Published online 4 June 2020

Nucleic Acids Research, 2020, Vol. 48, No. 13 ***e77***
doi: 10.1093/nar/gkaa453

LocalSTAR3D: a local stack-based RNA 3D structural alignment tool

Xiaoli Chen, Nabila Shahnaz Khan and Shaojie Zhang*

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

Received February 23, 2020; Revised May 15, 2020; Editorial Decision May 18, 2020; Accepted May 27, 2020

ABSTRACT

A fast-growing number of non-coding RNA structures have been resolved and deposited in Protein Data Bank (PDB). In contrast to the wide range of global alignment and motif search tools, there is still a lack of local alignment tools. Among all the global alignment tools for RNA 3D structures, STAR3D has become a valuable tool for its unprecedented speed and accuracy. STAR3D compares the 3D structures of RNA molecules using consecutive base-pairs (stacks) as anchors and generates an optimal global alignment. In this article, we developed a local RNA 3D structural alignment tool, named LocalSTAR3D, which was extended from STAR3D and designed to report multiple local alignments between two RNAs. The benchmarking results show that LocalSTAR3D has better accuracy and coverage than other local alignment tools. Furthermore, the utility of this tool has been demonstrated by rediscovering kink-turn motif instances, conserved domains in group II intron RNAs, and the tRNA mimicry of IRES RNAs.

INTRODUCTION

Many non-coding RNAs play important catalytic and regulatory roles by folding into versatile three-dimensional (3D) structures. Comparing RNA 3D structures yields valuable information on their functional conservation and evolutionary relations (1). Currently, many powerful tools have been developed to compare the RNA 3D structures. According to the utilities of these tools, they can be categorized into three groups, which are global alignment, motif search, and local alignment tools. The first group, global alignment tools, aim to align the whole RNAs and generate optimal alignments by using different techniques, which include SARA (2,3), LaJolla (4), R3D Align (5,6), SETTER (7), R-click (8), STAR3D (9), Elastic Shape Analysis (ESA) (10,11), RNA-align (12) and RMalign (13). STAR3D and SARA maximize the number of matched nucleotides or atoms within a distance cutoff between a pair of RNA structures, while recently developed tools RNA-align and RMalign were designed to maximize a size-independent alignment score.

The second group, motif search tools, focus on recurrent 3D components in the RNA structures (14–16). RNA Bricks (17) is a database that provides RNA 3D structure motifs and their interaction information along with an algorithm for 3D motif search and comparison. Almost all of these motifs are single hairpin loops, internal loops, or multi-loops. Most of the motif search tools that are designed to align single loops cannot be used to detect the larger conserved 3D structures between two RNAs.

The third group, tools that are capable of local alignment, are designed to report multiple conserved substructures for a pair of RNAs. We define the 'conserved substructure' as a collection of loops and helices that are adjacent in sequence. The conserved substructures can be as small as RNA structural motifs or as large as whole conserved domains. For distantly related RNAs, although the overall architectures of their 3D structures are different, the conserved substructures can still be found among them (18). In this case, the global RNA 3D structural alignment may not be able to generate meaningful results, and the local alignment would be a better solution. Most of the tools in this group are not designed specifically for local alignment, but offer options for both global and local alignment. ARTS (18) is the first tool that tackles the RNA 3D structural alignment problem. It handles global and local alignment at the same time. ARTS first searches for all matched successive base pairs between the input RNA structures as seeds. It then conducts a global extension from each matched seed that maximizes a weighted sum of matched phosphate atoms and base pairs between RNA structures. For each alignment, ARTS uses one single seed and doesn't apply any constraint for adjacency either in sequence or in 3D space during the global extension. Therefore the connectivity in the alignment is not guaranteed. The matched phosphate atoms and base pairs can be sparse for some RNAs that are distantly related. In these cases, ARTS may generate fragmented alignments. An alternative approach to local alignment is to encode the RNA 3D structures into one-dimensional (1D) sequences and apply the local sequence alignment algorithms to it. Tools adopting this approach include DIAL (19),

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

SARSA (20), iPARTS (21) and iPARTS2 (22), etc. DIAL incorporates the pseudo-dihedral and/or dihedral angle, sequence and base-pairing similarity into the scoring function of their dynamic programming algorithms. In addition to the global and local alignment, DIAL offers a semiglobal mode for motif searching. Web servers SARSA, iPARTS and iPARTS2 use a structural alphabet (SA) based approach to reduce RNA 3D structures into 1D sequences. While SARSA and iPARTS use a SA of 23 letters to decode the RNA backbone conformations for each nucleotide, the recent tool iPARTS2 utilizes a more sophisticated SA that consists of 92 elements carrying both base and backbone geometric information. Although iPARTS2 always generates consecutive local alignments, the structural similarity is low in some instances. A thorough comparison of ARTS, iPARTS2, and LocalSTAR3D is represented in the results section.

To solve the issues in existing local RNA structural alignment tools, we propose LocalSTAR3D, a local stack-based RNA 3D structural alignment tool that optimally searches for connected and conserved substructures. LocalSTAR3D is extended from STAR3D, a global alignment tool developed by Ge and Zhang (9). STAR3D constructs a consensus of stacks by searching the maximum clique in a compatible graph of stack pairs defined by 2D topology and 3D geometry, and uses it to generate an optimal global alignment. To find the local alignments, LocalSTAR3D uses the distance between stack pairs as a new constraint in the compatible graph and searches for non-overlapping subgraphs corresponding to the maximal local and compatible stack pair sets. The maximal local and compatible stack pair sets are used as anchors to guide the alignment. LocalSTAR3D can generate more than one local alignment. The non-overlapping local alignments are assembled from the stacks and the loops and then sorted by the alignment scores.

To illustrate the necessity of local RNA 3D structural alignment tools, we compared LocalSTAR3D with two global alignment tools, SARA and STAR3D. The alignment results show that LocalSTAR3D can find more conserved substructures than those in optimal global alignments. LocalSTAR3D was then benchmarked with other state-of-the-art local RNA 3D structural alignment tools, ARTS and iPARTS2. The results show that LocalSTAR3D identifies conserved substructures with better accuracy and coverage. We further examined some interesting examples demonstrating the utility of LocalSTAR3D, including searching for kink-turn motifs on a 23s rRNA, detecting the conserved substructures in different classes of group II introns, and rediscovering the tRNA mimicry of IRES RNA.

MATERIALS AND METHODS

Preprocessing

The inputs of LocalSTAR3D are a pair of RNA 3D structures. In addition to the PDB format used by STAR3D, LocalSTAR3D also accepts RNA structures in PDBx/mmCIF format. Users can provide atomic coordinates in PDB or PDBx/mmCIF format files, or the PDB IDs along with chain IDs and LocalSTAR3D will search and download them from the PDB website automatically. There are several RNA base pair annotation tools, including MC-Annotate (23,24), RNAView (25), FR3D (16), DSSR (26), ClaRNA (27) and CompAnnotate (28). LocalSTAR3D uses DSSR to identify the base-pairing interactions as it supports both PDB and PDBx/mmCIF formats. After obtaining the base pair annotations, similar to STAR3D, LocalSTAR3D removes the crossing base pairs from the base-pair annotations by using RemovePseudoknots (29).

Detecting the conserved stacks

We use a method similar to STAR3D to detect the conserved stacks. Conserved stacks are the double-helical regions with root-mean-square deviation (RMSD) lower than a certain threshold. Similar to STAR3D, LocalSTAR3D calculates the RMSD based on the geometric center of six backbone atoms C3′, C4′, C5′, O3′, O5′, and P of each nucleotide by using the Kabsch method (30). The default threshold is 4 Å in STAR3D and LocalSTAR3D. Ge and Zhang indicated that the structural similarity of conserved stacks is statistically significant to distinguish them from the random stack pairs (9). To detect the conserved stacks, all the stacks of size k (the default value of k is 3) allowing overlap in the input RNAs formed by consecutive nested Watson–Crick and wobble base pairs, called ‘k-stacks’, are identified. If the RMSD of a pair of k-stacks is lower than 4 Å, it will be considered as a ‘k-stack pair’. After that, the consecutive k-stack pairs are merged into extended stack pairs, called ‘e-stack pairs’. These e-stack pairs are used as anchors for the following loop alignment.

Modifications are made in LocalSTAR3D to facilitate the local alignment. One of the major modifications is that LocalSTAR3D fixes the stacks that DSSR may fail to annotate by searching for potential canonical base pairs around the annotated ones. LocalSTAR3D provides an option to fill the gap between annotated base pairs, if the gap is an internal loop with one nucleotide in each helix and the two nucleotides in this small internal loop can form canonical base pair based on their nitrogenous bases. The reason for providing this option is that the e-stacks generated from original annotated base pairs can be too sparse for some RNA structures. Filling the small gaps between annotated base pairs improves the coverage of the local alignments, while the RMSD does not change significantly.

Constructing the local and compatible e-stack pair sets

Similar to STAR3D, LocalSTAR3D uses a consensus of e-stack pairs as anchors to guide the alignment. The e-stack pairs in the consensus for local alignment should be compatible with each other in topology and adjacent in sequence. To follow the definitions in STAR3D, for two RNAs A and B, the sets of e-stacks in A and B are denoted as Q^A and Q^B respectively. The set of e-stack pairs is denoted as S. Each e-stack pair $s_i(\in S)$ contains an e-stack $q_i^A(\in Q^A)$ and $q_i^B(\in Q^B)$. STAR3D defines three types of relations between two e-stacks in an RNA, overlapping, juxtaposing, and enclosing (9). The overlapping e-stacks are prohibited to be present in the same stack consensus. Take two e-stack pairs s_i and s_j as an example, where s_i is formed by (q_i^A, q_i^B) and s_j is formed by (q_j^A, q_j^B). For q_i^A and q_j^A in RNA A, there are four possible relations, which are q_i^A juxtaposing and preceding q_j^A, q_i^A juxtaposing and succeeding q_j^A, q_i^A enclosing q_j^A, and q_j^A enclosing q_i^A. It is the same for q_i^B and q_j^B in RNA B. The e-stack pair s_i is considered as compatible with s_j, if and only if q_i^A and q_j^A have the same relation as q_i^B and q_j^B. For example, if q_i^A is enclosing q_j^A and q_i^B is enclosing q_j^B, e-stack pairs s_i and s_j are considered compatible. STAR3D uses the maximum clique of compatible e-stack pairs as the anchors to guide the alignment. Ge and Zhang proved that for a set of e-stack pairs, if any two of the e-stack pairs are compatible, the corresponding two e-stack sets have the same tree structure (9). This theorem indicates that the 3D structural consensus of the e-stack pairs can be inferred by the maximum cliques in the compatible graph, which is built by using the e-stack pairs as the vertices, and their compatible relations as edges.

This approach works efficiently to maximize the matched nucleotides between a pair of RNAs, but it can not be applied to local alignment directly. Because the optimal global alignment may not be connected in sequence and does not contain any multiple alignment cases. To get the optimal local alignments, LocalSTAR3D enforces the connectivity of the alignment by applying a new adjacency constraint to the compatible graph. We call the new graph 'local compatible graph'. To construct the local and compatible e-stack pair sets from the local compatible graph, a new method was developed to compute the subgraphs from the local compatible graph. While STAR3D used the standard Bron-Kerbosch algorithm to compute the maximum clique from the compatible graph, we designed a modified version of the Bron-Kerbosch algorithm to search for maximal subsets of e-stack pairs in the local compatible graph.

Two compatible e-stack pairs are *adjacent* if the distance in sequence between two pair of aligned strands, one from each e-stack pairs, are lower than a user-set cutoff. The examples of *adjacent* e-stack pairs are shown in Figure 1. We denote the left and the right strand of the e-stack in e-stack pair s_i and RNA A as l_i^A and r_i^A and similarly for RNA B as l_i^B and r_i^B. Without loss of generality, only the cases where the e-stack pair s_i comes before s_j is shown for the juxtaposing relation, and s_i enclosing s_j is shown for the enclosing relation. For juxtaposing relation, e-stack pairs s_i and s_j are *adjacent* if and only if (r_i^A, l_j^A) and (r_i^B, l_j^B) are within the cutoff distance. For enclosing relation, e-stack pairs are *adjacent* if their left strands (l_i^A, l_j^A) and (l_i^B, l_j^B) or right strands (r_i^A, r_j^A) and (r_i^B, r_j^B) are within the cutoff distance. The default value of this cutoff is 15 nucleotides. A higher cutoff will include more e-stack pairs into the conserved e-stack sets, but will also increase the possibility to have long gaps in the alignment. We checked the distances between the neighboring stacks in a 23s rRNA (the minimum stack size is 3), and found all of these distances are shorter than this default value. In the results section, we show that LocalSTAR3D generates better local and conserved substructures than other state-of-the-art tools by using this default parameter.

We formulate the problem of finding the maximal local and compatible e-stack pair sets as a graph problem. An example is shown in Figure 2. In the compatible graph, the vertices represent the e-stack pairs. The maximal local and compatible e-stack pair sets can be inferred by the subgraphs that are cliques formed by solid edges and connected graphs formed by the dashed edges at the same time. STAR3D uses the standard Bron-Kerbosch algorithm (31) to recursively search for maximum cliques. Within each recursive call, it adds a vertex to the current subset from the candidate set while preserving complete compatibility. In our modified Bron–Kerbosch algorithm, we maintain an extra set, called 'adjacent set', containing all the vertices that are immediately adjacent to the vertices in the current subset. In each recursive call, we add a vertex to the current subset from the intersection of the candidate set and adjacent set. Since the adjacent set is much smaller than the candidate set, the search space dramatically shrinks in LocalSTAR3D.

LocalSTAR3D iteratively retrieves the maximal local and compatible e-stack pair set which contains the most base-pairing nucleotides. This step was demonstrated by the example in Figure 2 that searches for the top two local and compatible e-stack pair sets. The maximal local and compatible e-stack sets are {1-I, 2-II, 3-III} and {1-II, 2-IV}. As each e-stack has six base-pairing nucleotides, the first maximal local and compatible e-stack pair set is {1-I, 2-II, 3-III} containing 18 nucleotides and the second is {1-II, 2-IV} containing 12 nucleotides. If there are multiple e-stack pair sets with the same number of base-pairing nucleotides, the one with the lowest RMSD will be selected first. After selecting an e-stack pair set, LocalSTAR3D removes all the e-stack pair sets that overlap with this e-stack pair set to avoid generating duplicated local alignments.

Assembling the local alignments

Since two RNAs may have multiple similar substructures, LocalSTAR3D was designed to report the top n local alignments (the default value of n is 5). We maintain a priority queue of local alignments ranked by their alignment scores and RMSD. The local alignments are added into the priority queue iteratively until reaching the user-set local alignment number n. After generating the local and compatible e-stack pair sets, the loop regions between the e-stacks are aligned under the guide of the e-stack pairs by using the methods developed in STAR3D (9). For each pair of loops, the optimal structural alignment is calculated by a dynamic programming algorithm. The alignment score for each pair of loops is the sum of scores for the matched, inserted and deleted nucleotides. The score of an e-stack is calculated by multiplying the matching score by the number of nucleotides in the e-stack. LocalSTAR3D uses the same score for each matched, inserted, and deleted nucleotide that was defined in STAR3D. The aligned loop regions and e-stacks are then concatenated into a local alignment. An alignment score is calculated for each candidate local alignment as the sum of the scores of the loop regions and the e-stacks. The candidate local alignments with RMSD larger than a user-set cutoff (the default value is 4 Å) are discarded. The resulting local alignments are first sorted in descending order by

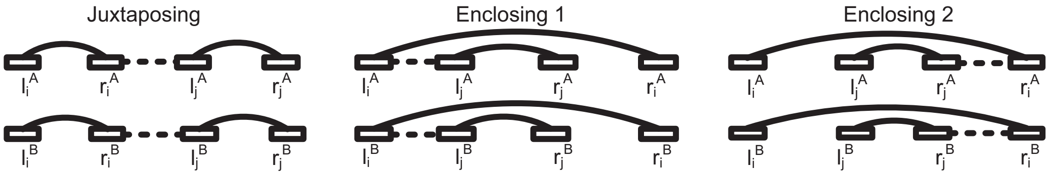

Figure 1. Three cases of adjacent e-stack pairs. For each case, the first row contains the e-stacks in RNA *A*, the second row contains the e-stacks in RNA *B*. The black boxes connected by the curve lines are the complementary strands of an e-stack. The dashed lines represent that the strands meet the adjacency requirement.

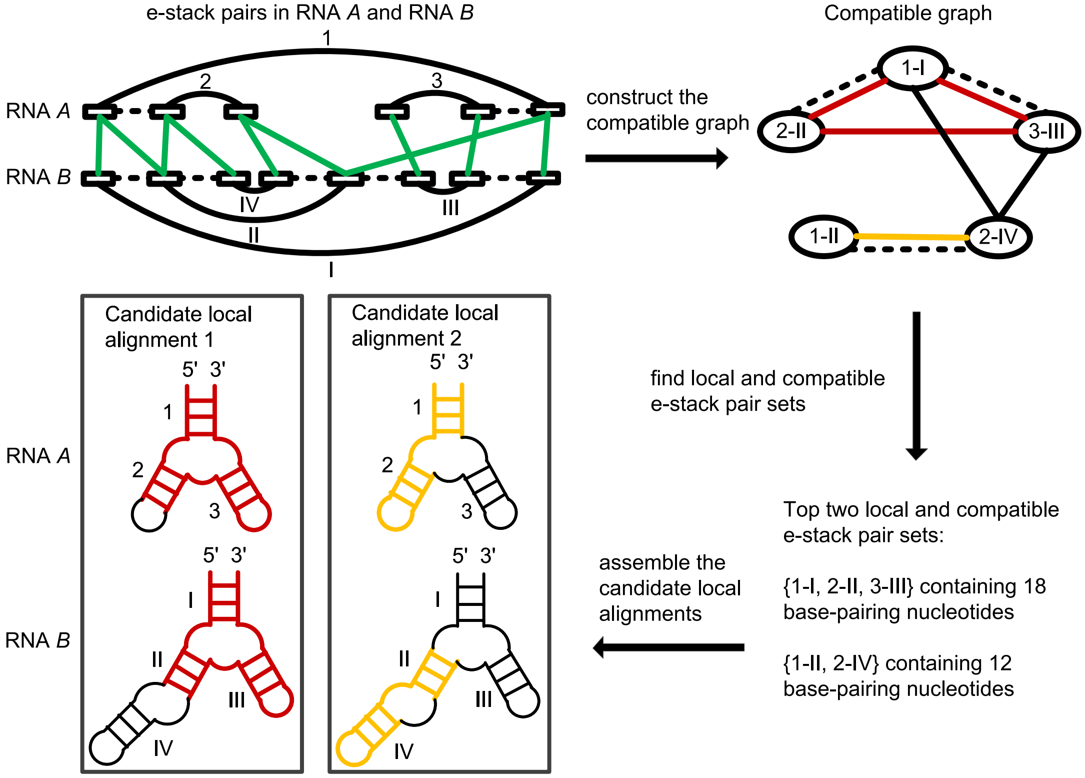
e-stack pairs in RNA A and RNA B
RNA A
RNA B
1
2
3
I
II
III
IV
construct the compatible graph
Compatible graph
1-I
2-II
3-III
1-II
2-IV
find local and compatible e-stack pair sets
Top two local and compatible e-stack pair sets:
{1-I, 2-II, 3-III} containing 18 base-pairing nucleotides
{1-II, 2-IV} containing 12 base-pairing nucleotides
assemble the candidate local alignments
Candidate local alignment 1
Candidate local alignment 2
5' 3'

Figure 2. Key steps in LocalSTAR3D's algorithmic framework demonstrated by aligning RNA *A* to RNA *B*. In e-stack pairs of RNA *A* and RNA *B*, the green solid lines show the matching between the e-stacks. The curve lines and dashed lines represent the complementary strands and the adjacency respectively. In the compatible graph, the vertices represent the e-stack pairs. The dashed lines represent the adjacent e-stack pairs. The solid lines represent the compatible e-stack pairs. The red and yellow solid lines represent two local and compatible e-stack pair sets. The candidate local alignments are indicated by using the same colors of the corresponding e-stack pair sets in the compatible graph.

their alignment scores and then in ascending order by their RMSD.

RESULTS

Benchmarking tools

To show that the local alignment tools can find more conserved substructures than that are included by optimal global alignments, LocalSTAR3D was compared with global alignment tools, STAR3D and SARA. The reason we choose STAR3D and SARA among many global alignment tools is that they aim to maximize the matched nucleotides and/or base pairs within a cutoff of RMSD, which are the similar metrics used in this study. Next, to assess the quality of local alignments, LocalSTAR3D was benchmarked with other RNA 3D structural local alignment tools, ARTS and iPARTS2. Default parameters were used for STAR3D, SARA, LocalSTAR3D and ARTS. The 'Local alignment' parameter was used in benchmarking for iPARTS2.

Metrics used in benchmarking

We evaluated the alignment quality by both accuracy and coverage. A better local RNA 3D structure alignment should be a connected structure with greater length and lower RMSD. Therefore, we parsed the output from each tool and extracted Aligned Connected Structures (ACSs) in each alignment for comparison. Later we compared the ACSs by using the Percentage of Connected Structural Identity (PCSI) and the Percentage of aligned Connected Secondary Structure (PCSS) values. Metrics ACS, PCSI, and PCSS are defined below.

To define ACS, we first define connected nucleotides. For nucleotides n_i^A and n_j^A in RNA A, n_i^A and n_j^A are considered as connected, if n_i^A is pairing with n_j^A, or n_i^A and n_j^A are within a certain distance in sequence (5 nucleotides in this study). In the alignment between two RNA structures, A and B, two aligned nucleotide pairs (n_i^A, n_i^B) and (n_j^A, n_j^B) are considered connected, if n_i^A is connected to n_j^A or n_i^B is connected to n_j^B. By using this definition of connected nucleotides, insertions and deletions are allowed in one of the input RNAs. ACSs are defined as the maximal connected components formed by connected nucleotides.

PSI (Percentage of Structural Identity) and PSS (Percentage of aligned Secondary Structure) have been used to evaluate the quality of global 3D structural alignments (9,32,33). Extended from PSI and PSS, we define PCSI and PCSS to evaluate the quality of local 3D structural alignments. PCSI is defined as the percentage of the number of aligned nucleotides in an ACS within 4 Å distance with respect to the length of the shorter RNA sequence. PCSS is defined as the percentage of the number of aligned base pairs in an ACS within 4 Å distance with respect to the smaller number of base pairs among those two RNAs.

Alignment quality assessment using R-FSCOR dataset

The R-FSCOR dataset (3) that contains 194 RNAs was used to assess the alignment quality in this study. The R-FSCOR dataset is a collection of the representatives of the RNA clusters grouped with at least 90% structural identity in the SCOR dataset (34). The five tools are compared by calculating PCSI and PCSS values of the all-to-all alignments for the R-FSCOR dataset. For global alignment tools STAR3D and SARA, the optimal global alignments were cut into non-overlapping ACSs. For the local alignment tools, the largest ACS was extracted from each of the top five local alignments. The overlapping ACSs were removed. The resulting ACSs were sorted by their PCSI and PCSS. The top two ACSs are used for the comparison. None of the tools can generate alignments for all the input RNA 3D structures. The combinations of 194 RNA chains in the R-FSCOR dataset generate 18 721 RNA pairs in total. By using the R-FSCOR dataset as the input, STAR3D generated 17 809 outputs, SARA generated 17 157 outputs, ARTS generated 11 980 outputs, and LocalSTAR3D generated 18 136 outputs. The cases where the LocalSTAR3D could not generate alignment were due to the failure of finding the e-stack pairs of default length 3. To detect the conserved structures in such cases, users can specify a smaller length (e.g. 2) of the e-stack pairs, with the trade-off of a potential longer run time. Since the standalone program of iPARTS2 is unavailable and it is impractical to run a large number of alignments on the iPARTS2 web server, a subset of 1000 RNA pairs were randomly chosen from all possible RNA pairs in the R-FSCOR dataset. The random RNA pairs were generated by using a fixed seed 12 345 for reproducibility. In the randomly selected subset, iPARTS2 generated 991 outputs. To make the comparison fair, the results were used in comparison only if both LocalSTAR3D and the corresponding benchmarking tool had outputs. Figure 3 shows the cumulative frequency curves of PCSI and PCSS for the top two ACSs to compare LocalSTAR3D and other tools. The mean PCSI and mean PCSS for each tool are shown in Table 1. The numbers of the cases where one tool is better than the other are shown in Supplementary Table S1. As shown in Figure 3, the cumulative frequency curves of the second ACS for global alignment tools are close to the origin point, indicating their second ACSs are very small in most cases. None of the five tools can always find the second ACS. So the cumulative frequency curves in the second ACS start at the points in which the y-axis values are smaller than 1. As shown in the Figure 3, stack-based tools ARTS, STAR3D, and LocalSTAR3D have better PCSS, compared to SARA and iPARTS2. STAR3D has slightly better PCSS and lower PCSI than LocalSTAR3D. One of the reasons is that LocalSTAR3D applies the adjacent constraint when searching for the e-stack pairs. Therefore LocalSTAR3D finds less but adjacent e-stack pairs compared to STAR3D. Another reason is that LocalSTAR3D applies an RMSD cutoff for the whole alignment (the default value is 4 Å). Users can specify a higher RMSD cutoff for alignments to get a higher coverage. Among the 18 136 outputs that LocalSTAR3D generated for the R-FSCOR dataset, 12 871 contain more than one local alignment. In addition, LocalSTAR3D was the only tool that generated alignments for which both the first and the second ACSs have PCSI and PCSS equal to 1. There are 239 such cases in the alignments that LocalSTAR3D generated for the R-FSCOR dataset. The second ACS having PCSI and PCSS equal to 1 is possible when one of the input RNAs is shorter than the half of the other one. In these cases, the full length of the shorter RNA is aligned to more than one location in the longer RNA. As shown in Figure 3D, although iPARTS2 generated a second ACS in more cases than LocalSTAR3D, the PCSI and PCSS values are smaller in most cases. It is also worth mentioning that the PCSS values of iPARTS2 are significantly smaller than those of LocalSTAR3D. The main reason is that iPARTS2 reduces the RNA 3D structure into a 1D sequence and loses most of the base pair information.

Kink-turn motifs on 23S rRNA

Kink-turn motifs are well studied recurrent internal loops that produce sharp turns (kinks) in its two supporting helices (35). There are multiple known instances of kink-turn motifs in the Haloarcula marismortui 23S rRNA, which makes it an ideal example to test the performance of the local alignment tools. The known instances summarized in a study of RNA structural motifs in ribosome RNAs (36) were used here as ground truth. While global alignment tools output an optimal alignment of the kink-turn motif against the 23S rRNA, local alignment tools are supposed to generate more than one local alignment. The kink-turn motif (PDB: 4bw0, chain A, 26 nucleotides) and the Haloarcula marismortui 23S rRNA (PDB: 1s72, chain 0, 2922 nucleotides) were used in this test. The 'no dangling end' setting was used in LocalSTAR3D. As summarized in Table 2, LocalSTAR3D found more known kink-turn motif instances than other local alignment tools. The locations overlapping with known kink-turn motif instances are marked with '*'. Some of the kink-turn motif instances are

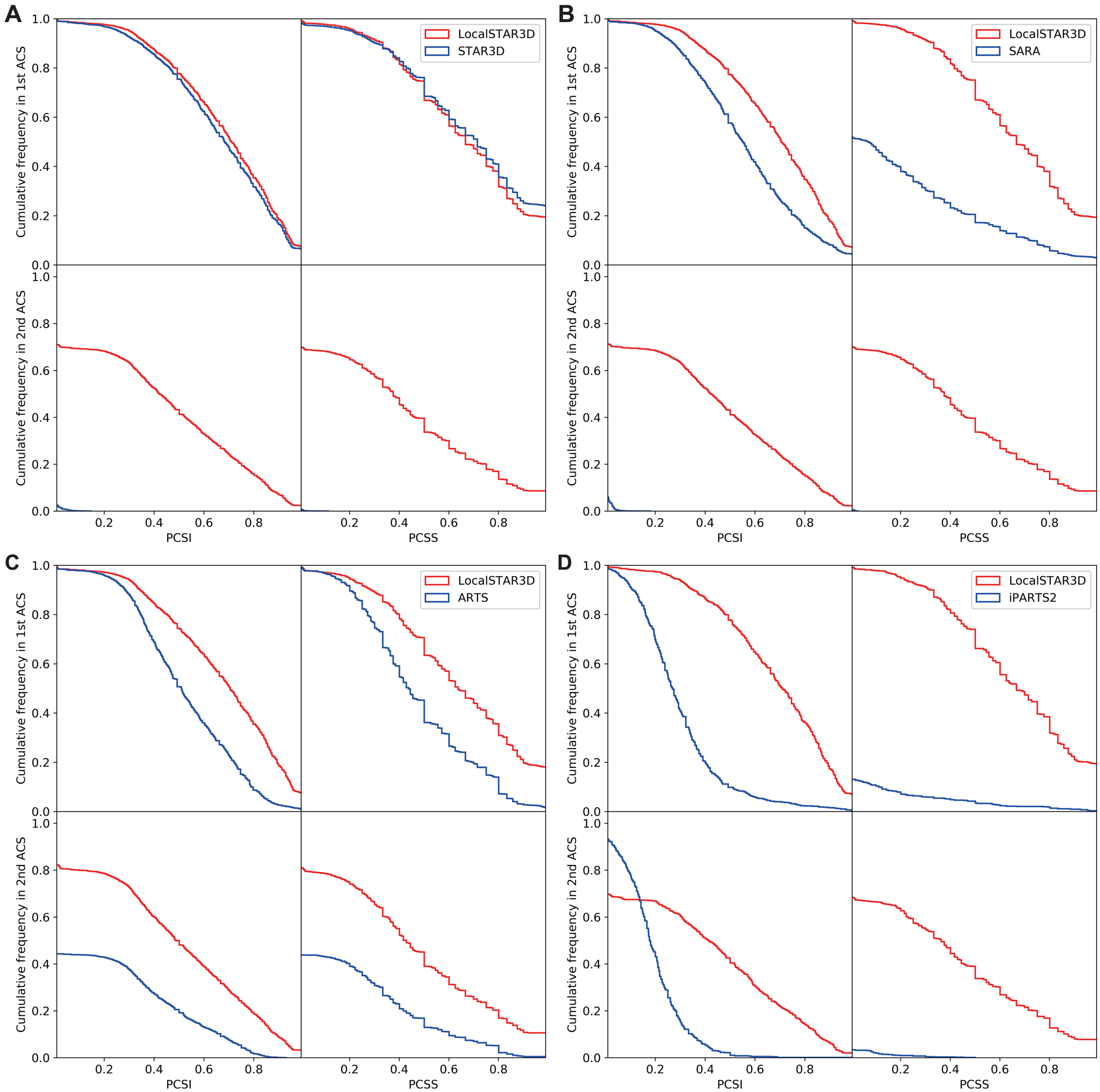

Figure 3. The cumulative frequencies of the PCSI and the PCSS values of LocalSTAR3D, STAR3D, SARA, ARTS and iPARTS2. (**A**) LocalSTAR3D versus STAR3D. (**B**) LocalSTAR3D versus SARA. (**C**) LocalSTAR3D versus ARTS. (**D**) LocalSTAR3D versus iPARTS2.

Table 1. The comparison of mean PCSI and mean PCSS values between LocalSTAR3D and four other tools by using the R-FSCOR dataset

	# of overlapped alignments	STAR3D		SARA		ARTS		iPARTS2		LocalSTAR3D	
		PCSI	PCSS	PCSI	PCSS	PCSI	PCSS	PCSI	PCSS	PCSI	PCSS
STAR3D vs. LocalSTAR3D	17 806	0.66/0	**0.68**/0							**0.69/0.41**	0.66/**0.39**
SARA vs. LocalSTAR3D	16 971			0.56/0	0.22/0					**0.68/0.41**	**0.66/0.39**
ARTS vs. LocalSTAR3D	11 976					0.53/0.22	0.47/0.19			**0.68/0.48**	**0.64/0.45**
iPARTS2 vs. LocalSTAR3D	978							0.31/0.19	0.05/0.01	**0.68/0.40**	**0.66/0.38**

Each cell has two values separated by '/'. The first value is corresponding to the first ACS. The second value is corresponding to the second ACS. The better mean values are set to bold. Values that are smaller than 0.01 are shown as 0.

Table 2. Kink-turn (PDB: 4bw0, chain A) motif search results on the Haloarcula marismortui 23s rRNA (PDB: 1s72, chain 0) by using LocalSTAR3D, iPARTS2 and ARTS

Ranking	Location	# of aligned nt	RMSD	Known
LocalSTAR3D				
1	75–83/91–102	21	2.21 Å	*
2	1585–1610	23	3.62 Å	*
3	934–942/1024–1036	22	3.73 Å	*
4	243–252/257–268	21	2.44 Å	*
5	1307–1319/1338–1347	20	3.52 Å	*
6	1142–1156/1211–1221	21	2.94 Å	*
iPARTS2				
1	76–101	23	4.06 Å	*
2	1592–1610	19	4.91 Å	*
3	742–749	8	6.07 Å	
4	683–706	24	14.11 Å	
5	803–811	9	0.67 Å	
ARTS				
1	553–554/1324–1335	14	1.16 Å	
2	250–262	13	1.05 Å	*
3	218/388–403	13	1.26 Å	
4	466–478	13	1.30 Å	
5	278–283/367–373	13	1.36 Å	

'*' indicates that this location overlaps with a known kink-turn motif instance.

located near a hairpin loop, which may be included in the local alignment. In this case, the region in the local alignment would be shown as a single-stranded region. LocalSTAR3D outputs the top five local alignments by default, which is the same as iPARTS2 in its local alignment mode. More alignments can be obtained from LocalSTAR3D by changing the default parameter. For the local alignments between the kink-turn motif and the Haloarcula marismortui 23S rRNA, all the top six hits from LocalSTAR3D are known instances. Both ARTS and iPARTS2 started to generate false positives from the third alignment. The alignments after the top five are not shown for ARTS and iPARTS2. Since iPARTS2 encodes the RNA 3D structures into 1D sequences, all the local alignments from iPARTS2 are single stranded. Both true positives from iPARTS2 overlap with the kink-turn motif instances that are located near a hairpin loop. The only true positive generated by ARTS only covers one strand of the kink-turn motif.

Self-splicing group II introns

Group II introns are well-studied RNAs that splice via two transesterification reactions. Group II introns are grouped into three classes: IIA, IIB and IIC (37,38). The sequences of group II introns in the different classes are diverse, but conserved substructures can still be found among them. Among all six domains, both domain V and domain VI are involved in the self-splicing, and domain V was reported as the most conserved domain (39,40). To test the local RNA 3D structural alignment tools, one representative RNA in each class was selected. We selected 5g2x for group IIA (39), 4r0d for IIB (40), and 3igi for IIC (41) from PDB. Three local alignment tools were compared by one-to-one alignments of these three group II intron RNAs. The top alignment generated by each tool was used for comparison. The alignments generated by all three tools for 5g2x and 4r0d are shown in Figure 4. The complete comparisons are sum-

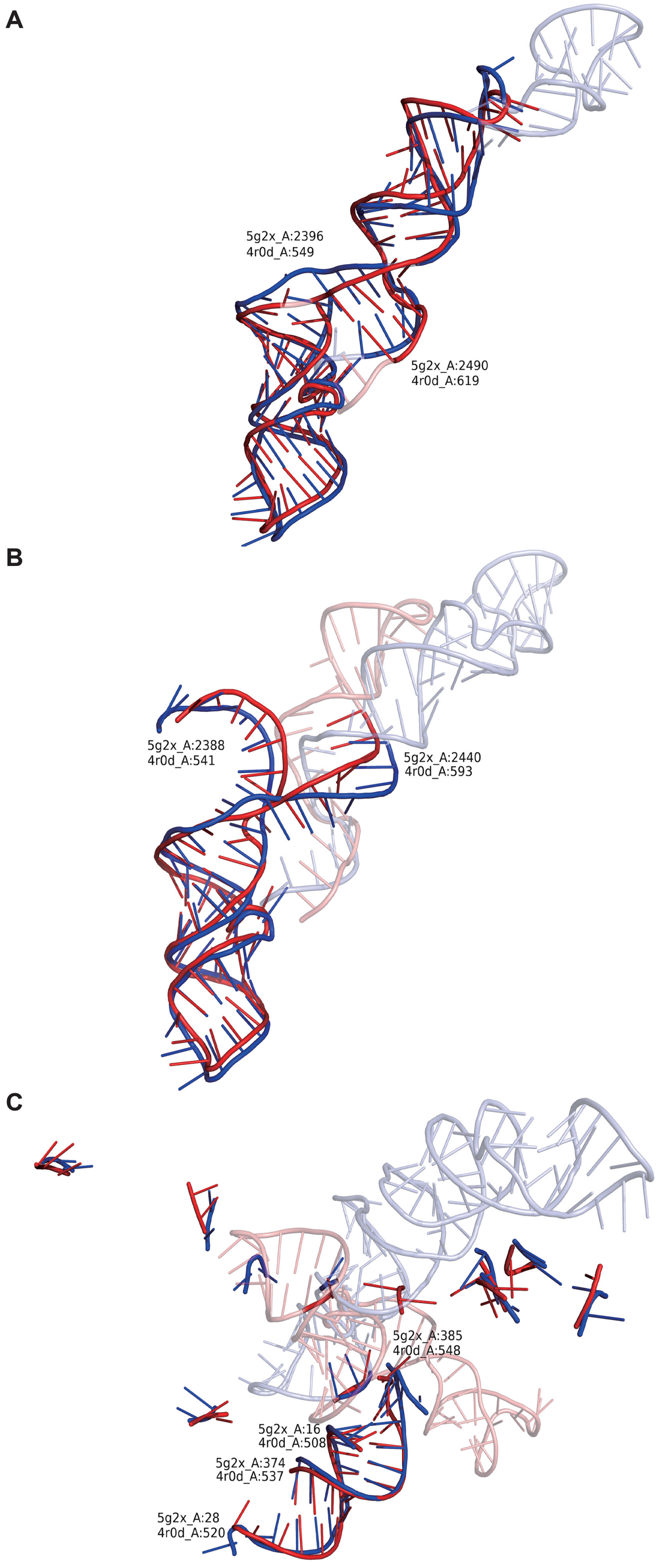

Figure 4. The best local alignment between self-splicing group II introns 5g2x and 4r0d generated by LocalSTAR3D, iPARTS2 and ARTS. The aligned parts are marked as blue (5g2x) and red (4r0d). The complete domain V and domain VI are displayed in half transparent light blue (5g2x) and salmon color (4r0d). (**A**) The result of LocalSTAR3D. (**B**) The result of iPARTS2. (**C**) The result of ARTS.

Table 3. Local alignment of Group II intron RNAs by using LocalSTAR3D, iPARTS2, and ARTS

RNAs	Tools	# of aligned nt	RMSD	# of aligned nt in ACS
5g2x, 4r0d	LocalSTAR3D	70	3.97 Å	**70**
	iPARTS2	53	4.42 Å	53
	ARTS	52	2.00 Å	39
5g2x, 3igi	LocalSTAR3D	36	3.23 Å	**36**
	iPARTS2	26	4.53 Å	26
	ARTS	45	1.90 Å	35
4r0d, 3igi	LocalSTAR3D	118	3.88 Å	**118**
	iPARTS2	57	11.10 Å	57
	ARTS	145	1.93 Å	66

The largest ACS for each pair of RNAs is set to bold.

marized in Table 3. The alignments are compared by the number of aligned nucleotides, RMSD, and the number of nucleotides in the largest ACSs. LocalSTAR3D generated ACSs whose lengths are larger than other tools. As shown in Figure 4, the alignment for 5g2x and 4r0d generated by LocalSTAR3D contains both Domain V and Domain VI that are involved in the self-splicing, while the alignment generated by iPARTS2 only contains Domain V with a few up/downstream nucleotides. In the alignment for 5g2x and 3igi, although the number of nucleotides in the largest ACS in LocalSTAR3D's alignment is slightly greater than that in ARTS's alignment, LocalSTAR3D's alignment overlaps with the Domain V of both group II intron structures, while ARTS's alignment missed this most conserved domain. The local alignments of the self-splicing group II introns show that LocalSTAR3D can identify the conserved RNA 3D substructures that have the same biological function.

The tRNA mimicry of Viral IRES RNAs

To further illustrate the potential application of LocalSTAR3D, we used LocalSTAR3D to study the tRNA mimicry of viral internal ribosome-entry site RNAs (IRESs). Canonical eukaryotic translation initiation is a highly complicated mechanism, in which a unique nucleotide structure at the 5$'$ end of the mRNAs, known as the cap structure, plays an important role. Instead of using the cap structure, the translation initiation of some viral RNAs use a cap-independent mechanism, which is driven by IRES RNAs (42). These viral RNAs thereby hijack the translation machinery in infected cells and efficiently outcompete canonical mRNAs (43). It was firmly established that their flexible structures are critical for IRESs to initiate translation. Studying the IRES structures may help us to understand the mechanistic principles of the ribosome, and to eventually have better control of virus infection through methods such as vaccine design.

IRESs are highly diverse in structure and mechanism but accomplish the same molecular tasks to form an elongation-component 80S ribosome (42). According to the factors they need to function, the viral IRES RNA were clustered into four types (44). Type IV IRESs are the most autonomous IRESs, which do not require the initiation factors and initiator tRNA (43). CrPV IRES is the most well-studied RNA in type IV IRESs, which was first discovered in Australian field crickets, and can cause high mortality in crickets and olive fruit flies (45). Jan *et al.* proved that it can mimic the function of a Met-tRNA$_i$, by examining a P-site-occupied CrPV IRES in a minimal reconstituted system (46). The first high-resolution crystal structure of CrPV IRES revealed that the stem-loop in its pseudo-knot I (PK I) has a highly similar structure with the tRNA's anticodon loop (47).

To study the tRNA mimicry of the type IV IRES, we aligned CrPV IRES (PDB: 6d90, chain 4) (48) to a canonical P site tRNA (PDB: 4v5c, chain AV) (49) by using LocalSTAR3D. An approximate superimposition of these two RNA 3D structures was shown in a previous study (48). The top alignment between CrPV IRES and the tRNA generated by LocalSTAR3D is at 6d90:4(6177–6199) and 4v5c:AV(25–45) with RMSD 2.67 Å. The aligned regions in two structures and the superimposition of these regions are shown in Figure 5. This alignment contains the tRNA anticodon loop mimicry region identified in previous report (47). In addition to the loop region, LocalSTAR3D revealed the similarity in the stacks that support the loops. Since ARTS does not accept atomic coordinates in PDBx format, it can not be used to align these two RNA 3D structures. The top local alignment generated by iPARTS2 for these two RNA 3D structures only contains four nucleotides and does not overlap with the anticodon loop in tRNA. This example shows that LocalSTAR3D can detect not only the conserved substructures between RNAs in the same class, but also the similar substructure between unrelated RNAs.

Run time

A comparison of the Run time of three local RNA 3D alignment tools are summarized in Table 4. LocalSTAR3D and ARTS were tested on an Ubuntu 16.04 system running on a desktop computer with an i7-8700 CPU. iPARTS2 was tested by using its web server. We first compare the run time by using the examples presented above and a pair of rRNAs. As shown in Table 4, for relatively small input RNAs (rows 1–2), ARTS is the fastest tool. For larger input RNAs, such as the Thermus thermophilu 16S rRNA (PDB:1j5e, chain A, 1522 nucleotides) and the Haloarcula marismortui 23S rRNA (PDB:1s72, chain 0, 2922 nucleotides), LocalSTAR3D is faster, because it only considers the top 10 000 e-stack pairs by default. The major aim of LocalSTAR3D is to identify the conserved substructures between a pair of non-homologous RNAs. For a pair of non-homologous RNAs, LocalSTAR3D generates less than 10 000 e-stack pairs during the alignment in most cases. In rare cases where the pair of input RNAs have very large conserved substructures, users can increase the cutoff of the number of e-stack pairs in LocalSTAR3D to obtain slightly improved local alignments. The run time per nucleotide for each tool is evaluated by using the R-FSCOR dataset. As shown in Supplementary Table S2, LocalSTAR3D generated aligned nucleotides faster than iPARTS2 but slower than ARTS. For each pair of RNA 3D structures, it is worth noting that ARTS may generate overlapping aligned nucleotides in different local alignments, while LocalSTAR3D generates non-overlapping local alignments. In conclusion, LocalSTAR3D can find local alignments for most of the RNAs

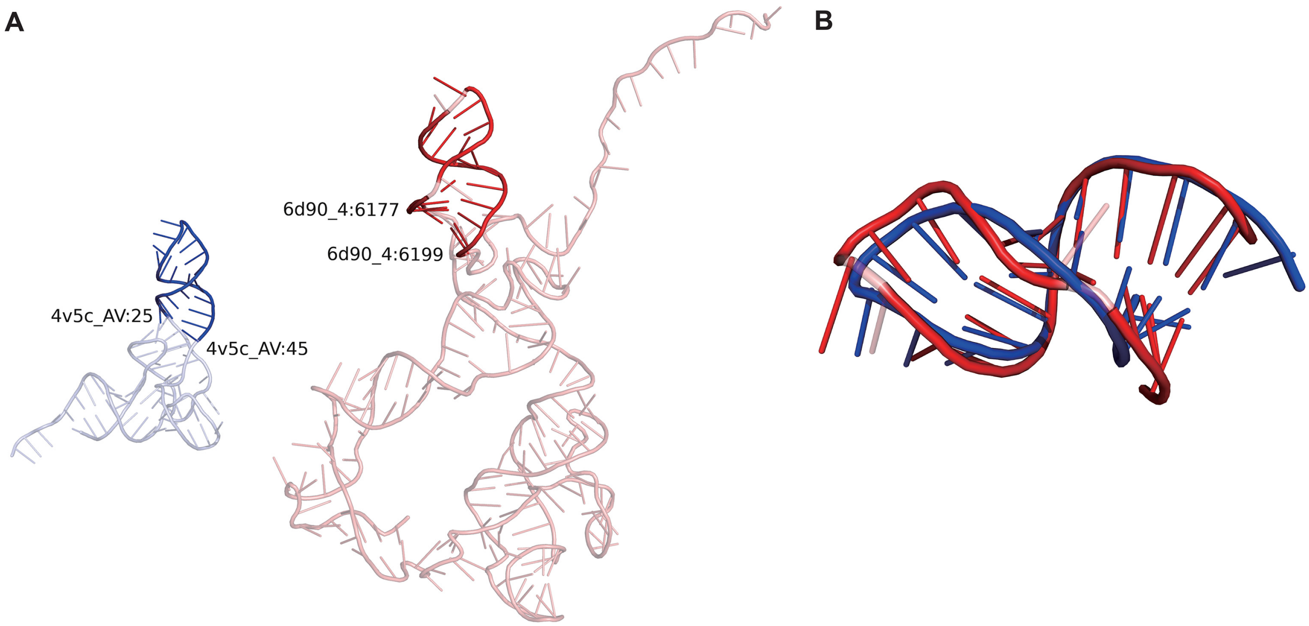

Figure 5. The structure alignment between CrPV IRES and tRNA. (**A**) The aligned regions are highlighted in tRNA (left) and CrPV IRES (right). The boundaries of the regions are labeled by PDB ids, chain ids, and the indices of the nucleotides. (**B**) The superimposition of the aligned regions.

Table 4. Run time of ARTS, iPARTS2 and LocalSTAR3D (in seconds)

RNAs	ARTS	iPARTS2	LocalSTAR3D
Kink-turn and 23S rRNA	**0.3**	14.4	0.9
Group II introns	**0.3**	23.4	17.8
tRNA and CrPV IRES	N/A	3.2	**1.4**
16S rRNA and 23S rRNA	53.1	375.3	**25.3**

The best performance is set to bold.

in one minute on a modern desktop. For large RNAs, LocalSTAR3D is faster than other state-of-the-art tools.

DISCUSSION

In this study, we developed a novel local RNA 3D structural alignment tool, LocalSTAR3D. The benchmark results show that LocalSTAR3D generates better local alignment than the other state-of-the-art tools. We have presented some interesting biology cases to illustrate the utility of LocalSTAR3D. LocalSTAR3D can find conserved RNA substructures from small RNA elements, such as kink-turn motif, to whole conserved domains. It can also detect the structure mimicry of viral RNAs, such as the tRNA mimicry of CrPV IRES RNA.

Further analysis of the results generated by LocalSTAR3D is still required. LocalSTAR3D can be integrated into a clustering pipeline, which will be similar to the one developed by our lab for motif clustering (50). The clustering pipeline equipped with LocalSTAR3D will be able to generate conserved local structures shared by multiple RNA molecules, which will provide further insight into their functional and evolutionary relations.

Another direction of future study is to take into account the long-range interaction. LocalSTAR3D considers local components in RNA molecules as a collection of stems and loops that are adjacent in sequence. This adjacent relation can be extended to the 3D space. We plan to calculate the spatial distance between nucleotides, which will be used to determine if these nucleotides are neighbors in 3D space. By utilizing the spatial adjacency, we will be able to detect the conserved RNA substructures involving long-range interaction.

DATA AVAILABILITY

http://genome.ucf.edu/LocalSTAR3D.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

The authors thank Chin-Lung Lu and Yu-Han Chen for the help to access the iPARTS2 server.

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. Doudna,J.A. (2000) Structural genomics of RNA. *Nat. Struct. Biol.*, **7**, 954–956.
2. Capriotti,E. and Marti-Renom,M.A. (2008) RNA structure alignment by a unit-vector approach. *Bioinformatics*, **24**, i112–118.
3. Capriotti,E. and Marti-Renom,M.A. (2009) SARA: a server for function annotation of RNA structures. *Nucleic Acids Res.*, **37**, W260–265.

4. Bauer,R., Rother,K., Moor,P., Reinert,K., Steinke,T., Bujnicki,J. and Preissner,R. (2009) Fast structural alignment of biomolecules using a hash table, N-grams and string descriptors Algorithms. *Algorithms*, **2**, 692–709.
5. Rahrig,R.R., Leontis,N.B. and Zirbel,C.L. (2010) R3D Align: global pairwise alignment of RNA 3D structures using local superpositions. *Bioinformatics*, **26**, 2689–2697.
6. Rahrig,R.R., Petrov,A.I., Leontis,N.B. and Zirbel,C.L. (2013) R3D Align web server for global nucleotide to nucleotide alignments of RNA 3D structures. *Nucleic Acids Res.*, **41**, 15–21.
7. Cech,P., Svozil,D. and Hoksza,D. (2012) SETTER: web server for RNA structure comparison. *Nucleic Acids Res.*, **40**, W42–W48.
8. Nguyen,M.N. and Verma,C. (2015) Rclick: a web server for comparison of RNA 3D structures. *Bioinformatics*, **31**, 966–968.
9. Ge,P. and Zhang,S. (2015) STAR3D: a stack-based RNA 3D structural alignment tool. *Nucleic Acids Res.*, **43**, e137.
10. Laborde,J., Robinson,D., Srivastava,A., Klassen,E. and Zhang,J. (2013) RNA global alignment in the joint sequence-structure space using elastic shape analysis. *Nucleic Acids Res.*, **41**, e114.
11. He,G., Steppi,A., Laborde,J., Srivastava,A., Zhao,P. and Zhang,J. (2014) RASS: a web server for RNA alignment in the joint sequence-structure space. *Nucleic Acids Res.*, **42**, W377–W381.
12. Gong,S., Zhang,C. and Zhang,Y. (2019) RNA-align: quick and accurate alignment of RNA 3D structures based on size-independent TM-scoreRNA. *Bioinformatics*, **35**, 4459–4461.
13. 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.
14. Zhong,C., Tang,H. and Zhang,S. (2010) RNAMotifScan: automatic identification of RNA structural motifs using secondary structural alignment. *Nucleic Acids Res.*, **38**, e176.
15. Zhong,C. and Zhang,S. (2015) RNAMotifScanX: a graph alignment approach for RNA structural motif identification. *RNA*, **21**, 333–346.
16. 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.
17. Chojnowski,G., Waleń,T. and Bujnicki,J.M. (2014) RNA Bricka database of RNA 3D motifs and their interactions. *Nucleic Acids Res.*, **42**, D123–D131.
18. Dror,O., Nussinov,R. and Wolfson,H. (2005) ARTS: alignment of RNA tertiary structures. *Bioinformatics*, **21**, 47–53.
19. Ferre,F., Ponty,Y., Lorenz,W.A. and Clote,P. (2007) DIAL: a web server for the pairwise alignment of two RNA three-dimensional structures using nucleotide, dihedral angle and base-pairing similarities. *Nucleic Acids Res.*, **35**, W659–W668.
20. Chang,Y.F., Huang,Y.L. and Lu,C.L. (2008) SARSA: a web tool for structural alignment of RNA using a structural alphabet. *Nucleic Acids Res.*, **36**, 19–24.
21. 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.
22. 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.
23. Gendron,P., Lemieux,S. and Major,F. (2001) Quantitative analysis of nucleic acid three-dimensional structures. *J. Mol. Biol.*, **308**, 919–936.
24. Lemieux,S. and Major,F. (2002) RNA canonical and non-canonical base pairing types: a recognition method and complete repertoire. *Nucleic Acids Res.*, **30**, 4250–4263.
25. Yang,H., Jossinet,F., Leontis,N., Chen,L., Westbrook,J., Berman,H. and Westhof,E. (2003) Tools for the automatic identification and classification of RNA base pairs. *Nucleic Acids Res.*, **31**, 3450–3460.
26. 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.
27. Waleń,T., Chojnowski,G., Gierski,P. and Bujnicki,J.M. (2014) ClaRNA: a classifier of contacts in RNA 3D structures based on a comparative analysis of various classification schemes. *Nucleic Acids Res.*, **42**, e151.
28. Islam,S., Ge,P. and Zhang,S. (2017) CompAnnotate: a comparative approach to annotate base-pairing interactions in RNA 3D structures. *Nucleic Acids Res.*, **45**, e136.
29. Smit,S., Rother,K., Heringa,J. and Knight,R. (2008) From knotted to nested RNA structures: a variety of computational methods for pseudoknot removal. *RNA*, **14**, 410–416.
30. Kabsch,W. (1978) A discussion of the solution for the best rotation to relate two sets of vectors. *Acta Crystallogr. A*, **34**, 827–828.
31. Bron,C. and Kerbosch,J. (1973) Algorithm 457: finding all cliques of an undirected graph. *Commun. ACM*, **16**, 575–577.
32. Capriotti,E. and Marti-Renom,M.A. (2008) RNA structure alignment by a unit-vector approach. *Bioinformatics*, **24**, i112–i118.
33. Hoksza,D. and Svozil,D. (2012) Efficient RNA pairwise structure comparison by SETTER method. *Bioinformatics*, **28**, 1858–1864.
34. Tamura,M., Hendrix,D.K., Klosterman,P.S., Schimmelman,N.R., Brenner,S.E. and Holbrook,S.R. (2004) SCOR: Structural Classification of RNA, version 2.0. *Nucleic Acids Res.*, **32**, D182–D184.
35. 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.
36. Zhong,C. and Zhang,S. (2012) Clustering RNA structural motifs in ribosomal RNAs using secondary structural alignment. *Nucleic Acids Res.*, **40**, 1307–1317.
37. Michel,F., Kazuhiko,U. and Haruo,O. (1989) Comparative and functional anatomy of group II catalytic intronsa review. *Gene*, **82**, 5–30.
38. Toor,N., Hausner,G. and Zimmerly,S. (2001) Coevolution of group II intron RNA structures with their intron-encoded reverse transcriptases. *RNA*, **7**, 1142–1152.
39. Qu,G., Kaushal,P.S., Wang,J., Shigematsu,H., Piazza,C.L., Agrawal,R.K., Belfort,M. and Wang,H.-W. (2016) Structure of a group II intron in complex with its reverse transcriptase. *Nat. Struct. Mol. Biol.*, **23**, 549.
40. Robart,A.R., Chan,R.T., Peters,J.K., Rajashankar,K.R. and Toor,N. (2014) Crystal structure of a eukaryotic group II intron lariat. *Nature*, **514**, 193–197.
41. Toor,N., Keating,K.S., Fedorova,O., Rajashankar,K., Wang,J. and Pyle,A.M. (2010) Tertiary architecture of the Oceanobacillus iheyensis group II intron. *RNA*, **16**, 57–69.
42. Kieft,J.S. (2008) Viral IRES RNA structures and ribosome interactions. *Trends Biochem. Sci.*, **33**, 274–283.
43. Yamamoto,H., Unbehaun,A. and Spahn,C.M. (2017) Ribosomal chamber music: toward an understanding of IRES mechanisms. *Trends Biochem. sci.*, **42**, 655–668.
44. Filbin,M.E. and Kieft,J.S. (2009) Toward a structural understanding of IRES RNA function. *Curr. Opin. Struct. Biol.*, **19**, 267–276.
45. Manousis,T. and Moore,N.F. (1987) Cricket paralysis virus, a potential control agent for the olive fruit fly, Dacus oleae Gmel. *Appl. Environ. Microbiol.*, **53**, 142–148.
46. Jan,E., Kinzy,T.G. and Sarnow,P. (2003) Divergent tRNA-like element supports initiation, elongation, and termination of protein biosynthesis. *Proc. Natl. Acad. Sci. U.S.A.*, **100**, 15410–15415.
47. Costantino,D.A., Pfingsten,J.S., Rambo,R.P. and Kieft,J.S. (2008) tRNA-mRNA mimicry drives translation initiation from a viral IRES. *Nat. Struct. Mol. Biol.*, **15**, 57–64.
48. Pisareva,V.P., Pisarev,A.V. and Fernandez,I.S. (2018) Dual tRNA mimicry in the cricket paralysis virus IRES uncovers an unexpected similarity with the Hepatitis C Virus IRES. *Elife*, **7**, e34062.
49. Voorhees,R.M., Weixlbaumer,A., Loakes,D., Kelley,A.C. and Ramakrishnan,V. (2009) Insights into substrate stabilization from snapshots of the peptidyl transferase center of the intact 70S ribosome. *Nat. Struct. Mol. Biol.*, **16**, 528–533.
50. 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.