Published online 10 January 2023

NAR Genomics and Bioinformatics, 2023, Vol. 5, No. 1 **1**
https://doi.org/10.1093/nargab/lqac102

Identification and characterization of RNA pentaloop sequence families

Md.Sharear Saon, Charles C. Kirkpatrick and Brent M. Znosko*

Department of Chemistry, Saint Louis University, Saint Louis, MO 63103, USA

Received October 05, 2022; Revised October 28, 2022; Editorial Decision December 01, 2022; Accepted December 12, 2022

ABSTRACT

One of the current methods for predicting RNA tertiary structure is fragment-based homology, which predicts tertiary structure from secondary structure. For a successful prediction, this method requires a library of the tertiary structures of small motifs clipped from previously solved RNA 3D structures. Because of the limited number of available tertiary structures, it is not practical to find structures for all sequences of all motifs. Identifying sequence families for motifs can fill the gaps because all sequences within a family are expected to have similar structural features. Currently, a collection of well-characterized sequence families has been identified for tetraloops. Because of their prevalence and biological functions, pentaloop structures should also be well-characterized. In this study, 10 pentaloop sequence families are identified. For each family, the common and distinguishing structural features are highlighted. These sequence families can be used to predict the tertiary structure of pentaloop sequences for which a solved structure is not available.

INTRODUCTION

In addition to transcription and translation, functionalities of RNA range from regulating gene expression (1–3) and catalysis (4,5) to acting as a potential therapeutic target (1,6,7). It is the tertiary folding which is responsible for the diverse functionalities of RNA (8–12). Therefore, to better understand RNA biology, a better understanding of RNA tertiary structure is required. However, current experimental methods, such as X-ray crystallography (13,14), cryogenic electron microscopy (cryo-EM) (15), and nuclear magnetic resonance (NMR) (16,17), are time consuming, expensive, and require extensive technical skill in order to solve RNA tertiary structure. Therefore, improving methods to predict tertiary structure from sequence will pave the way toward improved understanding of the structure-function relationships of RNA.

One of the current methods for predicting tertiary structure is fragment-based homology. In this method, the secondary structure of the whole molecule is fragmented into smaller motifs; these motifs are then searched for within a library of fragments extracted from previously solved tertiary structures of RNA (18–28). Finally, all the 3D fragments for each motif are assembled to predict the structure of the whole molecule. Examples of software programs that utilize fragment-based methods are RNAComposer (19,23,28), 3dRNA (24), Vfold3D (25), FARNA (26) and MC-Fold/MC-Sym (27). The accuracy of the prediction of tertiary structure using fragment-based methods can be improved by identifying and characterizing sequence families that all adopt the same structural features for all the secondary structural motifs. These sequence families can provide 3D templates for secondary structure components when no structure is available for a given sequence.

Currently, GNRA (N = any nucleotide and R = G or A) (29–31), UNCG (32–34) and RNYA (Y = C or U) (35) are the common tetraloop sequence families described extensively in the literature. The Znosko lab has used a protocol similar to the one used in this study to identify a few additional tetraloop sequence families, including YGAR, UGGU and RMSA (M = A or C; R = A or G; S = G or C and Y = C or U) (22).

Although pentaloops are not as common as tetraloops, they are abundant and serve important biological roles. For example, in *Escherichia coli*, 13% and 24% of the total hairpins in 16S rRNA (31) and large subunit rRNA (36), respectively, are pentaloops. Pentaloops serve a variety of roles, including acting as nucleation sites for higher order folding (37,38) and recognition sites for other biomolecules (39–43). Despite their abundance and roles, a comprehensive analysis of pentaloop structures is lacking. The only proposed sequence family for pentaloops is GNRNA (43–45). GNRNA pentaloops adopt a structure similar to that of GNRA tetraloops. In both GNRA tetraloops and GNRNA pentaloops, the N, R and A nucleotides stack (29–31,43–45), and the first G and last A form a sheared G·A pair (29–31,43–45). To allow this three base stack and G·A sheared pair, the second N of GNRNA pentaloops is extruded from the hairpin (43–45).

*To whom correspondence should be addressed. Tel: +1 314 977 8567; Fax: +1 314 977 2521; Email: brent.znosko@slu.edu

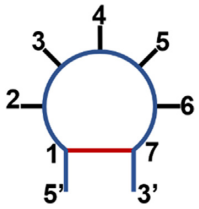

Figure 1. Secondary structure of a pentaloop with nucleotides numbered.

Here, we complete a comprehensive structural analysis of previously solved pentaloops. We identify 10 pentaloop sequence families and list their structural features as well as their distinguishing features. These sequence families can provide more insight into RNA 3D structure and sequence-structure relationships.

MATERIALS AND METHODS

Data preparation

The RNA Characterization of Secondary Structure Motifs (RNA CoSSMos) database (46,47) was used to locate all RNA three-dimensional structures solved by X-ray crystallography which contain pentaloops. RNA CoSSMos requires all secondary structure motifs to have at least two canonical closing base pairs. Python scripts (22) were used to download all the PDB structures containing pentaloops from the Research Collaboratory for Structural Bioinformatics (RCSB) Protein Data Bank (PDB) (48). Pentaloops were then clipped from the structures, and residues were renumbered starting with the 5′ closing nucleotide and ending with the 3′ closing nucleotide (Figure 1). Because coordinates for hydrogen atoms were missing from some structures, all hydrogen atom coordinates were removed from all the structures. The clipped structures were then checked for missing atoms, multiple coordinates for the same atom, or empty structural files. The residue renumbering, removal of hydrogen atom coordinates, and quality check were performed according to the protocol developed previously (22).

Identifying sequence representative structures

Sequence representative structures were identified for sequences with multiple solved structures (Table 1). The reason for using sequence representative structures is two-fold. First, when identifying structural features for a sequence family, finding sequence representative structures helped to prevent bias towards the sequences with more available structures. Second, using sequence representative structures allows for the prediction of the mostly likely structural conformation from sequence, which could aid in 3D structure prediction from sequence. It has been noted previously (22,32,49) that a given sequence does not always correspond to one 3D structure. While the approach of using a sequence representative structure has its advantages, we acknowledge that we may be missing out on some structural idiosyncrasies displayed by some instances of a particular pentaloop sequence.

To identify the sequence representative structures, first, all structures with the same sequence were grouped together. Information (such as PDB ID, RNA type, length, chain ID, and residue index) was collected from the RCSB PDB (48) for each structure. If structures appear to be the same motif repeated in multiple PDB files, these structures were condensed to one structure before the sequence representative structure was determined. As a simple example, let us consider an imaginary pentaloop sequence that has 10 structures. Of these 10, four start at residue 60 of $tRNA^{Phe}$, five start at residue 1500 in *E. coli* 16s rRNA, and one is from a signal recognition particle. The four loops in $tRNA^{Phe}$ are the same loop in multiple PDB structures, so these were collapsed to one structure. Similarly, the five loops in 16S rRNA are the same loop in multiple PDB structures, so these were also collapsed to one structure. Ultimately, the sequence representative structure for this pentaloop was determined from just three structures, the one $tRNA^{Phe}$ loop, the one 16S rRNA loop and the one loop from the signal recognition particle. One average structure of these three structures would be calculated, and the structure closest to the average would serve as the sequence representative structure.

Comparing and clustering sequence representative structures

Sequence representative structures for all the unique sequences were compared using the sugar atoms, phosphate atoms, and three atoms from the nitrogenous base (C4, C8 and N9 for purines and C2, C6 and N1 for pyrimidines), and all-against-all root mean square deviation (RMSD) values were calculated. Using the RMSD values, a distance matrix was calculated. Biopython's unweighted pair group method with arithmetic mean (UPGMA) (50) modified previously (22) was used to generate a distance tree. As was done previously (22), branches within 1.0 Å RMSD cut-off were clustered together. The decision to use a 1.0 Å cut-off for clustering was based on previous work with tetraloops (22). During that project, tetraloops were clustered using 0.2 Å intervals with a range of 0.8–2.2 Å. Because the GNRA and UNCG sequence families were already well-characterized, the clustering results were analyzed with a focus on these two families. Based on that analysis, a 1.0 Å cut-off was used for tetraloops. A similar analysis was done for pentaloops; however, well-established pentaloop families are not available for comparison. Ultimately, a 1.0 Å cut-off was also used for pentaloops.

Each cluster (Table 1) was assigned a degenerate consensus sequence based on the Cavener rules (51). In short, for each position of the consensus sequence, a nucleotide identity was assigned if the frequency of that nucleotide is >50% and greater than twice the frequency of the next most frequent nucleotide. Multiple nucleotides were coassigned to a position if the sum of their frequency was >75% and if neither of these nucleotides met the requirement for assigning a single nucleotide. If none of the two rules mentioned above can be applied to assign a nucleotide set to a position, then 'N' was assigned.

For better visualization of the clusters and their structural features (see below), representative structures for each cluster were identified. To identify the cluster representative structure, an average structure was calculated from all the sequence representative structures within a cluster. Then,

Table 1. Clustered unique sequences with representative structures

Clusters[a,b]	Unique sequence[a]	Total number of structures	Sequence representative structure[c]
CGYAYAG	CGUAUAG	4	2XLK_CGUAUAG_D_11*
	CGCACAG	2	5O7H_CGCACAG_A_33
KGARYAV	GGAGUAC	1	1XJR_GGAGUAC_A_22
	UGAAUAG	4	4WF9_UGAAUAG_X_687
	UGAACAA	1	6VMY_UGAACAA_A_224*
CGCAAYG	CGCAAUG	39	1KC8_CGCAAUG_A_196
	CGCAACG	27	3CME_CGCAACG_0_196*
CSGCGAG	CCGCGAG	66	1YHQ_CCGCGAG_0_218*
	CGGCGAG	49	4WF9_CGGCGAG_X_250
GMYKGRC	GACUGGC	28	5FJC_GACUGGC_A_24
	GCUGGAC	2	6MWN_GCUGGAC_B_651*
GSUSRUC	GCUGAUC	66	1YHQ_GCUGAUC_0_1431
	GGUCGUC	43	2ZJP_GGUCGUC_X_1338
	GCUCGUC	7	4WFB_GCUCGUC_X_1362*
GUAAMKC	GUAAAUC	1	3P49_GUAAAUC_A_42*
	GUAACGC	4	3Q1Q_GUAACGC_B_78
UUSMMRA	UUGCAAA	8	1U0B_UUGCAAA_A_33*
	UUGAAGA	1	4JYA_UUGAAGA_Y_33
	UUCACAA	1	5T83_UUCACAA_A_58
WGAAADK	UGAAAGG	66	2QEX_UGAAAGG_0_873*
	UGAAAAG	33	2ZJP_UGAAAAG_X_793
	AGAAAUU	7	4WFA_AGAAAUU_X_825
YUGUUCG	UUGUUCG	24	1K73_UUGUUCG_A_2587
	CUGUUCG	51	4WFB_CUGUUCG_X_2579*

[a]The direction of sequences is 5′-3′, and hairpin residues are underlined.
[b]D = A, G or U; K = G or U; M = A or C; R = A or G; S = G or C; V = A, C or G, W = A or U, and Y = C or U.
[c]The name of each structure includes the PDB ID, sequence, chain ID, and residue index of the first residue (nucleotide 2 in Figure 1) separated by an underscore '_'. The asterisk identifies the cluster representative structure.

the structure that most closely resembled the average structure (the structure that had lowest RMSD to the average structure) was identified as the representative structure for that cluster.

Characterization of clusters

Using Dissecting the Spatial Structures of RNA (DSSR) (52), all sequence representative structures within each cluster were characterized to identify structural features of that cluster. Three Python scripts (22) were used to annotate and tally five structural features: stacking interactions, base conformation, sugar pucker, hydrogen bonds (not part of any base pair), and base pairs (non-canonical) between loop nucleotides. All structural features are defined by DSSR (52). While characterizing structures, tertiary interactions between the pentaloop and any residue outside of the pentaloop were not considered. In addition to the sequence representative structures (also referred to as dataset 1), two additional datasets of structures were compiled: all structures in the cluster (not just the sequence representative structures) and all other representative structures not in the cluster (also referred to as datasets 2 and 3, respectively). Dataset 2 was used because the number of unique sequences (n) that make up dataset 1 was very small. In a small dataset, one structure having/missing a feature drastically alters the percentage of structures with that feature. Dataset 2 (all structures in the cluster) contained a significantly larger n, resulting in percentages that were not as sensitive to individual structures with or missing a certain feature. Identity and frequency of the structural features within the first two datasets were compared with the identity and frequency of the same structural features within the third dataset. Any structural feature present in $\geq$75% of structures in datasets 1 or 2 was included in the cluster analysis tables. If the percentage of occurrence of a structural feature is $\geq$75% for the first and/or second dataset, but <75% in the third dataset and at least 20% below what is found in the first and/or second dataset, that structural feature was considered a distinguishing feature of that cluster.

Comparison to NMR structures

Although NMR structures were not included in the main analysis, a comparison was made between pentaloops found in NMR structures and the pentaloops found in X-ray crystal structures. PDB entries determined by NMR that contain a pentaloop were identified. For each ensemble of structures, one representative structure was identified by averaging all structures in the ensemble and finding the ensemble structure that was closest to the average. NMR pentaloops were clustered separately and together with crystal structures. The resulting clusters were compared to the clusters determined from crystal structures only.

RESULTS AND DISCUSSION

Sequence representative structures

As of June 2021, 1486 pentaloop structures were identified in RNA three-dimensional structures solved by X-ray crystallography. Nine structures did not pass the quality check (Supplementary Table S1). From the 1477 clipped pentaloop structures (Supplementary Table S2), 86 unique sequences (Supplementary Table S3) were identified. Disproportionate redundancy of solved tertiary structures for

unique sequences was observed within the dataset of pentaloop structures. For example, the 5′-CAAAAUG-3′ pentaloop sequence was found only twice in the dataset, whereas the 5′-GCUCAAC-3′ pentaloop sequence was found 105 times in the dataset. Therefore, to avoid bias toward structures with overrepresented sequences, a representative structure was calculated for each sequence with multiple structures (Supplementary Table S3).

Cluster identification and degenerate consensus sequence assignment

Twenty-four sequence representative structures were clustered into 10 sequence families (Table 1 and Figure 2). The remaining 62 sequence representative structures were unclustered (Figure 2). Degenerate consensus sequences were assigned to represent the sequences within each cluster. Details of the analysis of clustered sequences while assigning a degenerate consensus sequence and distances between clusters can be found in Supplementary Tables S4 and S5, respectively. Four clusters, GSUSRUC (S = G or C and R = A or G), KGARYAV (K = G or U; Y = C or U and V = A, C, or U), UUSMMRA (M = A or C), and WGAAADK (W = A or U and D = A, G or U), each contained three unique sequences, and the rest of the six clusters contained two unique sequences.

Comparison to NMR structures

While NMR structures contain a wealth of information, only crystal structures were used in the subsequent analysis. The decision to exclude NMR structures was three-fold. First, and most importantly, studies have shown that NMR-derived structures exhibit more steric clashes and conformational ambiguities than their crystallographic counterparts (53). Second, the inclusion of only X-ray structures for this pentaloop analysis mirrors what was done previously for hairpin structure analyses by the Major lab for triloops (54) and the Znosko lab for tetraloops (22). Third, the submission of an ensemble of structures for NMR-derived RNA structures significantly complicates analysis for NMR structures.

A preliminary analysis was completed to show that including NMR structures would have little effect on the reported clusters. Sixty-two PDB entries determined by NMR that contain a pentaloop were identified. These 62 PDB entries consist of 791 total structures due to multiple structures submitted for each ensemble. In order to identify one representative structure from each ensemble, all structures in an ensemble were averaged, and the ensemble structure closest to the average was taken forward in the analysis. When clustering the NMR structures only, only two unique sequences were placed into a cluster; all other sequences were unclustered. When using all of the pentaloop structures in the PDB (those solved by NMR and crystallography), the NMR structures only represent ∼4% of the total structures. When this combined set of structures was clustered, again, only two NMR unique sequences were clustered, with all other NMR structures being unclustered. These results suggest that there is likely a better protocol to analyze NMR ensembles. Due to the steric clashes and conformational ambiguities raised in the literature, historical precedence, the small percentage of NMR-derived pentaloop structures in comparison to X-ray-derived structures, and only two clustered NMR unique sequences if they were included in the analysis, only X-ray structures were included in the following analysis.

Structural analysis for sequence families

Two clusters, CGYAYAG (Y = C or U; Figure 3A; and Table 2 and Supplementary Table S6) and KGARYAV (K = G or U; R = A or G; Y = C or U; V = A, C or G; Figure 3B; and Table 2 and Supplementary Table S7), were identified with sequences that fall within the previously identified GNRNA sequence family. For CGYAYAG, there were two unique sequences, 5′-CGUAUAG-3′ and 5′-CGCACAG-3′, representing four and two total structures, respectively. The sequence representative structure for 5′-CGUAUAG-3′ was identified as the cluster representative structure for CGYAYAG. The cluster representative structure was clipped from the stem-loop repeat of clustered regularly interspaced short palindromic repeats (CRISPR)-derived RNAs bound to endoribonuclease (55). No residues in the cluster representative structure were involved in tertiary interactions. Four stacking interactions were identified between nucleotides (nts) 1 and 2, 3 and 4, 4 and 6, and 6 and 7. For all residues in CGYAYAG, base conformation and sugar pucker were identified as anti and C3′-endo, respectively. In addition to the closing base pair between nts 1 and 7, a base pair between nts 2 and 6 was identified. Three additional hydrogen bonds, between O2′ (hydroxyl) of nt 2 and N6 (amino) of nt 4, N2 (amino) of nt 2 and OP1 of nt 5, and O2′ (hydroxyl) of nt 2 and N7 of nt 4, were identified. These three hydrogen bonds, along with the four stacking interactions, sugar pucker for nts 2–6, and the base pair between nts 2 and 6 were identified as the distinguishing structural features of CGYAYAG (Table 2).

For KGARYAV, three unique sequences, 5′-GGAGUAC-3′, 5′-UGAAUAG-3′ and 5′-UGAACAA-3′, were clustered, where only the second sequence has multiple structures (four structures). For KGARYAV, the sequence representative structure for 5′-UGAACAA-3′ was identified as the cluster representative structure. The cluster representative structure was clipped from a cobalamin riboswitch of *Bacillus subtillis* (56). No tertiary interactions were identified between the loop residues and other residues of the riboswitch. Four stacking interactions were identified between nts 1 and 2, 3 and 4, 4 and 6, and 6 and 7. The base conformation for all residues was found to be anti. Except for nts 4 and 5, the sugar pucker for the rest of the nucleotides was C3′-endo. No base pair between loop nucleotides was identified. One hydrogen bond between O2′ (hydroxyl) of nt 2 and N7 of nt 4 was identified. Along with this hydrogen bond, all four stacking interactions and the sugar pucker for nts 2, 3 and 6 were identified as KGARYAV cluster distinguishing structural features (Table 2).

According to the literature, in GNRNA pentaloops, there are four stacking interactions between nts 1 and 2, 3 and 4, 4 and 6, and 6 and 7, while nt 5 stays outward (43–45). There is also a sheared base pair between nts 2 and

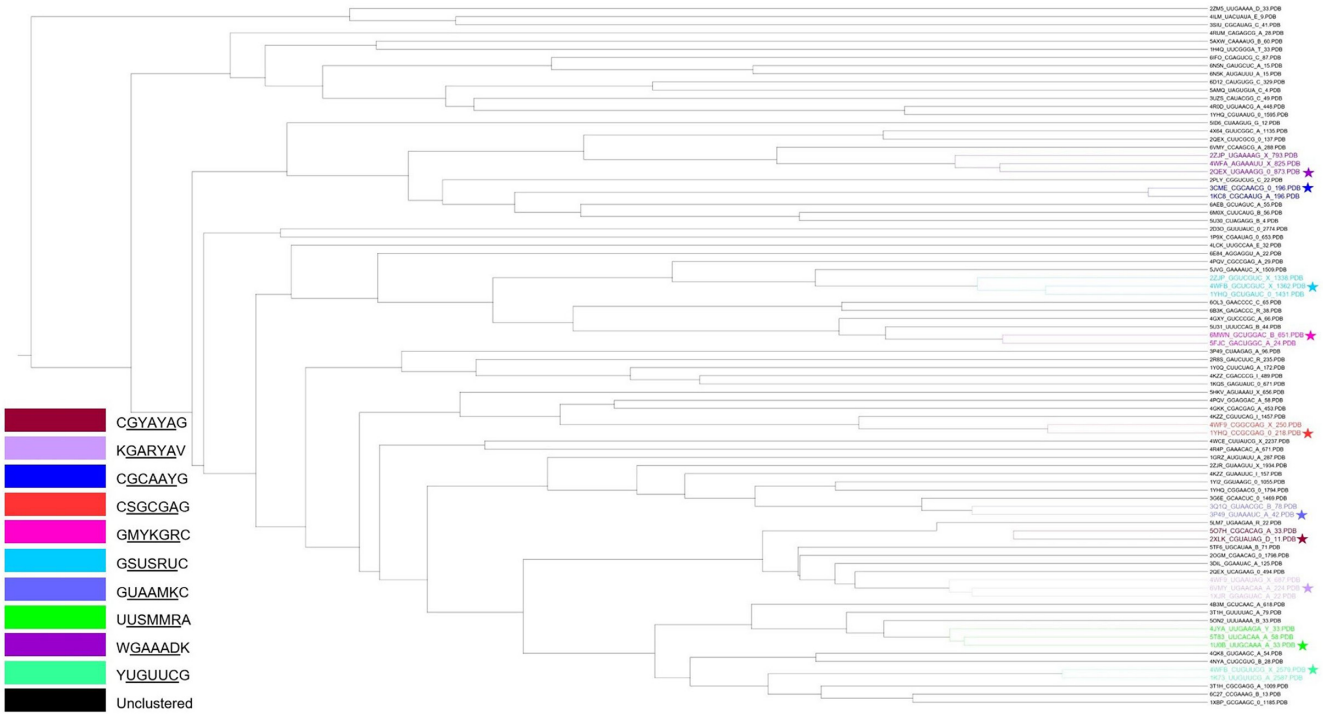

Figure 2. Distance tree (generated by Archaeopteryx 0.9901 beta (65)) created from an all-against-all sequence representative structure analysis. Each branch on the right is indicating a sequence representative structure (cluster representative structures are asterisked) where the structures within a cluster are magnified and color-coded according to the legend.

6 (43–45). For CGYAYAG and KGARYAV, all the stacking interactions identified in the literature were also identified in this study. However, the sheared base pair between nts 2 and 6 was only identified in CGYAYAG. In addition to the presence/absence of this base pair, additional differences were also identified. For example, the sugar pucker for all residues in CGYAYAG is C3′-endo, whereas the sugar pucker for nts 4 and 5 in KGARYAV were not C3′-endo. In addition, CGYAYAG had two additional hydrogen bonds that were not found in KGARYAV. The structural data collected here suggests that these two sequence families adopt different structural features and should be considered different families (and not grouped together under one GNRA pentaloop family).

To our knowledge, the following eight clusters are all newly identified sequence families for pentaloops. The first newly identified cluster in this study was CGCAAYG (Y = C or U; Figure 3C; and Table 2 and Supplementary Table S8). The total number of hits for the two unique sequences in this cluster, 5′-CGCAAUG-3′ and 5′-CGCAACG-3′, were 39 and 27, respectively. The cluster representative structure for this family was identified from the sequence representative structure of 5′-CGCAACG-3′. The cluster representative structure was clipped from the large ribosomal subunit of *Haloarcula marismortui* bound with a peptidyl-tRNA analog (57). The G of the hairpin loop has tertiary interactions with an internal loop and a hairpin loop. Three stacking interactions between nts 1 and 2, 2 and 4, and 5 and 7 were identified in all sequence representative structures. All bases are in the anti conformation. The sugar puckers for nts 1, 3, 5, and 7 were C3′-endo, and the sugar puckers for nts 2, 4 and 6 were C2′-endo. In addition to the closing base pair between nt 1 and 7, a base pair between nt 2 and 5 was identified in all structures and representative structures of the cluster. Two additional hydrogen bonds, between O2′ (hydroxyl) of nt 2 and N7 of nt 7 and between O2′ (hydroxyl) of nt 5 and O4′ of nt 7, were identified. These two hydrogen bonds, along with the three stacking interactions, sugar puckers for nts 2–6, and base pair between nts 2 and 5 were identified as distinguishing structural features of this cluster (Table 2).

The second new cluster identified was CSGCGAG (S = G or C; Figure 3D; and Table 2 and Supplementary Table S9). For the two unique sequences in this cluster, 5′-CCGCGAG-3′ and 5′-CGGCGAG-3′, the total number of structures were 66 and 49, respectively. The sequence representative structure for 5′-CCGCGAG-3′ was identified as the cluster representative structure. This pentaloop structure was clipped from the *Haloarcula marismortui* large (50S) ribosomal subunit bound with azithromycin (58). The cluster representative structure did not have any tertiary interactions. Five stacking interactions were identified between nts 1 and 2, 2 and 4, 3 and 5, 5 and 6, and 6 and 7. The base conformation for all residues is anti. Except for nt 4 (C2′-endo), the sugar pucker for all other residues is C3′-endo. A base pair was identified between nts 2 and 6. Two additional hydrogen bonds between O2′ (hydroxyl) of nt 2 and N7 of nt 5 and O2′ (hydroxyl) of nt 2 and O6 (carbonyl) of nt 5 were identified in all sequence representative structures. All the stacking interactions, hydrogen bonds, base pair, and sugar pucker for nts 2- 6 were identified as the distinguishable structural features for this cluster (Table 2).

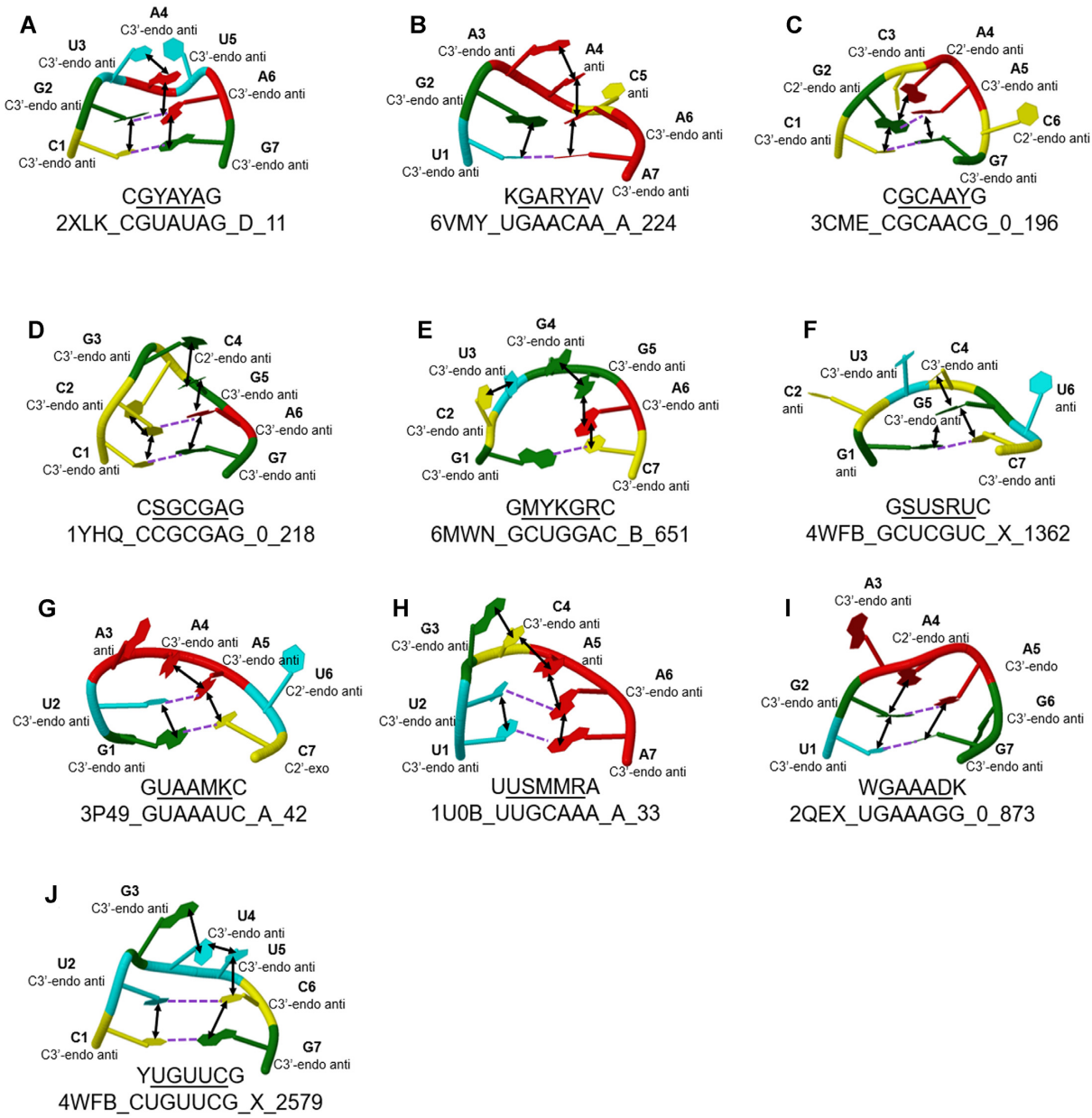

Figure 3. Cluster representative structure for each of the 10 clusters (**A**–**J**). 3D structures were created by DSSR-Jmol (66). If determined to be a feature of the cluster, stacking interactions (black arrows), glycosidic conformation, sugar pucker and base pairing (dashed lines) are shown for each cluster. Additional hydrogen bonds are not shown for clarity. Degenerate sequence (underlined section of the sequence represents the hairpin residues) and name of the representative structure for each cluster are shown below each cluster representative structure.

The next new cluster of pentaloops is GMYKGRC (M = A or C, Y = C or U, K = G or U and R = A or G; Figure 3E; and Table 2 and Supplementary Table S10). This cluster contains two unique sequences, 5′-GACUGGC-3′ and 5′-GCUGGAC-3′, where the total number of structures was 28 and 2, respectively. The sequence representative structure for 5′-GCUGGAC-3′ was identified as the cluster representative structure. This pentaloop was found in the hepatitis A virus internal ribosome entry sites (IRES) domain V (dV) complexed with a crystallization chaperone (59). The cluster representative structure did not have any tertiary interactions. Four stacking interactions were identified between nts 2 and 3, 4 and 5, 5 and 6, and 6 and 7. The base conformation and sugar pucker for all nts were identified as anti and C3′-endo, respectively. No additional base pairs or hydrogen bonds within the motif residues were identified. All stacking interactions and the sugar pucker for nts 2–6 were identified as the distinguishable structural features for the cluster (Table 2).

For the new cluster GSUSRUC (S = G or C and R = A or G; Figure 3F; and Table 2 and Supplementary Table S11), there were three unique sequences, 5′-GCUGAUC-3′, 5′-GGUCGUC-3′ and 5′-GCUCGUC-3′, with a total number of structures of 66, 43 and 7, respectively. The sequence rep-

Table 2. Summary of structural features of clusters[a]

	Interaction				
Cluster name	Stacking interaction between	Base conformation	Sugar pucker	Base pair within the loop residues	Additional hydrogen bonds
CGYAYAG	nt 1 and nt 2 nt 3 and nt 4 nt 4 and nt 6 nt 6 and nt 7	All anti	All C3′-endo	Between nt 2 and nt 6	O2′ (hydroxyl) of nt 2 and N7 of nt 4 O2′ (hydroxyl) of nt 2 and N6 (amino) of nt 4 N2 (amino) of nt 2 and OP1 of nt 5
KGARYAV	nt 1 and nt 2 nt 3 and nt 4 nt 4 and nt 6 nt 6 and nt 7	All anti	nts 1–3 and 6–7 are C3′-endo	-	O2′ (hydroxyl) of nt 2 and N7 of nt 4
CGCAAYG	nt 1 and nt 2 nt 2 and nt 4 nt 5 and nt 7	All anti	nts 1, 3, 5 and 7 are C3′-endo nts 2, 4, and 6 are C2′-endo	Between nt 2 and nt 5	O2′ (hydroxyl) of nt 2 and N7 of nt 7 O2′ (hydroxyl) of nt 5 and O4′ of nt 7
CSGCGAG	nt 1 and nt 2 nt 2 and nt 4 nt 3 and nt 5 nt 5 and nt 6 nt 6 and nt 7	All anti	nts 1–3 and 5–7 are C3′-endo nt 4 is C2′-endo	Between nt 2 and nt 6	O2′ (hydroxyl) of nt 2 and N7 of nt 5 O2′ (hydroxyl) of nt 2 and O6 (carbonyl) of nt 5
GMYKGRC	nt 2 and nt 3 nt 4 and nt 5 nt 5 and nt 6 nt 6 and nt 7	All anti	All C3′-endo	-	-
GSUSRUC	nt 1 and nt 5 nt 4 and nt 5 nt 5 and nt 7	All anti	nts 3–5, and 7 are C3′-endo	-	OP2 of nt 6 and N4 (amino) of nt 7
GUAAMKC	nt 1 and nt 2 nt 4 and nt 5 nt 5 and nt 7	All anti	nts 1–2 and 4–5 are C3′-endo nt 6 is C2′-endo nt 7 is C2′-exo	Between nt 2 and nt 5	O2′ (hydroxyl) of nt 2 and N7 of nt 4
UUSMMRA	nt 1 and nt 2 nt 3 and nt 4 nt 4 and nt 5 nt 5 and nt 6 nt 6 and nt 7	All anti	nts 1–4 and 6–7 are C3′-endo	Between nt 2 and nt 5	-
WGAAADK	nt 1 and nt 2 nt 2 and nt 4 nt 5 and nt 7	nts 1–4 and 6–7 are anti	nts 1–3 and 5–7 are C3′-endo	Between nt 2 and nt 5	O2′ (hydroxyl) of nt 2 and OP1 of nt 4 O2′ (hydroxyl) of nt 5 and OP1 of nt 7
YUGUUCG	nt 1 and nt 2 nt 3 and nt 4 nt 4 and nt 5 nt 5 and nt 6 nt 6 and nt 7	All anti	All C3′-endo	Between nt 2 and nt 6	-

[a]See Supplementary Tables S6-S15 for a more detailed analysis of each cluster, including statistics, the identification of distinguishing features, and a comparison between each cluster and all other pentaloops.

resentative structure for 5′-GCUCGUC-3′ was identified as the cluster representative structure. This pentaloop structure was located in the *Staphylococcus aureus* large ribosomal subunit complexed with a pleuromutilin derivative (60). One tertiary interaction was identified between the first C residue in the hairpin loop and another hairpin of the ribosomal subunit. Three stacking interactions were identified between nts 1 and 5, 4 and 5, and 5 and 7. For all residues, the base conformation was anti. For nts 3–5 and 7, the sugar pucker is C3′-endo. Only one hydrogen bond was identified, between OP2 of nt 6 and N4 (amino) of nt 7. All stacking interactions, sugar puckers for nts 3–5, and hydrogen bond between nts 6 and 7 are the distinguishing features for GSUSRUC (Table 2).

Another new cluster, GUAAMKC (M = A or C and K = G or U; Figure 3G; and Table 2 and Supplementary Table S12), contains two unique sequences, 5′-GUAAAUC-3′ and 5′-GUAACGC-3′, where the first sequence has only one available structure, and the second has four available structures. The sequence representative structure for 5′-GUAAAUC-3′ was identified as the cluster representative structure. This pentaloop structure was clipped from the glycine riboswitch of *Fusobacterium nucleatum* (61). The last U residue of the hairpin loop has tertiary interactions with an internal loop within the riboswitch. Three stacking interactions were identified between nts 1 and 2, 4 and 5, and 5 and 7. All bases are in the anti-conformation. The sugar pucker for nts 1, 2, 4, and 5 were identified as C3′-endo. For nts 6 and 7 in the sequence representative structures, the sugar puckers were found to be C2′-endo and C2′-exo, respectively. One base pair was identified between nts 2 and 5. One additional hydrogen bond was identified between O2′ (hydroxyl) of nt 2 and N7 of nt 4. Along with this base pair and hydrogen bond, all three stacking interactions and sugar conformations for nts 2 and 4–7 were found to be distinguishable features for GUAAMKC (Table 2).

The new cluster UUSMMRA (S = G or C, M = A or C and R = A or G; Figure 3H; and Table 2 and Supplementary Table S13) contains three unique sequences: 5′-UUGCAAA-3′, 5′-UUGAAGA-3′, and 5′-UUCACAA-3′, where the first sequence has eight available structures, and each of the next two sequences has only one structure. 5′-UUGCAAA-3′ is the sequence of the cluster representative structure. This pentaloop was found in *Escherichia coli* tRNA (Cys) complexed with cysteinyl-tRNA synthetase (62). The cluster representative structure did not have any tertiary interactions. Five stacking interactions were identified between nts 1 and 2, 3 and 4, 4 and 5, 5 and 6, and 6 and 7. Bases of all nucleotides were in the anti-conformation. Except for nt 5, all sugar puckers are C3′-endo. In addition to the closing base pair between nts 1 and 7, a base pair between nts 2 and 5 was identified. No additional hydrogen bonds within the loop residues were identified. In addition to the stacking interactions, the sugar pucker for nts 2–6, and the base pair between nts 2 and 5 were identified as distinguishing structural features for UUSMMRA (Table 2).

Three unique sequences were clustered into the new cluster WGAAADK (W = A or U; D = A, G or U; K = G or U; Figure 3I; and Table 2 and Supplementary Table S14), 5′-UGAAAGG-3′, 5′-UGAAAAG-3′, and 5′-AGAAAUU-3′, with the number of available structures being 66, 33 and 7, respectively. The representative structure for the sequence 5′-UGAAAGG-3′ was identified as the cluster representative structure. The pentaloop structure representing this cluster was clipped from *Haloarcula marismortui* 23S ribosomal RNA complexed with negamycin (63). The cluster representative structure did not have any tertiary interactions. Three stacking interactions between nts 1 and 2, 2 and 4, and 5 and 7 were identified. All residues except nt 5 were identified to form anti glycosidic angles. Except for nt 4, all other residues had a C3′-endo sugar pucker. A base pair between nts 2 and 5 was identified along with the closing base pair between nts 1 and 7. Two additional hydrogen bonds were identified between the O2′ (hydroxyl) of nt 2 and OP1 of nt 4 and the O2′ (hydroxyl) of nt 5 and OP1 of nt 7. Along with these two hydrogen bonds, all three stacking interactions, sugar puckers for nts 2, 3, 5 and 6, and the base pair between nts 2 and 5 were found to be distinguishable structural features for this cluster (Table 2).

The last new cluster identified for pentaloops was YUGUUCG (Y = C or U; Figure 3J; and Table 2 and Supplementary Table S15) with two unique sequences, 5′-UUGUUCG-3′ and 5′-CUGUUCG-3′, representing 24 and 51 structures, respectively. The sequence for the cluster representative structure was 5′-CUGUUCG-’3. The cluster representative structure was found in the *Staphylococcus aureus* large ribosomal subunit complexed with a pleuromutilin derivative (60). The C residue of the hairpin loop has a tertiary interaction with another hairpin within the ribosomal subunit. Five stacking interactions were found between nts 1 and 2, 3 and 4, 4 and 5, 5 and 6, and 6 and 7. The base conformation for all residues was anti. The sugar pucker for all residues was C3′-endo. A base pair between nts 2 and 6 was identified in addition to the closing base pair between nts 1 and 7. No hydrogen bonds between hairpin residues were identified. All five stacking interactions, sugar pucker for nts 2–6, and base pair between nts 2 and 6 were identified as the distinguishable structural features for YUGUUCG (Table 2).

Within all clusters, stacking interactions were found to be more prevalent than hydrogen bond interactions. All clusters have at least three stacking interactions whereas three clusters (GMYKGRC, UUSMMRA and YUGUUCG) did not exhibit any hydrogen bonding between the motif residues. Except for nt 1 in GMYKGRC, all residues forming closing base pairs have stacking interactions. However, GMYKGRC has a stacking interaction between nts 2 and 3, which is not present in any other cluster. Seven out of 10 clusters were found to have at least one base pair between the hairpin nucleotides.

Comparison to previously identified clusters

The Zirbel and Leontis labs have developed a valuable online RNA database, the RNA 3D Motif Atlas (64). In this database, they collect internal loops and hairpins extracted from RNA 3D structures and cluster the loops into motif groups. There are several major differences between the RNA 3D Motif Atlas and the analysis reported here. One major difference is how ‘pentaloop’ is defined. Here, a pentaloop is strictly limited to hairpin loops closed by at least two adjacent canonical base pairs. Also, the first mismatch

of the hairpin cannot be a canonical pair. Lastly, a hairpin cannot contain two adjacent canonical pairs within the loop residues. A second major difference is that this work only considers pentaloops, not hairpins of other sizes. The RNA 3D Motif Atlas compares all hairpin sizes to each other in order to generate their motif groups. A third major difference is that different clustering protocols were used in the two projects. Due to these differences, it is not surprising that none of the clusters identified here match any of the motif groups identified by the RNA 3D Motif Atlas.

Although the two projects are ultimately different, comparisons can be made. For example, the GACUGGC pentaloop from PDB ID 5FJC and the GCUGGAC pentaloop from PDB ID 6MWN were clustered into the GMYKGRC cluster here. In the RNA 3D Motif Atlas, the same pentaloops can be found as HL_32512.8 in cluster KNBBVNS. The GMYKGRC cluster is more limited in sequence space and does fall within the KNBBVNS sequence. Also, the CGGCGAG pentaloop from PDB ID 4WF9 was clustered into the CSGCGAG cluster here. In the RNA 3D Motif Atlas, the same pentaloop can be found as HL_67216.5 in cluster YMKCRAG. As seen in the example above, the YMKCRAG cluster is more limited in sequence space and does fall within the YMKCRAG sequence.

In summary, 10 sequence families, their representative structural features, and their distinguishing features were identified from a dataset of 1477 structures solved by X-ray crystallography. In 3D structure prediction using a fragment-based homology method, these sequence families can provide 3D templates for a pentaloop sequence when no structure is available. For example, to date there is no solved structure available for the pentaloop sequence 5′-CGUACAG-3′. However, this sequence is a member of the CGYAYAG sequence family identified in this study. Therefore, the structural features such as four stacking interactions, base and sugar conformation, base pair within the pentaloop nucleotides, and additional hydrogen bonds in CGYAYAG can be assigned as structural features for 5′-CGUACAG-3′. In addition to contributing to tertiary structure prediction from sequence, these sequence families along with tertiary interaction information can provide insights into RNA-ligand binding interactions.

SUPPLEMENTARY DATA

Supplementary Data are available at NARGAB Online.

FUNDING

National Institutes of Health [2R15GM085699-04]. Funding for open access charge: National Institutes of Health.
Conflict of interest statement. None declared.

REFERENCES

1. Fu,D., Shi,Y., Liu,J.-B., Wu,T.-M., Jia,C.-Y., Yang,H.-Q., Zhang,D.-D., Yang,X.-L., Wang,H.-M. and Ma,Y.-S. (2020) Targeting long non-coding RNA to therapeutically regulate gene expression in cancer. *Mol. Ther. Nucleic Acids*, **21**, 712–724.
2. Brookes,E. and Pombo,A. (2009) Modifications of RNA polymerase II are pivotal in regulating gene expression states. *EMBO Rep.*, **10**, 1213–1219.
3. Prasanth,K.V., Prasanth,S.G., Xuan,Z., Hearn,S., Freier,S.M., Bennett,C.F., Zhang,M.Q. and Spector,D.L. (2005) Regulating gene expression through RNA nuclear retention. *Cell*, **123**, 249–263.
4. Strulson,C.A., Molden,R.C., Keating,C.D. and Bevilacqua,P.C. (2012) RNA catalysis through compartmentalization. *Nat. Chem.*, **4**, 941–946.
5. Noller,H.F., Hoffarth,V. and Zimniak,L. (1992) Unusual resistance of peptidyl transferase to protein extraction procedures. *Science*, **256**, 1416–1419.
6. Sczepanski,J.T. and Joyce,G.F. (2013) Binding of a structured D-RNA molecule by an L-RNA aptamer. *J. Am. Chem. Soc.*, **135**, 13290–13293.
7. Ren,S., Liu,Y., Xu,W., Sun,Y., Lu,J., Wang,F., Wei,M., Shen,J., Hou,J., Gao,X. *et al.* (2013) Long noncoding RNA MALAT-1 is a new potential therapeutic target for castration resistant prostate cancer. *J. Urol.*, **190**, 2278–2287.
8. Campagnola,G., McDonald,S., Beaucourt,S., Vignuzzi,M. and Peersen,O.B. (2015) Structure-function relationships underlying the replication fidelity of viral RNA-dependent RNA polymerases. *J. Virol.*, **89**, 275–286.
9. Dyson,M.R., Mandal,N. and RajBhandary,U.L. (1993) Relationship between the structure and function of Escherichia coli initiator tRNA. *Biochimie*, **75**, 1051–1060.
10. Lee,K., Varma,S., SantaLucia,J. and Cunningham,P.R. (1997) In vivo determination of RNA structure-function relationships: analysis of the 790 loop in ribosomal RNA. *J. Mol. Biol.*, **269**, 732–743.
11. McCarthy,N. (2005) Form and function. *Nat. Rev. Cancer*, **5**, 669–669.
12. Travers,A. and Muskhelishvili,G. (2015) DNA structure and function. *FEBS J.*, **282**, 2279–2295.
13. Reyes,F.E., Garst,A.D. and Batey,R.T. (2009) Strategies in RNA crystallography. *Methods Enzymol.*, **469**, 119–139.
14. Westhof,E. (2015) Twenty years of RNA crystallography. *RNA*, **21**, 486–487.
15. Fernandez-Leiro,R. and Scheres,S.H.W. (2016) Unravelling biological macromolecules with cryo-electron microscopy. *Nature*, **537**, 339–346.
16. Varani,G., Aboul-ela,F. and Allain,F.H.T. (1996) NMR investigation of RNA structure. *Prog. Nucl. Magn. Reson. Spectrosc.*, **29**, 51–127.
17. Fürtig,B., Richter,C., Wöhnert,J. and Schwalbe,H. (2003) NMR spectroscopy of RNA. *ChemBioChem*, **4**, 936–962.
18. Popenda,M., Blazewicz,M., Szachniuk,M. and Adamiak,R.W. (2008) RNA FRABASE version 1.0: an engine with a database to search for the three-dimensional fragments within RNA structures. *Nucleic Acids Res.*, **36**, D386–D391.
19. Popenda,M., Szachniuk,M., Antczak,M., Purzycka,K.J., Lukasiak,P., Bartol,N., Blazewicz,J. and Adamiak,R.W. (2012) Automated 3D structure composition for large rnas. *Nucleic Acids Res.*, **40**, e112.
20. Popenda,M., Szachniuk,M., Blazewicz,M., Wasik,S., Burke,E.K., Blazewicz,J. and Adamiak,R.W. (2010) RNA FRABASE 2.0: an advanced web-accessible database with the capacity to search the three-dimensional fragments within RNA structures. *BMC Bioinform.*, **11**, 231–231.
21. Pucci,F. and Schug,A. (2019) Shedding light on the dark matter of the biomolecular structural universe: progress in RNA 3D structure prediction. *Methods*, **162-163**, 68–73.
22. Richardson,K.E., Adams,M.S., Kirkpatrick,C.C., Gohara,D.W. and Znosko,B.M. (2019) Identification and characterization of new RNA tetraloop sequence families. *Biochemistry*, **58**, 4809–4820.
23. Biesiada,M., Pachulska-Wieczorek,K., Adamiak,R.W. and Purzycka,K.J. (2016) RNAComposer and RNA 3D structure prediction for nanotechnology. *Methods*, **103**, 120–127.
24. Zhao,Y., Huang,Y., Gong,Z., Wang,Y., Man,J. and Xiao,Y. (2012) Automated and fast building of three-dimensional RNA structures. *Sci. Rep.*, **2**, 734–734.
25. Zhao,C., Xu,X. and Chen,S.-J. (2017) Predicting RNA structure with Vfold. *Methods Mol. Biol.*, **1654**, 3–15.
26. Das,R. and Baker,D. (2007) Automated de novo prediction of native-like RNA tertiary structures. *Proc. Natl. Acad. Sci. U.S.A.*, **104**, 14664–14669.
27. Parisien,M. and Major,F. (2008) The MC-Fold and MC-Sym pipeline infers RNA structure from sequence data. *Nature*, **452**, 51–55.
28. Biesiada,M., Purzycka,K.J., Szachniuk,M., Blazewicz,J. and Adamiak,R.W. (2016) Automated RNA 3D structure prediction with rnacomposer. *Methods Mol. Biol.*, **1490**, 199–215.

29. Correll,C.C. and Swinger,K. (2003) Common and distinctive features of GNRA tetraloops based on a GUAA tetraloop structure at 1.4 Å resolution. *RNA*, **9**, 355–363.
30. Heus,H. and Pardi,A. (1991) Structural features that give rise to the unusual stability of RNA hairpins containing GNRA loops. *Science*, **253**, 191–194.
31. Woese,C.R., Winker,S. and Gutell,R.R. (1990) Architecture of ribosomal RNA: constraints on the sequence of "tetra-loops". *Proc. Natl. Acad. Sci. U.S.A.*, **87**, 8467–8471.
32. Bottaro,S. and Lindorff-Larsen,K. (2017) Mapping the universe of RNA tetraloop folds. *Biophys. J.*, **113**, 257–267.
33. Varani,G., Cheong,C. and Tinoco,I. (1991) Structure of an unusually stable RNA hairpin. *Biochemistry*, **30**, 3280–3289.
34. Ennifar,E., Nikulin,A., Tishchenko,S., Serganov,A., Nevskaya,N., Garber,M., Ehresmann,B., Ehresmann,C., Nikonov,S. and Dumas,P. (2000) The crystal structure of UUCG tetraloop. *J. Mol. Biol.*, **304**, 35–42.
35. Rowsell,S., Stonehouse,N.J., Convery,M.A., Adams,C.J., Ellington,A.D., Hirao,I., Peabody,D.S., Stockley,P.G. and Phillips,S.E.V. (1998) Crystal structures of a series of RNA aptamers complexed to the same protein target. *Nat. Struct. Biol.*, **5**, 970–975.
36. Gutell,R.R. and Fox,G.E. (1988) A compilation of large subunit RNA sequences presented in a structural format. *Nucleic Acids Res.*, **16** (Suppl), r175–r269.
37. Fedorova,O. and Pyle,A.M. (2005) Linking the group II intron catalytic domains: tertiary contacts and structural features of domain 3. *EMBO J.*, **24**, 3906–3916.
38. Cate,J.H., Gooding,A.R., Podell,E., Zhou,K., Golden,B.L., Szewczak,A.A., Kundrot,C.E., Cech,T.R. and Doudna,J.A. (1996) RNA tertiary structure mediation by adenosine platforms. *Science*, **273**, 1696–1699.
39. Lapouge,K., Perozzo,R., Iwaszkiewicz,J., Bertelli,C., Zoete,V., Michielin,O., Scapozza,L. and Haas,D. (2013) RNA pentaloop structures as effective targets of regulators belonging to the RsmA/CsrA protein family. *RNA Biol.*, **10**, 1031–1041.
40. Liu,S., Ghalei,H., Lührmann,R. and Wahl,M.C. (2011) Structural basis for the dual U4 and U4atac snRNA-binding specificity of spliceosomal protein hPrp31. *RNA*, **17**, 1655–1663.
41. Skrisovska,L., Bourgeois,C.F., Stefl,R., Grellscheid,S.-N., Kister,L., Wenter,P., Elliott,D.J., Stevenin,J. and Allain,F.H.T. (2007) The testis-specific human protein RBMY recognizes RNA through a novel mode of interaction. *EMBO Rep.*, **8**, 372–379.
42. Stefl,R. and Allain,F.H.T. (2005) A novel RNA pentaloop fold involved in targeting ADAR2. *RNA*, **11**, 592–597.
43. Legault,P., Li,J., Mogridge,J., Kay,L.E. and Greenblatt,J. (1998) NMR structure of the bacteriophage λ N peptide/boxB RNA complex: recognition of a GNRA fold by an arginine-rich motif. *Cell*, **93**, 289–299.
44. Huppler,A., Nikstad,L.J., Allmann,A.M., Brow,D.A. and Butcher,S.E. (2002) Metal binding and base ionization in the U6 RNA intramolecular stem-loop structure. *Nat. Struct. Biol.*, **9**, 431–435.
45. Schärpf,M., Sticht,H., Schweimer,K., Boehm,M., Hoffmann,S. and Rösch,P. (2000) Antitermination in bacteriophage λ. *Eur. J. Biochem.*, **267**, 2397–2408.
46. Vanegas,P.L., Hudson,G.A., Davis,A.R., Kelly,S.C., Kirkpatrick,C.C. and Znosko,B.M. (2012) RNA CoSSMos: characterization of secondary structure motifs–A searchable database of secondary structure motifs in RNA three-dimensional structures. *Nucleic Acids Res.*, **40**, D439–D444.
47. Richardson,K.E., Kirkpatrick,C.C. and Znosko,B.M. (2020) RNA CoSSMos 2.0: an improved searchable database of secondary structure motifs in RNA three-dimensional structures. *Database*, **2020**, baz153.
48. Burley,S.K., Bhikadiya,C., Bi,C., Bittrich,S., Chen,L., Crichlow,G.V., Christie,C.H., Dalenberg,K., Di Costanzo,L., Duarte,J.M. *et al.* (2020) RCSB Protein Data Bank: powerful new tools for exploring 3D structures of biological macromolecules for basic and applied research and education in fundamental biology, biomedicine, biotechnology, bioengineering and energy sciences. *Nucleic Acids Res.*, **49**, D437–D451.
49. Lemieux,S. and Major,F. (2006) Automated extraction and classification of RNA tertiary structure cyclic motifs. *Nucleic Acids Res.*, **34**, 2340–2346.
50. Cock,P.J.A., Antao,T., Chang,J.T., Chapman,B.A., Cox,C.J., Dalke,A., Friedberg,I., Hamelryck,T., Kauff,F., Wilczynski,B. *et al.* (2009) Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics*, **25**, 1422–1423.
51. Cavener,D.R. (1987) Comparison of the consensus sequence flanking translational start sites in Drosophila and vertebrates. *Nucleic Acids Res.*, **15**, 1353–1361.
52. Lu,X.-J., Bussemaker,H.J. and Olson,W.K. (2015) DSSR: an integrated software tool for dissecting the spatial structure of RNA. *Nucleic Acids Res.*, **43**, e142.
53. Bermejo,G.A., Clore,G.M. and Schwieters,C.D. (2016) Improving NMR structures of RNA. *Structure*, **24**, 806–815.
54. Lisi,V. and Major,F. (2007) A comparative analysis of the triloops in all high-resolution RNA structures reveals sequence structure relationships. *RNA*, **13**, 1537–1545.
55. Haurwitz,R.E., Jinek,M., Wiedenheft,B., Zhou,K. and Doudna,J.A. (2010) Sequence- and structure-specific RNA processing by a CRISPR endonuclease. *Science*, **329**, 1355–1358.
56. Chan,C.W. and Mondragón,A. (2020) Crystal structure of an atypical cobalamin riboswitch reveals RNA structural adaptability as basis for promiscuous ligand binding. *Nucleic Acids Res.*, **48**, 7569–7583.
57. Simonović,M. and Steitz,T.A. (2008) Peptidyl-CCA deacylation on the ribosome promoted by induced fit and the O3′-hydroxyl group of A76 of the unacylated A-site tRNA. *RNA*, **14**, 2372–2378.
58. Tu,D., Blaha,G., Moore,P.B. and Steitz,T.A. (2005) Structures of MLSBK antibiotics bound to mutated large ribosomal subunits provide a structural explanation for resistance. *Cell*, **121**, 257–270.
59. Koirala,D., Shao,Y., Koldobskaya,Y., Fuller,J.R., Watkins,A.M., Shelke,S.A., Pilipenko,E.V., Das,R., Rice,P.A. and Piccirilli,J.A. (2019) A conserved RNA structural motif for organizing topology within picornaviral internal ribosome entry sites. *Nat. Commun.*, **10**, 3629.
60. Eyal,Z., Matzov,D., Krupkin,M., Wekselman,I., Paukner,S., Zimmerman,E., Rozenberg,H., Bashan,A. and Yonath,A. (2015) Structural insights into species-specific features of the ribosome from the pathogen *Staphylococcus aureus*. *Proc. Natl. Acad. Sci. U.S.A.*, **112**, E5805–E5814.
61. Butler,EthanB., Xiong,Y., Wang,J. and Strobel,ScottA. (2011) Structural basis of cooperative ligand binding by the glycine riboswitch. *Chem. Biol.*, **18**, 293–298.
62. Hauenstein,S., Zhang,C.-M., Hou,Y.-M. and Perona,J.J. (2004) Shape-selective RNA recognition by cysteinyl-tRNA synthetase. *Nat. Struct. Mol. Biol.*, **11**, 1134–1141.
63. Schroeder,S.J., Blaha,G. and Moore,P.B. (2007) Negamycin binds to the wall of the nascent chain exit tunnel of the 50S ribosomal subunit. *Antimicrob. Agents Chemother.*, **51**, 4462–4465.
64. Petrov,A.I., Zirbel,C.L. and Leontis,N.B. (2013) Automated classification of RNA 3D motifs and the RNA 3D motif atlas. *RNA*, **19**, 1327–1340.
65. Han,M.V. and Zmasek,C.M. (2009) phyloXML: XML for evolutionary biology and comparative genomics. *BMC Bioinform.*, **10**, 356.
66. Hanson,R.M. and Lu,X.-J. (2017) DSSR-enhanced visualization of nucleic acid structures in Jmol. *Nucleic Acids Res.*, **45**, W528–W533.