Published online 8 March 2021

Nucleic Acids Research, 2021, Vol. 49, No. 11 ***e61***
https://doi.org/10.1093/nar/gkab131

RNAMotifContrast: a method to discover and visualize RNA structural motif subfamilies

Shahidul Islam, Md Mahfuzur Rahaman and Shaojie Zhang*

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

Received October 11, 2020; Revised February 16, 2021; Editorial Decision February 17, 2021; Accepted February 18, 2021

ABSTRACT

Understanding the 3D structural properties of RNAs will play a critical role in identifying their functional characteristics and designing new RNAs for RNA-based therapeutics and nanotechnology. While several existing computational methods can help in the analysis of RNA properties by recognizing structural motifs, they do not provide the means to compare and contrast those motifs extensively. We have developed a new method, RNAMotifContrast, which focuses on analyzing the similarities and variations of RNA structural motif characteristics. In this method, a graph is formed to represent the similarities among motifs, and a new traversal algorithm is applied to generate visualizations of their structural properties. Analyzing the structural features among motifs, we have recognized and generalized the concept of motif subfamilies. To asses its effectiveness, we have applied RNAMotifContrast on a dataset of known RNA structural motif families. From the results, we observed that the derived subfamilies possess unique structural variations while holding standard features of the families. Overall, the visualization approach of this method presents a new perspective to observe the relation among motifs more closely, and the discovered subfamilies provide opportunities to achieve valuable insights into RNA's diverse roles.

INTRODUCTION

Non-coding RNAs have been one of the central focuses of biological and medical research due to its connection to many cellular functions (1–3) and diseases, including cancer (4) and Alzheimer's (5). One of the key features that dictate non-coding RNA functions is their 3D structures (6–8). As a result, understanding the characteristics of RNA 3D structures becomes a critical element of biological research. With >5000 RNA 3D structures being available in the PDB (9) database and growing, the opportunities to gain deep insight into the architecture and the functionality of RNA are greatly expanding.

One approach to identify the characteristics of RNA is finding recurring structural components across various types of RNAs. Those recurring structural components are called RNA structural motifs and considered as building blocks of the RNA architectures (10,11). The importance of RNA structural motifs is shown in many contexts, including how different molecules engage with RNA through interactions in the known motif regions (12–14). Moreover, it has been shown that RNA motifs can be used in building structural elements in nanotechnology (15). Finding RNA structural motifs of similar characteristics and identifying variations of them can help in understanding the basics of RNA functions and aid in the rising new directions of RNA-inspired research.

There are methods, such as RNAMotifScan (16), RNAMotifScanX (17) and FR3D (18), that utilize well-defined motifs to search for new instances in different locations and organisms. These methods have discovered many instances of known motifs, such as kink-turn (12), reverse kink-turn (19), sarcin-ricin (13), C-loop (20) and E-loop (21). There are also approaches that can find similarities among motifs through clustering, such as, RNAMotifScan alignment-based clustering (22,23) and FR3D alignment-based RNA 3D Motif Atlas (24). In addition to finding instances of known motif families, these computational methods also have identified new motif families. These methods attempt to put all the instances of a motif family into the same cluster. But they cannot manage to do so in many cases due to the inherent variations of structural features in the motif instances. If these methods allow too much flexibility to encompass these variations, they run the risk of putting instances of different motif families together. As a result, they choose the option to be rigid to some extent. Consequently, the instances of one family get separated into multiple groups in these clustering results. The details of these separations into groups and the corresponding implications is an area worth investigating, but there is not much work focusing in this direction yet.

A few research were conducted to extensively analyze the variations in a couple of well-known motif families, such as kink-turn and sarcin-ricin. Leontis *et al.* system-

*To whom correspondence should be addressed. Tel: +1 407 8236095; Fax: +1 407 8235835; Email: shzhang@cs.ucf.edu

© The Author(s) 2021. Published by Oxford University Press on behalf of Nucleic Acids Research.
This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial License (http://creativecommons.org/licenses/by-nc/4.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact journals.permissions@oup.com

atically evaluated isosteric relationships among different base-pairing interactions (25), and consequently showed that similar structural features of motifs can be achieved with a variety of base-pairings (26). The extent of acceptable sequence variations, especially the isosteric variants, to achieve desired structural features of known motif families is also shown in simulation-based experiments (27,28). Substantial work on the kink-turn motif family has been done by Lilley (29,30) to identify different structural features and corresponding functional behaviors. The analysis shown in all these studies presents the importance of recognizing the variations of sequences and structures to understand the functionalities of motifs. However, no existing computational method provides the capability to comprehensively compare and contrast the variations in a given set of motif family instances.

In this work, we have the generalized goal to address the variations in the motif families and categorize them into subfamilies based on their similar and unique structural features. We have designed a *de novo* computational method, RNAMotifContrast, which provides a comprehensive insight into RNA structural motifs through its analysis and visualization of structural features in a way that was not possible before. RNAMotifContrast first creates a structural similarity-based graph with relational properties that utilize (i) the flexible alignment of RNAMotifScanX and (ii) a new similarity measurement, based on the alignment length and RMSD. Then, it uses a novel traversal algorithm that guides the ordering and superimposition of structures to visualize the contrasting features of motif families. Applying this method on a newly prepared data set, we have specified and discovered RNA structural motif subfamilies for all the well-known motif families. By analyzing the known motif families such as kink-turn, reverse kink-turn, sarcin-ricin, C-loop, E-loop and T-loop, we have identified key structural characteristics and corresponding subfamilies for all of them. Instances of subfamilies can be found at the Supplementary Table S5, along with all the images and corresponding details of subfamily features in the Supplementary Website (http://genome.ucf.edu/RNAMotifContrast).

METHODS

In RNAMotifContrast, we take all the instances of an RNA structural motif family from our curated dataset as input and identify structural features to analyze the variations among these motifs (details of the dataset is provided in the Results section). From the input motif instances, we first consider the annotations of base-base interactions and generate pairwise alignments. With the help of those alignments, we determine the most similar structure for each motif using alignment-length-restricted RMSD comparisons (explained in the 'Alignment-length-restricted RMSD comparison' subsection). Based on the most similar pairings, we create a directed graph that provides the platform to assess the overall relations among the motifs to identify the groups with similar structural features. We apply a merging algorithm to combine the smaller groups of similar structural motifs into forming subfamilies. We then execute a three-layer hierarchical traversal algorithm that connects and orders all the motifs in the motif family through building parent–child relationships. These connections guide the superimposition of the motifs and generate the coveted images that give the most visually explicit comparison among the RNA structural motifs. The steps of this method are illustrated in Figure 1, and the descriptions are given in the following sections.

Extract annotations and coordinates from PDB data

For the RNA motif instances in the input, we download corresponding structure files from the PDB database. For the PDB files, we collect the FR3D (18) annotations from their webpage and also generate the DSSR (31) annotations. We then generate a merged annotation for each PDB combining the annotations from these two tools to increase the probability of recognizing more interactions in a motif than any of these tools can do individually. While merging, conflict of annotations happens in some cases. For those cases, two annotation tools give different annotations for the same pair of bases. We address these conflicts, based on the likelihood of interactions occurring between a given pair of bases. We counted the frequencies of interactions for all possible pairs of bases in RNAs from the annotations of all PDBs in the nonredundant PDB list (32) release 3.57 at resolution 4.0 Å. The frequency and the corresponding ranking of annotation for each pair of bases are given in Supplementary Table S3. We consider one straightforward rule - the higher the frequency, the higher the likelihood of that interaction between a given pair of bases. As a result, from the conflicting annotations, we choose the one interaction which has a higher number of occurrences in RNAs. Examples of such conflict resolves are shown in the Supplementary Figures S3 and S4. Moreover, for each structural motif, we generate a customized PDB file. It includes only the coordinates of the local region, which contains the given motif. This reduction of coordinates from the original PDB reduces the memory usage significantly and improves the run-time while generating images using PyMOL.

Pairwise alignment of motifs

The next step in RNAMotifContrast is to generate pairwise alignments among all the motifs. To generate better pairwise alignments efficiently, we have developed a new, improved version of the existing alignment tool RNAMotifScanX (17). We are calling this version RNAMotifScanX 2.0. Major modifications made in this version are: (i) adding a new module to align structures that do not have any common base-pairing annotations but may have shared features based on sequences and base-stacking interactions (33), (ii) choosing better alignment based on various structural features when alignment scores are equal among multiple options of aligning two motifs, (iii) including possible triple-interactions for improved heuristics in the branch-and-bound approach to ensure finding the optimal alignment, (iv) balancing the penalty for missing base pairs in the aligned core region for both query and target structure (which was different in the original RNAMotifScanX), (v) implementing a basic Genetic Algorithm for clique finding (34) when RNAMotifScanX is likely to take

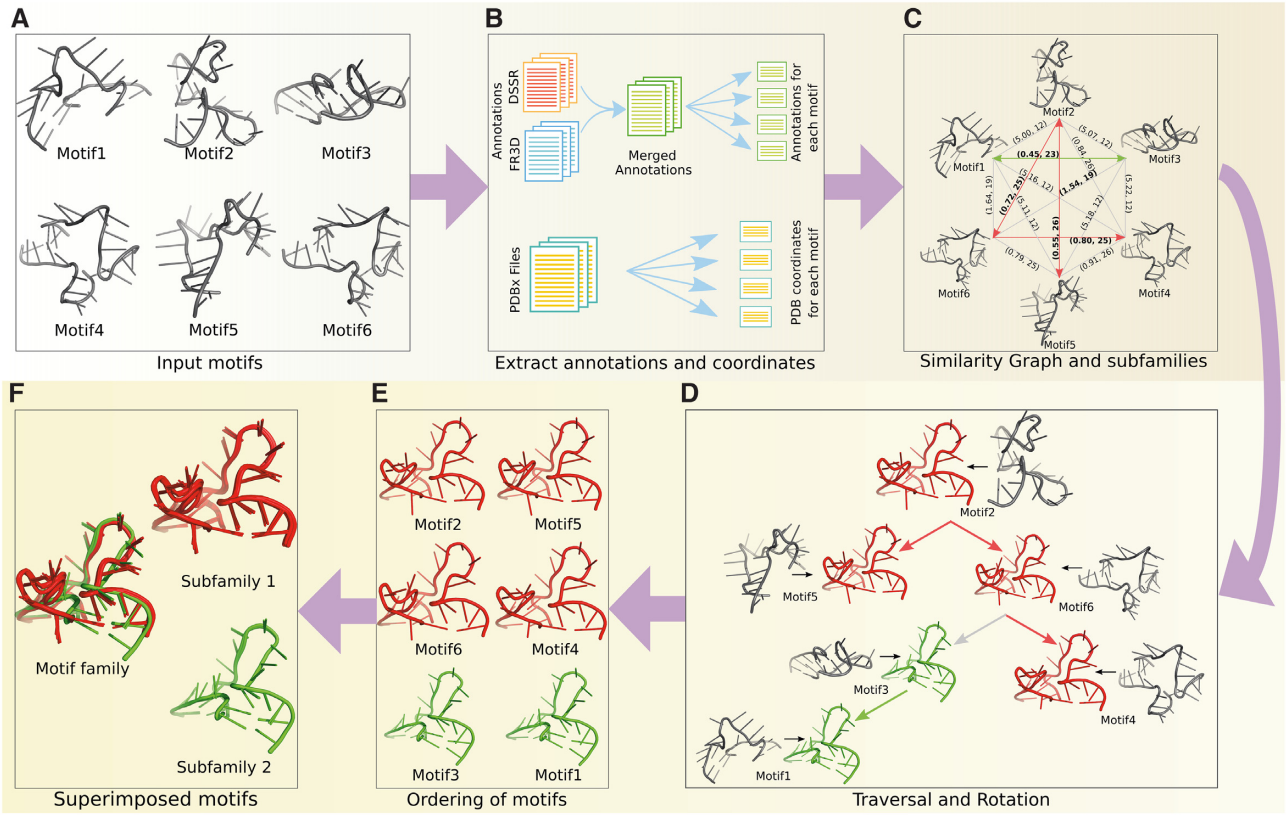

Figure 1. The pipeline of RNAMotifContrast. (**A**) PyMOL images of the input motifs in the default orientation. (**B**) Visual representation of annotation and coordinate extraction of the regions around each motif. (**C**) Generating Similarity Graph from the pairwise alignments and identifying subfamilies. Edges are labeled with '(RMSD, alignment length)'. Edges connecting the instances of a subfamily are colored the same. (**D**) The parent-child relationships generated from the traversal algorithm and the corresponding rotation of structures to superimpose each motif to its parent. The root node is rotated to get a better orientation to observe the features. (**E**) Motifs ordered according to the traversal. (**F**) Subfamily and family-wise superimposition of motifs.

too much time to find the clique and the alignment accordingly and (vi) redefining the boundary of the region representing core features of aligned structural motifs. Previously, the boundaries of core aligned regions were defined based on the outer-most aligned base pairs. However, with the extensive analysis in this project, we identified that the boundary could be extended to make longer alignment and include more structural features. For each segment in the alignment, we extended them in both directions as long as it improves the alignment score. The details of these modifications are discussed in the Supplementary Data. We use the aligned residues from RNAMotifScanX 2.0 to superimpose motifs and calculate RMSDs. The alignment lengths correspond to the number of aligned residues in the pairwise alignments. These alignment lengths and RMSDs are used in the alignment-length-restricted RMSD comparison to find the most similar pair of motifs accordingly.

Alignment-length-restricted RMSD comparison

A commonly used measurement to evaluate a set of RNA 3D structure alignments is Root-Mean-Square Deviation (RMSD) (35). However, RMSD may not provide the best assessment while comparing aligned structures of various lengths. When we have a smaller number of nucleotides to align, it would be more likely to get a better RMSD. Nevertheless, achieving a better RMSD with a small number of nucleotides may be a less desired option compared to a longer alignment with a relatively worse RMSD. To address this issue, we compare two alignments based on RMSD only if the alignment lengths are above a threshold. Otherwise, we consider that the longer alignment is better, regardless of its RMSD. The alignment length thresholds depend on the properties of a given family (see Supplementary Table S1). Further explanation with an example is provided in the Supplementary Data.

Generating similarity graph and subfamilies

The relations among motifs are represented in a directed graph, which we call the Similarity Graph. In the Similarity Graph, the motifs represent the nodes. For each motif A, we consider the alignments with all other motifs and apply the alignment-length-restricted RMSD comparison to find the most similar structure with it. If motif B is the most similar structure for A, we add an incoming edge to A from B. The incoming edge to motif B may come from A or another motif C. As we consider exactly one most similar motif for each of the motifs in building this Similarity Graph, each node will have exactly one incoming edge. With this one particular restriction, the Similarity Graph evolves into a graph with interesting properties. While all the motifs in a Simi-

larity Graph may not be connected, one or more mutually exclusive subsets of them will be connected to each other. We call such subsets Connected Motif Groups (CMGs). The graph, along with the CMGs, has these two important properties: (i) it consists of one or more CMGs, where each CMG has at least two nodes, (ii) each CMG has exactly one cycle in it. The formal proofs of these properties are provided in the Supplementary Data along with an example of Similarity Graph in Supplementary Figure S1.

A directed edge from B to A in a CMG implies that B is the most similar structure for A. As a result, if we put A in a group, it should be in the same group as B. By applying the transitivity rule for all edges, we can deduce that all nodes in a CMG can be placed in the same group. However, we also consider the fact that there might be similarities among the CMGs. We address this situation with a merging algorithm. For any pair of CMGs, we test if they can be merged together. We check how many motif instances of one CMG have strong similarities with the instances of the other CMG. An alignment is considered well-matched and represents strong enough similarity if it passes the alignment length threshold for the given family and has a good RMSD (1.0 Å by default, and it can be configured to a different value by the user). If the number of well-matched pairs of motifs passes a given percentage or count threshold from both CMGs, we merge them. If multiple pair of CMGs passes the threshold, we merge the pair with the best average RMSD first. We continue this process in a guided-tree based approach until no further merging is possible. As the end result, each merged set of CMGs represents a subfamily with certain structural properties of their own. An example Similarity Graph is shown in Figure 1C, where each motif represents a node, and the red/green colored edges represent the best similarity connection among the instances of CMGs. In this case, there are two CMGs, and they cannot be merged further. So, from these six motif instances, we get two subfamilies.

Motif traversal, ordering and visualization

In this section, we address how to superimpose and visualize multiple motifs that can show the structural features in the most comparable and contrasting way. One straightforward and commonly used method for the superimposition of multiple motifs is the 'align-to' approach. With this method, one motif is considered as a reference, and all other motifs are superimposed with the reference. While this approach works fine for the highly similar structures, it does not show the structural characteristics well for superimposing motifs with different types of variations. The situation with the 'align-to' approach for an example set of motifs is shown in Figure 2A–D. For the first three cases, we used the 'align-to' method of PyMOL with three different algorithmic parameters. We then addressed the fact that this method in PyMOL uses sequence alignment based superimposition. So, we developed a version of the 'align-to' method, where the alignment from RNAMotifScanX 2.0 is used to determine the residues to superimpose. The outcome of the RNAMotifScanX based 'align-to' is shown in Figure 2D. It improves the result significantly over the 'align-to' in PyMOL, but it is not the best possible option. We designed a new approach that utilizes the properties of the Similarity Graph to determine the relationship and ordering in superimposition.

Instead of superimposing all the motifs with a fixed reference motif, we dynamically choose the reference of superimposition for each motif. For a given CMG, we utilize the directed edges among motifs to represent the parent-child relationships for the superimposition, where the parent is used as the reference to superimpose the child. For example, an edge from motif A to B implies A can be used as the reference to superimpose B. We choose a starting motif from the motifs in the cycle of CMGs, depending on two criteria. If it is the first CMG to superimpose, the starting motif is selected based on the higher connectivity and better alignments with other motifs. From the second CMG, the selection of the starting motif depends on the already superimposed CMGs. Starting with the first motif, we follow the directed edges to superimpose the subsequent motifs and progressively superimpose them on top of the existing superimposed motifs. The traversal of a CMG makes sure that a motif is superimposed with the most similar motif through the parent-child relationship. However, in order to superimpose all instances of a subfamily or the whole family, we need to address multiple CMGs that are not connected with each other. We developed a three-layer traversal algorithm to create the parent-child relations among the CMGs in the subfamilies and the subfamilies of a family. The first goal of the three-layer traversal is to generate the best possible parent-child relations among the subfamilies and components. The second goal is to achieve a comparative visualization by generating the side-by-side images to keep similar motifs as close as possible. The definition and details of the traversal algorithm are provided in the Supplementary Data. An example of this traversal approach is also given in Supplementary Figure S2.

For the visualization, we provide users the option to choose the orientation of the first motif in the superimposition. From there, we traverse all the motifs and rotate accordingly to align and superimpose with its parent structurally. This parent–child relationship to define the rotation reference is one of the key features of RNAMotifContrast that enables us to generate the improved comparative visualization of the motifs. One straightforward outcome we get from this traversal and rotation is the superimposed images of each input family and their subfamilies. The superimposed images provide an overview of the structural features and variations. For detailed observations of the subfamily characteristics, we also generate side-by-side images of the rotated motifs according to the ordered list of the traversal. Figure 1E shows the example of side-by-side images, and Figure 1F shows an example of the superimposed images. Besides, we generate progressive images that display the changes for adding each structure to the pool of superimposed motifs. The progressive images are provided in the Supplementary Website. Moreover, we provide an option to save the PyMOL sessions of our superimposed motifs, which allows users further flexibility for observing the structural features of subfamilies from different orientations. Overall, the outcome of this traversal algorithm and the visualization provides us the opportunity to analyze the

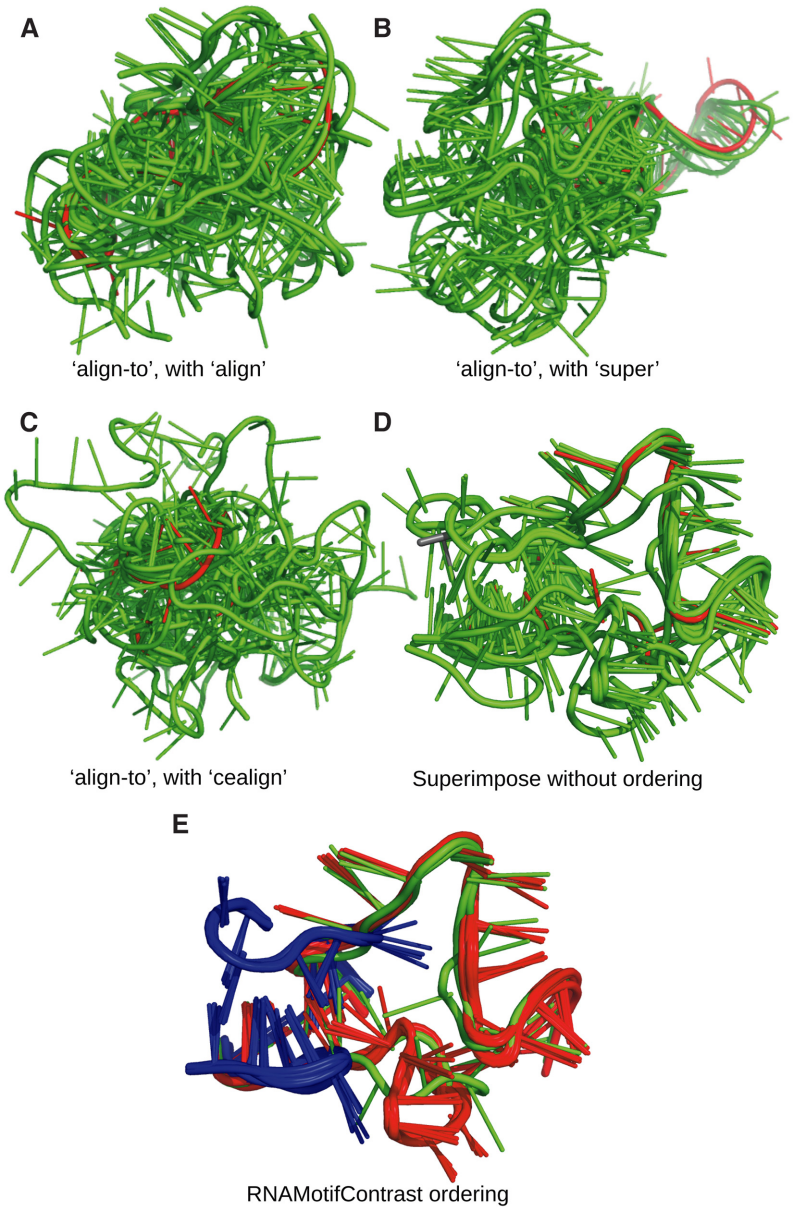

Figure 2. Comparison of different superimposition approaches for an example dataset. The first loop in all the superimposition is colored red. (**A**–**C**) Superimposition using three different algorithms of 'align-to' approach in PyMOL. (**D**) Superimposition for our implementation of 'align-to' using RNAMotifScanX alignment. (**E**) Superimposition using traversal of RNAMotifContrast. In (A–D), the same color (green) is used to represent all the superimposed loops to the first loop (in red). In (E), three different colors (red, green and blue) are used in this superimposition to represent three different subfamilies identified for this set of loops.

fine details of similarities and dissimilarities among the motif instances.

RESULTS

Dataset with instances of known motif families

As a means to show its effectiveness, we have applied RNAMotifContrast on a dataset of known motifs of internal and hairpin loops. To prepare this data, we first considered the instances of well-known motif families annotated in the clustering work of Ge *et al*. (23), which is built

Table 1. Number of instances in two different sources and in the filtered merged list for known motif families

	Number of motif instances			
Loop type	Ge *et al*.	RNA 3D M. atlas	Merged list	Filtered list
IL	250	256	368	347
HL	209	364	415	397
Total	459	620	783	744

IL, internal loop; HL, hairpin loop.

Table 2. Subfamilies of the known motif families with their average RMSDs and aligned lengths for three different superimposition

Loop type	Motif family	No. of Motifs	No. of subfamilies (Sizes)	Superimposition avg. RMSD/aligned length		
				No ordering	Ordered	Subfamilies
Internal loop (IL)	Kink-turn (KT)	67	4 (45,7,10,5)	2.804 / 6	0.678 / 10	0.468 / 10
	reverse Kink-turn (rKT)	8	3 (4,2,2)	1.490 / 19	0.910 / 22	0.702 / 23
	Sarcin-ricin (SR)	72	4 (55,7,7,3)	1.726 / 7	0.913 / 10	0.419 / 10
	C-loop (CL)	41	6 (11,8,5,7,3,7)	2.389 / 5	0.726 / 7	0.589 / 7
	E-loop (EL)	47	3 (24,19,4)	1.521 / 7	0.506 / 8	0.474 / 8
	Hook-turn (HT)	34	3 (27,5,2)	1.483 / 8	0.830 / 8	0.762 / 8
	Tandem-shear (TS)	44	3 (39,2,3)	1.016 / 5	0.491 / 6	0.482 / 6
	Tetraloop-receptor (TR)	19	2 (17,2)	0.992 / 5	0.534 / 6	0.536 / 6
	L1-complex (L1C)	6	2 (4,2)	2.400 / 11	1.701 / 13	0.833 / 13
	Rope-sling (RS)	9	1 (9)	0.632 / 8	0.409 / 9	0.409 / 9
Hairpin loop (HL)	GNRA	291	6 (254,12,10,5,3,7)	0.561 / 5	0.272 / 5	0.252 / 5
	T-loop (TL)	106	6 (81,5,3,8,4,5)	0.827 / 8	0.538 / 8	0.443 / 9

'Sizes' represents number of motifs in each subfamily. 'No ordering' represents superimposition with single reference using RNAMotifScanX alignment. 'Ordered' represents superimposition using parent-child relation from traversal. 'Subfamilies' represents separate superimposition of subfamilies.

upon the RNAMotifScan alignment. Then, using these motifs, we identified corresponding clusters in RNA 3D Motif Atlas (24) (Release 3.2), which include some of those instances, along with additional motif occurrences discovered through their 3D structure comparison. For both these clustering results, we mapped the loops to the RNAs of the nonredundant list (version 3.57). This mapping makes them compatible to find overlap of loops. While processing these loops, we have identified and removed some of them, which have a significant number of missing residues. By merging the results of these two clustering methods, we have created a combined list of instances for each known motif family. We additionally applied filtering based on the RNAMotifScanX alignment score and the corresponding RMSD value to exclude instances that do not have good enough alignment with any other instance in the family. The final list of motif instances incorporated (i) the feature of FR3D, which is built upon matching 3D structural properties and (ii) the flexibility of structural variations allowed by RNAMotifScan. Consequently, this data set provides a new platform to do an extensive analysis of the motif properties with the help of RNAMotifContrast. The number of instances for the filtered list, as well as instances from each source, is given in Table 1, and the instances of the merged list for each family is provided in Supplementary Table S4. A more detailed description of the curation process is provided in the Supplementary Data along with the numbers of instances in Supplementary Table S2.

Analysis of subfamily extraction and traversal based superimposition

We apply RNAMotifContrast on each motif family to extract key features to compare and contrast the instances. As the outcome of merging the CMGs of the similarity graph, we discover the subfamilies. However, defining how many exact subfamilies are there, depends on the merging threshold. To merge two connected motif groups (CMGs), we have used the criteria that combines the required thresholds: at least 50% of instances from each CMG need to have alignment with the instances of other CMG that passes the threshold of the alignment length for that family and the RMSD 1.0 Å. The alignment length threshold we found for the dataset is given in Supplementary Table S1. It is worth mentioning here that we provide the additional feature in our tool for users to modify all these thresholds. The number of subfamilies along with the instance count (sizes) for each family is given in Table 2.

From Table 2, we can notice that the number of subfamilies varies dramatically from one family to another. There are families, such as Rope-sling (22), for which the instances in the family are structurally very similar. With the merging thresholds we have used, all the instances of this family are grouped together. On the other hand, some motif families are distributed into many subfamilies. We found six subfamilies for hairpin T-loop and six subfamilies for C-loop. In accordance with our criteria, these separations into subfamilies imply that even half of the instances in any subfamily pair do not have enough similarity with each other. Most of the alignments among the instances of those subfamilies were unacceptable for having RMSD worse than 1.0 Å or alignment length less than the threshold. For example, the alignment length threshold we found for C-loop is 5, and it implies that for an alignment to be capable of representing the features of the C-loop, it has to be at least five residues long. For the hairpin T-loop, the threshold is 8, and accordingly, an alignment can be acceptable only if the length of the alignment is at least 8.

Additionally, Table 2 shows the comparison of average RMSDs and average alignment length for three options of superimposition: (i) when all the motifs are superimposed with a single reference motif (using the RNAMotifScanX alignment based 'align-to' approach), (ii) when the motifs are ordered and superimposed using the traversal of the similarity graph and (iii) when the motifs are separated to superimpose as subfamilies. The changes from option 1 (no ordering) to option 2 (ordered) shows the effect of parent–child relationship based superimposition, guided by the traversal of subfamilies and CMGs. The change in option 3 (subfamilies) compared to option 2 shows the impact of separating the motifs into subfamilies and superimposing among subfamily instances (effectively the average by excluding the inter-subfamily superimposition). In most of the cases, average RMSDs for the option 3 are expected to be better than the option 2. However, in some cases the inter-subfamily best edge might have a lower RMSD than

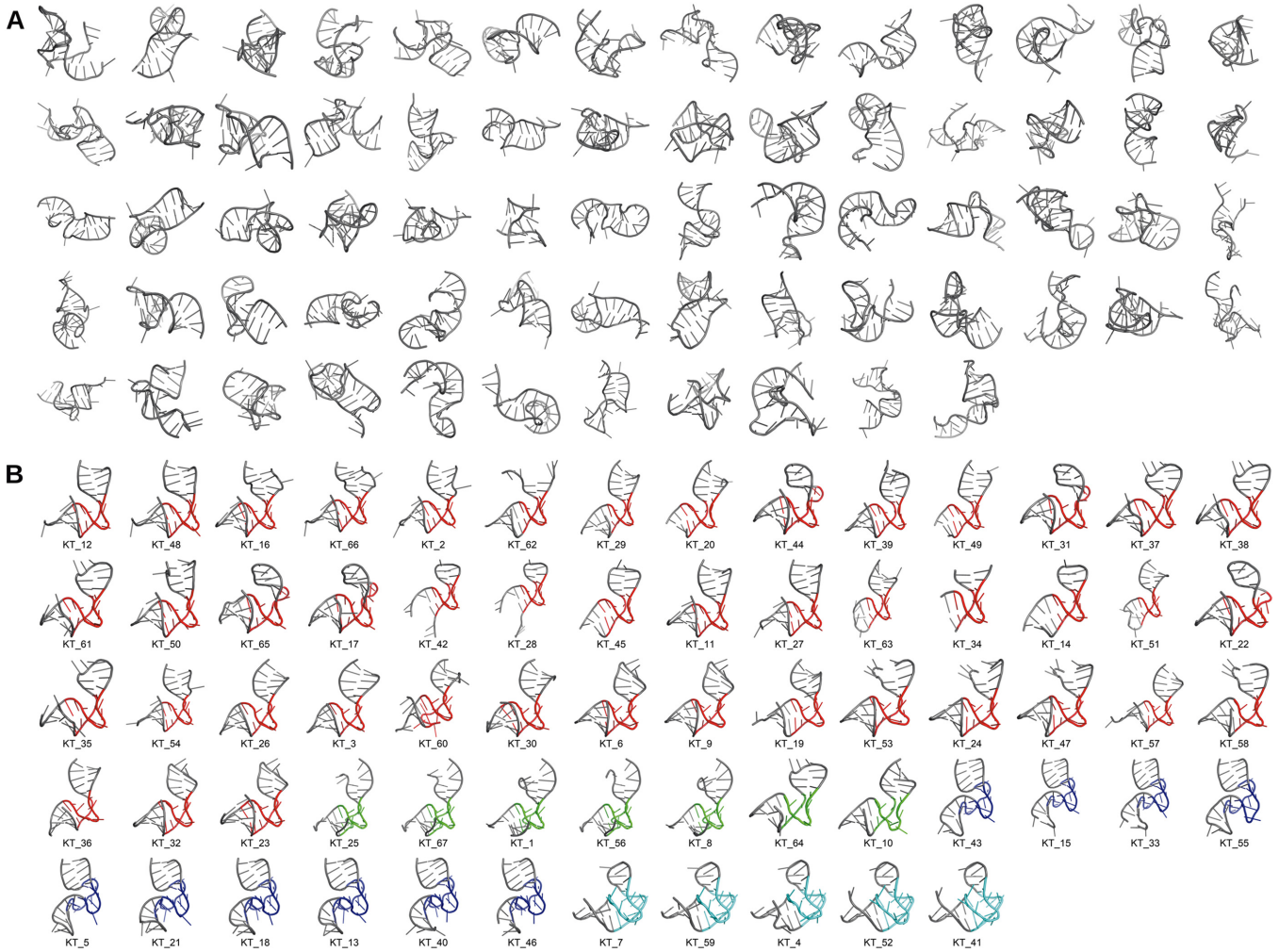

Figure 3. RNAMotifContrast input/output for kink-turn motif family data. (**A**) Input motif instances from the kink-turn family in the default orientation of PDB data. (**B**) The rotated and ordered output motifs according to the traversal algorithm of RNAMotifContrast. The aligned parts of the loops are colored. The non-aligned part of the loops and the helices around them are gray. The different colors represent instances of different subfamilies. The motif ID in (B) corresponds to the input order of the motif in (A).

average. It represents strong similarity between a single pair of instances in two subfamilies while the overall similarities among the instances are not good enough to merge them together. One such example is the case for the Tetraloop-receptor shown in Table 2. In summary, the similarity score (average RMSD and average alignment length) improves significantly with the use of traversal based ordering of motif for the superimposition.

Properties of the discovered motif subfamilies

The traversal based superimposition and side-by-side images provides an explicit visualization of the similarities and differences of the structural features among the motifs and their subfamilies. We additionally identify a representative motif from each subfamily for further comparative analysis on the subfamilies. The representatives are selected based on the overall alignment quality of a motif with other members of the subfamily. Color-annotated images (see Figure 4 as an example) for all the representatives motifs and the corresponding details of sequence and interactions are provided in the Supplementary Website. We utilized the representatives to recognize the source of structural differences among motif subfamilies. Based on the representative analysis of our result, we identified the association of structural variations with the following sources: (i) the variation in base-pairing and base-stacking interactions, (ii) different bulge lengths and (iii) varying nucleotide sequence. For a given motif family, some subfamilies have more interactions than other subfamilies, while some of them have different types of interactions. For example, CL-Sub1 and CL-Sub3 representatives have six base-pairing interactions compared to CL-Sub4 representative having only four. Similarly, EL-Sub1 has six base-pairing interactions, and EL-Sub3 has four. Some of the subfamilies have longer or shorter bulges compared to the more frequent type of instances of a given family. The motifs in L1C-Sub2 has a longer bulge that creates an interesting structural extension. Both L1C-Sub1 and L1C-Sub2 representative motifs have six base-pairing interactions, but two of those interactions are different, while

four of them are similar. The additional residues of the longer bulges also cause additional base-stacking interactions. The subfamily representatives of the hairpin T-loop shows how the bulge length and the additional stack interactions can affect the structural features of the motifs to the extent where they can be considered as different subfamilies (while most of them have a similar set of annotated base-pairing interactions). In some cases, the differences in participating nucleotides in the interactions cause the structural variation. The representative of subfamilies GNRA–Sub5 and GNRA–Sub6 have similar type of interactions, but the variation of sequences caused the instances to be partitioned into subfamilies.

Overall, the structural variations of subfamilies in a family stem from one or more of these sources of variations. Visualization and the corresponding description of properties for all subfamilies are provided in the Supplementary Website. In the remaining section, we present the analysis of three very well-known motif families, kink-turn, reverse kink-turn, and sarcin-ricin, to show the characteristics of the subfamilies in depth.

Kink-turn

Kink-turn (12) is a well-studied motif with very distinctive features and known to play an essential role in RNA structural architecture along with serving as a binding site for proteins. It is an asymmetric internal loop that changes the direction of the helix with the characteristics kink and turns. In our kink-turn motif family, there are 67 instances. From Figure 3A, we can observe the fact that it is difficult to assess the structural features in the default orientation of the motifs even when they belong to the same family and placed side by side. On the other hand, the parent-child relationship based ordered and rotated representation of the instances in Figure 3B provides significantly improved means to make comparison and have a better understanding on the properties of motif instances in the family. For the kink-turn motif family, we have found four subfamilies with 45, 7, 10 and 5 instances accordingly. By observing the subfamilies in Figure 3 and the corresponding representatives in Figure 4, we can identify the characteristics that differentiate them.

The structural features of kink-turn subfamily 1 (KT-Sub1) shows the characteristics of the traditionally addressed kink-turn (12) and its variations studied by many works, including Lescoute *et al.* (26) and Lilley (29,30). From the interactions of the representative motif shown in Figure 4A, we can observe several features that define kink-turn properties. The representative motif has the bulge followed by two A/G and G/A interactions, which corresponds to the definition of traditionally known kink-turn (12). The motif also shares three 'H/S trans' and one 'S/S trans' interactions with the consensus of kink-turn reported by Lescoute *et al.* (26). These key features of this motif evidently show KT-Sub1 to be the representative of the well-known kink-turn. Annotations of other instances of this subfamily also show there are some variations of sequences, annotations, and location of interactions among them. However, the structural similarities among the instances of this subfamily are evident from the visualization outcome generated by RNAMotifContrast, which is presented in the side-by-side images in Figure 3B and the superimposition image in Figure 4A. It is worth mentioning here that the side-by-side image facilitates additional opportunities to observe and analyze the structural features of the whole family compared to only using the superimposition of structures.

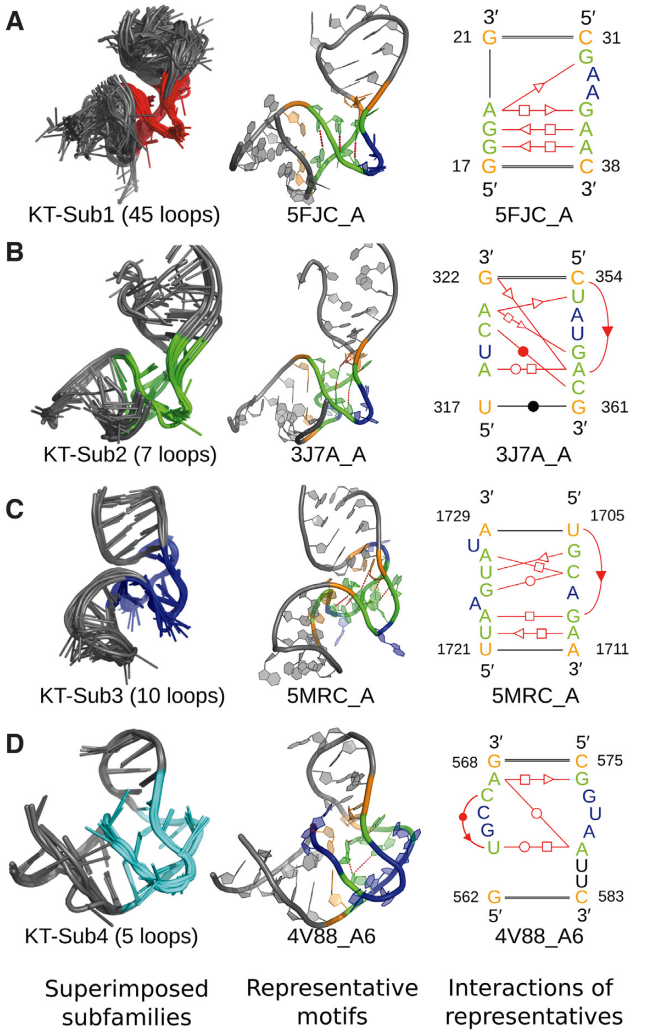

Figure 4. Superimposed motifs and analysis of the kink-turn subfamily properties through the representative motifs along with their interactions (**A**) 45 motif instances of KT-Sub1 (the representative motif of this subfamily is similar to the well known kink-turn instances), (**B**) seven motif instances of KT-Sub2, (**C**) 10 motif instances of KT-Sub3 and (**D**) five motif instances of KT-Sub4. The residues on the loop boundaries that form 'W/W cis' interactions are marked orange. The noncanonical interactions are colored red, and the residues associated with them are colored green. The additional residues in the loop are marked blue. The representatives show that the kink and turns of the subfamilies are created and supported by different sets of noncanonical interactions.

On the other hand, KT-Sub2, KT-Sub3, and KT-Sub4 have the kink and turn, but they have some unique characteristics compared to KT-Sub1. Their interaction sets and structural features are quite different, which is shown in Figure 4. For the KT-Sub2 representative, the structure of the longer strand is almost identical to the KT-Sub1, even though the base-pairing interactions in these motifs are not

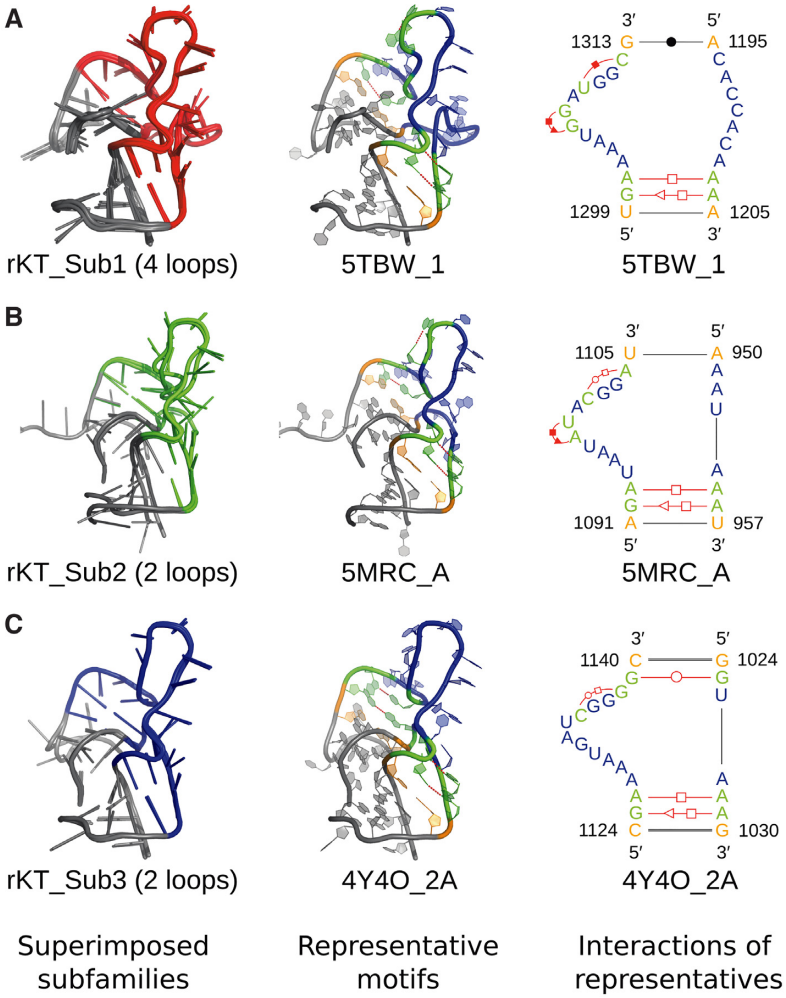

Figure 5. Subfamily result analysis for the reverse kink-turn motif family. The coloring scheme for the representative motifs is the same as Figure 4. Superimposed motifs, along with the 3D images and interactions of the subfamily representatives for (**A**) four motif instances of rKT-Sub1, (**B**) two motif instances of rKT-Sub2 and (**C**) two motif instances of rKT-Sub3.

similar. The structure of the shorter strand is different and has an S-shape twist, which is similar to the sarcin-ricin motif. This twist is present in most instances of this subfamily. Both the subfamilies KT-Sub3 and KT-Sub4 have additional structural features, including kinks in both strands. KT-Sub3 has three kinks, one in the short strand and two in the longer strand. A more biological investigation into the function of these complex kink is likely to bring interesting insight. The set of interactions in some of these subfamilies are significantly different from the traditionally defined kink-turn. However, given a subfamily, the interactions are very conserved, and those interactions help them to achieve the structural features of kink-turn. The details of interactions for all the instances of these subfamilies are provided in the Supplementary Website.

Reverse kink-turn

Reverse kink-turn has similar structural properties to the kink-turn motifs with its kink and the turn, but the direction of the kink is different (19). It turns toward the major groove while the kink-turn motif turns toward the minor groove. For this family, there are eight instances. We have identified three reverse kink-turn subfamilies with four, two and two instances. While all the subfamilies have the common kink and turn features of reverse kink-turn, each subfamily has some structural characteristics of their own. The superimposed motifs, along with the structural details of the representative motifs for each subfamily, are shown in Figure 5. It gives an explicit perspective on the visualization aspect of RNAMotifContrast. The structures here not only clearly shows the common structural features of the instances but also provides the variations among them explicitly.

By comparing rKT-Sub1 with rKT-Sub2 and rKT-Sub3, we can observe that the sequence and structure of the longer strands are very similar to each other. However, the shorter strand in rKT-Sub1 is very different in terms of the length and shape of the bulge. It has significantly more nucleotides than other two. As can be seen from Figure 5A, these extra nucleotides create an extended kink of the motif. This addition of structural features can be expected to have some additional functional implications.

All the subfamily representatives have two common non-canonical interactions - 'A/A H/H trans' and 'G/A S/H

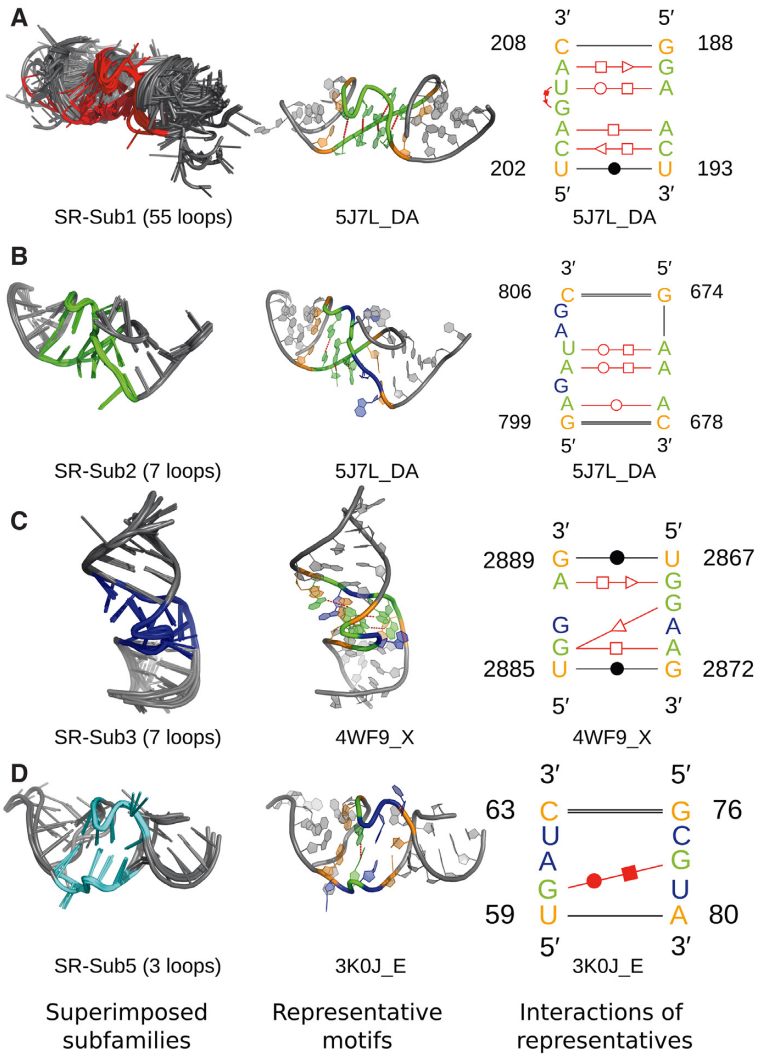

Figure 6. Subfamily result analysis for the sarcin-ricin motif family. The coloring scheme for the representative motifs is the same as Figure 4. Superimposed motifs, along with the 3D images and interactions of the subfamily representatives for (**A**) 55 motif instances of SR-Sub1, (**B**) seven motif instances of SR-Sub2, (**C**) seven motif instances of SR-Sub3 and (**D**) three motif instances of SR-Sub4.

trans', which corresponds to the consensus defined by Leontis *et al.* (36). rKT-Sub1 has one interaction that is in a similar location as rKT-Sub2 and rKT-Sub3, but the interaction is annotated as 'H/H *cis*' instead of 'W/H *trans*'. Both rKT-Sub1 and rKT-Sub2 have another common 'H/S *cis*' interaction which is absent in rKT-Sub3. Further biological analysis on these subfamilies can provide insight into the various roles the reverse kink-turn may play in living cells.

Sarcin-ricin

Sarcin-ricin is another very-well studied motif family and recognized site for critical functional interactions with proteins (13). It is an asymmetric internal loop with its characteristics S-like twist. For the 72 instances of the sarcin-ricin family, we have discovered four subfamilies with 55, 7, 7 and 3 instances. The visualization of the subfamilies is shown in Figure 6. The SR-Sub1 is the largest subfamily with 55 instances and represents traditionally well-known instances of the sarcin-ricin family. The representative motif of this subfamily contains the G/U/A base triplets that have been emphasized highly in (28) as strictly conserved. Overall, the motif shows two 'H/S *trans*', one 'W/H *trans*', one 'H/S *cis*' and one 'H/H *trans*' interactions. These interactions, along with the sequence of this motif, also match exactly with the known sarcin-ricin consensus (21). The other subfamilies (SR-Sub2, SR-Sub3, SR-Sub4) forms the characteristics S-turn of sarcin-ricin motifs but deviates from the well-known instances in terms of interactions and sequence properties. While some of the features correspond to the expected variations in the known flexible region of sarcin-ricin (28), some features are quite unique and show flexibility options in other regions too.

SR-Sub3 contains one 'H/S trans' and one 'H/H trans' interaction, which corresponds with the interactions in SR-Sub1. Compared to SR-Sub1 interactions, it is missing three interactions and contains an additional 'S/S *trans*' interaction. SR-Sub2 has a 'W/H *trans*' interaction, which corresponds with SR-Sub1, but that is the only interaction

match. SR-Sub2 additionally has another 'W/H trans' and a 'W/W trans' interaction. SR-Sub4 also has some interesting features. First of all, it only has one interaction—a 'W/H cis'. However, it has symmetric sarcin-ricin like structural twists in both strands, which is not the case for other subfamilies. Sarcin-ricin being a well-studied motif and known to have very conserved sequence and structure, the variations discovered through these subfamilies provide opportunities to explore new directions to understanding the properties of this motif family.

CONCLUSION AND DISCUSSION

In this manuscript, we presented RNAMotifContrast, a novel computational approach to discover RNA structural motif subfamilies. It takes the instances of a RNA structural motif family as input and generates one or more subfamilies as output. Additionally, optimized superimpositions of the motifs and side-by-side images are produced to visualize the variations among the motif instances of a family. The similarity graph to determine the relationship among the instances and the traversal algorithm to generate the superimposition are the two major features of RNAMotifContrast. The subfamilies we identified provide significant insights into the characteristics of RNA structural motif families compared to the resources of existing motif databases.

We have identified the sources of the structural variations among the subfamilies in terms of sequence and interactions. While our identified sources can explain most of them, there is a possibility that the variation for some instances come from the coordinate error due to resolution problems in the PDB data, or problems in interaction annotations, or the clustering result. Those are some inherent characteristics of the data we are using as input and not addressed in our method. However, only a small percentage of instances are likely to be characterized in those categories, and the effects on results are expected to be insignificant.

RNAMotifContrast provides the infrastructure for extensive future research to analyze the characteristics of RNA structural motifs. We have taken a significant step in utilizing this method and showing its effectiveness by focusing on a dataset which considers the union of motif instances from the work of Ge *et al*. and the RNA 3D Motif Atlas. In the future, more instances of motif families will improve the result even further, as more data will provide a better understanding of other possible variations and improve the estimate of the alignment length threshold. Additionally, this method can be configured to apply on any structure to compare and contrast, including longer RNA components and even proteins. The configuration required for this extension includes providing pairwise alignments as input and defining the set of atoms in the residues to be used for superimposition.

DATA AVAILABILITY

RNAMotifContrast source codes are available at the GitHub (https://github.com/ucfcbb/RNAMotifContrast). The detailed results are available on the Supplementary Website (http://genome.ucf.edu/RNAMotifContrast).

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

FUNDING

National Institute of General Medical Sciences of the National Institutes of Health (NIH NIGMS) [R01GM102515]. Funding for open access charge: NIH NIGMS [R01GM102515].
Conflict of interest statement. None declared.

REFERENCES

1. Eddy,S.R. (2001) Non-coding RNA genes and the modern RNA world. *Nat. Rev. Genet.*, **2**, 919–929.
2. Storz,G. (2002) An expanding universe of noncoding RNAs. *Science*, **296**, 1260–1263.
3. Rinn,J.L. and Chang,H.Y. (2012) Genome regulation by long noncoding RNAs. *Annu. Rev. Biochem.*, **81**, 145–166.
4. Chan,J.J. and Tay,Y. (2018) Noncoding RNA:RNA regulatory networks in cancer. *Int J Mol Sci*, **19**, 1310.
5. Adams,B.D., Parsons,C., Walker,L., Zhang,W.C. and Slack,F.J. (2017) Targeting noncoding RNAs in disease. *J. Clin. Invest.*, **127**, 761–771.
6. Mortimer,S.A., Kidwell,M.A. and Doudna,J.A. (2014) Insights into RNA structure and function from genome-wide studies. *Nat. Rev. Genet.*, **15**, 469–479.
7. Cruz,J.A. and Westhof,E. (2009) The dynamic landscapes of RNA architecture. *Cell*, **136**, 604–609.
8. Warf,M.B. and Berglund,J.A. (2010) Role of RNA structure in regulating pre-mRNA splicing. *Trends Biochem. Sci.*, **35**, 169–178.
9. 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.
10. Moore,P.B. (1999) Structural motifs in RNA. *Annu. Rev. Biochem.*, **68**, 287–300.
11. Leontis,N.B., Lescoute,A. and Westhof,E. (2006) The building blocks and motifs of RNA architecture. *Curr. Opin. Struct. Biol.*, **16**, 279–287.
12. Klein,D.J., Schmeing,T.M., Moore,P.B. and Steitz,T.A. (2001) The kink-turn: a new RNA secondary structure motif. *EMBO J.*, **20**, 4214–4221.
13. 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. U.S.A.*, **90**, 9581–9585.
14. García-Ortega,L., Álvarez Garía,E., Gavilanes,J.G., Martínez-del Pozo,A. and Joseph,S. (2010) Cleavage of the sarcin-ricin loop of 23S rRNA differentially affects EF-G and EF-Tu binding. *Nucleic Acids Res.*, **38**, 4108–4119.
15. Yesselman,J.D., Eiler,D., Carlson,E.D., Gotrik,M.R., d'Aquino,A.E., Ooms,A.N., Kladwang,W., Carlson,P.D., Shi,X., Costantino,D.A. *et al.* (2019) Computational design of three-dimensional RNA structure and function. *Nat. Nanotechnol.*, **14**, 866–873.
16. Zhong,C., Tang,H. and Zhang,S. (2010) RNAMotifScan: automatic identification of RNA structural motifs using secondary structural alignment. *Nucleic Acids Res.*, **38**, e176.
17. Zhong,C. and Zhang,S. (2015) RNAMotifScanX: a graph alignment approach for RNA structural motif identification. *RNA*, **21**, 333–346.
18. Sarver,M., Zirbel,C.L., Stombaugh,J., Mokdad,A. and Leontis,N.B. (2008) FR3D: finding local and composite recurrent structural motifs in RNA 3D structures. *J. Math. Biol.*, **56**, 215–252.
19. 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.
20. Leontis,N.B. and Westhof,E. (2003) Analysis of RNA motifs. *Curr. Opin. Struct. Biol.*, **13**, 300–308.
21. 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.
22. Zhong,C. and Zhang,S. (2012) Clustering RNA structural motifs in ribosomal RNAs using secondary structural alignment. *Nucleic Acids Res.*, **40**, 1307–1317.

23. Ge,P., Islam,S., Zhong,C. and Zhang,S. (2018) De novo discovery of structural motifs in RNA 3D structures through clustering. *Nucleic Acids Res.*, **46**, 4783–4793.
24. 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.
25. Leontis,N.B., Stombaugh,J. and Westhof,E. (2002) The non-Watson-Crick base pairs and their associated isostericity matrices. *Nucleic Acids Res.*, **30**, 3497–3531.
26. 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.
27. Šponer,J., Bussi,G., Krepl,M., Banáš,P., Bottaro,S., Cunha,R.A., Gil-Ley,A., Pinamonti,G., Poblete,S., Jurečka,P. *et al.* (2018) RNA structural dynamics as captured by molecular simulations: a comprehensive overview. *Chem. Rev.*, **118**, 4177–4338.
28. Havrila,M., Réblová,K., Zirbel,C.L., Leontis,N.B. and Šponer,J. (2013) Isosteric and nonisosteric base pairs in RNA motifs: molecular dynamics and bioinformatics study of the sarcin-ricin internal loop. *J. Phys. Chem. B*, **117**, 14302–14319.
29. Lilley,D.M. (2012) The structure and folding of kink turns in RNA. *Wiley Interdiscip. Rev. RNA*, **3**, 797–805.
30. Lilley,D.M. (2014) The K-turn motif in riboswitches and other RNA species. *Biochim. Biophys. Acta*, **1839**, 995–1004.
31. 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.
32. Leontis,N.B. and Zirbel,C.L. (2012) Nonredundant 3D structure datasets for RNA knowledge extraction and benchmarking. In: Leontis,N. and Westhof,E. (eds). *RNA 3D Structure Analysis and Prediction*, chapter 13, Springer-Verlag Berlin Heidelberg. pp. 281–298.
33. Major,F. and Thibault,P. (2007) RNA Tertiary Structure Prediction. In: Lengauer,T. (ed). *Bioinformatics – From Genomes to Therapies*, Vol. **1**, chapter 15, Wiley-VCH Verlag GmbH & Co. KGaA. pp. 491–539.
34. Huang,B. (2002) Finding maximum clique with a genetic algorithm. In: Master's thesis, Penn State Harrisburg.
35. 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.
36. Leontis,N.B., Lescoute,A. and Westhof,E. (2006) The building blocks and motifs of RNA architecture. *Curr. Opin. Struct. Biol.*, **16**, 279–287.