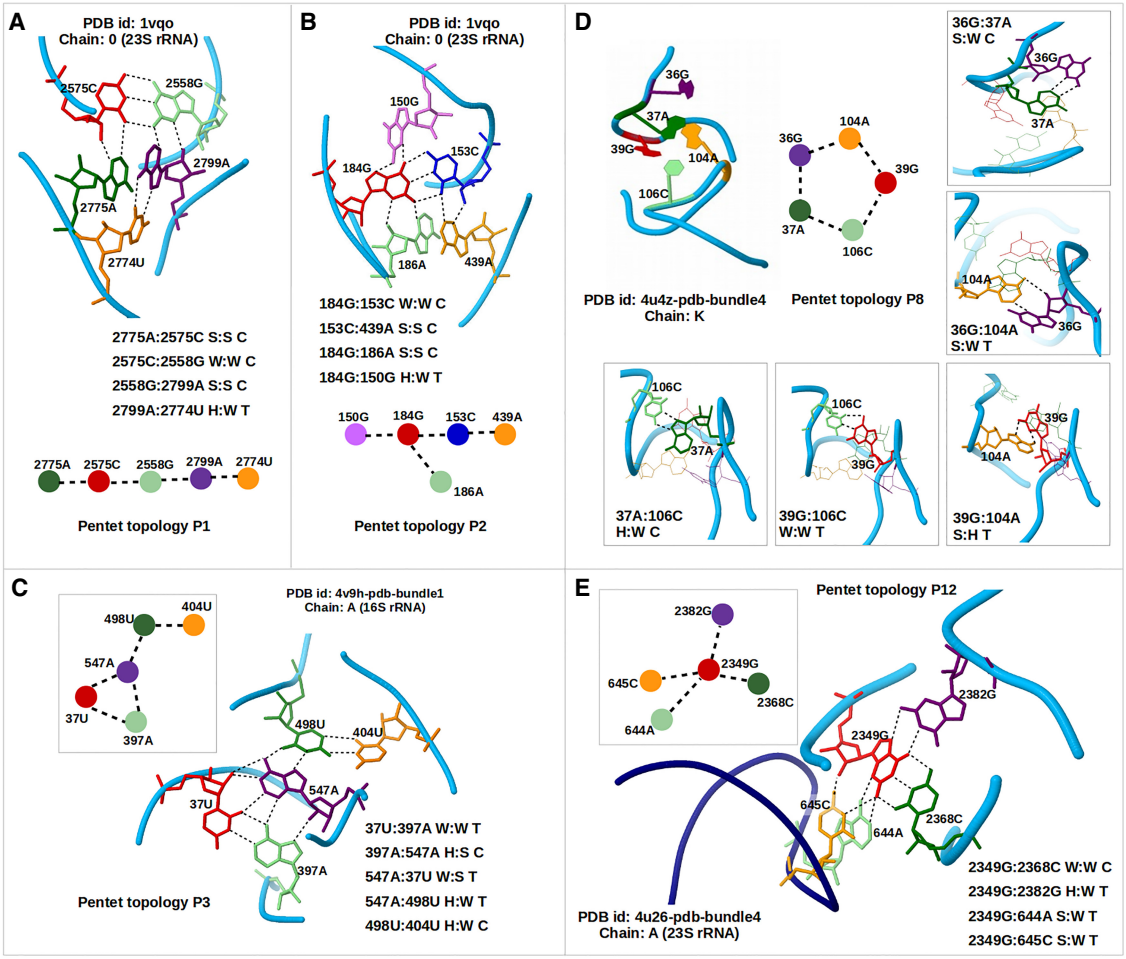


FIGURE 6. (*A*) Example of a pentet having P1 or linear topology. 2775A-2575C-2558G-2799A-2774U cSS-cWW-cSS-tHW pentet observed in *H. marismortui* 23S rRNA. (*B*) Example of P2 pentet topology. 184G[153C[439A] 186A 150G] [cWW-cSS]/cSS/tHW pentet present in *H. marismortui* 23S rRNA. (*C*) The conserved structure of the P3 pentet topology, 547A[<397A 37U> 498U[404U]] cSH/tWS/[tHW-cHW], tWW pentet observed in *T thermophilus* 16S rRNA. (*D*) P8 or cyclic pentet 37A-106C-39G-104A-36G-> cHW-tWW-tSH-tWS-cSW-> is observed in 5.8S rRNA of *S. cerevisiae*. Constituent base-pairs are separately shown in boxes. (*E*) Example of P12 or star pentet- 2349G[644A 645C 2368C 2382G] tSW/tSW/cWW/tHW observed in *E. coli* 23S rRNA, which mediates interactions between nucleotides of domain-II (backbone shown in dark blue color) and domain-VII (backbone shown in sky blue color).

located at the large nine-way junction in domain-I (junction where helices 4, 5, 8, 9, 10, 11, 12, 13, and 14 meet), which probably helps in maintaining the structural fold of that junction region. Again, another P1 pentet 1542A-1528A-1464C-1447G-1544A tHH-cSS-cWW-cSW is present near a four-way junction of domain-III (junction, where helix 56, 57, 58, and 59 meet). Pentets having P1 topology are also observed at many conserved locations of eukaryotic 25S rRNA of *S. cerevisiae*.

The second most frequent pentet topology is P2 (one node having degree = 3, one node having degree = 2, and three nodes having degree = 1). P2 pentets are also found at various structural contexts of both the 16S rRNA and 23SrRNA of *T. thermophilus*. Thus in 16S rRNA, one variant of P2 pentet, annotated as "68G [152A[19C] 101A 64G], [tSS-tHW]/cWW/cHW" (where 68G is the central base which is pairing with 152A, 101A, and 64G and where 152A is also pairing with 19C), mediates long distance interaction and brings together two distantly placed internal loop regions (internal loops present in helix 6 and helix 8) in the 5′-domain. On the other hand, in 23S rRNA, a conserved P2 pentet 2335A[2332U[2323G] 2322A 2296U], [cSS-cWW]/tWW/tHW is observed in a three-way junction in domain-V (junction between helices 83, 84, and 85), where it probably stabilizes the folded structure. Apart from *T. thermophilus*, conserved P2 pentets are also found in *H. marismortui* 23S rRNA. Figure 6 B demonstrates one of the P2 pentets present in a large multihelix junction in domain-I of *H. marismortui* 23S rRNAs. The nomenclature conventions of P2 pentets are explained in Section S3 of the Supplemental Material and are illustrated in Supplemental Figure S3.

In contrast with recurrent occurrences of P1 and P2 topologies in multiple structural contexts and with various base compositions and geometries, P3 and two other rarely observed topologies P8 and P12 are conserved in specific structural contexts. For example, a P3 pentet is observed in a conserved multihelix junction of the 5′-domain of the *T. thermophilus* 16S rRNA (Fig. 6C) and a P8 or cyclic pentet (all five nodes having degree = 2, cyclic) is observed in yeast 5.8S rRNA (the conserved nucleotide composition, interaction patterns and the occurrence context of P8 pentets are described in Fig. 6D). Another rare variant of pentet topology is the P12 or star topology (one node having degree = 4, four nodes having degree = 1), where one single central nucleotide forms base-pairing interaction with four other nucleotides at the same time, thereby making the structure quite crowded. P12 pentet is observed in a conserved structural context of *E. coli* 23S rRNA, where it mediates a long range interaction between domain-II and domain-VII (Fig. 6E).

Such conserved occurrences of different base-multiplets at specific structural contexts of functional RNAs, indicate a strong correlation between the geometry and stability of these multiplets and the function of the RNA molecule. However, mapping out the exact relationship between formation of a specific multiplet and its functional consequences is a formidable task and is beyond the scope of this work. Instead, our analysis provides a general idea about the contribution of different multiplets in the structure and function of RNA. For example, we have studied the occurrence contexts of all the pentets that occur in *T. thermophilus* (prokaryote) and *S. cerevisiae* (eukaryote) rRNAs in the HDRNAS data set. As shown in Supplemental Table S1, the occurrence context of all pentets can be classified broadly in two types of structural contexts—(i) multiway (or multihelix) junctions and (ii) mediation of long-range interaction between distantly placed secondary structural elements including loop–loop and loop–helix interaction. As we know, shape and size of a multiway junction determines the spatial arrangement of stem regions with respect to each other (Laing et al. 2009, 2012; Chen et al. 2018). At the same time, structure of a long-range interaction determines the spatial organization of secondary structural elements. Therefore, it may be inferred that formation of a multiplet is crucial for bringing together different secondary structural elements in folded RNA, where their spatial arrangement is primarily determined by the unique shape and geometry of such multiplets.

The possible topological variants of sextet and septet are even higher than that of pentets and naturally most of the topologies are ruled out due to steric complications. However, these multiplets have typical context dependent functional roles and are found in some conserved positions at 16S and 23S rRNA across different species. Supplemental Figures S4 and S5 illustrate different conserved occurrences of sextet and septet elements, respectively, in rRNAs of different species. Interestingly, apart from in rRNAs, linear sextet elements are found in important structural contexts of crystal structures of different riboswitches, which include the class-I preQ1 riboswitch (Klein et al. 2009), SAM-II riboswitch bound to *S*-adenylmethionine (Gilbert et al. 2008), and divalent metal sensing M-box riboswitch, involved in Mg^{2+} homeostasis (Dann et al. 2007). Although the base composition and interaction geometries are not identical, in both preQ1 and SAM-II riboswitches, the sextet elements stabilize the pseudoknot formation between the 5′ terminal helix (S1 in preQ1 and P1 in SAM-II) and loop 3 (L3 in both the cases). The base composition, geometry, and structural context of the above mentioned two sextet elements are described in Figure 7 (A1 and A2). In the M-box riboswitch the linear sextet (Fig. 7A3) 162A-23U-72A-105A-68C-73G cWW-tSS-tWW-cSS-cWW mediate long-range interaction between the P2 helix, P4 helix, L4, and L5 loop of the aptamer domain.

In Table 1 it is observed that 8-mer structural elements or octets are the highest order elements which are found in all the three crystal structure data sets studied. Interestingly, these octets occur in conserved structural folds. For example, Figure 7 (B1 and B2) depicts the conserved structural fold present in 23S rRNA of different species, like *H. marismortui*, *T. thermophilus* and *D. radiodurans*, where octets are observed. Note that, interaction geometry of octets observed in *T. thermophilus* and *H. marismortui* are identical (Fig. 7B3), whereas, despite the structural similarity of the overall fold, the nucleotide composition, interaction geometry and topology are quite different in the case of *D. radiodurans* (Fig. 7B4,B5). This phenomenon again underscores the importance of the consideration of topology in describing the higher order multiplets.

As shown in Table 2, the 9-mer structural elements (nonets) are not observed in any of the nonredundant data sets and found to occur only 11 times in the larger X-ray data set. Interestingly, though limited in number, all these occurrences are in important structural contexts. For example, a nonet element, which is observed in domain-I of *T. thermophilus* 23S rRNA, provides compactness to the structure by mediating tertiary contacts between three distantly placed nonhelical stretches and one small helical region. Figure 7 (C1 and C2), illustrates the occurrence context of this nonet element.

In summary, context analysis of different RNA base-multiplets reveals that most of these are conserved in specific structural contexts, and are possibly catering to specific functional roles.

Conclusions and future work

The abundance of multiplets and their conserved occurrences in different important structural and functional

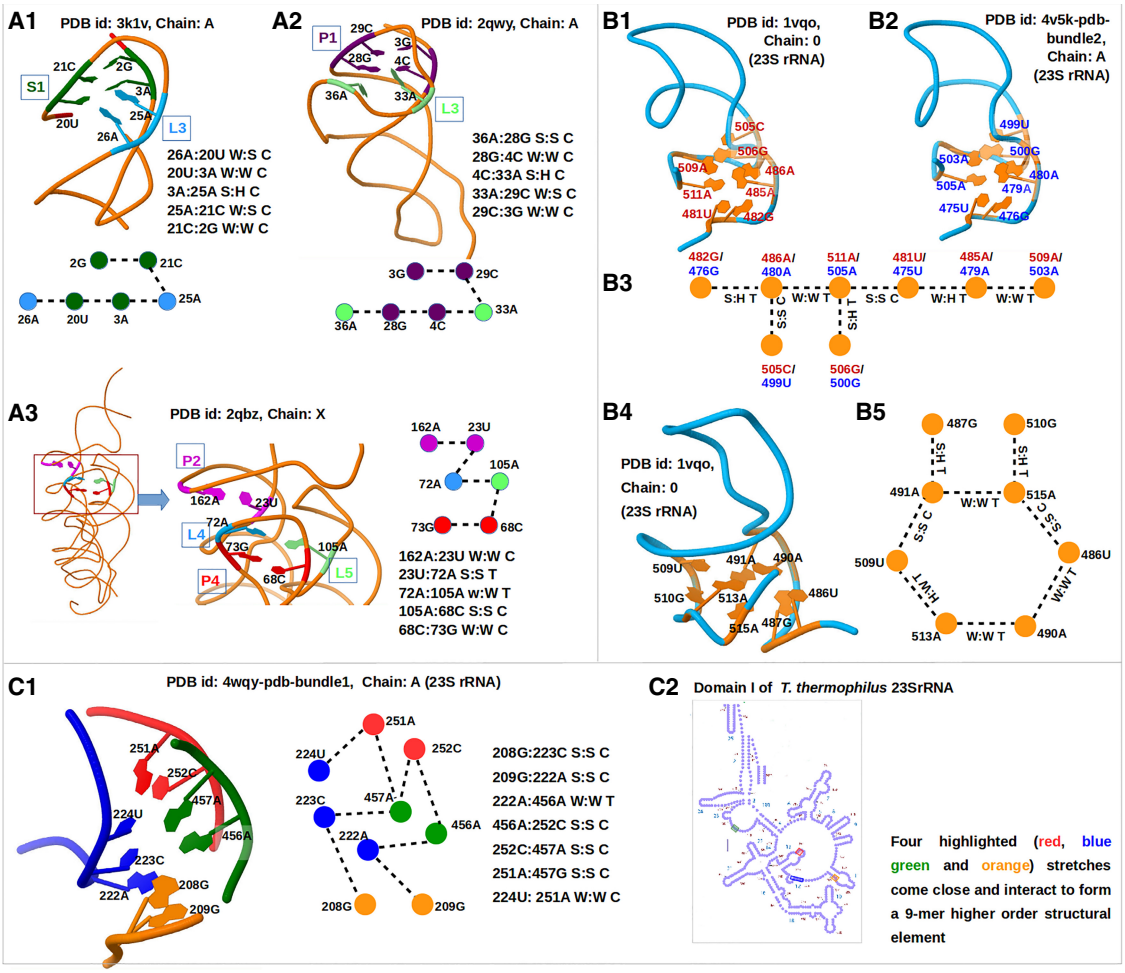


FIGURE 7. Structural context of RNA base-multiplets beyond base-pentets. (*A1*) Sextet element found in class-I preQ1 riboswitch, which mediates pseudoknot formation between S1 helix (green) and L3 loop (sky blue). (*A2*) Sextet element found in SAM-II riboswitch bound to *S*-adenosylmethionine, which mediates pseudoknot formation between P1 helix (purple) and L3 loop (light green). (*A3*) Sextet element found in aptamer domain of M-box riboswitch. This sextet element mediates interaction between P2 helix (pink), P4 helix (red), L4 loop (sky blue), and L5 loop (light green). (*B1*) Structural context of octet element in domain-I of *H. marismortui* 23S rRNA. (*B2*) Structural context of octet element in domain-I of *T. thermophilus* 23S rRNA. (*B3*) The conserved topology of octet elements shown in *B1* and *B2*. (*B4*) Structural context of octet element in *D. radiodurans* 23S rRNA. (*B5*) Topology of the octet element shown in *B4*. (*C1*) The molecular view of the structural context of a nonet in *T. thermophilus* 23S rRNA (Domain-I). Four different colors are used to differentiate between four distantly placed RNA stretches and corresponding schematic representation of the interaction pattern of constituent nucleotides. (*C2*) Secondary structure map of domain-I of *T. thermophilus* 23S rRNA, where interacting regions are highlighted in colored boxes.

contexts in functional RNA molecules underlines the importance of their roles in RNA structural biology. Here we have presented a detailed and systematic bioinformatics study of multiplets detected by using a graph-based approach, in available RNA structure. To facilitate this, we have also developed and proposed a topology-based characterization and nomenclature scheme for RNA base-multiplets.

Our investigations reveal that, of the several possible topologies in base-multiplets, only some are actually observed, and that the most frequently observed ones are the linear topologies followed by relatively complex topologies. The occurrence frequencies also decrease noticeably as the multiplet order increases, and the highest order elements observed being nonets. Higher multiplets, and multiplets with complex topologies, though observed with lower frequencies, were nearly always found to have conserved geometries and were found in specific important structural contexts. The role of quartets in maintaining the preformation of binding pockets of riboswitches (Sharma et al. 2009), and in stabilizing the tetraloop-receptor interactions (Wu et al. 2012) has already been reported in literature. It has been observed in this study that in ribosome structure most of the multiplets were involved either in connecting the junction loops in multihelix junctions or in bringing together sequentially distant secondary structural domains. Therefore, as a whole it can be said that multiplets play important roles in compacting

the RNA 3D structures by organizing helical stems and by providing a typical structural framework to flexible loop regions.

An additional outcome of our study is the identification of multiple instances of bifurcated base-triplets as components of higher order multiplets. In a bifurcated-triplet, a central base simultaneously pairs with two other bases from the same edge, and both the base-pairs, respectively, involve two hydrogen bonds. Such bifurcated geometries are particularly important for complex topologies which involve a single central base-paired with four other bases, using only its three edges. These findings lead to detailed characterization of such bifurcated triplets, which will be addressed in our future work.

Functional RNA molecules have complex 3D structures, and any attempt at performing a systematic analysis brings out several complicating features. While developing our systematic approach, we have avoided some of these complications by limiting our scope of study. For example, the multiplets studied here were restricted to only those where all base-pair units are proper base-pairs with at least two hydrogen bonds and have significantly good geometries. To be specific, unlike as in some earlier reports where intermediate triplets were reported (Abu Almakarem et al. 2012) or as in a coplanar search of DSSR software, we have not included multiplets involving base-pairs mediated by single hydrogen bonds.

The abundance of the multiplets and their conserved occurrences in different important structural contexts of functional RNA molecules underlines the necessity for a proper characterization scheme. Topology-based characterization of RNA base-multiplets proposed here can supplement this requirement and can hence enrich RNA structural bioinformatics research.

Another aspect, which we have not addressed, is the question of energetics. Our investigations reveal that base-multiplets often involve weak C-H…O/N mediated base-pairs, flexible sugar mediated base-pairs, and base-pairs with bifurcated geometries. To understand the role of multiplets, therefore, it is necessary to investigate the interaction energies and stabilities of such base-pairs and their contribution toward structural stability of higher order multiplets. An interaction energy-based characterization of RNA multiplets is likely to provide important molecular level insights into the world of RNA structure and function. We plan to include these and other open issues, in our future studies.

MATERIALS AND METHODS

Data sets used for study

To explore the varieties of different multiplets we have considered two nonredundant (nr) data sets of available RNA structures, obtained from HDRNAS and Nucleic Acid Database (NDB) databases respectively. For more detailed survey, we also have listed all the quartets observed in a larger data set obtained from the RCSB-PDB database considering both X-ray and NMR structures. The details of the data sets are as follows:

Nonredundant data set from HDRNAS

This data set (Ray et al. 2012) contains 167 pdb files, having nucleotide chain length >30 nt and resolution 3.5 Å or better. HDRNAS (http://www.saha.ac.in/biop/www/HD-RNAS.html) classifies all the RNA structures available in the RCSB-PDB database having resolution cutoff 3.5 Å or better and at least one RNA chain longer than 9 nt according to their functional roles, e.g., transfer RNA (tRNA), messenger RNA (mRNA), ribosomal RNA (rRNA), ribozymes, riboswitches, ribonucleases, SRP-RNAs, etc. The tRNAs and ribosomal RNAs are further classified according to the amino acid it may carry and sedimentation coefficient, respectively. In order to exclude small synthetic RNA constructs, we have used a chain length cutoff of 30 nt or larger for the classification procedure. The structures are then classified according to the source organism from which the RNA molecules were isolated and crystallized. The nonredundant data set consists of the best representative RNA structures from each of these subclasses, determined by the best resolution and *R*-factor (free *R*-value). Names of the 167 pdb files are mentioned in Section S5 of the Supplemental Material.

Nonredundant data set from NDB

The nonredundant list of RNA structures created by the BGSU server (release id:1.89) contains 839 RNA structures (Leontis and Zirbel 2012), which have a resolution of 3.5 Å or better. In this list no restriction on chain length has been implemented, therefore it contains both synthetic and functional RNA molecules. The NDB nonredundant list was created by removing uninteresting redundancy by clustering all the RNA structures based on sequence comparison, structural superposition, and geometric analysis and by selecting one representative from each cluster. The names of 839 pdb files present in the NDB nonredundant list are mentioned in the Supplemental Material (Section S5).

Larger crystal structure data set from RCSB-PDB

As we have observed, both the nr data sets have some unique examples of multiplets, although their frequencies are low, therefore, to understand the actual varieties of multiplets we have taken a larger data set of 1873 RNA crystal structures available in the RCSB-PDB database (Westbrook et al. 2003) until 21 September, 2015 (listed in Section S5 of the Supplemental Material). Here a resolution cutoff of 3.5 Å or better has been considered to maintain the quality of the data set.

NMR structure data set from RCSB-PDB

This data set contains all RNA containing NMR structures (591 pdb files) available at the RCSB-PDB database until 21 September, 2015 (listed in Section S5 of the Supplemental Material) (Westbrook et al. 2003).

Base-pair representation scheme

In the Introduction section, we have mentioned the Leontis and Westhof scheme for characterizing and classifying base-pairs. In principle, each nucleotide base has three distinct edges viz. the Watson–Crick edge (W), the Hoogsteen edge (H) (C-H edge in pyrimidines) and the Sugar edge (S) (Fig. 1A). A given edge of one base can potentially pair up with any one of the three edges of a second base, which have compatible hydrogen bond donors and acceptors, to form a base-pair. Two bases can approach in either cis or trans orientation of the glycosidic bonds around the axis defined by drawing a line parallel to and between the hydrogen bonds joining the edges (Fig. 1B). Based on the interaction edge and cis/trans orientation, with respect to the glycosidic bonds, base-pairs are classified into 12 distinct geometric families (W:W, W:H, W:S, H:H, H:S, S:S—each in both cis and trans geometry). In addition to this, to demonstrate the ability of a base to participate in higher order interactions like triplets and quartets, they defined a way of schematic representation of all base-pair families using right angle triangle notation. These notations are very useful to define spatial arrangement and local strand orientation of base-pair components in complex structural motifs. However, the base-pair families are most conveniently represented in an abbreviated format, consisting of the initials of the interacting edge names and the glycosidic bond orientation. For example, if the Watson–Crick edge of a uracil interacts with the Hoogsteen edge of an adenine in trans fashion, then the abbreviated name of the base-pair will be written as U:A W:H T, where W:H T is the base-pair family name. Such representation convention of base-pairs is used in output files of the BPFind program. In this manuscript, in order to describe base-pair geometries we have used a similar convention, as used in the BPFind output. However, in nomenclature of multiplets, to make the notation even more compact, further abbreviation is adopted where the ":" symbol between two interacting bases and their respective edges has been omitted and glycosidic bond orientation ("c" for cis and "t" for trans) is mentioned before the interaction edge. For example, W:H T geometry is written as tWH in the multiplet nomenclature syntax. A similar abbreviation was first used by Stombaugh et al. (2009).

Base-pair identification using BPFind

A base-multiplet is composed of two or more base-pair units. Our program requires base-pairing information as the input to find connected components from the large RNA graph. Here, we have used the BPFind program (Das et al. 2006) for this purpose. BPFind implements a hypothesis driven algorithm to detect base-pairs that are significantly planar and are stabilized by at least two good hydrogen bonds. It defines donor and acceptor pairs for all possible base-pair geometries as per the LW classification scheme. It also predefines precursor atoms for each donor and acceptor atom. Then the program checks for the following geometric criteria for detecting base-pairs—(i) distance between the hydrogen bonded heavy atoms $\leq$3.8 Å (default) and (ii) pseudo angle between the hydrogen bonding atoms and their precursors $\geq$120° (default). In order to identify only those base-pairs that have significantly good geometries BPFind also calculates a composite goodness parameter called E_value and reports only those pairs which have E_value <1.8 [default] (Das et al. 2006). BPFind identifies bifurcated triplets, where both the constituent base-pairs satisfy the above mentioned criteria and also explicitly identifies and annotates protonated base-pairs. However, it excludes base-pairs involving a single hydrogen bond, which are also detected by some of the other methods. In the present work we have primarily considered only those base-multiplets, where each of the constituent base-pairing interactions meet default detection criteria of the BPFind program. Thus, all the instances of RNA multiplets reported in this study, have reasonably good geometry.

Method for detection of RNA base-multiplets

Higher order multiplets are detected by using an in-house developed python program, which implements depth first search algorithm to identify connected subgraph components from large RNA graphs. As the input, this program uses the list of base-pairs detected by BPFind software. This program also determines the degree of each node present in the connected subgraphs. As already discussed in the Results and Discussion section, connected subgraph components correspond to the base-multiplets and the information about degrees of constituent nodes partially helps to understand the topology. However, the exact connectivities and interaction geometries are only identified by checking base-pairing between each component node explicitly from the BPFind output files. A separate python program is used for such detailed annotation.

Methods for studying occurrence contexts

In order to study the structural context of different higher order multiplets, we have investigated the secondary structure map of ribosomal subunits from different domains of life, available at Ribovision suite (http://apollo.chemistry.gatech.edu/RibosomeGallery) (Bernier et al. 2014). Helix numbers mentioned to describe the occurrence contexts of different base-multiplets are written according to the numbering given in this suite. The information gathered from these 2D maps are validated by examining available crystal structures using JMol molecular visualization software. The probable structural roles are hypothesized based on the context analysis studies of multiple instances of RNA base-multiplets.

SUPPLEMENTAL MATERIAL

Supplemental material is available for this article.

ACKNOWLEDGMENTS

S.B. acknowledges UGC and A.H. acknowledges CSIR for SRF support.

Received August 31, 2018; accepted February 18, 2019.

REFERENCES

Abu Almakarem AS, Petrov AI, Stombaugh J, Zirbel CL, Leontis NB. 2012. Comprehensive survey and geometric classification of

base triples in RNA structures. *Nucleic Acids Res* **40:** 1407–1423. doi:10.1093/nar/gkr810

Agarwala P, Pandey S, Maiti S. 2015. The tale of RNA G-quadruplex. *Org Biomol Chem* **13:** 5570–5585. doi:10.1039/C4OB02681K

Bernier CR, Petrov AS, Waterbury CC, Jett J, Li F, Freil LE, Xiong X, Wang L, Migliozzi BLR, Hershkovits E, et al. 2014. RiboVision suite for visualization and analysis of ribosomes. *Faraday Discuss* **169:** 195–207. doi:10.1039/C3FD00126A

Biffi G, Tannahill D, Balasubramanian S. 2012. An intramolecular G-quadruplex structure is required for binding of telomeric repeat-containing RNA to the telomeric protein TRF2. *J Am Chem Soc* **134:** 11974–11976. doi:10.1021/ja305734x

Cech TR, Zaug AJ, Grabowski PJ. 1981. In vitro splicing of the ribosomal RNA precursor of tetrahymena: involvement of a guanosine nucleotide in the excision of the intervening sequence. *Cell* **27:** 487–496. doi:10.1016/0092-8674(81)90390-1

Chawla M, Sharma P, Halder S, Bhattacharyya D, Mitra A. 2011. Protonation of base pairs in RNA: context analysis and quantum chemical investigations of their geometries and stabilities. *J Phys Chem B* **115:** 1469–1484. doi:10.1021/jp106848h

Chen Y-L, Sutton JL, Pollack L. 2018. How the conformations of an internal junction contribute to fold an RNA domain. *J Phys Chem B* **122:** 11363–11372. doi:10.1021/acs.jpcb.8b07262

Clemson CM, Hutchinson JN, Sara SA, Ensminger AW, Fox AH, Chess A, Lawrence JB. 2009. An architectural role for a nuclear noncoding RNA: NEAT1 RNA is essential for the structure of paraspeckles. *Mol Cell* **33:** 717–726. doi:10.1016/j.molcel.2009.01.026

Dann CE, Wakeman CA, Sieling CL, Baker SC, Irnov I, Winkler WC. 2007. Structure and mechanism of a metal-sensing regulatory RNA. *Cell* **130:** 878–892. doi:10.1016/j.cell.2007.06.051

Das J, Mukherjee S, Mitra A, Bhattacharyya D. 2006. Non-canonical base pairs and higher order structures in nucleic acids: crystal structure database analysis. *J Biomol Struct Dyn* **24:** 149–161. doi:10.1080/07391102.2006.10507103

Djelloul M, Denise A. 2008. Automated motif extraction and classification in RNA tertiary structures. *RNA* **14:** 2489–2497. doi:10.1261/rna.1061108

Doudna JA, Cech TR. 2002. The chemical repertoire of natural ribozymes. *Nature* **418:** 222–228. doi:10.1038/418222a

Gan HH, Fera D, Zorn J, Shiffeldrim N, Tang M, Laserson U, Kim N, Schlick T. 2004. RAG: RNA-As-Graphs database—concepts, analysis, and features. *Bioinformatics* **20:** 1285–1291. doi:10.1093/bioinformatics/bth084

Gilbert SD, Rambo RP, Van Tyne D, Batey RT. 2008. Structure of the SAM-II riboswitch bound to *S*-adenosylmethionine. *Nat Struct Mol Biol* **15:** 177–182. doi:10.1038/nsmb.1371

Grabow W, Jaeger L. 2013. RNA modularity for synthetic biology. *F1000Prime Rep* **5:** 46. doi:10.12703/P5-46

Guerrier-Takada C, Gardiner K, Marsh T, Pace N, Altman S. 1983. The RNA moiety of ribonuclease P is the catalytic subunit of the enzyme. *Cell* **35:** 849–857. doi:10.1016/0092-8674(83)90117-4

Halder A, Halder S, Bhattacharyya D, Mitra A. 2014. Feasibility of occurrence of different types of protonated base pairs in RNA: a quantum chemical study. *Phys Chem Chem Phys* **16:** 18383–18396. doi:10.1039/C4CP02541E

Harrison A-M, South DR, Willett P, Artymiuk PJ. 2003. Representation, searching and discovery of patterns of bases in complex RNA structures. *J Comput Aided Mol Des* **17:** 537–549. doi:10.1023/B:JCAM.0000004603.15856.32

Havrila M, Réblová K, Zirbel CL, Leontis NB, Šponer J. 2013. Isosteric and nonisosteric base pairs in RNA motifs: molecular dynamics and bioinformatics study of the sarcin-ricin internal loop. *J Phys Chem B* **117:** 14302–14319. doi:10.1021/jp408530w

Henkin TM. 2008. Riboswitch RNAs: using RNA to sense cellular metabolism. *Genes Dev* **22:** 3383–3390. doi:10.1101/gad.1747308

Hornby D. 1993. Hydrogen bonding in biological structures. *FEBS Lett* **323:** 295. doi:10.1016/0014-5793(93)81362-4

Kim S, Suddath F, Quigley G, McPherson A, Sussman J, Wang A, Seeman N, Rich A. 1974a. Three-dimensional tertiary structure of yeast phenylalanine transfer RNA. *Science* **185:** 435–440. doi:10.1126/science.185.4149.435

Kim SH, Sussman JL, Suddath FL, Quigley GJ, McPherson A, Wang AH, Seeman NC, Rich A. 1974b. The general structure of transfer RNA molecules. *Proc Natl Acad Sci* **71:** 4970–4974. doi:10.1073/pnas.71.12.4970

Kim N, Zheng Z, Elmetwaly S, Schlick T. 2014. RNA graph partitioning for the discovery of RNA modularity: a novel application of graph partition algorithm to biology. *PLoS One* **9:** e106074. doi:10.1371/journal.pone.0106074

Klein DJ, Schmeing TM, Moore PB, Steitz TA. 2001. The kink-turn: a new RNA secondary structure motif. *EMBO J* **20:** 4214–4221. doi:10.1093/emboj/20.15.4214

Klein DJ, Edwards TE, Ferré-D'Amaré AR. 2009. Cocrystal structure of a class-I preQ1 riboswitch reveals a pseudoknot recognizing an essential hypermodified nucleobase. *Nat Struct Mol Biol* **16:** 343–344. doi:10.1038/nsmb.1563

Laing C, Jung S, Iqbal A, Schlick T. 2009. Tertiary motifs revealed in analyses of higher-order RNA junctions. *J Mol Biol* **393:** 67–82. doi:10.1016/j.jmb.2009.07.089

Laing C, Wen D, Wang JTL, Schlick T. 2012. Predicting coaxial helical stacking in RNA junctions. *Nucleic Acids Res* **40:** 487–498. doi:10.1093/nar/gkr629

Lemieux S, Major F. 2006. Automated extraction and classification of RNA tertiary structure cyclic motifs. *Nucleic Acids Res* **34:** 2340–2346. doi:10.1093/nar/gkl120

Leontis NB, Westhof E. 2001. Geometric nomenclature and classification of RNA base pairs. *RNA* **7:** 499–512. doi:10.1017/S1355838201002515

Leontis NB, Westhof E. 2002. The annotation of RNA motifs. *Comp Funct Genomics* **3:** 518–524. doi:10.1002/cfg.213

Leontis NB, Westhof E. 2003. Analysis of RNA motifs. *Curr Opin Struct Biol* **13:** 300–308. doi:10.1016/S0959-440X(03)00076-9

Leontis NB, Zirbel CL. 2012. Nonredundant 3D structure datasets for RNA knowledge extraction and benchmarking. In *Nucleic acids and molecular biology RNA 3D structure analysis and prediction* (ed. Leontis NB, Westhof E), pp. 281–298. Springer, Berlin, Heidelberg.

Lu X-J, Olson WK. 2008. 3DNA: a versatile, integrated software system for the analysis, rebuilding and visualization of three-dimensional nucleic-acid structures. *Nat Protoc* **3:** 1213–1227. doi:10.1038/nprot.2008.104

Lu X-J, Olson WK, Bussemaker HJ. 2010. The RNA backbone plays a crucial role in mediating the intrinsic stability of the GpU dinucleotide platform and the GpUpA/GpA miniduplex. *Nucleic Acids Res* **38:** 4868–4876. doi:10.1093/nar/gkq155

Lu X-J, Bussemaker HJ, Olson WK. 2015. DSSR: an integrated software tool for dissecting the spatial structure of RNA. *Nucleic Acids Res* **43:** e142.

Malgowska M, Czajczynska K, Gudanis D, Tworak A, Gdaniec Z. 2016. Overview of the RNA G-quadruplex structures. *Acta Biochim Pol* **63:** 609–621. doi:10.18388/abp.2016_1335

McPhee SA, Huang L, Lilley DMJ. 2014. A critical base pair in k-turns that confers folding characteristics and correlates with biological function. *Nat Commun* **5:** 5127. doi:10.1038/ncomms6127

Mládek A, Sharma P, Mitra A, Bhattacharyya D, Šponer J, Šponer JE. 2009. Trans Hoogsteen/sugar edge base pairing in RNA. Structures, energies, and stabilities from quantum chemical calculations. *J Phys Chem B* **113:** 1743–1755. doi:10.1021/jp808357m

Montero JJ, de Silanes IL, Graña O, Blasco MA. 2016. Telomeric RNAs are essential to maintain telomeres. *Nat Commun* **7:** 12534. doi:10.1038/ncomms12534

Nagaswamy U, Voss N, Zhang Z, Fox GE. 2000. Database of non-canonical base pairs found in known RNA structures. *Nucleic Acids Res* **28:** 375–376. doi:10.1093/nar/28.1.375

Nagaswamy U, Larios-Sanz M, Hury J, Collins S, Zhang Z, Zhao Q, Fox GE. 2002. NCIR: a database of non-canonical interactions in known RNA structures. *Nucleic Acids Res* **30:** 395–397. doi:10.1093/nar/30.1.395

Nissen P, Ippolito JA, Ban N, Moore PB, Steitz TA. 2001. RNA tertiary interactions in the large ribosomal subunit: the A-minor motif. *Proc Natl Acad Sci* **98:** 4899–4903. doi:10.1073/pnas.081082398

Parisien M, Major F. 2008. The MC-Fold and MC-Sym pipeline infers RNA structure from sequence data. *Nature* **452:** 51–55. doi:10.1038/nature06684

Ray SS, Halder S, Kaypee S, Bhattacharyya D. 2012. HD-RNAS: an automated hierarchical database of RNA structures. *Front Genet* **3:** 59.

Rázga F, Špačková N, Réblová K, Koča J, Leontis NB, Šponer J. 2004. Ribosomal RNA kink-turn motif—a flexible molecular hinge. *J Biomol Struct Dyn* **22:** 183–194. doi:10.1080/07391102.2004.10506994

Réblová K, Šponer JE, Špačková N, Beššeová I, Šponer J. 2011. A-minor tertiary interactions in RNA kink-turns. Molecular dynamics and quantum chemical analysis. *J Phys Chem B* **115:** 13897–13910. doi:10.1021/jp2065584

Sashital DG, Butcher SE. 2006. Flipping off the riboswitch: RNA structures that control gene expression. *ACS Chem Biol* **1:** 341–345. doi:10.1021/cb6002465

Schneemann A. 2006. The structural and functional role of RNA in icosahedral virus assembly. *Annu Rev Microbiol* **60:** 51–67. doi:10.1146/annurev.micro.60.080805.142304

Serganov A, Yuan Y-R, Pikovskaya O, Polonskaia A, Malinina L, Phan AT, Hobartner C, Micura R, Breaker RR, Patel DJ. 2004. Structural basis for discriminative regulation of gene expression by adenine- and guanine-sensing mRNAs. *Chem Biol* **11:** 1729–1741. doi:10.1016/j.chembiol.2004.11.018

Serganov A, Huang L, Patel DJ. 2008. Structural insights into amino acid binding and gene control by a lysine riboswitch. *Nature* **455:** 1263–1267. doi:10.1038/nature07326

Sharma P, Mitra A, Sharma S, Singh H, Bhattacharyya D. 2008. Quantum chemical studies of structures and binding in noncanonical RNA base pairs: the trans Watson-Crick:Watson-Crick family. *J Biomol Struct Dyn* **25:** 709–732. doi:10.1080/07391102.2008.10507216

Sharma M, Bulusu G, Mitra A. 2009. MD simulations of ligand-bound and ligand-free aptamer: molecular level insights into the binding and switching mechanism of the add A-riboswitch. *RNA* **15:** 1673–1692. doi:10.1261/rna.1675809

Sharma P, Chawla M, Sharma S, Mitra A. 2010a. On the role of Hoogsteen:Hoogsteen interactions in RNA: ab initio investigations of structures and energies. *RNA* **16:** 942–957. doi:10.1261/rna.1919010

Sharma P, Šponer JE, Šponer J, Sharma S, Bhattacharyya D, Mitra A. 2010b. On the role of the cis Hoogsteen:sugar-edge family of base pairs in platforms and triplets-quantum chemical insights into RNA structural biology. *J Phys Chem B* **114:** 3307–3320. doi:10.1021/jp910226e

Šponer JE, Leszczynski J, Sychrovský V, Šponer J. 2005a. Sugar edge/sugar edge base pairs in RNA: stabilities and structures from quantum chemical calculations. *J Phys Chem B* **109:** 18680–18689. doi:10.1021/jp053379q

Šponer JE, Špačková N, Kulhánek P, Leszczynski J, Šponer J. 2005b. Non-Watson-Crick base pairing in RNA. quantum chemical analysis of the cis Watson-Crick/sugar edge base pair family. *J Phys Chem A* **109:** 2292–2301. doi:10.1021/jp050132k

Šponer JE, Špačková N, Leszczynski J, Šponer J. 2005c. Principles of RNA base pairing: structures and energies of the trans Watson-Crick/sugar edge base pairs. *J Phys Chem B* **109:** 11399–11410. doi:10.1021/jp051126r

Stombaugh J, Zirbel CL, Westhof E, Leontis NB. 2009. Frequency and isostericity of RNA base pairs. *Nucleic Acids Res* **37:** 2294–2312. doi:10.1093/nar/gkp011

Westbrook J, Feng Z, Chen L, Yang H, Berman HM. 2003. The Protein Data Bank and structural genomics. *Nucleic Acids Res* **31:** 489–491. doi:10.1093/nar/gkg068

Wu L, Chai D, Fraser ME, Zimmerly S. 2012. Structural variation and uniformity among tetraloop-receptor interactions and other loop-helix interactions in RNA crystal structures. *PLoS One* **7:** e49225. doi:10.1371/journal.pone.0049225

Xin Y, Olson WK. 2009. BPS: a database of RNA base-pair structures. *Nucleic Acids Res* **37:** D83–D88. doi:10.1093/nar/gkn676

Zheng G, Lu X-J, Olson WK. 2009. Web 3DNA—a web server for the analysis, reconstruction, and visualization of three-dimensional nucleic-acid structures. *Nucleic Acids Res* **37:** W240–W246. doi:10.1093/nar/gkp358

Going beyond base-pairs: topology-based characterization of base-multiplets in RNA

Sohini Bhattacharya, Ayush Jhunjhunwala, Antarip Halder, et al.

RNA 2019 25: 573-589 originally published online February 21, 2019
Access the most recent version at doi:10.1261/rna.068551.118