A comprehensive survey of long-range tertiary interactions and motifs in non-coding RNA structures

Davyd R. Bohdan¹, Valeria V. Voronina², Janusz M. Bujnicki³ and Eugene F. Baulin ³,*

¹Department of Innovation and High Technology, Moscow Institute of Physics and Technology, Dolgoprudny 141701, Russia, ²Department of Information Systems, Ulyanovsk State Technical University, Ulyanovsk 432027, Russia and ³Laboratory of Bioinformatics and Protein Engineering, International Institute of Molecular and Cell Biology in Warsaw, Warsaw 02-109, Poland

Received December 14, 2022; Editorial Decision June 30, 2023; Accepted July 07, 2023

ABSTRACT

Understanding the 3D structure of RNA is key to understanding RNA function. RNA 3D structure is modular and can be seen as a composition of building blocks of various sizes called tertiary motifs. Currently, long-range motifs formed between distant loops and helical regions are largely less studied than the local motifs determined by the RNA secondary structure. We surveyed long-range tertiary interactions and motifs in a non-redundant set of non-coding RNA 3D structures. A new dataset of annotated LONg-RAnge RNA 3D modules (LORA) was built using an approach that does not rely on the automatic annotations of non-canonical interactions. An original algorithm, ARTEM, was developed for annotation-, sequence- and topology-independent superposition of two arbitrary RNA 3D modules. The proposed methods allowed us to identify and describe the most common long-range RNA tertiary motifs. Along with the prevalent canonical A-minor interactions, a large number of previously undescribed staple interactions were observed. The most frequent long-range motifs were found to belong to three main motif families: planar staples, tilted staples, and helical packing motifs.

INTRODUCTION

Non-coding RNAs play important roles in many cellular processes (1). Like that of proteins, the functions of non-coding RNAs are largely determined by their spatial structure (2). RNA 3D structure is modular and can be seen as a composition of building blocks of various sizes called tertiary motifs (3,4). Such recurrent motifs are often functionally important and structurally conserved (5–8). Generally, all the motifs are defined based on hydrogen bonding interactions (6,9–12), hydrophobic stacking interactions (13–15), or both of them (16–19). The structural context matters for RNA tertiary motifs formation and can be leveraged for the prediction of their location (20–22).

In the context of RNA secondary structure, tertiary motifs can be divided into two groups - local motifs and long-range motifs. Local motifs include interactions within a secondary structure element or a number of adjacent elements, e.g., a GNRA tetraloop motif is a 4-nucleotide hairpin of a particular sequence and geometric configuration (16); a kink-turn motif is a sharp turn in phosphodiester backbone formed by two stems and an internal loop between them.

¹To whom correspondence should be addressed. Tel: +48 225970700 Email: ebaulin@ijmch.gov.pl; efbaulin@gmail.com

© The Author(s) 2023. 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
(17); a coaxial helix motif is a coaxial arrangement of two stems that usually share a common multiple junction (23). Therefore, local motifs are subject to a natural comprehensive classification based on five standard secondary structure elements: a hairpin, a bulge, an internal loop, a multiple junction, and a stem. Such classification is used in many databases that list local tertiary motifs annotated in known RNA 3D structures (24–29).

Long-range motifs are formed by interactions between distant structure elements and, as of today, they lack a natural classification. A few specific cases of such motifs were characterized, e.g., GAAA/11nt motif formed by a hairpin of GAAA sequence and an 11-nucleotide internal loop (30), D-loop/T-loop - an intricate network of interactions between two hairpins (7,31–33), and A-patch - a continuous stack of A-minor interactions (6). Various reports described motifs that can be either local or long-range (34), e.g., A-minor motifs (20), ribose zippers (9), ribose-phosphate zippers (35), and base-intercalated or base-wedged elements (bie/bwe) (13).

Unlike local motifs, long-range motifs do not have clear demarcations of where they begin and end. Therefore, the common strategy of their identification is to first annotate the involved interactions and then assemble the connected components of the participating residues (36). The majority of existing approaches are limited to base-pairing interactions (21,36–38). The annotation of the base-pairing interactions is usually performed with one of the five standard tools (39–43). A pseudoknot removal procedure is often used at the RNA secondary structure annotation step (36,38), e.g., using K2N algorithm (44). Furthermore, long-range motifs often overlap and share nucleotides with each other, e.g., GAAA/11nt motifs always include A-minor interactions (20), and A-minor interactions very often intersect with ribose zippers (31). Thus far, few studies analyzed this kind of inter-motif relationships (20,31).

There are different tools for the pairwise superposition of RNA 3D structures, which can be divided into two main groups based on the specific purpose they were developed to fulfill. The majority of the tools focus on producing a structure-based RNA sequence alignment (45–52) and therefore usually require single RNA chains as input (46,47,50). Another large group of tools is focused on the identification of local tertiary motifs in sets of RNA 3D structures and involves local superposition of single or multiple RNA chain fragments (53–56). Hence, these tools rely on some of the structural features of the local motifs, e.g., on sequence conservation, annotation of canonical and non-canonical interactions, and backbone topology. Among the least constrained of the currently available tools which can be considered of general use are SETTER (57) and CLICK (58). However, SETTER relies on RNA secondary structure elements, and CLICK is designed to produce RNA sequence alignment. To the best of our knowledge, up to the present moment, there was no tool specifically designed for the comparison of two arbitrary RNA 3D modules, able to find local and global structural similarities simultaneously, and independent of RNA sequence, RNA backbone topology, and annotation of any interactions.

In this work, we performed a comprehensive survey of long-range RNA tertiary interactions and motifs. We selected a dataset of long-range nucleotide doublets using a capacious <6.0 Å threshold and a generalized RNA secondary structure description model that allows pseudo-knots of any complexity (59). The doublets were grouped into long-range RNA 3D modules using a newly developed approach that does not rely on the annotation of any interaction types, only on spatial proximity. The assembled 3D modules were then used to derive motif instances and identify common long-range tertiary motifs via a new approach to the pairwise superposition of RNA 3D modules. The results confirm that A-minor-like motifs are the most populated types of long-range contacts in non-coding RNA 3D structures, but also discover a number of previously undescribed interactions. Furthermore, the results show large disagreements between the standard automatic RNA structure annotation tools.

Our software tools based on novel algorithms make it possible to perform sequence-, topology- and annotation-independent description, analysis, and superposition of the RNA structural modules and are not limited to long-range motifs. They can be used to analyze RNA structures as well as RNA-containing complexes.

MATERIALS AND METHODS
Dataset of long-range nucleotide doublets

To prepare a dataset of long-range nucleotide doublets from experimentally determined RNA 3D structures we used a custom non-redundant set of RNA structures from the Protein data bank (PDB, (60)) that included one RNA chain per non-coding RNA family from Rfam (61). The non-redundant set of RNA 3D structures was built as follows. From Rfam (version 14.7) we downloaded the non-coding RNA family identifiers matching RNA chains from PDB. In case of multiple hits per RNA chain, we considered only the family with the lowest E-value. The further choice of non-redundant RNA structures was limited to the RNA chains from the BGSU representative set (62) (version 3.215 with 4.0 Å resolution cutoff) in order to eliminate structures of low quality and facilitate direct comparisons between the sets. Finally, to construct a non-redundant set of RNA 3D structures, we manually selected one representative RNA chain per Rfam family based on three features: resolution, number of resolved residues, and number of base pairs of any type, according to the annotation using the DSSR tool (40). By default, we selected the chains with the best resolution, but in several cases, we did otherwise, e.g., when the best-resolution chain was a short fragment of the complete molecule, or when the second-best resolution chain was significantly longer and the difference in resolution was negligible. The resulting set included the structures of 97 non-coding RNA families from 89 PDB entries (Supplementary Table S1).

We used a non-redundant set of RNA structures with one structure per non-coding RNA family from Rfam (61) rather than the more commonly used representative set of RNA structures with one structure per pair of a source organism and an RNA molecule type (62). This allowed us to significantly reduce the share of ribosomal RNA (rRNA) residues within the dataset. Thus, among the representative structures linked to Rfam families the share of
non-rRNA chains is 71.5% and the share of non-rRNA residues is only 18.26%, whereas in the used non-redundant subset the share of non-rRNA chains is 88.7% and the share of non-rRNA residues is 41.62% (Supplementary Table S1).

To annotate RNA secondary structure elements and nucleotide doublets we used the urslib2 python library (https://github.com/febos/urslib2) along with DSSR annotations of canonical (i.e., Watson–Crick and wobble) base pairs. The library includes an implementation of a generalized RNA secondary structure description model (59) and can take as input a coordinate file in PDB or mmCIF format along with a tool-independent list of canonical base pairs. The description model handles pseudoknots of any complexity allowing us to skip any pseudoknot removal procedures.

We defined a nucleotide doublet as any two ribonucleotides that have any non-hydrogen atoms within 6.0 Å of each other. The threshold of 6.0 Å was chosen to guarantee the inclusion of all possible direct and indirect (e.g., water- or ion-mediated) interactions that would allow the correct assembly of the entire interacting interfaces. The minimal distance between the atoms of two nucleotides in a doublet was assigned as the distance between the doublet nucleotides. Then, we defined a long-range nucleotide doublet as a nucleotide doublet of two ribonucleotides that belong to distant RNA secondary structure elements, stems or loops. For that, we used the definitions from (59) (Figure 1).

According to these definitions, a stem is defined as a maximal set of $N \geq 2$ consecutive Watson–Crick (A–U, G–C) or wobble (G–U) base pairs. An RNA chain is then represented as an alternating sequence of stem and non-stem regions (Figure 1), with non-stem regions possibly being of zero length (e.g., absent non-stem region b on Figure 1). Each stem (S) confines a set of regions that form a loop. Dangling ends are treated as a loop of a dummy stem that confines the entire sequence. A stem and a loop are called adjacent if at least a pair of their regions are neighbors in the RNA sequence. In relation to adjacent stems, each loop belongs to one of four types: a hairpin (H), a bulge (B), an internal loop (I) or a multiple junction (J). A bulge is a special case of an internal loop that involves a region of zero length. In relation to pseudoknots, each loop belongs to one of three classes: classical loop (C, distant from pseudoknots), isolated loop (I, adjacent to a pseudoknot), or pseudoknotted loop (P, part of a pseudoknot). Then, two RNA secondary structure elements, a stem and a loop, two stems, or two loops, are called distant from each other if they are not adjacent and the sets of their adjacent elements do not overlap, i.e., they do not share a third element that is adjacent to both. For example, on Figure 1 HC2 and S3 are distant, but HP3 and J10 are not distant as they are both adjacent to S3. For the strict formal description of the used definitions see (59) or Supplementary Text 1 in (20).

In case of more than one RNA chain forming together an RNA secondary structure, an intermolecular long-range nucleotide doublet was added to the dataset if at least one of its nucleotides belonged to an RNA chain from the non-redundant set of RNA structures. Overall, the formed dataset included 17272 long-range nucleotide doublets formed by 70 RNA chains from 62 PDB entries of the non-redundant set (Supplementary Table S2).

### Annotating motifs and interaction types

To annotate the long-range doublets with the commonly known motifs and interaction types, we used five conventional tools: DSSR (40), FR3D (39), MC-annotate (41), RNAView (43) and ClaRNA (42) (Table 1). For a number of common motifs, we also used the urslib2 library. Unfortunately, we could not use the NASSAM tool (63), as there was only a web server version available and it failed to work on some of the PDB entries (e.g., 6HIW).

Where applicable, we also annotated motifs and interaction types involving only one of the nucleotides of a doublet. In the case of FR3D, we parsed the annotations available in the RNA Structure Atlas (24) without running the tool explicitly. To run PDB-only tools on mmCIF-only PDB entries we used pdb-bundles provided by RCSB PDB (60). We applied the default score threshold ≥0.5 for all ClaRNA annotations used in this work. Separately, we also stored ClaRNA base pair annotations of any score.

### Assembly of long-range RNA 3D modules

To assemble long-range RNA 3D modules from the long-range nucleotide doublets we used the following approach: two doublets were considered to belong to the same module if (i) they shared one common residue and (ii) the minimal distance between the non-hydrogen atoms of the other two residues was less than an ad hoc threshold of 6.0 Å. The second rule prevents forming of large clumps composed of close but independent interaction networks. Then, the modules were defined as connected components of a graph with the doublets as vertices and belonging to the same module as edges.

Overall, the 17272 long-range doublets were grouped into 1264 long-range RNA 3D modules. The modules formed a dataset called LORA (LOnge-RAnge). The LORA modules are stored both in PDB and mmCIF formats and
| Tool          | Reference       | Link                                                                 | Version / access date | Allowed format | Command                                                                 | Dependencies | Motif / Interaction type | Involving the long-range doublet | Involving each residue separately |
|---------------|-----------------|----------------------------------------------------------------------|-----------------------|----------------|--------------------------------------------------------------------------|--------------|--------------------------|-----------------------------------|----------------------------------|
| MC-Annotate   | Gendron et al. 2001 | [https://major.iric.ca/MajorLabEn/MC-Tools.html](https://major.iric.ca/MajorLabEn/MC-Tools.html) | standalone 1.6.2 PDB  | MC-Annotate inpfile > outpfile | -                                                   | base pair              | yes                      | yes                                |                                  |
| RNAView       | Yang et al. 2003 | [http://ndbserver.rutgers.edu/ndbmodule/services/download/rnaview.html](http://ndbserver.rutgers.edu/ndbmodule/services/download/rnaview.html) | Standalone June 2022 PDB | maval inpfile | -                                                                      | base pair             | yes                      | yes                                |                                  |
| FR3D          | Sarver et al. 2008 | [http://rna.bgsu.edu/ma3dbh/pdb/1XJR/interactions/fr3d/all/csv](http://rna.bgsu.edu/ma3dbh/pdb/1XJR/interactions/fr3d/all/csv) | web-database June 2022 | -              | -                                                                      | base pair             | yes                      | yes                                |                                  |
| NASSAM        | Hamdani et al. 2012 | [http://211.25.251.163/nassam/](http://211.25.251.163/nassam/) | web-server June 2022 PDB/mmCIF | -              | -                                                                     | -                             | -                        | -                                  |                                  |
| ClRNA         | Wale et al. 2014 | [http://genesilico.pl/clarna/](http://genesilico.pl/clarna/) | Standalone July 2022 PDB | python27 clarna.py -i inpfile > outpfile | Simplejson networkx scapy biopthon==1.76 | base pair                      | yes                      | yes                                |                                  |
| DSSR          | Lu et al. 2015 | [http://forum.x3dna.org/rna-structures/](http://forum.x3dna.org/rna-structures/) | Standalone v2.0.0–2020aug01 PDB/mmCIF | x3dna-dssr-2 -i = inpfile -format = mmCIF -dsdr = long – u-turn –more – non-pair –po4 -tminor = N -o = outpfile | -         | residue conformation (syn/anti + sugar pucker) | -                                  |                                  |
| urssl2        | -               | [https://github.com/febos/urssl2](https://github.com/febos/urssl2) | Standalone May 2022 PDB/mmCIF | see [https://github.com/febos/urssl2/blob/main/playground.ipynb](https://github.com/febos/urssl2/blob/main/playground.ipynb) | DSSR         | bik/bwe                  | yes                                | yes                                |                                  |

Table 1. Overview of the tools and motifs and interaction types used to annotate the dataset of long-range nucleotide doublets.
available in the form of a GitHub repository (https://github.com/fbos/LOR.A). Each module is assigned an identifier composed of a PDB entry and an ordinal number separated by an underscore, e.g., ‘707y_2’.

**Pairwise superposition of RNA 3D modules**

To superimpose two arbitrary RNA 3D modules we developed a new algorithm called ARTEM (Aligning RNA Tertiary Motifs). The algorithm relies on an assumption that in the best possible superposition of two similar RNA 3D modules at least one residue has a counterpart with RMSD close to zero, and works as described in Figure 2.

For two RNA modules X and Y of sizes N and M, the procedure is performed for each possible pair of matched residues. First, we superimpose the two modules using the Kabsch algorithm (64) based on the 5-atom representations of the two residues. Second, a subset of mutually closest residues at a distance < MATCH RANGE Å from each other is defined based on the centers of mass of their 5-atom representations. Third, superposition by the Kabsch algorithm is performed based on the 3-atom base representations of the residues from the mutually closest subset. At the last step of the procedure, three objects are yielded: the superimposed module Y (Y’), the size of the mutually closest subset (SIZE), and RMSD of the atoms used for the second superposition.

Here, module X is considered a reference, and module Y is a query, hence the module X coordinates are constant. For the 5-atom representation, we use a scheme close to the one used in SimRNA (65): the representation includes the P atom, the center of mass of the ribose atoms, and three base atoms—N9, C2 and C6 for purines, and N1, C2 and C4 for pyrimidines. For the 3-atom base representation, the same three base atoms are used. The MATCH RANGE threshold was chosen manually and set to 3.0 Å to eliminate random ‘noise’ matchings. The 3-atom representation is used in step 3 instead of the 5-atom representation to facilitate the superposition of similar base arrangements having different backbone topologies. For all the modified residues present in the used set of RNA structures we assigned 5-atom representations manually by analogy with their closest standard residues.

ARTEM is able to identify similar 3D arrangements of bases in RNA 3D structures regardless of their sequence, backbone topology, and annotated interaction types. The theoretical time complexity of ARTEM is \( O(n^4) \), where \( n \) is the total number of atoms in the two modules, as the Kabsch algorithm of complexity \( O(n^2) \) (66) and the search for mutually closest residues of complexity \( O(n^3) \) are being run \( O(N^2M) \approx O(n^2) \) times, giving \( O(n + n^2)O(n^2) = O(n^4) \). A parallelized Python implementation of ARTEM is available in the form of a GitHub repository (https://github.com/david-bogdan-r/ARTEM). The implementation uses k-d trees (67) for the mutually closest residues search which on average works better than the calculation of all the pairwise distances running \( O(n^2) \) in time. The tool reads and writes both PDB and mmCIF formats. The user is able to specify the particular residues of interest or run ARTEM for the entire coordinate files. Also, the user can choose to save the superimposed query structures in the preferred format.

The implementation with default parameters takes around 1 min to run an entire 5970-residue eukaryotic ribosome (PDB entry 707Y) against a 160-residue TPP riboswitch (PDB entry 2GDI) on 32 cores, taking under 2Gb RAM at peak on an AMD Ryzen 9 5950X machine with 128Gb RAM (command: ‘python artem.py r = 707y.cif q = 2gdi.cif threads = 32’). On the same machine on 32 cores, a run of a 2828-residue LSU rRNA (PDB entry 1FFK) against itself requires 20 min in time and 70 Gb of RAM.

**Identification of long-range RNA tertiary motifs**

We used a modified version of ARTEM to build a dataset of long-range motif instances derived from the LORA modules. The dataset of motif instances was then used to identify the recurrent long-range motifs.

In the modified version of ARTEM, we employed a value of RESRMSDMAX, which is the maximum of RMSD values calculated separately for each pair of matched residues of a given superposition. In step three of ARTEM, the 5-atom residue representation was used instead of the 3-atom representation. We also utilized an additional trimming procedure during step three of ARTEM, where we iteratively remove residues from the matched subset one by one until either a RESRMSDMAX threshold is satisfied or the subset is empty. The usage of 5-atom representation along with a low RESRMSDMAX threshold allows to avoid the hits between close motifs with different interaction networks, and the usage of the trimming procedure eliminates missing a smaller hit because of a larger hit that includes the smaller one but does not satisfy the thresholds. The modifications are incorporated in the ARTEM GitHub package as non-default resrmsdmax and trim parameters (https://github.com/david-bogdan-r/ARTEM).

We defined a **tertiary motif instance** as a LORA module sub-structure (a subset of residues) that involves at least one long-range doublet and produces at least one good structural superposition with any other sub-structure of another LORA module using the ARTEM modification. A good structural superposition was defined as a superposition of RESRMSDMAX < 1.0 Å. Thus, a tertiary motif instance according to the above definition can be considered a reflection of local structural similarity between two LORA modules. The threshold was chosen based on the visual examination of the results from preliminary experiments. Stricter thresholds resulted in an increase in the number of false negatives, i.e., missing relevant structural similarities, e.g., between GC and CG Watson–Crick base pairs with RESRMSDMAX ≈ 0.9 Å. In contrast, the relaxed thresholds significantly increased false positives, i.e., accidental close residue arrangements of completely different interaction networks, e.g., Type I and Type II A-minors. Therefore, the threshold RESRMSDMAX < 1.0 Å was chosen to minimize the number of false positives and keep the number of false negatives as limited as possible.

We then defined a **long-range tertiary motif instance** as a sub-structure of a tertiary motif instance obtained by removing the residues not involved in any long-range doublet within the motif instance. Such truncation helps to ignore the differences between the motif instances caused solely by non-relevant peripheral residues. With the optimized
threshold value, we performed a long-range motif instance search for each LORA module versus all the LORA modules. Overall, we identified almost half a million (489999) long-range tertiary motif instances. Each motif instance was assigned a unique identifier comprising a source module identifier and an ordinal number; e.g., ‘6ih6_18.261’ is an instance number 261 derived from module 6ih6_18.

To derive the most frequent long-range motifs from the long-range motif instances of sizes from 2 to 20 we applied the following clustering procedure separately for each motif size. For each motif instance of size N having at least ten hits found during the motif instance search, we counted its matches, i.e., the other instances of size N that produce a good superposition of size N with the current instance using the ARTEM modification. The motif instances were then sorted in the decreasing number of their matches. A motif instance along with its matches were considered a cluster center and its cluster. We should note here that a particular motif instance can be a match for more than one cluster center. Then, a cluster was removed from the final set if it included at least one motif instance matching at least one center of a bigger cluster. This was done to eliminate highly overlapping clusters. The centers of the resulting set of clusters were considered candidate long-range tertiary motifs. From the candidate motifs, recurrent long-range tertiary motifs were then selected manually. And finally, we used the ARTEM modification to annotate the recurrent motifs in LORA modules with manually identified motif-specific RMSD thresholds. The LORA GitHub package (https://github.com/febos/LORA) includes the 489999 motif instances in the form of a table listing the residue sets, all the cluster members in PDB format, and the recurrent motifs in both PDB and mmCIF formats.

RESULTS

Summary statistics of long-range nucleotide doublets

In this work, we defined a long-range nucleotide doublet as two ribonucleotides from distant RNA secondary.

Figure 2. Pseudocode and illustration of the ARTEM algorithm for the pairwise superposition of arbitrary RNA 3D modules. The three-step procedure is performed for all possible single-residue matchings. Step 1: superimpose the modules based on the two residues. Step 2: identify the set of mutually closest residues. Step 3: superimpose the modules based on the mutually closest residues. The two RNA 3D modules are local A-minors from an across-bulged motif (beige) and long-range A-minors from a GNRA-tetraloop/receptor motif (light blue) described in (20).
structure elements, stems or loops, with any non-hydrogen atoms within 6.0 Å of each other, see the Materials and Methods section. The collected dataset included 17272 long-range doublets (Supplementary Table S2) formed by 11204 nucleotides from 70 RNA chains of the non-redundant set of RNA structures (Supplementary Table S1). 16078 (93%) doublets were formed by 10714 (95.6%) nucleotides of a standard residue type (32.3% G, 26.3% A, 23.4% C, 18% U) that were assigned with a particular conformation (anti/syn ~C3-end/~C2-end) using DSSR (40). The dataset included 221 (2%) modified residues. 1554 (9%) of the long-range doublets are intermolecular doublets of nucleotides from different RNA chains forming a shared RNA secondary structure.

The nucleotide residues in all the doublets analyzed are almost equally split between stems (48%) and loops (52%). However, more than 80% of the adenosines belong to loops, whereas 69% of the cytidines and 61% of the guanosines belong to stems, and uridines show the most equal split with 57% of loop residues. Pseudoknotted stems and loops contribute only 16% of the residues in the doublet, with an almost two-fold difference between stems (5.5%) and loops (10.5%). Among the 84% of classical (pseudoknot-free) stems and loops the largest share belongs to stems (42%), followed by hairpins (18%), internal loops and bulges (15% combined), and multiple junctions (9%).

Stem-loop doublets form 55% of the dataset, and loop-loop and stem-stem doublets form 30% and 15% respectively. A(loop)–G(stem) and A(loop)–C(stem) are the most common long-range doublets representing together 20.3% of the dataset (11.8% + 8.5%). The third most common doublet is A(loop)–G(loop) (5.8%). The most frequent pairs of RNA secondary structure elements among the long-range doublets are stem-hairpin (16.8%), stem-internal loop (12.4%), stem-stem (11.7%), stem-multiple junction (7.6%), and hairpin-hairpin (5.4%).

Among the closest atoms of the long-range doublets, the most frequent are O2’ (20%) and OP1 (14.3%), followed by O4’ (8.8%). Among base-specific atoms, the most frequent are N2G (5.3%), C2A (3.9%), N1A (3.3%) and N6A (3.2%). The most frequent pairs of closest atoms in the long-range doublets are O2’–OP1 (7%) and O2’–O2’ (5.6%), and C2A-N2G (1.3%) is the most frequent base-specific pair of atoms.

Automatic annotation tools perform poorly for assembling long-range RNA 3D modules

The long-range doublets were annotated with the previously defined interaction and motif types using the five conventional automatic annotation tools: DSSR (40), FR3D (39), ClaRNA (42), RNAView (43) and MC-Annotate (41), see the Materials and Methods section.

DSSR annotated 3244 (18.8%) doublets as non-pairing H-bonds (H-bonds that are not a part of canonical or non-canonical base pair) and 1345 (7.8%) doublets as A-minor interactions. The five tools annotated on average 1153 (6.7%) doublets as base pairs, ClaRNA being the most conservative tool (553 bps) and RNAView the least conservative (1625 bps). Simultaneously, RNAView annotated as base stacking just 26 doublets, whereas FR3D annotated 819 (4.7%) doublets, and the other three tools on average annotated 446 doublets.

Each of the annotated interaction and motif types is dominated by loop adenosines composing from 25% (non-pairing H-bonds) to 56.5% (base-intercalated elements (13)) of the interacting residues. Among all the 11204 nucleotides forming the long-range doublets loop adenosines compose 21%. Base stacking interactions are dominated with A(loop)–A(loop) doublets (17%–36%). Almost all of the other interaction and motif types are dominated by A(loop)–G(stem) doublets (12%–38%), whereas A(loop)–C(loop) doublets prevail in DSSR ribose zippers (36%) and non-pairing H-bonds (13%).

Strikingly, among 2919 long-range doublets annotated as a base pair by at least one tool, only 13% are consensus five-tool annotations and almost 49% are single-tool annotations.

We further examined the base pair annotations separately for each Leontis–Westhof type (11). The Leontis–Westhof type (LW type) is a three-letter code specifying the first letters of the two interacting base edges (Watson–Crick/Hoogsteen/Sugar) along with their relative configuration (cis/trans), e.g., tHS is an interaction between Hoogsteen edge of one base and Sugar edge of another base with the bases being in trans configuration. Among the base pairs of determined LW type annotated in the long-range doublets, four types cSS/tSS/cSW/tSW compose from 60% to 83% depending on the tool, and types that involve a Hoogsteen edge compose just 7%–16%. Surprisingly, single-tool annotations dominate ten of the twelve possible LW types, composing more than 50% in eight of the ten (Supplementary Figure S1). Two exceptions, cWW and tWW types are dominated by four-tool (50.6%) and five-tool (38.6%) annotations respectively.

We discovered 180 long-range doublets annotated with different base pair LW types by different tools (Supplementary Table S3, Figure 3). A single example of a doublet annotated with three different LW types was found (Figure 3D) and the other 179 doublets were annotated with two LW types. 148 (82%) of such doublets involve adenosines (66 Å–G, 53 Å–C, 17 Å–A, 12 Å–U), 143 (79%) are SS/SW disagreements (Figure 3A, B), and for 91 of them, among other tools, FR3D annotates an SS type and DSSR annotates an SW type. Notably, cSS/tSS/cSW/tSW base pairs are common among ubiquitous A-minor interactions, hence their annotation is of the highest importance for the assembly of long-range RNA 3D modules.

Furthermore, 10379 (60%) of the long-range doublets were not annotated with any common interaction type using any of the tools. It means they would be totally missed in the annotation-based assembly of RNA 3D modules, and that includes 2095 doublets (12%) that involve at least one pair of non-hydrogen atoms under 4.0 Å.

An overview of the long-range RNA 3D modules dataset

The dataset of 17272 long-range doublets was used to assemble the LORA (LOng-RAnge) dataset of 1264 long-range RNA 3D modules. For that, we developed an original
approach that does not rely on the automatic annotations of non-canonical interactions, see the Materials and Methods section. The LORA dataset included 389 modules of size two (i.e., composed of two residues), 183 modules of size three, 686 modules of sizes from 4 to 81 (Supplementary Figure S2), and six rRNA modules larger than 100 residues in size. A unique identifier made of a PDB entry and an ordinal number was assigned to each module, e.g., ‘7o7y_2’ (Supplementary Table S2).

To assess the suitability of the module assembly we examined instances of the well-known ‘named’ long-range RNA tertiary motifs (Figure 4). Module 6ugg_0 (15 residues) is a canonical instance of the D-loop/T-loop interaction motif from tRNA (Figure 4A). Module 6ugg_0 includes two long-range base pairs G17-U55(tWS) and G18-C56(cWW), both annotated with all five tools, and an independent long-range base stacking G14-G59 annotated with all the tools except RNAView.

Module 7m4y_45 (21 residues) is a double symmetrical A-patch formed between 23S and 5S rRNAs (Figure 4B). The A-patch is the only conserved tertiary interface between 23S and 5S rRNAs, and it is conserved throughout all the kingdoms of life (6). Our dataset includes three instances of this A-patch - modules 7m4y_45, 6th6_57 and 7o7y_11. In module 7m4y_45 5S rRNA purines A76, A96 and G97 interact with the minor groove of 23S rRNA near base pair G860-C913(cWW), and in turn, 23S rRNA purines A859, A915 and A916 interact with the minor groove of 5S rRNA near base pair G77-U94(cWW). Here, DSSR and RNAView annotate the only long-range base pair G77-A915(tSS) and DSSR annotates zero long-range A-minors. In turn, FR3D annotates three base pairs of ‘near’ geometry (i.e., interactions close to a base pair but not meeting the strict criteria): G77-A859(ncSS), A96-G860(ntSS), and G77-A915(ntSS), MC-Annotation annotates four ‘pairings’ between an O2’ atom and a Sugar edge (G77-A859, A76-G860, U78-A916, and G93-A916), and ClaRNA annotates G77-A915(tSS) base pair with score 0.28. Hence, no automatic annotation tool identified the entire A-patch involving six purine bases.

Module 2z75_1 (24 residues) is an A-patch from glmS ribozyme, which involves a stack of six purine bases (A129, G128, A127, A104, A105, A106) interacting with the minor groove of canonical base pairs A28-U51, G37-C52, C36-G53 and U59-A54 (Figure 4C). Here, DSSR annotates three long-range base pairs, but simultaneously it annotates five A-minors involving five of the six purine bases, except A106. FR3D annotates seven long-range base pairs of ‘near’ geometry which involve five of the six purine bases, except A129. MC-Annotation too annotates long-range interactions involving five of the six purines but not A129. ClaRNA annotates three of the purines involved in long-range base pairs with a score above the threshold and three more as part of base pairs with scores below the threshold. RNAView is the only tool able to annotate here long-range base pairs involving all six purine bases.

Module 6th6_41 (16 residues) is an instance of the GNRA-tetraloop/tetraloop receptor interaction motif (Figure 4D). Here a GAAA tetraloop (G769, A770, A771, A772) interacts with the minor groove of canonical base pairs U2451-A2475 and G2452-C2474. Here all the tools annotate the long-range interaction involving A770, but RNAView and MC-Annotation fail to annotate the long-range interaction involving A771.

Overall, none of the five automatic annotation tools is able to fully annotate more than two of the four described RNA 3D modules, and with all of the tools simultaneously still it would be possible to annotate only three of the four modules.
Exploring sequence-, topology-, and annotation-independent structural similarities of D-loop/T-loop-like motifs

To compare the assembled long-range RNA 3D modules we developed a new algorithm for pairwise structural superposition, ARTEM (Aligning RNA TErtiary Motifs). The ARTEM algorithm is designed to find structural similarities between two arbitrary RNA 3D modules of possibly different sizes and without a priori knowledge of the nucleotide matchings between the modules, see the Materials and Methods section.

To demonstrate the capabilities of the ARTEM algorithm we performed a search for RNA 3D modules structurally similar to the D-loop/T-loop interaction motif from tRNA represented in the LORA dataset with module 6ugg_0, (Figure 5A, C). We searched for the similarities with RMSD ≤ 3.0 Å, SIZE ≥ 7 residues, and RMSD/SIZE ratio ≤ 0.25 that involve at least one long-range nucleotide doublet within the superimposed subset of residues. These thresholds were chosen to filter out non-relevant hits, e.g., two- or three-residue hits with RMSD close to zero, or hits with no long-range doublets. Overall we found 32 hits (Supplementary Table S4, Figure 5B), among which the D-loop/T-loop-like motifs were found in tRNA (modules 3rg5_0, 3q1q_0), Y RNA (6cul_0), virus tRNA-like UTR (6ni0_0), Hatchet ribozyme (6jq5_0), RNase P (2a64_0), several riboswitches (e.g., 4ffu_1 and 3f2x_1), and multiple hits were found in archaeal, bacterial and eukaryotic rRNAs.

Three of the 32 motifs found are tetraloops (GAAA in 6th6_41, GCAA in 7m4y_3 and UAAC in 3q1q_3) forming a nested non-canonical base pair in place of the long-range base pair U55-G17(tSW) in the D-loop/T-loop motif (Figure 5D). The found tetraloops show that ARTEM can find similar RNA base arrangements in a topology-independent manner. Also, out of 29 non-tetraloop hits, only 14 are hairpin-hairpin interfaces as is the reference module 6ugg_0. Two previously known D-loop/T-loop-like motifs from tmRNA and group II intron (7) are absent in the LORA dataset, as they are formed with two loops having in common one adjacent stem, hence classified as not long-range but local interactions according to the used definitions. We then confirmed ARTEM’s capability to find these two motifs by running it for module 6ugg_0 against the entire tmRNA and group II intron molecules (Supplementary Table S4). However, a D-loop/T-loop-like motif could not be identified in the group II intron representative structure from the used non-redundant subset of RNA structures (PDB entry 7D1A, chain A), and its absence was confirmed by visual examination, but ARTEM identified the motif in

Figure 4. Examples of the assembled long-range RNA 3D modules belonging to ‘named’ tertiary motifs. (A) D-loop/T-loop interaction motif, module 6ugg_0, tRNA(Asp), PDB entry 6UGG, chain B; (B) Double symmetric A-patch, module 7m4y_45, 23S and 5S rRNA, PDB entry 7M4Y, chains A (light pink) and B (beige); (C) A-patch, module 2z75_1, glmS ribozyme RNA, PDB entry 2Z75, chain B; (D) GNRA-tetraloop/tetraloop-receptor, module 6th6_41, 23S rRNA, PDB entry 6TH6, chain BA.
solving the conserved structural motifs in RNA structure modeling.

**Canonical A-minor interactions and previously undescribed staple interactions form the most common long-range RNA tertiary motifs**

To identify commonly occurring long-range RNA tertiary motifs we applied the ARTEM algorithm to the dataset of LORA modules. First, we identified nearly half a million motif instances and then clustered them to derive the most populated motif types. We assigned unique identifiers to motif instances, e.g., motif instance ‘7m4y_128_455’ of module ‘7m4y_128’, see the Materials and Methods section.

Overall, we identified 489999 motif instances of sizes from 2 to 356 (Supplementary Figure S3). The further analysis was limited to motif sizes from 2 to 20, as only a negligible quantity (4147) of instances were larger than 20 residues in size. After applying the heuristic clustering procedure we identified 26932 clusters of motif instances of sizes from 2 to 20 residues. Among the derived cluster centers we manually selected 45 motifs of sizes from 2 to 10 that are most likely to be ‘units of recurrence’ (Supplementary Table S5), i.e., do not involve non-interacting residues and are not only found in ribosomal RNAs. Among the 45 motifs 29 explicitly represent interactions of the minor groove of a helical region and all the other 16 motifs include interactions involving Sugar edges of the residues. Such interactions are known to be by far the most common long-range interactions in non-coding RNA 3D structures (6,8).

Analyzing the recurrent motifs, we noted a number of motifs with several (from one to four) consecutive residues interacting simultaneously with both helical strands at several (from one to four) base pair layers (Figure 6). Out of the 45 long-range motifs we counted 27 such motifs we called ‘staples’. In every staple, at least one atom (most often O2') of each helical strand is involved in an interaction with the third strand, the staple strand. If the interactions involved only ribose atoms of a staple residue, we included ‘N’ in its name to emphasize the lack of base specificity, e.g., UAN4-staple (motif 44) is an interaction of a helix region with three bases and a ribose spanning four base pair layers of the helix (Figure 6A). If staple residues involved cross-strand stacking, we included a pipe symbol in its name, e.g., GA|AG3-staple (motif 43, Figure 6B).

According to the proposed naming. Type I A-minor is usually a case of A1-staple, as here interactions are limited to a single base pair layer and are highly specific to the adenine base. However, in some structural contexts Type I A-minor interaction can be an A2-staple, see e.g., module 7m4y_82. In total, 19 out of the 45 motifs included A-minor interactions, and 18 of them are also staples with a single exception of isolated Type II A-minor (motif 10). One of the most populated staple motifs not including A-minor interactions of any type is GN3-staple (Figure 6C), and a much less common but interesting example of such motifs is G3-staple formed with a single guanosine residue interacting with three base pair layers simultaneously (Figure 6D). Thus, we believe A-minor interactions are a special (and most common) case of a more general interaction type. Moreover, out of the 18 non-staple motifs, 15 are
staple fragments involving only one of the two helical strands, e.g., base pairs and ribose-ribose interactions. And in the other 3 motifs staple interactions form only a fragment and not the entire motif, e.g., motif 26 formed by two canonical base pairs interacting via their minor grooves (Figure 6E).

**Frequency and variability of the identified long-range tertiary motifs**

To quantitatively assess the frequency of the selected recurrent motifs among LORA modules we performed a search using the ARTEM algorithm modification. For each motif, we manually defined the annotation threshold parameter, either RMSD or RESRMSDMAX (maximum per-residue RMSD), and the threshold value, to minimize the number of false positive matches. Overall, 3880 long-range motif matches were identified (Figure 7A, Supplementary Table S6).

As expected, the most frequent motif was found to be cSS base pair (motif 1) with 515 matches (Supplementary Table S7), the base pair commonly found in both Type I and Type II A-minor interactions, and in ribose zippers. For verification purposes, we analyzed annotations of the five conventional tools for the 515 matches. FR3D annotated 277 matches as cSS base pair, 96 matches as ncSS base pair (base pair of ‘near’ geometry), and 2 matches as ncSW base pair (375 matches in total). The second most inclusive annotation is the DSSR A-minor annotation - 368 matches. The most strict annotations are base pairs by Clarna (120 matches) and base pairs by DSSR (154 matches in total, 116 matches of cSS type). Overall, only 20 matches of motif 1 (4%) were not annotated as anything by any tool.

The top 17 most frequent long-range motifs are related (include or are fragments of) to A-minor interactions and ribose zippers, and the most common motif of type undescribed previously is GN3-staple with 75 matches (motif 33, ranked 18th, Figure 6C). Interestingly, a large motif of 10 residues, UAAN4-staple (motif 44) was found to be unexpectedly frequent with 68 matches (ranked 21st, Figure 6A). The two rarest motifs are G3-staple (motif 36) and an unusual NAA4-staple involving spliced-apart staple residues (motif 45), both having only four matches found solely in ribosomal RNAs.
Then we clustered the 45 motifs by their pairwise overlap coefficients to identify ‘families’ of largely overlapping motifs (Figure 7B). We considered two motif instances overlapping if one instance is fully nested in the other. Then, for each pair of motifs X and Y, we defined the overlap coefficient as the maximum of two ratios - the ratio of X instances overlapping with Y instances and the ratio of Y instances overlapping with X instances. In the clustermap, we observed three main motif families—(i) planar-staple family with UAA4-staple motif (motif 44, Figure 6A) as the best family representative (Figure 7B, the largest square in the center); the family can be further divided into two sub-families, AAN3-staple sub-family (3’-part of the UAA4-staple) and UGA3-staple sub-family (5’-part of the UAA4-staple); (ii) tilted-staple family with GAIAG3-staple motif (motif 43, Figure 6B) as the best family representative (Figure 7B, bottom right square); (iii) helical-packing family (Figure 7B, top left square) with GN3-staple motif (motif 33, Figure 6C) and cWW(GC)-cWW(GU)-packing motif (motif 26, Figure 6E) as family representatives. Interestingly, even though cWW(GC)-cWW(GU)-packing motif (Figure 6E) seems to be symmetrical (indeed, out of its 38 matches, 28 matches are actually 14 pairs of the same instance in different orientations), it is actually specific to wobble-like base pairs with one of the residues shifted to the major groove. Out of the 38 matches, only a single match is a truly symmetric cWW(GC)-cWW(GC)-packing motif (module 6z5u_24).

Apart from the three main families, we can also see the cSS base pair (motif 1) and tWS/tSS base pair (motif 2) highly overlapping with the planar-staple and tilted-staple families, a two-motif family (motif 13 and motif 25) representing Type 1-A minor with additional tSH/GA base pair formed by the Hoogsteen edge of the staple adenosine, and three isolated motifs - NAA4-staple (motif 45), cWS base pair (motif 8), and tWS/tHS base pair (motif 7). Surprisingly, the cWS base pair and tWS/tHS base pair motifs did not overlap with other motifs by any of their 55 (25 + 30) instances.

Overall, we observed two interesting trends. First, sometimes larger motifs have more matches than their smaller fragments, and second, the overlapping between the motifs of the same family is often imperfect. We see this as the feature of structural variability of the motifs - some variations can completely disrupt the local fragment whereas the global characteristic arrangement of the entire motif can be still preserved, see Figure 5E for an example of such a variation of the D-loop/T-loop-like motif.

Classification of LORA modules

To analyze the structural context of the identified long-range tertiary motifs we visually examined all 1264 LORA modules and assigned a descriptive type to each module (Supplementary Table S8). If a module didn’t involve any apparent interactions we assigned a pair of the closest atoms and their pairwise distance in angstroms as the descriptive type of the module, e.g., ‘OP1-O2’, ‘S’ and considered such modules as ‘false positive hits’. We counted 585 such false positive hits, including 364 two-residue modules (94% of the two-residue modules), 133 three-residue modules (73%), 47 four-residue modules (45%), and 41 modules of larger sizes (7%). Thus, the LORA dataset included 679 ‘true positive hits’, i.e., reasonable long-range RNA 3D modules involving long-range tertiary interactions.

Among the 679 LORA modules, 357 (53%) involved interaction of the minor groove of a helical region, 274 (40%) involved staple interactions (77% of the 357 modules with minor groove interactions), 51 (7.5%) involved the helical packing interaction (two helical regions interacting via
their minor grooves), and only 17 (2.5%) involved interaction of the major groove of a helical region. In total, we counted 349 staple interactions. The staples spanned from one (1-staples, 24 instances) to six (6-staples, 2 instances) base pair levels. The most frequent were 3-staples (98 instances), followed by 2-staples (81 instances) and 4-staples (85 instances), whereas 5-staples were less frequent (29 instances). In 70% of the staple interactions the staple strand included from two to four residues, and the single largest staple included eight residues and two cross-strand stacks (module 6me0\_15). In a number of modules staple interactions were additionally stabilized by base stacking (e.g., module 6th6\_18) and base pairing (e.g., module 6me0\_5) interactions with the helical strands. Among the rare exotic modules, we can mention module 6d9j\_2 with a stem unrecognized by DSSR due to the wobble-like arrangement of the canonical Watson–Crick base pairs and module 7o7y\_29 formed by the minor grooves of two helical regions interacting via a mediating guanosine residue, which we arbitrarily named ‘helix-base-helix’ interaction.

We further examined the modules involving the cWS base pair (motif 8) and tWS/tHS base pair (motif 7) motifs. Overall we counted 41 such modules, 36 of them involving a staple, 11 with a staple formed by a hairpin loop, and five more with the D-loop/T-loop-like motif. In turn, the LORA dataset in total included 25 modules with a staple formed by a hairpin loop. Hence, from the data, we can conclude that the long-range cWS and tWS/tHS base pairs are specific to hairpin loop interactions, however, are not a part of any conserved larger motif.

**Comparison with the CaRNAval database**

We compared the assembled LORA modules dataset with the CaRNAval database of RINs (recurrent interaction networks) (36). The intersection of the two sets of PDB entries consists of 10 RNA structures. Within the 10 structures, we observe 17 RINs (counting only RINs not completely nested in any other RINs) and 24 LORA modules (Supplementary Table S9). Among the 17 RINs, we see five RINs completely nested into a LORA module, two RINs partially nested into a LORA module, seven RINs completely absent in the LORA dataset as being local according to the used definitions, and three RINs being isolated stems treated as long-range interactions as the result of pseudoknot removal procedure (36) and also absent from the LORA dataset. Among the 24 LORA modules we observe four modules fully involving a RIN, four modules partially involving a RIN, ten modules completely absent in the CaRNAval database, and six small (five two-residue and one three-residue) modules with no apparent interactions that are also absent in the CaRNAval database.

If we loosely define a set of RNA residues involved in tertiary interactions as a true positive hit, and therefore treat isolated stems from CaRNAval along with short LORA modules with no interactions as false positive hits, we can measure F1-score of the two approaches on the set of 41 objects (17 RINs and 24 LORA modules). F1-score of LORA is then 75% and F1-score of CaRNAval is 65% (Supplementary Table S9). The LORA dataset by design misses the local tertiary interactions, i.e., interactions formed by neighboring RNA secondary structure elements, and simultaneously produces a number of small modules with no apparent interactions. Whereas, the CaRNAval database by design misses the RINs that do not include at least two base pairs annotated with FR3D along with the intermolecular RINs, and simultaneously produces a number of RINs being isolated stems resulting from the pseudoknot removal procedure. Overall, the two approaches can be considered complementary.

**DISCUSSION**

We surveyed long-range tertiary interactions and motifs in a non-redundant set of non-coding RNA 3D structures. We identified long-range nucleotide doublets based on a liberal 6.0 Å threshold and the annotation of RNA secondary structure. We grouped the 17272 long-range doublets into the 3D modules using an original approach based purely on the proximity of the doublets. The assembled 1264 long-range modules composed the LORA dataset. And finally, we applied ARTEM, a novel algorithm for RNA tertiary motifs pairwise superposition, to derive nearly half a million long-range tertiary motif instances from the 3D modules of the LORA dataset, which were later clustered into tertiary motifs. The 45 most common long-range tertiary motifs were then picked manually and analyzed in detail.

We found that among the long-range doublets annotated as base pairs, ten Leonitis-Westhof base pair types are most often identified by one of the five conventional automatic annotation tools (Clara, DSSR, FR3D, MC-Annotate or RNAView), indicating the lack of consensus between these annotations. Using the well-known ‘named’ RNA tertiary motif examples we showed the high quality of the assembled long-range RNA 3D modules, which could not be achieved based on the automatic annotation of the involved interactions with any of the five conventional tools. Searching for D-loop/T-loop-like motifs we demonstrated the capability of the ARTEM algorithm to find structural similarities with different backbone topology or interaction network variations and even potential modeling errors involving structurally corresponding base pairs. Along with the well-known Type I and Type II A-minor interactions, we identified a multitude of previously undescribed interactions of one or more residues with the minor groove of a helical region simultaneously with both helical strands and at different base pair layers, which we called staples. Analyzing the overlap between the identified long-range motifs we observed three main long-range motif families—planar staples, tilted staples and helical packing motifs.

In comparison with the other databases of long-range RNA 3D motifs (36) LORA dataset has the advantage of including the entire structural context of the long-range interactions, as it does not rely on automatic annotation of non-canonical interactions. LORA only relies on the RNA secondary structure annotation used to separate the long-range doublets from the local ones, but as it was shown previously (42) and also shown in this work, the automatic annotation tools agree well in the annotation of canonical cWW base pairs, i.e., base pairs forming the RNA secondary structure.
The most important limitation of the ARTEM algorithm is simultaneously its most important advantage. It produces structural superpositions based purely on RMSD in annotation-, sequence- and topology-independent manner. For example, judging solely by RMSD, ARTEM cannot distinguish between Type I and Type II A-minor interactions, especially in large structural contexts, as these two interactions are almost identical in their base arrangement and differ only in a small shift of the adenosine residue. Therefore, in the analyses performed in this work, we also used the RESRMSDMAX value, the maximal per-residue RMSD, where a more strict selection was needed. By default, ARTEM identifies matches between spatially similar motifs that differ in topology, yet simultaneously it almost completely eliminates false negatives. Searching with ARTEM for a given reference motif in a given query RNA structure with no RMSD or SIZE thresholds it’s impossible to miss any of the motif occurrences unless an occurrence has zero residues that perfectly align with their reference counterpart, but in such a case it’s unclear why the occurrence is treated as the motif occurrence in the first place. For example, ARTEM will not identify a stack of two cSS base pairs as a motif structurally similar to a stack of two cWW base pairs.

The output of the ARTEM algorithm can be subject to additional filtering based on any required sequence specificity, interaction types or backbone topology. Therefore, since ARTEM can be used for two arbitrary RNA structures or structural fragments, we believe the algorithm will find wide use in many applications. Furthermore, ARTEM can easily process any modified residues (69) provided their 3-atom and 5-atom representations are specified by the user.

Among the 37 long-range tertiary motifs larger than two residues in size, only the canonical A-minor interactions were comprehensively studied to date. We believe the other motifs were largely understudied due to the general approach bias toward base pair annotations, as ribose-ribose interactions are too simplistic and not base-specific if considered in isolation, and most staples by their nature are not coplanar and cannot be decomposed into a set of base pairs. And only the canonical A-minor interactions can be naturally decomposed into a number of $cSS/\bar{t}SS/cSW/tSW$ base pairs. Moreover, the not coplanar nature of staples can also explain the disagreements between the standard annotation tools in their base pair annotations of the long-range doublets as the tools are focused mainly on coplanar base arrangements.

We observed a large number of cases of imperfect overlapping between the motifs of the same family, which we attribute to local variations disrupting the local fragment but preserving the global arrangement of the entire motif. The opposite phenomenon is much better known - smaller fragments are often found in different structural environments. Therefore, one often cannot conclude the presence of a motif judging by either the presence of its fragment or the presence of a larger motif usually involving the current motif as its fragment. Hence, to catch local structural variations of the RNA 3D motifs it is important to not rely on the standard components of the motifs for their identification. And this can be achieved by assembling high-quality reference libraries of motifs of various sizes.

Summarizing, we performed a comprehensive analysis of the long-range tertiary interactions and motifs in non-coding RNA 3D structures. A dataset of annotated long-range RNA 3D modules was built using the original approach, which goes beyond the automatic annotations of non-canonical interactions. A new algorithm was developed for annotation-, sequence- and topology-independent superposition of two arbitrary RNA 3D modules. The proposed methods were used to identify and describe the recurrent long-range RNA tertiary motifs. We believe the results obtained in this work as well as the proposed methods will have a great impact in the field of RNA structural studies, especially in comparative analyses of RNA structures and RNA-containing complexes.

DATA AVAILABILITY
The LORA dataset and all the supplementary materials are available at https://github.com/febos/LORA (DOI: 10.5281/zenodo.8009542). An implementation of the ARTEM algorithm is available at https://github.com/david-bogdan-r/ARTEM (DOI: 10.5281/zenodo.8009615). urslib2 python library used for the annotation of long-range nucleotide doublets is available at https://github.com/febos/urslib2 (DOI: 10.5281/zenodo.8009519).

SUPPLEMENTARY DATA
Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS
E.F.B. thanks his former supervisors Mikhail Roytberg and Ivan Kulakovskiy, who taught him priceless principles for conducting research that led to this work. Also, E.F.B. thanks his daughter Leia for a walk in the park during which he came up with the idea of the ARTEM algorithm.

FUNDING
E.F.B. was supported by the Polish National Science Center [NCN; grant 2017/26/A/NZ1/01083 to J.M.B.]; European Molecular Biology Organization [fellowship ALTF 525-2022 to E.F.B.]; J.M.B. was supported by the NCN [2017/25/B/NZ2/01294 to J.M.B.]. Funding for open access charge: IIMCB statutory funds. Conflict of interest statement. J.M.B. is an Executive Editor of Nucleic Acids Research.

REFERENCES
1. Santosh, B., Varshney, A. and Yadava, P.K. (2015) Non-coding RNAs: biological functions and applications. Cell Biochem. Funct., 33, 14–22.
2. Johnson, P., Lipovich, L., Grandé, D. and Morris, K.V. (2014) Evolutionary conservation of long non-coding RNAs: sequence, structure, function. Biochim. Biophys. Acta (BBA)-Gen. Subj., 1840, 1063–1071.
3. Leontis, N.B., Lesucute, A. and Westhof, E. (2006) The building blocks and motifs of RNA architecture. Curr. Opin. Struct. Biol., 16, 279–287.
4. Hendrix, D.K., Brenner, S.E. and Holbrook, S.R. (2005) RNA structural motifs: building blocks of a modular biomolecule. Q. Rev. Biophys., 38, 221–243.
5. Butcher, S.E. and Pyle, A.M. (2011) The molecular interactions that stabilize RNA tertiary structure: RNA motifs, patterns, and networks. Acco Chem. Res., 44, 1302–1311.

6. 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. U.S.A., 98, 4899–4903.

7. Chan, C.W., Chetnani, B. and Mondragon, A. (2013) Structure and function of the T-loop structural motif in noncoding RNAs. Wiley Interdiscip. Rev. RNA, 4, 507–522.

8. Batey, R.T., Rambo, R.P. and Doudna, J.A. (1999) Tertiary motifs in RNA structure and folding. Angew. Chem. Int. Ed., 38, 2326–2343.

9. Tamura, M. and Holbrook, S.R. (2002) Sequence and structural conservation in RNA ribozymes. J. Mol. Biol., 320, 455–474.

10. Zirbel, C.L., Sponer, J.E., Sponer, J., Stombaugh, J. and Leonis, N.B. (2009) Classification and energetics of the base-pseudohapase interactions in RNA. Nucleic Acids Res., 37, 4898–4918.

11. Leonis, N.B. and Westhof, E. (2001) Geometric nomenclature and classification of RNA base pairs. RNA, 7, 499–512.

12. Mladek, A., Sponer, J.E., Kuhlman, P., Lu, X.J., Olson, W.K. and Sponer, J. (2012) Understanding the sequence preference of recurrent RNA building blocks using quantum chemistry: the intrastrand RNA dinucleotide platform. J. Chem. Theory Comput., 8, 335–347.

13. Baulin, E., Metelev, V. and Bogdanov, A. (2020) Base-intercalated and base-wedged stacking elements in 3D-structure of RNA and RNA –protein complexes. Nucleic Acids Res., 48, 8675–8685.

14. Shahykova, A.A., Mikhaliova, D.S., Lukovskii, I.V., Fakhranurova, I.I. and Baulin, E.F. (2021) Annotation of the local context of RNA secondary structure improves the classification and prediction of A-minors. RNA, 27, 907–919.

15. Gianfratta, C., Reinhart, V., Barth, D. and Denise, A. (2021) A graph-based similarity approach to classify recurrent complex motifs from their context in RNA structures. 19th Symposium on Experimental Algorithms.

16. Gianfratta, C., Reinhart, V., Lepinet, O., Barth, D. and Denise, A. (2022) On the predictability of A-minors from their local contexts. RNA Biol., 19, 1208–1227.

17. Laing, C., Chen, D., Wang, S., Li, L., Zheng, L., Zhang, C. and Li, H. (2014) Vernal: a tool for mining fuzzy network motifs in RNA. Bioinformatics, 30, 3450–3455.

18. Zirbel, C.L., Sponer, J.E., Sponer, J., Stombaugh, J. and Leonis, N.B. (2009) C3RNA: a classifier of contacts in RNA 3D structures based on a comparative analysis of various classification schemes. Nucleic Acids Res., 42, e151.

19. Wang, T., Chojnowski, G., Gierski, P. and Bujnicki, J.M. (2014) C3RNA: a classifier of contacts in RNA 3D structures based on a comparative analysis of various classification schemes. Nucleic Acids Res., 42, e151.

20. Zheng, J., Xie, J., Hong, X. and Liu, S. (2019) RMalign: an RNA structural alignment tool based on a novel scoring function RMscore. BMC Genomics, 20, 276.

21. Zhao, Y.F., Huang, X.L. and Lu, C.L. (2020) SARsa: a web tool for structural alignment of RNA using a structural alphabet. Nucleic Acids Res., 36, W19–W24.

22. Zukowski, M., Antczak, M. and Szczuciński, M. (2023) High-quality, customizable heuristics for RNA 3D structure alignment. Bioinformatics, 39, btad315.
51. Wang, C.W., Chen, K.T. and Lu, C.L. (2010) iPARTS: an improved tool of pairwise alignment of RNA tertiary structures. *Nucleic Acids Res.*, 38, W340–W347.
52. Yang, C.H., Shih, C.T., Chen, K.T., Lee, P.H., Tsai, P.H., Lin, J.C., Yen, C.Y., Lin, T.Y. and Lu, C.L. (2016) iPARTS2: an improved tool for pairwise alignment of RNA tertiary structures, version 2. *Nucleic Acids Res.*, 44, W328–W332.
53. Islam, S., Rahaman, M.M. and Zhang, S. (2021) RNAMotifContrast: a method to discover and visualize RNA structural motif subfamilies. *Nucleic Acids Res.*, 49, e61–e61.
54. Roll, J., Zirbel, C.L., Sweeney, B., Petrov, A.I. and Leontis, N. (2016) JAR3D Webserver: scoring and aligning RNA loop sequences to known 3D motifs. *Nucleic Acids Res.*, 44, W320–W327.
55. 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.
56. Chen, X., Khan, N.S. and Zhang, S. (2020) LocalSTAR3D: a local stack-based RNA 3D structural alignment tool. *Nucleic Acids Res.*, 48, e77.
57. Čech, P., Svozil, D. and Hoksza, D. (2012) SETTER: web server for RNA structure comparison. *Nucleic Acids Res.*, 40, W42–W48.
58. Nguyen, M.N., Sim, A.Y., Wan, Y., Madhusudhan, M.S. and Verma, C. (2017) Topology independent comparison of RNA 3D structures using the CLICK algorithm. *Nucleic Acids Res.*, 45, e5.
59. Builin, E., Yazovey, V., Khachko, D., Spirin, S. and Roytberg, M. (2016) URS DataBase: universe of RNA structures and their motifs. *Database*, 2016, baw085.
60. Burley, S.K., Berman, H.M., Bhikadiya, C., Bi, C., Chen, L., Di Costanzo, L., Christie, C., Dalenberg, K., Duarte, J.M., Dutta, S. *et al.* (2019) RCSB Protein Data Bank: biological macromolecular structures enabling research and education in fundamental biology. *Biomed. Biotechnol. Energy Nucleic Acids Res.*, 47, D464–D474.
61. Kalvari, I., Nawrocki, E.P., Ontiveros-Palacios, N., Argasinska, J., Lamkiewicz, K., Marz, M., Griffiths-Jones, S., Toffano-Nioche, C., Gautheret, D., Weinberg, Z. *et al.* (2021) Rfam 14: expanded coverage of metagenomic, viral and microRNA families. *Nucleic Acids Res.*, 49, D192–D200.
62. Leontis, N.B. and Zirbel, C.L. (2012) Nonredundant 3D structure datasets for RNA knowledge extraction and benchmarking. In: *RNA 3D structure analysis and prediction*. Springer, Berlin, Heidelberg, pp. 281–298.
63. Hamdani, H.Y., Appasamy, S.D., Willett, P., Artymiuk, P.J. and Firdaus-Raih, M. (2012) NASSAM: a server to search for and annotate tertiary interactions and motifs in three-dimensional structures of complex RNA molecules. *Nucleic Acids Res.*, 40, W35–W41.
64. Kabsch, W. (1976) A solution for the best rotation to relate two sets of vectors. *Acta Crystallogr. A*, 32, 922–923.
65. Boniecki, M.J., Lach, G., Dawson, W.K., Tomala, K., Lukasz, P., Soltysinski, T., Rother, K.M. and Bujnicki, J.M. (2016) SimRNA: a coarse-grained method for RNA folding simulations and 3D structure prediction. *Nucleic Acids Res.*, 44, e63.
66. Dolezal, R., Froncova, K., Kirntmat, A. and Krejcar, O. (2020) Computational Complexity of Kabsch and Quaternion Based Algorithms for Molecular Superimposition in Computational Chemistry. In: *International Conference on Engineering Applications of Neural Networks*. Springer, Cham, pp. 473–486.
67. Bentley, J.L. (1975) Multidimensional binary search trees used for associative searching. *Commun. ACM*, 18, 509–517.
68. 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.
69. Boccaletto, P., Stefaniak, F., Ray, A., Cappannini, A., Mukherjee, S., Purta, E. and Bujnicki, J.M. (2022) MODOMICS: a database of RNA modification pathways. 2021 update. *Nucleic Acids Res.*, 50, D231–D235.