2340–2346 *Nucleic Acids Research, 2006, Vol. 34, No. 8*
doi:10.1093/nar/gkl120

Automated extraction and classification of RNA tertiary structure cyclic motifs

Sébastien Lemieux and François Major*

Institute for Research in Immunology and Cancer, Department of Computer Science and Operations Research, Université de Montréal, PO Box 6128, Downtown station, Montreal, Quebec H3C 3J7, Canada

Received October 4, 2005; Revised November 12, 2005; Accepted March 13, 2006

ABSTRACT

A minimum cycle basis of the tertiary structure of a large ribosomal subunit (LSU) X-ray crystal structure was analyzed. Most cycles are small, as they are composed of 3- to 5 nt, and repeated across the LSU tertiary structure. We used hierarchical clustering to quantify and classify the 4 nt cycles. One class is defined by the GNRA tetraloop motif. The inspection of the GNRA class revealed peculiar instances in sequence. First is the presence of UA, CA, UC and CC base pairs that substitute the usual sheared GA base pair. Second is the revelation of GNR(X_n)A tetraloops, where X_n is bulged out of the classical GNRA structure, and of GN/RA formed by the two strands of interior-loops. We were able to unambiguously characterize the cycle classes using base stacking and base pairing annotations. The cycles identified correspond to small and cyclic motifs that compose most of the LSU RNA tertiary structure and contribute to its thermodynamic stability. Consequently, the RNA minimum cycles could well be used as the basic elements of RNA tertiary structure prediction methods.

INTRODUCTION

The determination of X-ray crystallographic structures of the large ribosomal subunit (LSU) at high-resolution represents a milestone and an opportunity to learn further about RNA structure and folding (1–3). Computer tools and mathematical formalisms to describe and analyze nucleotide conformations and interactions are in place (4–10), and practical and theoretical attempts to tackle the next step of structural organization revealed the presence of repeated fragments (11–17). Theoretical studies either use general sub-graph enumerations, which have serious algorithmic and characterization problems (11,12), or require a pre-selection of the elements to be studied (13).

An RNA tertiary structure can be divided into fragments in many different ways. In the context of computer modeling, we were most interested in overlapping fragments that are easy to build and assemble in 3D space. From graph theory, we found most appropriate to determine a minimum cycle basis (18) of an RNA tertiary structure, which defines indivisible and cyclic fragments. In comparison to arbitrary fragments, indivisible and cyclic ones are both algorithmically and conceptually easier to manipulate, as they introduce special cases of enumeration and comparison.

Consider the LSU hairpin 2555–2580 shown in Figure 1. The 3D structure (Figure 1A) was obtained from the Protein Data Bank (PDB) (19). The tertiary structure graph of the hairpin is shown in Figure 1B. The graph contains 26 nt vertices and 36 interaction arcs: 11 are explicitly shown in Figure 1B and correspond to base pairing interactions and 25 are not shown and correspond to the phosphodiester linkage and base stacking. Figure 1C shows the twelve cycles of a minimum cycle basis of the hairpin graph.

The cycles of the minimum cycle basis include all nucleotides of the graph and are indivisible: if you choose any cycle and start at any nucleotide in the cycle, following the arcs of the cycle will get you back, after visiting all other nucleotides of the cycle, to the starting nucleotide and there is no alternative shorter path. Note that any given nucleotide of the graph can be involved in more than one cycle, and thus the minimum cycle basis does not partition the nucleotides. Finally, note that any pair of adjacent cycles shares at least one common edge.

Here, we computed the tertiary structure graph and a minimum cycle basis of a high-resolution LSU X-ray crystal structure. We analyzed the cycle space of the LSU using hierarchical clustering. We found that the notion of RNA cycles corresponds to the notion of RNA ‘motifs’, as employed in the RNA literature. The cycles of the LSU: (i) are repeated across the tertiary structure, (ii) have been described unambiguously by base pairing and base stacking interactions, (iii) are short, with a majority composed of 3- to 5 nt,

*To whom correspondence should be addressed. Tel: 514 343 6752; Fax: 514 343 5839; Email: francois.major@umontreal.ca

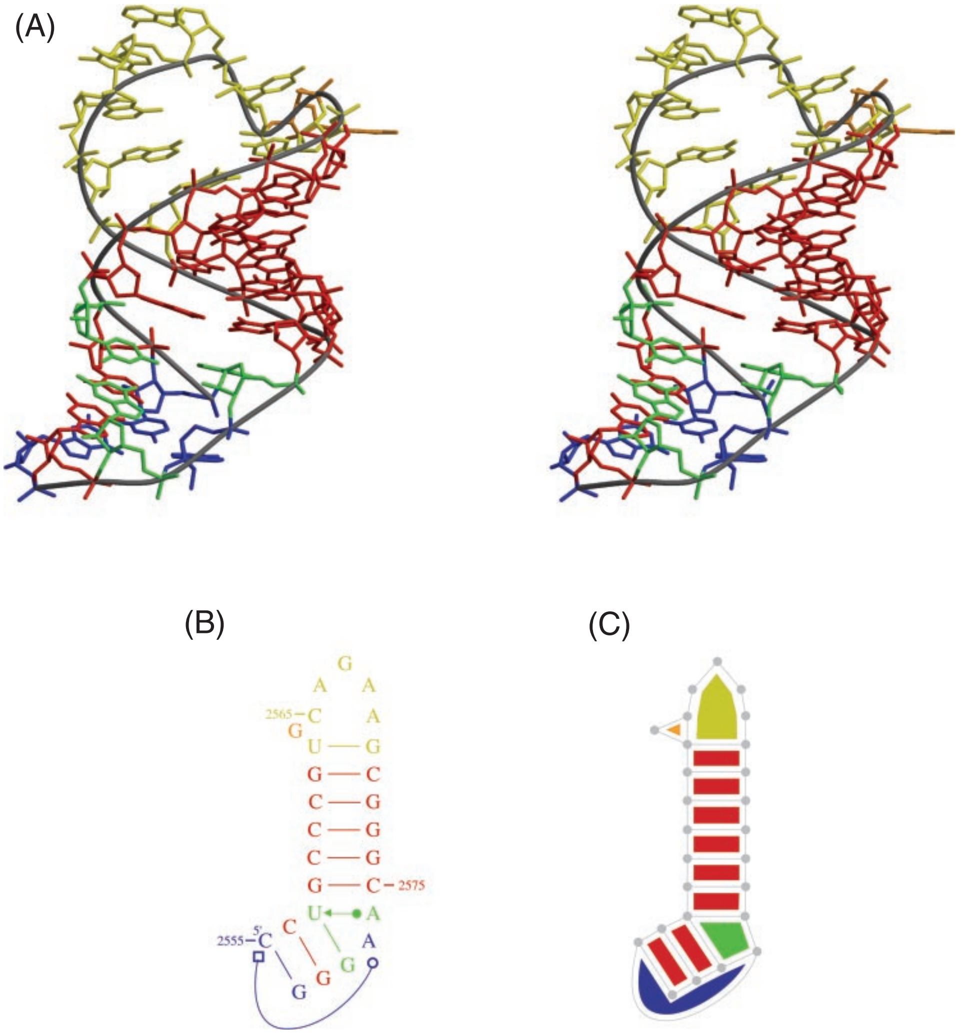


Figure 1. The large ribosomal subunit hairpin 2555–2580. (**A**) Stereo-view of the 3D structure. The phosphodiester chain is shown as a gray wire. The structure has a hairpin of 7 nt, closed by the U2563●G2570 base pair (yellow). Adjacent to the loop, U2563 and C2565 stack and G2564 bulges out of the hairpin (orange). The higher stem region (red) is formed by six Watson–Crick base pairs and ends by the non Watson–Crick *cis* U2557◂●A2576 base pair (green) (◂ sugar edge; ● Watson–Crick edge; the nomenclature for naming the non Watson–Crick base pairs was taken from Leontis and Westhof (4); see Materials and Methods). U2557 forms a Wobble base pair with G2577 (green); thus G2577●U2557◂●A2576 form a base triple (green). The stem is interrupted by the *trans* A2577○□C2555 base pair (blue) (□ Hoogsteen edge). C2555●G2580 (blue) is the last Watson–Crick base pair of the lower stem region; thus A2577○□C2555●G2580 form a base triple (blue). The core of the lower region is a short stem made of three Watson–Crick base pairs. (**B**) Tertiary structure graph. The colors match those in (A). The symbols shown for the base pairing types are the same as in (A). The sequence (phosphodiester linkage) and base stacking information is not shown. (**C**) Minimum cycle basis. Twelve indivisible cycles are shown. The nucleotides are translated in only one vertex type (gray dots). All interactions are shown, including the sequence and base stacking (gray lines). The colors match those in (A) and (B).

(iv) correspond, in some cases, to previously studied motifs and (v) capture the isosteric base pair phenomenon (instances that cluster structurally can differ in sequence).

MATERIALS AND METHODS

PDB structure

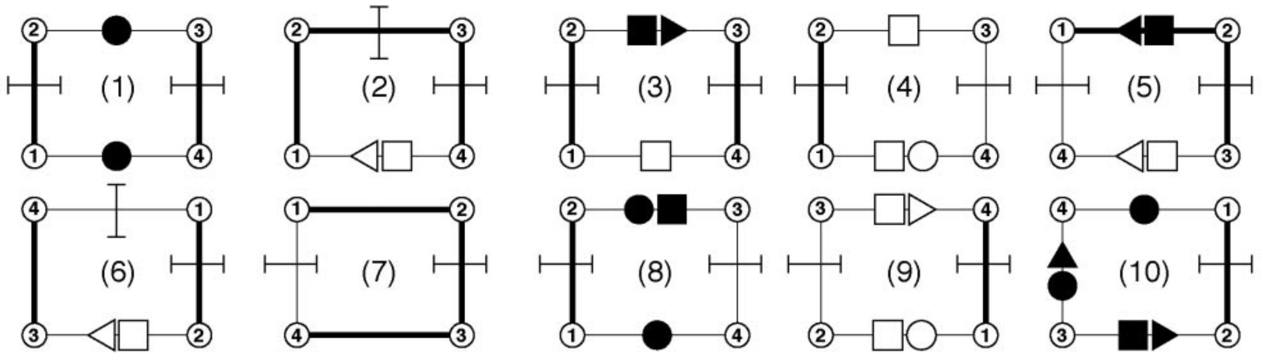
The 2.4 Å resolution X-ray crystal structure of *Haloarcula marismortui* LSU (1) (PDB code 1FFK) was obtained from the PDB (19).

LW nomenclature

We use the symbols suggested by Leontis and Westhof (4): ◂, ■ and ● to indicate respectively the Sugar, Hoogsteen and Watson–Crick edges of the bases involved in the formation of H-bonds. The symbols filled with black indicate the *cis* conformation of the base pairs; the empty symbols: ◁, □ and ○ indicate the *trans* conformation.

RNA graph

The base pairing and base stacking interactions of the structure were annotated using the *MC-Annotate* computer program (5,6). If we consider the phosphodiester link information, five major classes of nucleotide interactions result from this annotation: link (L), link-stack (LS), link-pair (LP), pair (P) and stack (S). In fact, we consider 12 base pairing types that are defined by combinations of the three interacting base edges (Watson–Crick, Hoogsteen and Sugar) and the two relative orientations of the backbones across the median line of the plane formed by the 2 bp, *cis* and *trans*, as



	#	Class	Base pairs	LSU index				Comment
(1)	637	LS-P-LS-P	(W/W,W/W)	**0**2562	**0**2563	**0**2570	**0**2571	Watson-Crick tandem
(2)	21	L-LS-LS-P	(H/S)	**0**2696	**0**2697	**0**2698	**0**2699	GNRA loop
(3)	19	LS-P-LS-P	(H/S,H/H)	**0**1532	**0**1533	**0**1658	**0**1659	Non Watson-Crick tandem
(4)	10	LS-P-S-P	(H/H,W/H)	**9**77	**9**79	**9**103	**9**104	Non Watson-Crick tandem
(5)	8	LP-LS-P-S	(S/H,H/S)	**0**1971	**0**1972	**0**1973	**0**2009	Non Watson-Crick tandem
(6)	7	LS-P-L-S	(H/S)	**0**1097	**0**1098	**0**1258	**0**1259	GNRA interior loop
(7)	6	L-LS-L-S		**0**1392	**0**1393	**0**1394	**0**1395	Double-stacked bulge
(8)	6	LS-P-S-P	(W/H,W/W)	**0**2118	**0**2276	**0**2277	**0**2470	Non Watson-Crick tandem
(9)	5	P-S-P-LS	(W/H,H/S)	**0**481	**0**485	**0**486	**0**482	Non Watson-Crick tandem
(10)	5	LS-P-P-P	(S/H,W/S,W/W)	**0**1231	**0**2498	**0**2522	**0**2523	Base triple

Figure 2. The ten most populated 4 nt cycle clusters of the LSU. The thick lines indicate adjacent nucleotides in the sequence (phosphodiester linkage). The numbering of the nucleotides is arbitrary. The numbers can be rotated clockwise and counter clockwise without modifying the properties of the cycles. The helical tandem, motif 1: LS-P-LS-P, is the most frequent with 637 instances, followed by the GNRA tetraloop motif (motif 2: L-LS-LS-P) with 21 instances. A motif similar to the GNRA tetraloop is motif 6: L-S-LS-P. It is composed of the same interactions as for the tetraloop, but nucleotides 4 and 1 (2 and 3 in motif 2) are not adjacent in the sequence. Several non Watson–Crick tandem clusters are shown (motifs 3–5, 8 and 9).

described in the Leontis and Westhof nomenclature (4). Consequently, a total of 27 interaction types are distinguished. Consider again the example of the LSU hairpin 2555–2580 (Figure 1). The hairpin PDB file (Figure 1A) was input to the *MC-Annotate* computer program, which produced the RNA tertiary structure graph shown in Figure 1B.

Minimum cycle basis

A minimum cycle basis of the RNA graph was computed using our implementation of the algorithm developed by Horton for general graphs (18). No distinction is made among the nucleotide or base interaction types. In our example, the annotated hairpin structure of Figure 1B is thus transformed in a theoretical graph made of one vertex type and one arc type, which translate in the gray dots and lines of the minimum cycle basis shown in Figure 1C.

The tandems of Watson–Crick base pairs (in red in Figure 1C) are cycles that can be described by the 'link-stack, pair (WC/WC), link-stack, pair (WC/WC)' interactions, or LS-P-LS-P for short (also shown as motif 1 in Figure 2). As we are dealing with cycles, the nucleotide we choose to start the description determines the 'phase' of the annotation. Therefore, we can equally describe the tandems of Watson–Crick stacked base pairs as 'pair (WC/WC), link-stack, pair (WC/WC), link-stack' or P-LS-P-LS for short, if we choose to start the annotation at the second nucleotide of the motif.

The cycles of the minimum cycle basis do not contain short-circuits. A cycle would contain a short-circuit if the shortest path between any pair of nucleotides in it was outside the cycle, resulting in the possibility to form two shorter (indivisible) cycles. Consider the hairpin loop in the example (Figure 1B). It contains 8 nt (if we include the nucleotides of the flanking base pair). In our notation, the loop would be described as L-L-L-LS-LS-LS-LS-P. However, the stacking interaction between U2563 and C2565 (see Figure 1C) introduces a short-circuit in the loop, making it possible to form two shorter cycles: L-L-S and S-L-LS-LS-LS-LS-P.

Clustering

In the last step of the Materials and Methods, we extract the 3D structures of the cycles of the minimum cycle basis, as returned by the application of Horton's algorithm. We submit them to hierarchical clustering using single linkage (the minimum distance between two groups is chosen as the clustering criterion). We chose single linkage for its simplicity of interpretation. However, note that different joining methods can be applied, and produce similar results (see Supplementary Figure M1A). To build the pairwise distance matrix necessary to the hierarchical clustering, we developed a cycle distance metric (see Supplementary Data) using a RNA fragment distance metric that we previously introduced (5). As can be seen in Supplementary Figure M1B, our distance metric is different, but correlates with the classical root-mean-square deviation (RMSD).

RESULTS AND DISCUSSION

LSU tertiary structure

The LSU tertiary structure contains 2826 nt and 4636 interactions, a mean of 3.3 interactions per nucleotide (2 × 4636/2826). The mean is over the three interactions needed to maintain the phosphodiester linkage of the sequence and a base pairing, and indicates that an important fraction of the nucleotides participates in the formation of tertiary interactions.

Minimal cycle basis

Our implementation of Horton's algorithm extracted a minimum cycle basis of the LSU much faster than expected from the algorithm's worst case running time, $O(n^7)$, confirming the LSU tertiary structure represents a particular type of graph. Since the LSU is currently the largest RNA in the PDB, we can conclude that our current version has a practical efficiency in the context of RNA tertiary structure.

Almost 90% of the LSU cycles contain five or less nucleotides (572 3 nt, 905 4 nt, 123 5 nt and 216 of 6 nt and more). These include the cycles in the canonical A-RNA double-helical regions, which are systematically represented by adjacent cycles of tandem Watson–Crick base pairs (see motif 1 in Figure 2; LS-P-LS-P). Almost 90% of the nucleotides in the LSU are involved in cycles of 3- and 4 nt (data not shown). Our analysis also indicates that the large cycles (size >5) are not repeated across the LSU (data not shown), and are more difficult to characterize geometrically.

Clustering

The hierarchical clustering results in a classification of all cycles of the minimum cycle basis of the LSU. The top ten populated classes of 4 nt are shown in Figure 2. As expected, the tandem of Watson–Crick base pairs is the most populated class (637 instances). Interestingly, the motif that ranks second with 21 instances is the GRNA tetraloop (motif 2), which is more frequent than any non Watson–Crick tandems. A total of 183 4 nt cycles were found as members of less-than-five instance clusters, which were not analyzed.

The number of thin arcs in Figure 2 indicates how many strands contribute to the formation of the cycles, or how many nucleotides in the cycle are not adjacent in the sequence. For instance, the tandem of Watson–Crick base pairs has two thin arcs, and thus is made of two strands. Motif 2 has one thin arc, and is made of one strand. Motif 4 has three thin arcs, and is made of three strands. In Figure 2, only two clusters are made of one-strand cycles: motifs 2 and 7. Four clusters are made of two-strand cycles (motifs 1, 3, 5 and 6), and four clusters are made of three-strand cycles.

Interestingly, motif 5, like motifs 3, 4, 8 and 9, is a non Watson–Crick tandem, but 1 bp involves two adjacent nucleotides in the sequence. In fact, if we make abstraction of the 'link' interaction, all base pair tandems shown in Figure 2, including motif 1, can be described with the S-P-S-P string. Motif 6 is also of particular interest, as it is made of the same base interactions as those found in motif 2, but is a two-strand cycle. The observation of cycles equivalent in interactions, but not in the number of participating strands, suggested that the 'link' interaction has only a minor role in the formation of the cycles. We thus decided to analyze motifs 2 and 6 more in details. Their 3D structures were extracted from the LSU X-ray crystal structure and superimposed for visualization (see Figures 3 and 4).

GNRA tetraloop

Motif 2 is the well-acknowledged and much studied GNRA tetraloop motif (L-LS-LS-P). The 21 occurrences of the GNRA cluster, as found in the LSU, share root-mean-square deviations (RMSD) below 3.0 Å (2.5 in the cycle distance). The GNRA tetraloop structure is thermodynamically stable (20), and is often found involved in stabilizing interactions with distant tetraloop receptors. The most frequent interaction is a S/S base pair between the A of the GNRA and the minor groove of an adjacent stem. However, we noted the following interactions as well. The 7th instance, in the list of Figure 3B, shows two interacting GNRA tetraloops through a W/W base pair and backbone interactions. The 10th instance shows a bifurcated H-bond involving the R of the GNRA. The 11th instance involves stacking on the N of the GNRA. The 13th and 21st instances are stabilized by S/O2′ H-bonds involving the A of the GNRA.

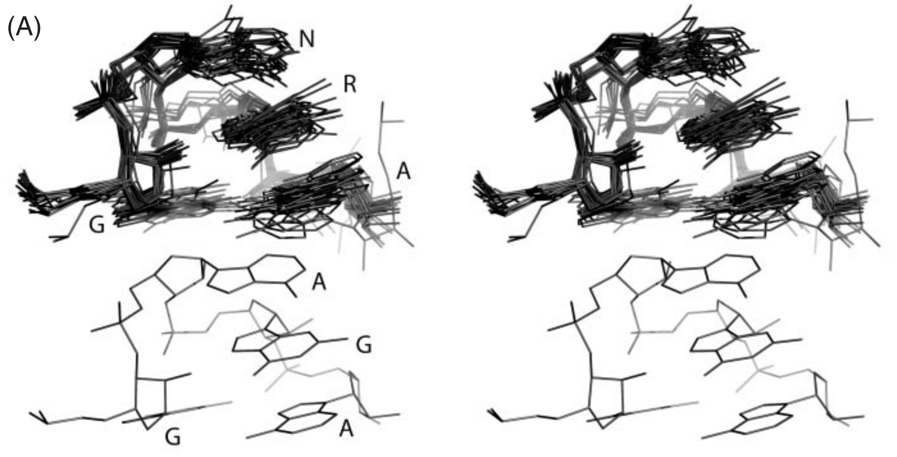
Four instances among the 21 identified do not conform to the GNRA sequence. Respectively, examples #9, #11, #19 and #20 (see Figure 3B) are constituted by the UCAA, CAGA, UCAC and CAAC sequences. The thermodynamic stability of these four non-standard GNRA tetraloops is preserved, respectively, by the base pairs U◁□A, C◁□A, U◁□C and C◁□C (see Figure 3C), which are said to be isosteric to the sheared G◁□A base pair present in all other occurrences. We say that base pairs are isosteric when they can be substituted in the structure without modifying the function. In general, base pairs that put in contact the same base edges and maintain relative glycosydic bond orientations (see Figure 3C) are isosteric.

A striking observation is that six instances in the GNRA cluster, 9–11, 13 and 21 do not have the A and R adjacent in the sequence, resulting in another variant of the GNRA tetraloop motif, as shown in Figure 3D. In these two-strand occurrences, note the conservation of the base pairing and stacking interactions, but one (instances 9–11), two (instance 13) or more (instance 21 has 21) nucleotides inserted between the A and R of the GNRA. In particular, instance 21 has the longest bulge and highest RMSD and cycle distance. This observation supports our above argument concerning the weak impact of the backbone, which may result from evolutionary events.

Note that lowering the cycle distance criterion of the cluster selection could have caught the distinction between the two GNRA variants of the cluster, as the six above examples are the only ones with cycle distances over 1.0 to instance 1 (Figure 3B). However, the RMSD had not allowed us to separate the instances.

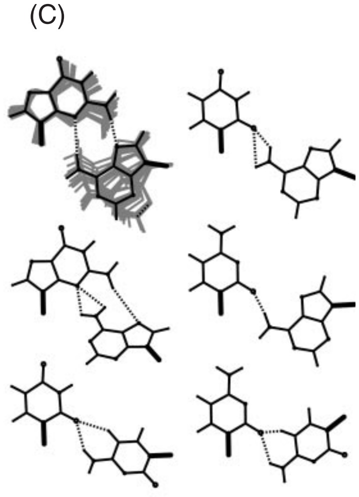
GNRA interior loop

The members of motif 6 adopt a structure that is very similar to the GNRA tetraloop with the G◁□A base pair and the NRA nucleotides stacked on the 3′ side (see Figure 4). Interestingly,



(B)

RMSD (Å)	Cycle distance	G	N	R	A
0.00	0.00	G-**0**2696	A-**0**2697	G-**0**2698	A-**0**2699
0.59	0.39	G-**0**1863	C-**0**1864	A-**0**1865	A-**0**1866
0.98	0.41	G-**0**1629	A-**0**1630	A-**0**1631	A-**0**1632
1.02	0.53	G-**0**1055	U-**0**1056	A-**0**1057	A-**0**1058
1.03	0.43	G-**0**2877	U-**0**2878	A-**0**2879	A-**0**2880
1.08	0.41	G-**0**691	A-**0**692	A-**0**693	A-**0**694
1.21	1.53	G-**0**482	C-**0**483	A-**0**484	A-**0**486
1.26	0.40	G-**9**90	C-**9**91	G-**9**92	A-**9**93
1.41	1.17	U-**0**1276	C-**0**1277	A-**0**1278	A-**0**1280
1.46	1.33	G-**0**1707	C-**0**1708	G-**0**1709	A-**0**1711
1.47	1.26	C-**0**494	A-**0**495	G-**0**496	A-**0**498
1.49	0.37	G-**0**2249	G-**0**2250	G-**0**2251	A-**0**2252
1.56	1.38	G-**0**506	A-**0**507	A-**0**508	A-**0**511
1.76	0.57	G-**0**469	U-**0**470	G-**0**471	A-**0**472
1.81	0.63	G-**0**2412	A-**0**2413	A-**0**2414	A-**0**2415
1.89	0.57	G-**0**2630	U-**0**2631	G-**0**2632	A-**0**2633
2.00	0.64	G-**0**1794	G-**0**1795	A-**0**1796	A-**0**1797
2.11	0.61	G-**0**577	C-**0**578	G-**0**579	A-**0**580
2.12	0.30	U-**0**253	C-**0**254	A-**0**255	C-**0**256
2.28	0.95	C-**0**1469	A-**0**1470	A-**0**1471	C-**0**1472
2.96	2.50	G-**0**314	G-**0**315	A-**0**316	A-**0**337



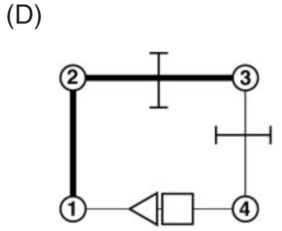
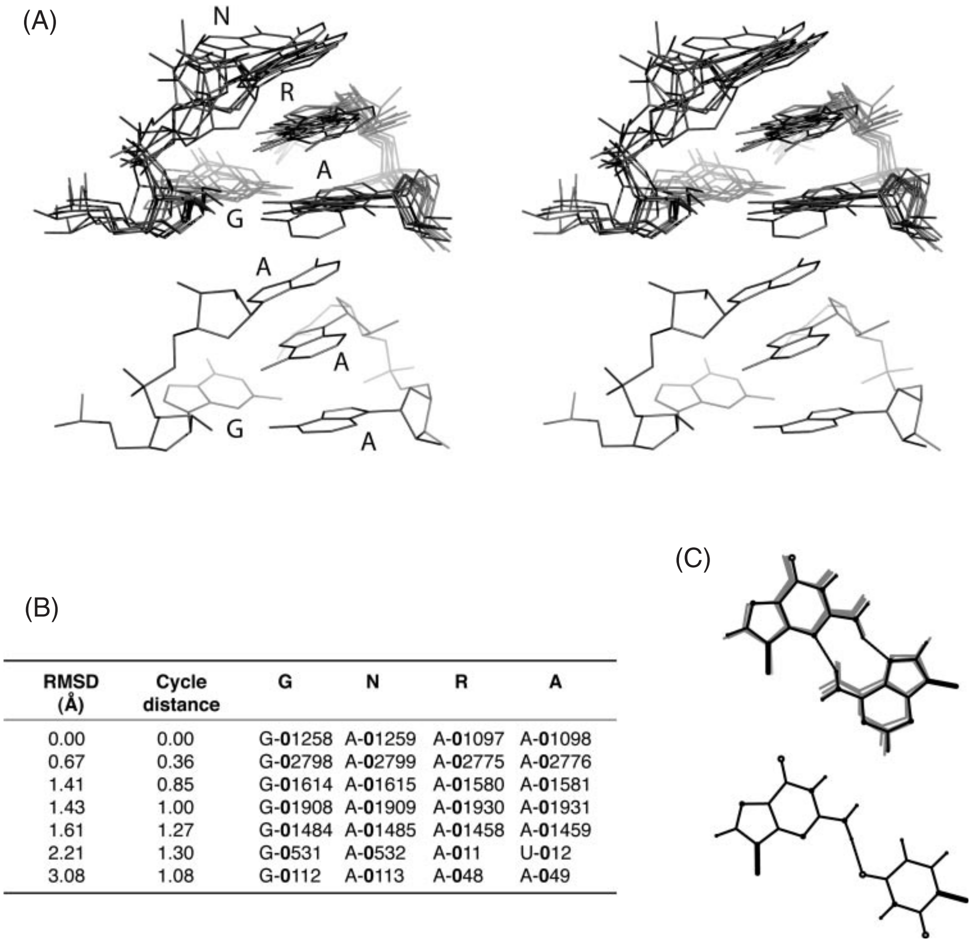


Figure 3. GNRA tetraloops from the LSU. (**A**) Stereo-view of 21 superimposed instances (above) and of one 'typical' (#1) GAGA instance (below) of the GNRA tetraloop 3D structures. The typical instance was chosen as the one minimizing the sum of squared distances to the others. The G◁□A base pair is shown at the bottom of the 3D structures. Superimposition of the cycles was made using three reference atoms and pseudo-atoms per nucleotide: N9 for purines and N1 for pyrimidines; a pseudo-atom at 1 Å of the N{1,9} atom in the direction of the C1′-N{1,9} vector, and another pseudo-atom at 1 Å of the N{1,9} in the direction of the normal to the base plane. These three atoms were selected to compare the relative positions of the bases of different sequences, independent of the backbone conformation. (**B**) Tetraloop sequence alignment, RMSD and cycle distances. The tetraloop sequences and their positions in the LSU are aligned according to G-N-R-A. Four instances do not conform to the GNRA sequence: #9, #11, #19 and #20, but adopt the GNRA conformation thank to isosteric U◁□A, C◁□A, U◁□C and C◁□C base pairs, respectively; rather than the usual G◁□A base pair. The symbols follow the nomenclature proposed by Leontis and Westhof (4). The RMSD and cycle distances were measured according to structure #1. (**C**) Isosteric sheared base pairs. The O atoms are shown using circles. Superimposition of 16 G◁□A sheared base pairs (upper-left) as compared to the five isosteric base pairs found in the GNRA tetraloops of the LSU: G◁^{b}□A (middle-left; from instance #21), U◁□A (upper-right), C◁□A (middle-right), U◁□C (lower-left), and C◁□C (lower-right). Note the conservation of the relative glycosyl bond orientations (shown in bold lines) and of the width covered by all isosteric sheared base pairs whether they are formed by purines, pyrimidines or of a purine and a pyrimidine. (**D**) Variant of the GNRA cycle. Six GNRA tetraloop cycles do not respect the loop structure by having R3 and A4 not adjacent in the sequence (not connected by phosphodiester linkages). Instances #7 and #9–11 have 1 nt inserted between the R and A of the GNRA; #13 has two; and #21 has 21 nt inserted. The instances of this variant are the only instances with cycle distances above 1.0 with instance #1, whereas no such selection criterion can be made using RMSD.



RMSD (Å)	Cycle distance	G	N	R	A
0.00	0.00	G-**0**1258	A-**0**1259	A-**0**1097	A-**0**1098
0.67	0.36	G-**0**2798	A-**0**2799	A-**0**2775	A-**0**2776
1.41	0.85	G-**0**1614	A-**0**1615	A-**0**1580	A-**0**1581
1.43	1.00	G-**0**1908	A-**0**1909	A-**0**1930	A-**0**1931
1.61	1.27	G-**0**1484	A-**0**1485	A-**0**1458	A-**0**1459
2.21	1.30	G-**0**531	A-**0**532	A-**0**11	U-**0**12
3.08	1.08	G-**0**112	A-**0**113	A-**0**48	A-**0**49

Figure 4. GNRA interior-loops from the LSU. (**A**) Stereo-view of seven superimposed instances (above) and of one 'typical' (#1) GAAA instance (below) of the GNRA interior-loop 3D structures. The typical instance was chosen as in Figure 3. The G◁□A base pair is shown at the bottom of the 3D structures. Superimposition of the cycles was made as indicated in Figure 3. Note the absence of a backbone between nucleotides N and R. (**B**) The interior-loop sequence alignment, RMSD and cycle distances. The interior-loop sequences and their positions in the LSU are aligned according to G-N-R-A. The sequence of all instances is GAAA, but #6 that contains a G◁□U base pair that is isosteric to the G◁□A sheared base pair of the other occurrences. RMSD are measured to structure #1. (**C**) Isosteric sheared base pairs. The O atoms are shown using circles. Superimposition of the six G◁□A sheared base pairs (above) and isosteric G◁□U base pair (below). Note the conservation of the relative glycosyl bond directions (shown in bold lines) and of the width covered by all isosteric sheared base pairs whether they are formed of purines or of a purine and a pyrimidine.

all instances are two-stranded, and are localized in interior loops instead of at the tip of a stem, as it is the case for the GNRA tetraloops. We refer to this variant of the GNRA as the GNRA interior-loop motif (L-S-LS-P).

The seven instances of this motif found in the LSU share a maximum RMSD near 3.1 Å (1.3 in the cycle distance). Like most of the 'regular' GNRA tetraloops, three instances of GNRA interior-loop bind to an adjacent stem by forming the tertiary S/S interaction involving the A of the GNRA (instances 1, 4 and 5). The S edge of the A of instance 2 is also involved in a tertiary interaction, but the H-bond with the adjacent stem is bifurcated. Several combinations of the N, R and A bases of all instances interact with the backbone of the adjacent stem. Finally, the N base of instance 6 interacts with the adjacent stem by a W/S interaction.

The GNRA interior-loop cycles share an average RMSD of ~3 Å with the instances of the GNRA cluster (data not shown). The structure of the GNRA interior-loop is almost identical to that of GNRA tetraloops, except for the N base that flips over, displacing the backbone on the other side of the structure and introducing most of the RMSD when compared to the 'regular' GNRA tetraloops.

Another observation about the GNRA interior-loop motif is the sequence conservation, as the sequence GA/AA appears in six out of the seven occurrences. In the outlying sequence, the sheared G◁□A base pair is substituted by an isosteric G◁□U base pair that is stabilized by the formation of one hydrogen bond (see Figure 4C).

CONCLUSION

We introduced a new RNA tertiary structural element, the cycle, which represents a formal step beyond the traditional base pair used as a first-class object in secondary structure. From the study of the LSU cycles, we learned that: (i) the backbone is not a determinant of RNA tertiary structure; (ii) small motifs are tolerant to nucleotide insertions and

(iii) the sequence-structure relationship is more complex than expected.

We were able to name and discuss these cycles in terms of base pairing and base stacking interactions, as the backbone was found to play a limited role in the cycle distance and RMSD clustering. This hypothesis is supported by finding cycles of similar fold and similar function, but involving different numbers of strands (cf. GNRA tetraloop versus GNRA interior-loops). We predict that many important motifs, not yet discovered, are independent of the sequence connectivity.

The observation of equivalent link-stack and stack interactions in instances of the GNRA tetraloop cluster revealed the ability of the cycles to 'eject' inserted nucleotides in favor of structure and function conservation. Such insertions could occur as well in the absence of interactions (cf. insertion between the G and N of the GNRA motif), creating larger cycles. However, our structural metric is currently limited to the comparison of cycles of the same size.

The 3D structures of the 4 nt cycles were extracted and classified using a nearest neighbor hierarchical clustering. By using a structural metric, the isosteric base pair phenomenon was captured, and the unusual instances were found, as they should be, among the members of the GNRA tetraloop and GNRA interior-loop clusters. The presence of these unusual instances and the structural variants of the GNRA cycle suggest the GNRA motif has more sequence and structure flexibility than was originally thought. In particular, the folding of the UCAC sequence in a structure very close to our representative GNRA tetraloop is indicative of a subtle sequence-structure relationship.

We will now exhaustively annotate and compute minimal cycle bases of other available high-resolution RNA structures. We will compare RNA tertiary structures in terms of the cycles in their minimal cycle bases. This approach should allow us to establish the recurrence of the cycles among different tertiary structures. We will also catalogue the RNA cycles, and employ them in a divide-and-conquer approach to RNA tertiary structure prediction and modeling.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR online.

ACKNOWLEDGEMENTS

The authors thank Eric Westhof, Neocles Leontis, Martin Larose and Patrick Gendron for useful discussions about RNA cycles and motifs. The authors thank Miklós Csürös, Philippe Thibault and Marc Parisien for reviewing and making suggestions about the manuscript. This work was supported by grants from the Canadian Institutes of Health Research (CIHR) (MT-14604) and from the Natural Sciences and Engineering Research Council (NSERC) of Canada (170165-01) to F.M. S.L. held a Ph.D. scholarship from the CIHR. F.M. is a member of the Centre Robert-Cedergren. Funding to pay the Open Access publication charges for this article was provided by NSERC.

Conflict of interest statement. None declared.

REFERENCES

1. Ban,N., Nissen,P., Hansen,J., Moore,P.B. and Steitz,T.A. (2000) The complete atomic structure of the large ribosomal subunit at 2.4 A resolution. *Science*, **289**, 905–920.
2. Wimberly,B.T., Brodersen,D.E., Clemons,W.M.Jr, Morgan-Warren,R.J., Carter,A.P., Vonrhein,C., Hartsch,T. and Ramakrishnan,V. (2000) Structure of the 30D ribosomal subunit. *Nature*, **407**, 327–339.
3. Carter,A.P., Celmons,W.M., Brodersen,D.E., Morgan-Warren,R.J., Wimberly,B.T. and Ramakrishnam,V. (2000) Functional insights from the structure of the 30S ribosomal subunit and its interactions with antibiotics. *Nature*, **407**, 340–348.
4. Leontis,N.B. and Westhof,E. (2001) Geometric nomenclature and classification of RNA base pairs. *RNA*, **7**, 499–512.
5. Gendron,P., Lemieux,S. and Major,F. (2001) Quantitative analysis of nucleic acid three-dimensional structures. *J. Mol. Biol*., **308**, 919–936.
6. Lemieux,S. and Major,F. (2002) RNA canonical and non-canonical base pairing types: a recognition metho and complete repertoire. *Nucleic Acids Res*., **30**, 4250–4263.
7. Yang,H., Jossinet,F., Leontis,N., Chen,L., Westbrook,J., Berman,H. and Westhof,E. (2003) Tools for the automatic identification and classification of RNA base pairs. *Nucleic Acids Res*., **31**, 3450–3460.
8. Harvey,S.C., Wang,C., Teletchea,S. and Lavery,R. (2003) Motifs in nucleic acids: molecular mechanisms restraints for base pairing and base stacking. *J. Comp. Chem*., **15**, 1–9.
9. Schneider,B., Moravek,Z. and Berman,H.M. (2004) RNA conformational classes. *Nucleic Acids Res*., **32**, 1666–1677.
10. Murray,L.J., Arendall,W.B.3rd, Richardson,D.C. and Richardson,J.S. (2003) RNA backbone rotamers-finding your way in seven dimensions. *Proc. Natl Acad. Sci. USA*, **100**, 13904–13909.
11. Gendron,P., Gautheret,D. and Major,F. (1998) Structural Ribonucleic Acid Motif Identification and classification. In Schaeffer,J. (ed.), *High Performance Computing Systems and Applications*. Kluwer Academic Press, Boston, MA, pp. 323–331.
12. Gan,H.H., Pasquali,S. and Schlick,T. (2003) Modular RNA architecture revealed by computational analysis of existing pseudoknots and ribosomal RNAs. *Nucleic Acids Res*., **31**, 2926–2943.
13. Huang,H.C., Nagaswamy,U. and Fox,G.E. (2005) The application of cluster analysis in the intercomparison of loop structures in RNA. *RNA*, **11**, 412–423.
14. Lescoute,A., Leontis,N.B., Massire,C. and Westhof,E. (2005) Recurrent structural RNA motifs, Isostericity Matrices and sequence alignments. *Nucleic Acids Res*., **33**, 2395–2409.
15. Leontis,N.B. and Westhof,E. (2003) Analysis of RNA motifs. *Curr. Opin. Struct. Biol*., **13**, 300–308.
16. Jaeger,L., Westhof,E. and Leontis,N.B. (2001) Tecto-RNA: modular assembly units for the construction of RNA nano-objects. *Nucleic Acids Res*., **29**, 455–463.
17. Lescoute,A. and Westhof,E. (2006) Topology of three-way junctions in folded RNAs. *RNA*, **12**, 83–93.
18. Horton,J.D. (1987) A polynomial-time algorithm to find the shortest cycle basis of a graph. *SIAM J. Comp*., **16**, 358–366.
19. Berman,H.M., Westbrook,J., Feng,Z., Gilliland,G., Bhat,T.N., Weissig,H., Shindyalov,I.N. and Bourne,P.E. (2000) The Protein Data Bank. *Nucleic Acids Res*., **28**, 235–242.
20. Jucker,F.M., Heus,H.A., Yip,P.F., Moors,E.H.M. and Pardi,A.A. (1996) A network of heterogeneous hydrogen bonds in GNRA tetraloops. *J. Mol. Biol*., **264**, 968–980.