Published online 5 October 2011 *Nucleic Acids Research, 2012, Vol. 40, No. 3* ***1307–1317***
doi:10.1093/nar/gkr804

Clustering RNA structural motifs in ribosomal RNAs using secondary structural alignment

Cuncong Zhong and Shaojie Zhang*

Department of Electrical Engineering and Computer Science, University of Central Florida, Orlando, FL 32816, USA

Received April 28, 2011; Revised September 10, 2011; Accepted September 12, 2011

ABSTRACT

RNA structural motifs are the building blocks of the complex RNA architecture. Identification of non-coding RNA structural motifs is a critical step towards understanding of their structures and functionalities. In this article, we present a clustering approach for *de novo* RNA structural motif identification. We applied our approach on a data set containing 5S, 16S and 23S rRNAs and rediscovered many known motifs including GNRA tetraloop, kink-turn, C-loop, sarcin–ricin, reverse kink-turn, hook-turn, E-loop and tandem-sheared motifs, with higher accuracy than the state-of-the-art clustering method. We also identified a number of potential novel instances of GNRA tetraloop, kink-turn, sarcin–ricin and tandem-sheared motifs. More importantly, several novel structural motif families have been revealed by our clustering analysis. We identified a highly asymmetric bulge loop motif that resembles the rope sling. We also found an internal loop motif that can significantly increase the twist of the helix. Finally, we discovered a subfamily of hexaloop motif, which has significantly different geometry comparing to the currently known hexaloop motif. Our discoveries presented in this article have largely increased current knowledge of RNA structural motifs.

INTRODUCTION

Thorough analysis of the three dimensional (3D) structures of non-coding RNAs (ncRNAs) is fundamental in understanding their versatile functionalities. A critical step for RNA structural analysis is identifying recurrent structural components, i.e. the RNA structural motifs, from the experimentally resolved structures. The recurrence of RNA structural motifs implies their high modularity and functional importance. Given the amazing speed in which RNA structures are being resolved, more efficient and accurate motif analysis tools are in urgent demand. As a result, developing computational tools for structural motif analysis is of great interest and can be expected to largely improve current ncRNA studies.

The RNA structural motifs are usually modeled by either their 3D geometries or their base pairing patterns. Traditionally, geometric discrepancy is used to assess the structural similarity between RNA structural motifs. While it is inefficient and overspecialized to consider all atoms, different tools usually select their own abstraction of the structural motifs. The RNA structural motif alignment or search tools such as NASSAM (1), PRIMOS (2), ARTS (3) and FR3D (4) can be classified into this category. Although these tools are extremely useful, recent study points out that the geometry-based approaches are restricted to motifs with rigid 3D structures, since local geometric deviations of the motif can be magnified when computing the global optimal superimposition (5). Therefore, base pairing information should also be considered when comparing RNA structural motifs. The recently developed RNA structural motif identification tool, RNAMotifScan, searches for structural fragments with similar base pairing pattern given by the query, and has uncovered new motif occurrences that have similar base pairing pattern but with relatively large geometric deviations (6).

The computational identification of RNA structural motifs can become much more difficult when there is no explicitly defined query. This problem, also referred as the *de novo* motif identification problem, is usually solved using clustering approaches that require no explicit query information. COMPADRES (7), a *de novo* clustering method developed based on PRIMOS (2), has successfully identified four new structural motif families from the resolved RNA 3D structures in Protein Data Bank (PDB) (8). However, the motif families identified by COMPADRES are mostly short motifs with rigid 3D topologies, while larger and more complicated motifs were not considered. In addition, the lack of conserved base interaction pattern for the newly identified motifs

*To whom correspondence should be addressed. Tel: +1 407 823 6095; Fax: +1 407 823 5835; Email: shzhang@eecs.ucf.edu

© The Author(s) 2011. Published by Oxford University Press.
This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/3.0), which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

makes further modeling, search and functional inference of these motifs rather difficult (9). As a result, base pairing patterns should also be considered in *de novo* structural motif identification.

Recently, Djelloul and Denise (9) have devised a clustering approach that purely considers base pairing pattern for *de novo* RNA structural motif identification. [In this article, we refer this method as the LENCS (Longest Extensible Non-Canonical Substructure) method.] They transformed each candidate structural motif instance into a base pairing graph, and applied graph isomorphism algorithm to identify maximum common subgraphs. After pairwise comparison, the structural fragments were organized using hierarchical clustering, and potential motif clusters were extracted by applying a universal cutoff. Although LENCS has successfully rediscovered many known motifs and suggested potential novel motifs, the graph isomorphism restriction makes it impossible to consider RNA structural motifs with base pair variations. Besides, the LENCS method completely ignored the sequences of the motifs, hence difficult to correctly incorporate base pair isostericity information (10).

We have developed RNAMotifScan to account for these problems (6), and expect to develop a more accurate clustering framework by incorporating RNAMotifScan. In addition, we also try to tackle three important issues in RNA structural motif clustering. First, it is well known that the annotation tools may make mistakes in base pair prediction due to inadequate resolution (6). Although this may not be an issue in model-based search application (as the query model is hand-curated and thus can represent the complete base pairing pattern), it can significantly affect clustering analysis since the erroneous base pair predictions may happen in both motif instances that are being compared. Second, the LENCS method only considers the fraction of matched base pairs between motif instances, but does not distinguish the importance of the matching. For example, the *trans* H/SE pair can be found in many motifs such as kink-turn, sarcin–ricin and tandem-sheared motifs, while the *cis* H/SE pair is much less frequent. In this case, matching *cis* H/SE pairs should be more informative than matching *trans* H/SE pairs. Finally, the hierarchical clustering approach applied by the LENCS method is not suitable for large-sized data sets, since it would be difficult to manually examine the huge hierarchical tree to determine the optimal cutting level.

To account for the first issue, we combined the base pair predictions made by two popular annotation tools: RNAVIEW (11) and MC-Annotate (12). In this way, we were likely to include all true base pairing interactions into the compiled candidate motif instances. RNAMotifScan is then responsible for identifying the optimal matching between these predictions and discarding additional base pairs with moderate penalty. To solve the second issue, we developed a statistical inference framework that can be used to measure the significance of the matchings. Each candidate motif instance was aligned to a set of artificial motif instances that simulate random structural segments from ribosomal RNAs. Consider the example in the previous paragraph, although we do not distinguish the alignment score between matching *trans* and *cis* H/SE pairs, we can expect lower *P*-value assigned to the matching of *cis* H/SE pairs. This is because *cis* H/SE pairs are much less frequently found, resulting in lower background alignment scores associated with the motif instances that contain this base pair and, therefore, more significant *P*-values for a match. Finally, to make the clustering analysis extensible to large-size data sets, we applied the Cluster Affinity Search Technique (CAST)-like (13) clique finding algorithm that can automatically generate individual clusters given only a universal *P*-value cutoff.

We applied our new clustering framework on two data sets (one for hairpin loop instances and the other for internal loop, bulge loop and junction loop instances, see 'Materials and Methods' section) that contain 5S (*Haloarcula marismortui*, PDBid: 1S72, chain '9'), 16S (*Thermus thermophilus*, PDBid: 1J5E, chain A) and 23S (*Haloarcula marismortui*, PDBid: 1S72, chain '0') ribosomal RNAs. We have identified totally 44 clusters (8 from the hairpin loop data set and 36 from the internal loop data set). These clusters define many known RNA structural motifs such as GNRA tetraloop (14), kink-turn (15), C-loop (16–19), sarcin–ricin (20–23), reverse kink-turn (24), hook-turn (25), E-loop (26,27) and tandem-sheared (28) motifs. The performance of our clustering framework shows significant improvement over the LENCS method. Specifically, the F-measure has been increased from 69.1% to 82.6%. Besides, we also identified several new occurrences of these known motifs. Finally, we also present three clusters corresponding to novel motif families that have not been characterized before. All clusters are sorted based on average *P*-values that indicate the in-cluster structural similarities and are available in the Supplementary Data.

MATERIALS AND METHODS

Data preparation

The resolved ribosomal RNA subunit structures (1S72 and 1J5E) were downloaded from PDB (8). The base pairs were annotated by RNAVIEW (11) and MC-Annotate (12). We combined (union) the annotations from both tools to generate the final annotation. The conflict predictions (different edge or orientation annotations for the same base pair) were resolved by taking the annotations from MC-Annotate. All non-canonical base pairs were temporarily discarded to reveal the general sketch of the A-form helices in the structures. Pseudoknots were then removed using K2N web server (29). Lone pairs were further removed to avoid accidental destruction of potential motifs. Finally, regions corresponding to hairpin loops, internal loops, bulge loops or junction loops (30) were identified from the resulting nested structures and all base pairs within these regions were recovered to construct candidate motif instances [similar to LENCS (9)]. The candidate instances that contain no non-canonical base pair were removed.

Candidate motif instances from 5S, 16S and 23S rRNAs were compiled into two data sets, one for hairpin loops and the other for internal loops, bulge loops and junction

loops (we will call this data set internal loop data set for short). Since sequence conservation in hairpin loop motifs is also very important in defining their functionalities, higher sequence weight should be applied for this data set. The hairpin loop data set contains 33 candidate instances and the internal loop data set contains 157 candidate instances. To account for different concatenation orders the strands (6), the symmetric counterpart of each motif instance in internal loop data set is also included.

Aligning structural components using RNAMotifScan

We applied RNAMotifScan (6) to measure the structural similarity between two candidate motif instances. RNAMotifScan matches two motif instances by a dynamic programming approach which takes into account base pair isostericity. For the internal loop data set, the sequence weight was set to 0.2 and the structure weight was set to 0.8. while for the hairpin loop data set, we raised the sequence weight to 0.4 and lowered the structure weight to 0.6. Because the hairpin loop motifs are usually defined by their lengths (e.g. tetraloop and hexaloop), we also doubled the default gap penalty for hairpin loop clustering. Other parameters were set to default.

Generating random structural motif instances

Given a candidate instance, we aim at generating a number of random motif instances that have similar length (allowing $\pm 20\%$ fluctuation) with the candidate instance and base pairing pattern with the ribosomal RNAs background. Our statistics indicate that in ribosomal RNAs, the base pair ratio (the number of canonical and non-canonical base pairs over the length of the sequence) is $\sim$50% (specifically, 51.7% for 5S rRNA, 50.0% for 16S rRNA and 50.2% for 23S rRNA). Among these base pairs, $\sim$15% of them correspond to non-nested base pairs (specifically, 15.5% for 5S rRNA, 14.6% for 16S rRNA and 16.9% for 23S rRNA), while the others form nested base pairs. (This statistic is solely based on MC-Annotate predicted base pairs.)

Since random sampling of existing structural segments from database may not result in enough randomness and sometimes introduce bias (31), we developed the following method to generate random motif instances. Given the base pair distribution for the ribosomal RNAs and assume the length of the random motif instance is n (predetermined based on the length of the candidate instance), we first build a perfectly stacked helix with $85\%*n/2$ base pairs (with the same base pair frequency as the background). Then we randomly insert $15\%*n$ unpaired nucleotides into the helix (with the same nucleotide frequency as the background). Finally, we add $15\%*n/2$ non-nested base pairs (also with the same base-pair frequency as the background) by randomly selecting two nucleotides from the constructed motif instance.

Extracting significant clusters

Upon the finishing of all-against-all pairwise alignments, a P-value was assigned for each alignment score. Alignment score distribution regarding each candidate instance was simulated by aligning it to a number of random instances generated using the method described above. The P-values were computed using optimal fitting that assumed general extreme value distribution (with MATLAB built-in function 'gevfit'). Since each alignment score is associated with two P-values (that are computed from both candidate instances' background score distributions), the higher P-value was assigned to ensure specificity.

After the computation of P-values, the all-against-all alignment scores were summarized into a graph, where the nodes represent the candidate motif instances and the edges indicate pairwise structural similarities (denoted by P-values). We extracted all strongly connected subgraphs by applying a CAST-like clique finding algorithm (13). The P-value cutoff was set to $10^{-3.5}$ (empirically determined) for both hairpin loop data set and internal loop data set.

RESULTS

We have identified 8 clusters from the hairpin loop data set and 36 clusters from the internal loop data set. (If two clusters are completely symmetric due to the inclusion of both strand orientations, only one of them is retained.) The clusters are sorted by their average P-values. To describe the results more clearly, we represent each cluster with a label of the data set ('CH' for the hairpin loop data set and 'CL' for the internal loop data set) followed by its ranking. For example, the kink-turn cluster, CL15, indicates that it was identified from the internal loop data set and ranked 15th by its average P-value. All naming and representation of base pairs follow the fashion proposed by Leontis and Westhof (32). The 3D structure figures were prepared using PyMol (http://www.pymol.org).

In this section, we will first discuss the clustering results regarding currently known motifs and present discovery of their new instances. We will then show three potential novel motif families revealed by our clustering analysis. Due to the limitation of space, many meaningful clusters were not discussed in this section. For instance, cluster CH2 represents the UUCG tetraloop motif (33), and cluster CL3 represents an extremely complex base pairing pattern where four base pairs are formed within only 4 nt. We anticipate that these clusters can also provide useful information for RNA structural motif studies. All clusters and detailed locations for each structural fragment can be found in Supplementary Table S1 (for internal loop motif instances) and Supplementary Table S2 (for hairpin loop motif instances).

Clustering of known motifs and their new instances

We have identified several clusters that correspond to known motifs including GNRA tetraloop, kink-turn, C-loop, sarcin–ricin, reversed kink-turn, hook-turn, E-loop and tandem sheared motifs. The clustering results of these known motifs and corresponding results generated by LENCS method are summarized in Table 1. Our clustering method, RNAMSC (RNAMotifScan based

Table 1. Comparison between two base pairing pattern based clustering methods: RNAMSC and LENCS

Motif	Cluster ID	Novel[a] instances	RNAMSC			LENCS		
			Sensitivity[b] (%)	Specificity[c] (%)	F-measure[d] (%)	Sensitivity (%)	Specificity (%)	F-measure (%)
GNGA Tetraloop	CH1	0	72.7 (8/11)	100 (8/8)	**84.2**	–	–	–
GNAA Tetraloop	CH3	1	63.6 (14/22)	93.3 (14/15)	**75.7**	–	–	–
Kink-turn	CL15	0	50.0 (5/10)	100 (5/5)	**66.7**	20.0 (2/10)	100 (2/2)	33.3
C-loop	CL24	0	75.0 (3/4)	100 (3/3)	**85.7**	50.0 (2/4)	100 (2/2)	66.7
Sarcin–ricin	CL13	3	100 (12/12)	100 (12/12)	**100**	66.7 (8/12)	100 (8/8)	80.0
Reverse Kink-turn	CL18	0	100 (3/3)	100 (3/3)	**100**	100 (3/3)	42.8 (3/7)	59.9
Hook-turn	CL17	0	66.7 (2/3)	100 (2/2)	**80.2**	100 (3/3)	60.0 (3/5)	75.0
E-loop	CL19	0	100 (4/4)	66.7 (4/6)	**80.0**	100 (4/4)	57.1 (4/7)	72.7
Tandem-sheared	CL23	1	33.3 (2/6)	100 (2/2)	49.6	100 (6/6)	75.0 (6/8)	**85.7**
Average performance[e]			73.8 (31/42)	93.9 (31/33)	**82.6**	66.7 (28/42)	71.8 (28/39)	69.1

[a]The novel instances are discussed in detail in corresponding sections. These instances are not counted for performance assessment.
[b]Expression in parenthesis corresponds to number of true positive over all known instances (see Supplementary Table S3 for the known instances).
[c]Expression in parenthesis corresponds to number of true positive over cluster size.
[d]F-measure = 2 * Sensitivity * Specificity / (Sensitivity + Specificity). The higher performances are in bold format.
[e]The average performance assessment does not include GNGA and GNAA tetraloop, since they were not identified by LENCS method.

Clustering), shows generally higher performance comparing to the LENCS method. The clustering results for these known motif families will be discussed separately below.

GNRA tetraloop. The GNRA tetraloop is an RNA structural motif in the hairpin loop region featured by its consensus sequence. The motif is found to interact with proteins (34) or other RNA structural elements (35,36). FR3D identified 21 GNRA tetraloop motif instances from 1S72 23S rRNA and 12 from 1J5E 16S rRNA. Our clustering method separates the GNRA tetraloop into two clusters: CH1 and CH3. The cluster CH1 contains tetraloops with consensus sequence 'GNGA' and the cluster CH3 contains tetraloops with consensus sequence 'GNAA'. The separation of the GNRA tetraloop motif is due to the strict universal *P*-value cutoff applied. The clustering performances of the two sets of GNRA tetraloop motif are summarized in Table 1. One potential novel GNAA tetraloop instance has been identified in cluster CH3. This novel instance and a well-established GNRA tetraloop instance are shown in Figure 1. The base pairing patterns and 3D geometries of these two instances are very similar.

Several GNRA instances were missed due to two major reasons: unusual base pair replacement and nucleotide insertion. For example, the GNRA tetraloop instance 1S72, chain '0', 1326-1331 was missed due to the fact that the G1327-A1330 sheared pair is replaced by *trans* W/H pair, while the instances 1S72, chain '0', 1706-1712 and 1J5E, chain A, 691-696 were missed because the closing canonical pair is replaced by sheared pairs. Furthermore, the instance 1J5E, chain A, 726-731 was missed due to the deletion of base pair G727-A729. The GNRA tetraloop instances 1S72, chain '0', 481-487, 493-499, 1054-1060, 1275-1281, 1468-1474 and 1793-1799 were missed due to one nucleotide insertion within the hairpin loop. The other missed instances, 1J5E, chain A, 1030A-1030D, was not included into the candidate set for its irregular nucleotide indexing.

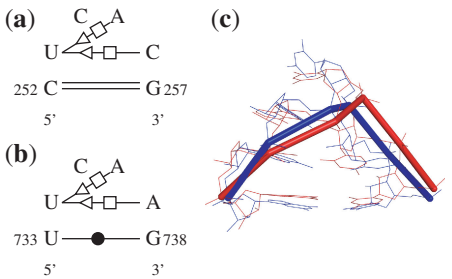

Figure 1. The base pairing patterns and superimposition of two GNRA tetraloop motif instances clustered in CH3. (**a**) A known GNRA tetraloop instance in 1S72, chain '0', 252-257. (**b**) The novel GNRA tetraloop instance in 1S72, chain '0', 733-738. (**c**) The superimposition between these two motif instances [red: (a); blue: (b)].

Kink-turn. The kink-turn motif is an asymmetric internal loop characterized by the 'kink' observed in its longer strand which causes a sharp turn between its two supporting helices (30,37). It is known to be an important recognition site for interaction with proteins or other RNA elements (15,38). We have identified four out of nine known kink-turn instances in 1S72 23S rRNA and the known instance in 1J5E 16S rRNA in cluster CL15 with no false positive prediction (Table 1). Base pair variations are frequently observed in kink-turn motif instances, making the sensitivity of both base pairing pattern based clustering methods relatively low. Therefore, some potential novel kink-turn instances can also be found in other clusters besides cluster CL15, as we will describe in details below.

One potential novel kink-turn motif instance is clustered with a known kink-turn motif instance in CL7. The highly conserved bulged nucleotides that correspond to the 'kink' can be found at U1149-A1150 in Figure 2a and A279-C280 in Figure 2b. Interestingly, two nucleotides (U244, C245) are inserted in the novel instance, which induces an 'S' shaped bend at the opposite strand of the 'kink' (see Figure 2b). The insertion has altered

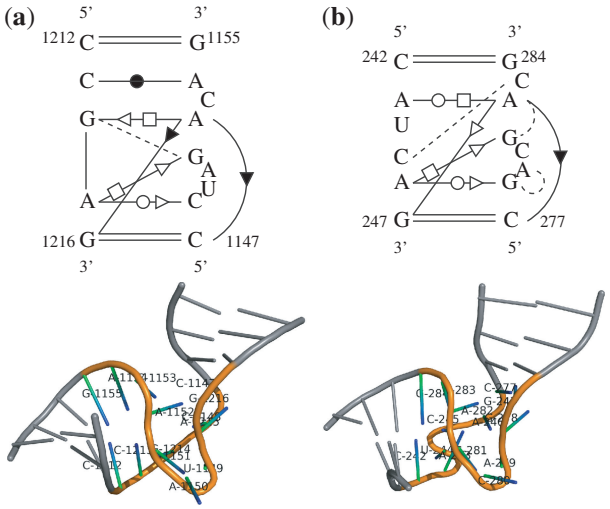

Figure 2. The base pairing patterns and structures of the two kink-turn motif instances clustered in CL7. (**a**) A known kink-turn instance found in 1S72, chain '0', 1147-1155/1212-1216. (**b**) The potential novel kink-turn instance found in 1J5E, chain A, 242-247/277-284. The dashed edges in the base pairing patterns (both in this figure and in the remaining figures of this article) correspond to additional base pairs annotated but not included into the consensus structure. The regions that are not part of the motif are colored gray (both in this figure and in the remaining figures of this article).

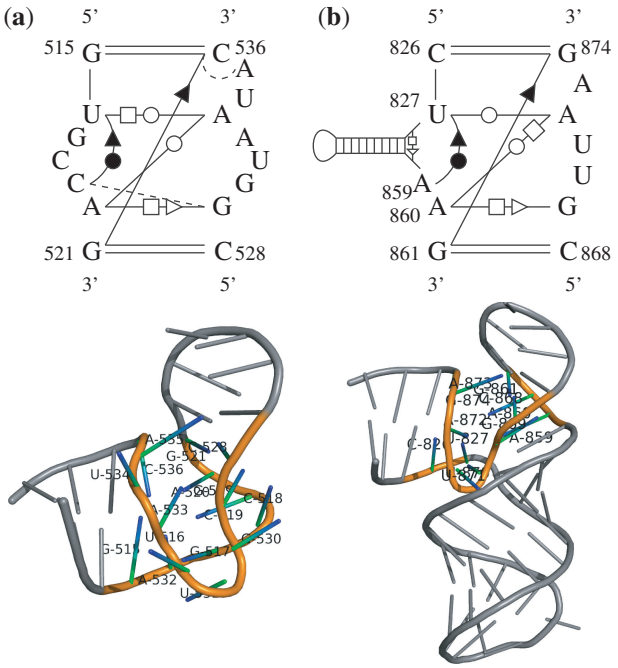

Figure 3. The base pairing patterns and structures of the two kink-turn motif instances clustered in CL6. (**a**) A novel kink-turn instance found in 1J5E, chain A, 515-521/528-536. (**b**) A novel kink-turn instance found in 1J5E, chain A, 826-861/868-874.

both base pairing pattern and geometry of the instance with unknown corresponding biological impact. However, we can still categorize this instance as kink-turn motif based on its base pairing and geometric similarity with the known kink-turn instance.

Another kink-turn cluster, CL6, contains two potential novel kink-turn instances. The base pairing patterns and 3D geometries of both instances are very similar to known kink-turn instances. However, both instances contain two pairs of cross-strand base-triples (Figure 3). These base-triples form two 'Z' shaped interactions (G515-C536-G521-C528, U516-A533-A520-G529 in Figure 3a and C826-G874-G861-C868, U827-A872-A860-G869 in Figure 3b). Unlike regular kink-turn instances, the two pairs of cross-strand base-triples extrude two bulge regions, one at each strand. In the first instance, G517-C519 are also bulged out in addition to G530-A532 that corresponds to the 'kink', making a much more severe turn at the companion strand comparing to regular kink-turn instances (Figure 3a). More interestingly, in the second instance, an A-form helix of 10 canonical base pairs is inserted at this region and interrupts the kink-turn instance (Figure 3b). These two motif instances reveal a potential new form of kink-turn motif where two bulges are extruded. It is also interesting to study the impact of the insertions on the binding activity of kink-turn motif.

C-loop. The C-loop motif is an asymmetric internal loop characterized by the base triple induced from the cytosine residue (37). We clustered two out of three known C-loop motif instances in 1S72 23S rRNA and the only known C-loop motif in 1J5E 16S rRNA in cluster CL24 (Table 1). We missed one known C-loop motif instance in 1S72, chain '0', 958-963/1005-1008 because of two nucleotide insertions, one at each strand (G960 and A1006). Also, four additional base pairs are annotated in this instance, which indicates unusual properties of this C-loop motif instance.

Sarcin–ricin. The sarcin–ricin motif (or sometimes referred as the G-bulge motif) is an asymmetric internal loop that is known to be involved in the interaction between the ribosomal RNA and elongation factors (23). There are 10 known sarcin–ricin motif instances in 1S72 (nine in 23S and one in 5S rRNA) and two in 1J5E. We have successfully clustered all 12 known sarcin–ricin instances in cluster CL13, while the LENCS method only clustered 8 of them (six in 1S72 and two in 1J5E, see Table 1). Three potential novel instances are also included in cluster CL13, which are presented in Figure 4.

Figure 4a shows a well-established sarcin–ricin motif instance in CL13. In its base pairing pattern, we can observe that the characterized bulged G1370 is interacting with its consecutive nucleotide U1371 using *cis* SE/H pair, followed by two non-canonical base pairs: *trans* W/H U1371-A2054 and *trans* W/SE A1372-G2053. These three base pairs have been used to characterize the sarcin–ricin motif (39–41). Figure 4b shows the first potential novel sarcin–ricin instance found in cluster CL13. This potential instance shows base pair variations in the two pairs before the bulged G (*cis* W/W C483-G450 and *cis* W/H G484-C459) comparing to the known instance. However, it is conserved for the three characteristic base pairs. The 3D geometry of this potential instance also

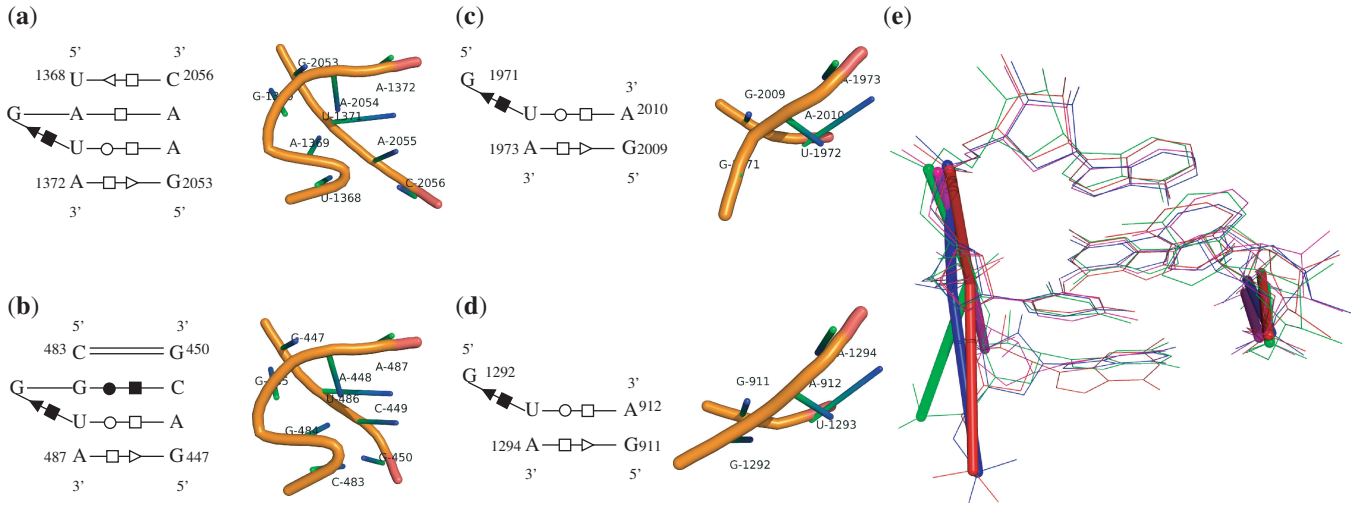

Figure 4. The base pairing patterns, structures and superimposition of the three base pairs formed near the bulged 'G' of four sarcin–ricin motif instances clustered in CL13. (**a**) A known sarcin–ricin instance found in 1S72, chain '0', 1368-1372/2053-2056. Three novel sarcin–ricin instances: (**b**) 1J5E, chain A, 483-487/447-450, (**c**) 1S72, chain '0', 1971-1974/2009-2010 and (**d**) 1S72, chain '0', 1251-1254/911-912. (**e**) The superimposition of three base pairs that characterize the sarcin–ricin motif in these four motif instances [red: (a); blue: (b); green: (c); magenta: (d)].

shows high similarity comparing to the known sarcin–ricin motif instance, where an 'S' shape turn can be observed.

The two potential sarcin–ricin motif instances, shown in Figure 4c and d, were identified from the junction loop regions instead of internal loop regions (where sarcin–ricin motif instances are usually found). It is worth noting that some known sarcin–ricin motif instances can also be found in the junction loop regions (e.g. the known sarcin–ricin motif instance at 1S72, chain '0', 380-384/405-408). These two potential sarcin–ricin instances are conserved in the three characteristic base pairs but without the other two base pairs. The absence of the other two base pairs makes the two instances smaller than regular sarcin–ricin motif instances and results in large geometric variations (i.e. the 'S' shape turn cannot be observed for these two instances). However, the local geometries associated with the three characteristic base pairs are still highly conserved in these two motif instances (Figure 4e), suggesting potential functional similarity between these two motif instances and regular sarcin–ricin motif instances. Nevertheless, the specific functions of these potential motif instances still need to be experimentally investigated.

Reverse kink-turn. The reverse kink-turn motif is also an asymmetric internal loop that produces a turn between two supporting helices such as kink-turn motif but towards the opposite direction (42). There are three known reverse kink-turn motif instances in 1S72. We have clustered all three known instances in cluster CL18 with no false positive predictions (Table 1). The LENCS method has also clustered these three known reverse kink-turn instances, however, with four unrelated instances. The reason for the false positive predictions is that the LENCS method does not consider nucleotide when determining base pair isostericity. For example, a false prediction made by LENCS in 1S72, chain '0', 2307-2310/2298-2300 contains a *trans* H/SE U2308-G2299 base pair. This base pair is matched to the *trans* H/SE A-C or A-G pair in the true reverse kink-turn instances. Although these base pairs have the same orientation and interacting edges, *trans* H/SE U-G pair is not isosteric with *trans* H/SE A-C or A-G pair. In our clustering framework, strict definition of base-pair isostericity is applied to avoid such unexpected false predictions.

Interestingly, two of the known reverse kink-turn instances (1S72, chain '0' 1527-1529/1662-1664 and 1531-1533/1658-1660) appear to be located close to each other, and manual inspection of the region suggests an instance of tandem reverse kink-turn (Figure 5). As there are only three known reverse kink-turn instances in the entire 23S rRNA, the chance of finding a tandem case is extremely low. Therefore, the tandem reverse kink-turn is likely to be required for certain biological functions. On the other hand, we investigated the other known reverse kink-turn instance (1S72, chain '0' 1132-1134/1228-1230) but did not found a tandem counterpart, which implies different functional roles played by single and tandem reverse kink-turn motif instances.

Hook-turn. The hook-turn motif is found at regular A-form helix regions, where one of the nucleotide chain sharply folds back toward the opposite direction (25). We identified two out of three known hook-turn motif instances in 1S72 23S rRNA with no false prediction (see Table 1). The LENCS method identified all three known hook-turn instances but include two unrelated motif instances. Figure 6 shows the two known hook-turn motif instances clustered in CL17, where conserved base-triples can be observed in both instances (G2267-C2243-A2244 and G2810-G2674-A2675). These two base-triples are both annotated solely by MC-Annotate, which indicates that these base-triples are likely to be real instead of being artifacts of combining RNAVIEW and MC-Annotate annotations (see Figure 6c for their superimposition). However, RNAVIEW does not predict these base-triples, making the LENCS method (which solely considers

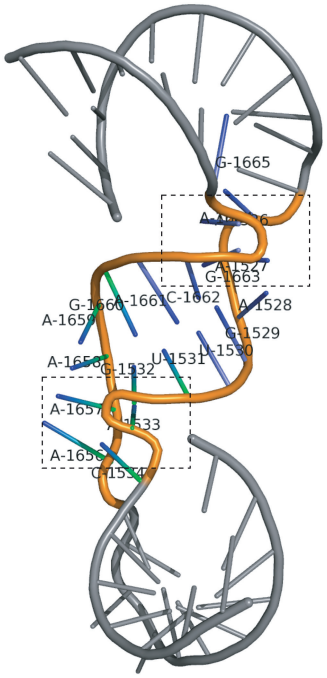

Figure 5. The tandem reverse kink-turn motif instance found in 1S72, chain '0', 1515-1540/1645-1670. The two reverse kink-turn instances are colored. The 'kink' regions are indicated by the two boxes.

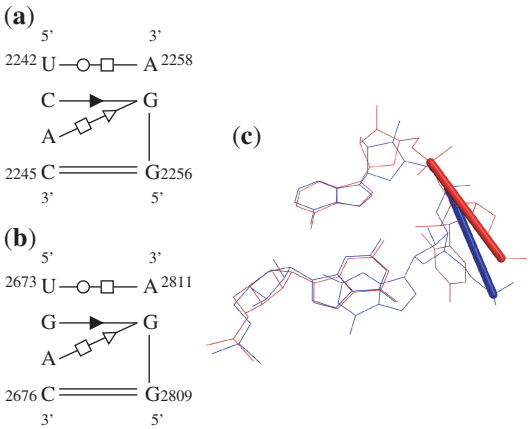

Figure 6. The base pairing patterns and superimposition of the base-triple interactions of the two known hook-turn instances identified in cluster CL17. (**a**) 1S72, chain '0', 2242-2245/2256-2258. (**b**) 1S72, chain '0', 2673-2676/2809-2811. (**c**) The superimposition of the base triples in these two motif instances shown in (a) and (b) [red: (a); blue: (b)].

RNAVIEW annotations) include the two unrelated motif instances. This base-triple is not predicted by either RNAVIEW or MC-Annotate in the other known hook-turn motif instance, 1S72, chain '0', 1457-1460/1483-1485, hence it was missed by our clustering method.

E-loop. The E-loop motif is a symmetric internal loop that contains the following base pairs: a *trans* H/SE base pair, a *trans* W/H or *trans* SE/H base pair, and a *cis* bifurcated or *trans* SE/H base pair as summarized by Leontis *et al.* (43). We notice that there are confusions in distinguishing E-loop and sarcin–ricin motifs since they share similar base pairing pattern (i.e. the three base pairs that define the E-loop motif). Another reason can be that bacterial 5S rRNA contains an E-loop motif while the corresponding region in *H. marismortui* appears to be sarcin–ricin motif. In this article, we consider an instance without the bulged G (and the base pair formed with its consecutive nucleotide) to be E-loop motif and otherwise sarcin–ricin motif.

Using this criterion, there are two E-loop motif instances in 1S72 23S rRNA and two in 1J5E 16S rRNA. We clustered all four instances in cluster CL19, with two false positive predictions that appear to be tandem-sheared motif instances (Table 1). The LENCS method has also successfully identified all four instances, but include three other unrelated motif instances, where one of them appears to be a sarcin–ricin motif instance (1J5E, chain A, 446-450/483-488) and the other two are kink-turn motif instances (1S72, chain '0', 241-244/267-270 and 1J5E, chain A, 683-687/703-707). The inclusion of false positive prediction by both methods even under strict *P*-value cutoff and graph isomorphism indicates that the universal cutoff which can optimize the overall clustering performance may not be strict enough for E-loop motif.

Tandem-sheared. The tandem-sheared motif consists of two consecutive sheared base pairs and is frequently observed in regular helix regions (28). There are four known tandem-sheared motif instances in 1S72 23S rRNA and two in 1J5E 16S rRNA. The LENCS method has identified all six known tandem-sheared motif instances but included two kink-turn motif instances. We identified two out of six known instances but with no false positive prediction in cluster CL23 (Table 1). The tandem-sheared instances identified by us are strictly closed by canonical base pairs at both ends, while the other missed instances are surrounded by additional non-canonical base pairs. We have also identified a potential novel tandem-sheared motif instance (also strictly closed by canonical base pairs) in cluster CL23. The base pairing patterns and structures of a known tandem-sheared motif instance and the potential novel instance are shown in Figure 7. The colored backbone regions correspond to the tandem-sheared base pairs, and slight inward turns can be observed in these regions from both instances.

Novel RNA structural motif families

The 'rope sling' motif. We have discovered a highly asymmetric bulge loop motif family that resembles the rope sling. The corresponding motif cluster (CL1), which has the lowest average *P*-value, consists of two motif instances: one from 1S72 23S rRNA and the other from 1J5E 16S rRNA. The base pairing patterns and structures of these two motif instances are shown in Figure 8. Both motif instances consist of two highly asymmetric strands, where the longer ones have 7–8 nt while the shorter ones have only two nucleotides. The first and last nucleotides of the longer strands form canonical base pairs with the two

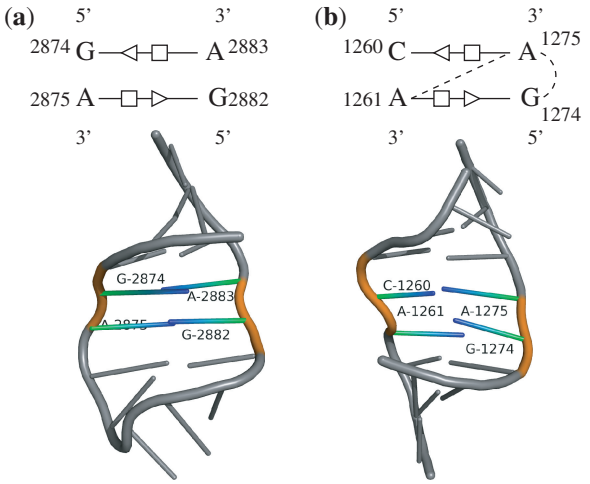

Figure 7. The base pairing patterns and structures of two tandem-sheared instances identified in cluster CL23. (**A**) A known tandem-sheared instance found in 1S72, chain '0', 2874-2875/2882-2883. (**b**) The novel tandem-sheared instance found in 1J5E, chain A, 1260-1261/1274-1275.

nucleotides in the shorter strands, leaving the other nucleotides in the longer strands bulged out from the main helix and resulting in a loop similar to rope sling (see Figure 8). Two consecutive nucleotides (C1105-A1106 and A572-A573) within the bulged chains form *cis* SE/H non-canonical interactions.

Several evidences indicate that the functionalities of the rope sling motif are carried out by its longer strand. First, a non-canonical *cis* SE/H base pair can be observed in the longer stand of each motif instance (C1105-A1106 and A572-A573). The nucleotide replacement (C1105 to A572) in this base pair is compensated by the isostericity between *cis* SE/H C-A and *cis* SE/H A-A base pairs. Second, two nucleotides in the longer strand of both motif instances also participate in non-nested canonical interactions (C1102-G1241 and C1103-G1240 in the first motif instance and G570-C866 and U571-A865 in the second motif instance, which are not shown in this figure). These conserved non-nested interactions also indicate the structural importance of these regions. Finally, high geometric similarity of the longer strands can also be observed from the superimposition between these two motif instances (Figure 8e). Therefore, we conjecture that the longer strands may determine the functionalities of the rope sling motif. Using RNAMotifScan, we also identified this motif from both 16S and 23S rRNA in *H. marismortui*, *T. thermophilus* and *E. coli*. The recurrence of this motif further indicates its structural or functional importance for ribosomal RNAs.

Motif that increases the twist at the helical region. Two internal loop motif instances, both closed by an A-U and a C-G canonical base pairs, were clustered in CL2. The conserved non-canonical base pairs between the two motif instances are the *cis* W/SE pairs formed at C1383-A935 and C36-A47 (see Figure 9a and b). The sequences are highly conserved at the left strand, where only one nucleotide substitution at the unpaired region can be observed (U1381 mutated to A34). On the other hand, a two-nucleotide deletion (between G933 and C934) is found at the right strand in the first motif instance. The nucleotide deletion alters the interaction between G933 and the left strand, violating the *trans* SE/H U33-G43 pair that can be observed in the second motif instance. The *trans* SE/W G43-C46 pair cannot be formed either, since G933 and C934 are too close to each other.

Superimposition of the two motif instances clearly reveals high structural similarity between the left strands (see Figure 9e), where two nucleotides (U1380, C1383 in the first motif and U33, C36 in the second motif) participate in the conserved base triple. The base triple indicates that the two nucleotides in the left strand are spatially close to each other. As a result, the left strand is likely to exhibit an unusual backbone conformation, such as a tight bend that can bring these two nucleotides together. Visualization of the local structures around the motif instances clearly shows increased twists at the corresponding regions (Figure 9c and d). The two strands of the motif instances are nearly parallel to each other and form planes that are perpendicular to the main helical axes, suggesting rather acute twists induced by this motif.

The functionalities of this motif family remain unclear without further experimental investigations. However, some evidences suggest potential binding activity of the motif. The twists deepen the groove where the potentially bound biomolecules can reside. At the same time, they also narrow down the helix, which can tightly clip the biomolecules that would have been embedded. Moreover, both motif instances are located at the surfaces of the ribosomal RNAs, which further suggests binding potentials.

New subfamily of hexaloop motif. We have identified two clusters that correspond to the hexaloop motif (CH6 and CH8). Cluster CH6 contains two hexaloop instances from 1S72 23S rRNA, both of which share the common base pairing pattern that two *trans* SE/H base pairs stack together. The nucleotide U1198 in the first motif instance is also annotated to be pairing with A1200, while this base pair is absent in the second motif instance (Figure 10a and b). This base pairing variation results in the geometric difference that A1199 in the first motif and A1919 in the second motif are extruded toward different directions (Figure 10c). Other than this difference, the backbones and the rest of nucleotides can be well superimposed, indicating true motif recurrence.

Cluster CH8 contains one motif instance from 1J5E 16S rRNA and one from 1S72 23S rRNA, both of which share the base pairing pattern that *trans* SE/H G-A pair (G314-A316 and G1316-A1318) stacks on *trans* W/H U-A pair (U313-A317 and U1315-A1319). The second motif instance contains two inserted cytosine residues between C1320 and G1323, which likely destruct the *trans* SE/H A-A pair (A317-A319) that can be observed in the first motif instance (Figure 10d and e). However, superimposition between the two motif instances reveals that the nucleotide insertions are well accommodated (see Figure 10f). Therefore, although the insertion increases

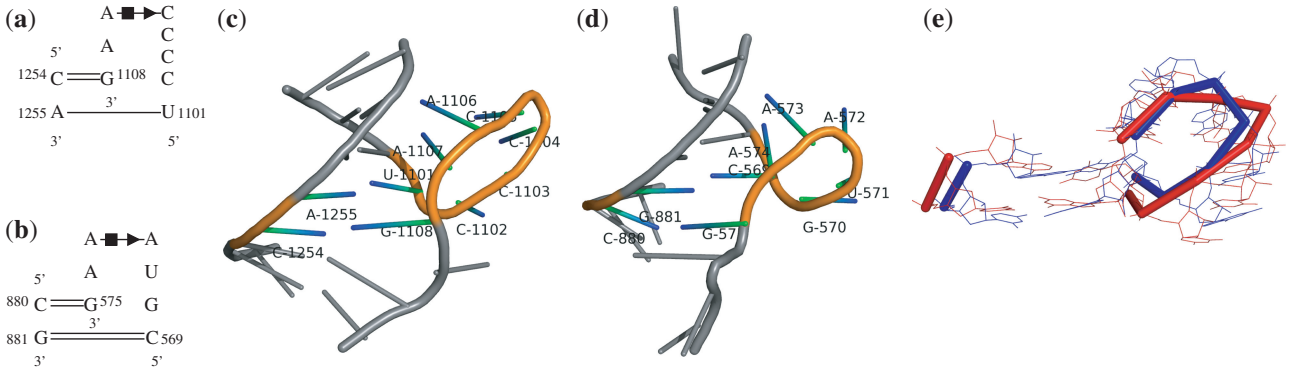

Figure 8. Potential novel motif family that resembles the rope sling. (**a**) and (**b**) The base pairing patterns of structural components found in 1S72, chain '0', 1254-1255/1101-1108 and 1J5E, chain A, 880-881/569-575, respectively. (**c**) and (**d**) Local structures around motif instances shown in (a) and (b), respectively. (**e**) The superimposition between these two motif instances [red: (a); blue: (b)].

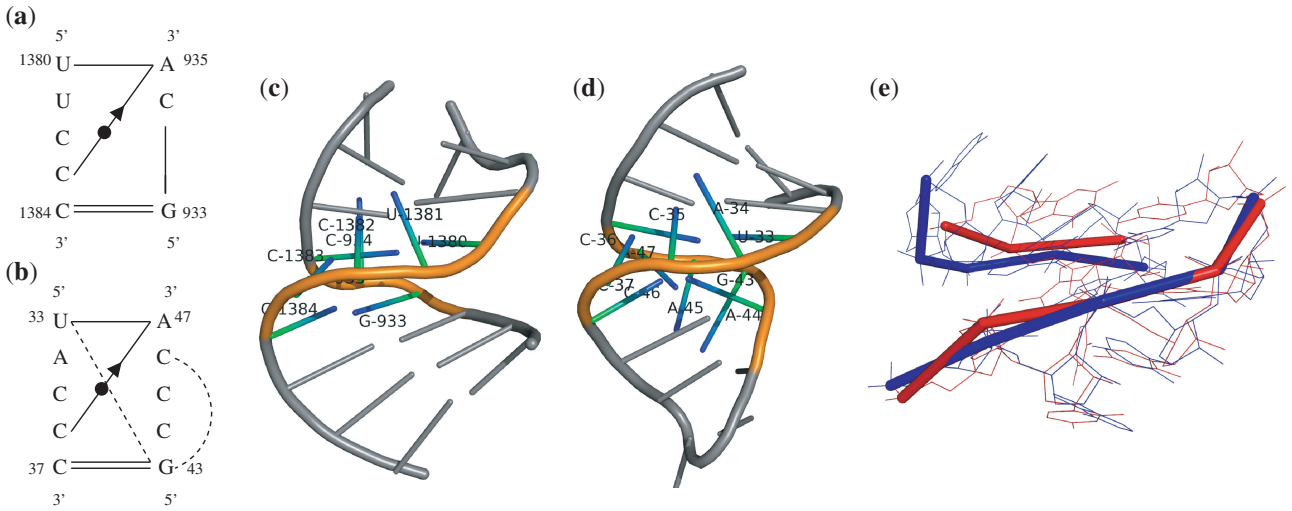

Figure 9. Potential novel motif family that increases the twists at the helical region. (**a**) and (**b**) The base pairing patterns of structural components found in 1J5E, chain A, 933-935/1380-1384 and 1S72, chain '9', 33-37/43-47, respectively. (**c**) and (**d**) Local structures around the motif instances shown in (a) and (b), respectively. (**e**) The superimposition between these two motif instances [red: (a); blue: (b)].

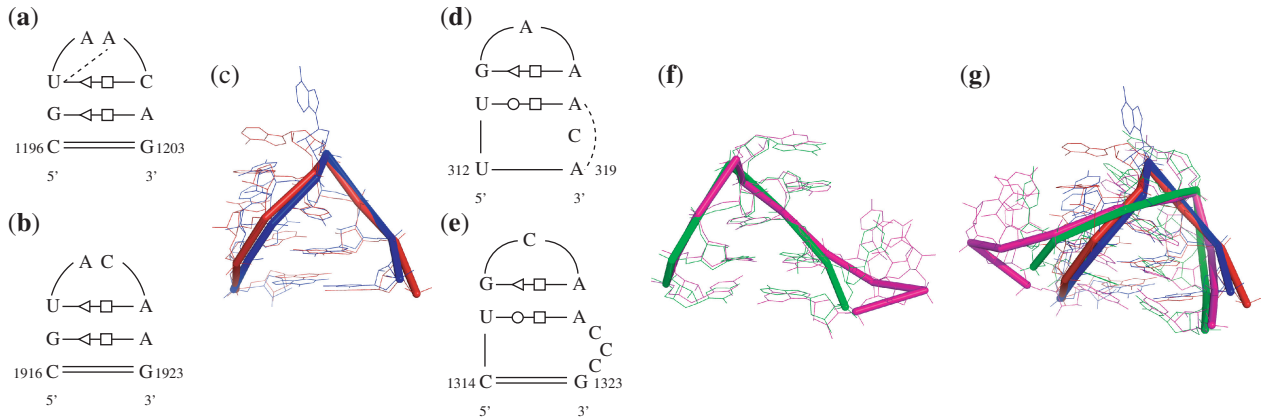

Figure 10. A novel type of hexaloop motif subfamily detected by our clustering method. (**a**) and (**b**) base pairing pattern of the two hexaloop motif instances identified in CH6: 1S72, chain '0', 1196-1203 and 1S72, chain '0', 1916-1923, respectively. (**c**) Superimposition between the motif instances shown in (a) and (b) [red: (a), blue: (b)]. (**d**) and (**e**) base pairing pattern of the two hexaloop motif instances identified in CH8: 1S72, chain '0', 312-319 and 1J5E, chain A, 1314-1323, respectively. (**f**) Superimposition between the motif instances shown in (d) and (e) [green: (d), magenta: (e)]. (g) Superimposition of the four hexaloop motif instances.

the hairpin loop length and the motif instance cannot be literally called 'hexaloop', we consider this instance to be true hexaloop motif due to its conservation in both base pairing pattern and 3D geometry.

The hexaloop motif family has been previously registered in the SCOR database (44), which defines only one hexaloop cluster in contrast to two subfamilies of hexaloop motif as suggested by our clustering results. SCOR identified all hexaloop motif instances found by us except the one with eight nucleotides. We consider that the two clusters of hexaloop motif have different sequence signatures and more importantly, different base pairing patterns (the *trans* SE/H G-A pair in CH6 comparing to the *trans* W/H U-A pair in CH8), therefore, should be classified into two different subfamilies. Indeed, superimposition of the four hexaloop motif instances clearly reveals two subfamilies of the motif that are consistent with our clustering predictions (Figure 10g). In this case, motif characterization should involve thorough consideration of both base pairing pattern and geometry, and classification of motif solely based on their sizes should be revised to incorporate such information.

DISCUSSION

In this article, we studied RNA structural motifs in ribosomal RNAs using a *de novo* clustering method based on base pairing patterns. The similarities between RNA structural motifs were evaluated by RNAMotifScan (6), which is a secondary structural alignment tool that considers non-canonical base pairs and their isostericity. We have significantly improved the existing clustering performance (Table 1) achieved by the LENCS method through addressing the three issues raised in the 'Introduction' section. The clustering framework can benefit future RNA structural motif analysis.

The newly identified motif instances were not discovered by previous base pairing pattern-based search methods since they contain base pair variations. The base pairs that are conserved in these instances can be critical in forming the motifs, and further studies should be conducted to elucidate their roles in maintaining proper functionalities of the motifs. On the other hand, the base pair variations should also be investigated to study functional evolution. Finally, more comprehensive consensus models can be built to facilitate future model-based searches by combining both information. The discoveries of novel motif families are also exciting. These new motifs may lead to the discovery of unknown structure–function relationships and define new building blocks for the RNA architecture, significantly improving our understanding of the RNA structural motifs.

Currently, the understanding of RNA structural motifs is limited even in well analyzed RNA structures. We plan to apply our new clustering framework on the entire PDB. The clustering framework is computationally efficient enough to handle this data set. However, the analysis step can be much more challenging when there are a large number of clusters being predicted. We are currently trying to develop methods that can automatically evaluate the output clusters and model corresponding structural motifs. We anticipate that more interesting discoveries can be made from the clustering analysis of the entire PDB.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online: Supplementary Tables S1–S3.

ACKNOWLEDGEMENTS

The authors thank the anonymous reviewers for their comments and suggestions.

FUNDING

Funding for open access charge: University of Central Florida.

Conflict of interest statement. None declared.

REFERENCES

1. Harrison,A.M., South,D.R., Willett,P. and Artymiuk,P.J. (2003) Representation, searching and discovery of patterns of bases in complex RNA structures. *J. Comput.-Aided Mol. Des.*, **17**, 537–549.
2. Duarte,C.M., Wadley,L.M. and Pyle,A.M. (2003) RNA structure comparison, motif search and discovery using a reduced representation of RNA conformational space. *Nucleic Acids Res.*, **31**, 4755–4761.
3. Dror,O., Nussinov,R. and Wolfson,H. (2005) ARTS: alignment of RNA tertiary structures. *Bioinformatics*, **21(Suppl. 2)**, 47–53.
4. Sarver,M., Zirbel,C., Stombaugh,J., Mokdad,A. and Leontis,N. (2008) FR3D: finding local and composite recurrent structural motifs in RNA 3D structures. *J. Math. Biol.*, **56**, 215–252.
5. Parisien,M., Cruz,J.A., Westhof,E. and Major,F. (2009) New metrics for comparing and assessing discrepancies between RNA 3D structures and models. *RNA*, **15**, 1875–1885.
6. Zhong,C., Tang,H. and Zhang,S. (2010) RNAMotifScan: automatic identification of RNA structral motifs using secondary structural alignment. *Nucleic Acids Res.*, **38**, 1–11.
7. Wadley,L.M. and Pyle,A.M. (2004) The identification of novel RNA structural motifs using COMPADRES: an automated approach to structural discovery. *Nucleic Acids Res.*, **32**, 6650–6659.
8. 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.
9. Djelloul,M. and Denise,A. (2008) Automated motif extraction and classification in RNA tertiary structures. *RNA*, **14**, 2489–2497.
10. Leontis,N., Stombaugh,J. and Westhof,E. (2002) The non-Watson-Crick base pairs and their associated isostericity matrices. *Nucleic Acids Res.*, **30**, 3497–3531.
11. 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.
12. Gendron,P., Lemieux,S. and Major,F. (2001) Quantitative analysis of nucleic acid three-dimensional structures. *J. Mol. Biol.*, **308**, 919–936.
13. Ben-Dor,A., Shamir,R. and Yakhini,Z. (1999) Clustering gene expression patterns. *J. Comput. Biol.*, **6**, 281–297.
14. 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. USA*, **87**, 8467–8471.

15. Klein,D., Schmeing,T., Moore,P. and Steitz,T. (2001) The kink-turn: a new RNA secondary structure motif. *EMBO J.*, **20**, 4214–4221.
16. Clemons,W.M., Brodersen,D.E., McCutcheon,J.P., May,J.L.C., Carter,A.P., Morgan-Warren,R.J., Wimberly,B.T. and Ramakrishnan,V. (2001) Crystal structure of the 30 s ribosomal subunit from thermus thermophilus: purification, crystallization and structure determination. *J. Mol. Biol.*, **310**, 827–843.
17. Wimberly,B.T., Brodersen,D.E., Clemons,W.M., Morgan-Warren,R.J., Carter,A.P., Vonrhein,C., Hartsch,T. and Ramakrishnan,V. (2000) Structure of the 30S ribosomal subunit. *Nature*, **407**, 327–339.
18. 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.
19. Torres-Larios,A., Dock-Bregeon,A.C., Romby,P., Rees,B., Sankaranarayanan,R., Caillet,J., Springer,M., Ehresmann,C., Ehresmann,B. and Moras,D. (2002) Structural basis of translational control by Escherichia coli threonyl tRNA synthetase. *Nat. Struct. Biol.*, **9**, 343–347.
20. Hausner,T.P., Atmadja,J. and Nierhaus,K.H. (1987) Evidence that the G2661 region of 23S rRNA is located at the ribosomal binding sites of both elongation factors. *Biochimie*, **69**, 911–923.
21. Moazed,D., Robertson,J.M. and Noller,H.F. (1988) Interaction of elongation factors EF-G and EF-Tu with a conserved loop in 23S RNA. *Nature*, **334**, 362–364.
22. Spackova,N. and Sponer,J. (2006) Molecular dynamics simulations of sarcin–ricin rRNA motif. *Nucleic Acids Res.*, **34**, 697–708.
23. Szewczak,A.A., Moore,P.B., Chang,Y.L. and Wool,I.G. (1993) The conformation of the sarcin/ricin loop from 28S ribosomal RNA. *Proc. Natl Acad. Sci. USA*, **90**, 9581–9585.
24. Strobel,S.A., Adams,P.L., Stahley,M.R. and Wang,J. (2004) RNA kink turns to the left and to the right. *RNA*, **10**, 1852–1854.
25. Szep,S., Wang,J. and Moore,P.B. (2003) The crystal structure of a 26-nucleotide RNA containing a hook-turn. *RNA*, **9**, 44–51.
26. Correll,C.C., Freeborn,B., Moore,P.B. and Steitz,T.A. (1997) Metals, motifs, and recognition in the crystal structure of a 5S rRNA domain. *Cell*, **91**, 705–712.
27. Leontis,N.B. and Westhof,E. (1998) The 5S rRNA loop E: chemical probing and phylogenetic data versus crystal structure. *RNA*, **4**, 1134–1153.
28. Cruz,J.A. and Westhof,E. (2011) Sequence-based identification of 3D structural modules in RNA with RMDetect. *Nat. Methods*, **8**, 513–519.
29. Smit,S., Rother,K., Heringa,J. and Knight,R. (2008) From knotted to nested RNA structures: a variety of computational methods for pseudoknot removal. *RNA*, **14**, 410–416.
30. Lescoute,A., Leontis,N., Massire,C. and Westhof,E. (2005) Recurrent structural RNA motifs, Isostericity Matrices and sequence alignments. *Nucleic Acids Res.*, **33**, 2395–2409.
31. Havgaard,J.H., Lyngsø,R.B. and Gorodkin,J. (2005) The FOLDALIGN web server for pairwise structural RNA alignment and mutual motif search. *Nucleic Acids Res.*, **33**, W650–W653.
32. Leontis,N. and Westhof,E. (2001) Geometric nomenclature and classification of RNA base pairs. *RNA*, **7**, 499–512.
33. 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.
34. Wool,I.G., Gluck,A. and Endo,Y. (1992) Ribotoxin recognition of ribosomal RNA and a proposal for the mechanism of translocation. *Trends Biochem. Sci.*, **17**, 266–269.
35. Doherty,E.A., Batey,R.T., Masquida,B. and Doudna,J.A. (2001) A universal mode of helix packing in RNA. *Nat. Struct. Biol.*, **8**, 339–343.
36. Nissen,P., Ippolito,J.A., Ban,N., Moore,P.B. and Steitz,T.A. (2001) RNA tertiary interactions in the large ribosomal subunit: the A-minor motif. *Proc. Natl Acad. Sci. USA*, **98**, 4899–4903.
37. Leontis,N.B. and Westhof,E. (2003) Analysis of RNA motifs. *Curr. Opin. Struct. Biol.*, **13**, 300–308.
38. Vidovic,I., Nottrott,S., Hartmuth,K., Luhrmann,R. and Ficner,R. (2000) Crystal structure of the spliceosomal 15.5kD protein bound to a U4 snRNA fragment. *Mol. Cell*, **6**, 1331–1342.
39. Correll,C.C., Freeborn,B., Moore,P.B. and Steitz,T.A. (1997) Metals, motifs, and recognition in the crystal structure of a 5S rRNA domain. *Cell*, **91**, 705–712.
40. Dallas,A. and Moore,P.B. (1997) The loop E-loop D region of Escherichia coli 5S rRNA: the solution structure reveals an unusual loop that may be important for binding ribosomal proteins. *Structure*, **5**, 1639–1653.
41. Seggerson,K. and Moore,P.B. (1998) Structure and stability of variants of the sarcin–ricin loop of 28S rRNA: NMR studies of the prokaryotic SRL and a functional mutant. *RNA*, **4**, 1203–1215.
42. Leontis,N.B., Lescoute,A. and Westhof,E. (2006) The building blocks and motifs of RNA architecture. *Curr. Opin. Struct. Biol.*, **16**, 279–287.
43. Leontis,N.B., Stombaugh,J. and Westhof,E. (2002) Motif prediction in ribosomal RNAs Lessons and prospects for automated motif prediction in homologous RNA molecules. *Biochimie*, **84**, 961–973.
44. Tamura,M., Hendrix,D., Klosterman,P., Schimmelman,N., Brenner,S. and Holbrook,S. (2004) SCOR: Structural Classification of RNA, version 2.0. *Nucleic Acids Res.*, **32**, D182–D184.