Published online 15 July 2015

Nucleic Acids Research, 2015, Vol. 43, No. 20 ***e137***
doi: 10.1093/nar/gkv697

STAR3D: a stack-based RNA 3D structural alignment tool

Ping Ge and Shaojie Zhang[*]

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

Received February 27, 2015; Revised June 18, 2015; Accepted June 26, 2015

ABSTRACT

The various roles of versatile non-coding RNAs typically require the attainment of complex high-order structures. Therefore, comparing the 3D structures of RNA molecules can yield in-depth understanding of their functional conservation and evolutionary history. Recently, many powerful tools have been developed to align RNA 3D structures. Although some methods rely on both backbone conformations and base pairing interactions, none of them consider the entire hierarchical formation of the RNA secondary structure. One of the major issues is that directly applying the algorithms of matching 2D structures to the 3D coordinates is particularly time-consuming. In this article, we propose a novel RNA 3D structural alignment tool, STAR3D, to take into full account the 2D relations between stacks without the complicated comparison of secondary structures. First, the 3D conserved stacks in the inputs are identified and then combined into a tree-like consensus. Afterward, the loop regions are compared one-to-one in accordance with their relative positions in the consensus tree. The experimental results show that the prediction of STAR3D is more accurate for both non-homologous and homologous RNAs than other state-of-the-art tools with shorter running time.

INTRODUCTION

Non-coding RNAs (ncRNAs) play diverse cellular functions in biological systems (1–4). Unlike mRNAs whose primary sequences are genetic codes for protein synthesis, the regulatory information of most ncRNAs is encoded in their architectures: the secondary structures defined by the hierarchical assembly of double-stranded stacks and higher-order three-dimensional (3D) structures consisting of packed secondary structure modules interlinked by tertiary interactions (5,6). Therefore, the structural alignments of such ncRNAs can provide essential insight to their functional and evolutionary relationships. However, compared to the development of the computational methods for RNA secondary structure analysis, the progress of RNA 3D structural alignment has been limited. Although the protein 3D structural alignment has been studied for years and many sophisticated methods have been proposed (7–11), it is hard to apply them directly to ncRNAs due to the different properties of their secondary structures.

Recently, with the rapid growth of RNA deposition in the Protein Data Bank (PDB) (12), a number of tools have been developed specifically for the alignments of RNA 3D structures. Generally, they can be categorized into two groups. In the first group, the base pairing interactions in the inputs are ignored or degraded into sequential information. Then the RNAs can be compared using the quadratic-time alignment algorithms. For example, both iPARTS (13) and LaJolla (14) represent RNA backbones as sequences of letters derived from the features of nucleotide torsion angles. iPARTS continues to apply conventional pairwise alignment methods to the encoded linear sequences, while LaJolla searches the similar ‘n-grams’ (substrings of length *n*) in the RNAs by using hash tables. Similar to LaJolla, FRIEs (15) also uses the matching of *k*-mer RNA fragments. In this method, a large set of training fragments from the PDB are clustered into tens of classes based on their structural properties. Each *k*-mer in an RNA can be labeled with the probabilities in these classes, and thus the similarity of two fragments can be measured with the dot product of their probability vectors. Rclick (16) is another RNA 3D structural alignment tool based on the detection of local similarity. The matches between *n*-body cliques (in which *n* member nucleotides satisfy that all pair-wise spatial distances are within a threshold) are determined by the superimposition of their atomic coordinates. With this local structural equivalence, the optimal global alignment is generated by using 3D least squares fitting. Unlike the previously mentioned tools, DIAL (17) incorporates base pairing interactions into its dynamic programming scoring function, which also accounts for sequence and torsion angle information. A penalty is assigned if the pairing attributes (paired or unpaired) of two aligned nucleotides are different. Elastic Shape Analysis (ESA) models the RNA 3D structures not as sequences but as curves in a four-dimensional space: the atomic coordinates are in 3D space and the sequence information is encoded as an additional dimension (18). Then

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

the similarity between two RNAs can be evaluated by minimizing their geodesic distance with a quadratic-time dynamic programming algorithm.

The other group of RNA 3D structural alignment tools relies on the comparison of base pairing interactions in the molecules. In ARTS (19), two successive base pairs are used together as a seed. The optimal matching of seeds in two RNAs is extended globally to the unpaired regions and the result is refined with the least squares fitting technique. Similar to ARTS, the final results of R3D Align (20) are assembled from the alignments of local neighborhoods. Neighbors are the spatially closest nucleotides in one RNA, which may imply interactions such as base pairs, tertiary interlinks and stacking contacts. The structurally similar neighbors in two RNAs are detected and the optimal combination of these local alignments is determined by employing maximum clique finding algorithm on a compatibility graph. SARA (21) does not discriminate the paired and unpaired regions in RNAs. Inspired by a protein 3D structural alignment method named MAMMOTH (22), SARA describes the backbone of an RNA as a series of unit-vectors. The distances between the unit spheres of inputs can be measured with URMS (unit-vector root mean square) and the corresponding global alignment is identified by using dynamic programming. The same procedure is applied only to base pairs if the pairing information is provided. The 3D structural alignment of entire RNAs can be optimized based on the mapping of pairing interactions. SETTER (23) integrates stacks and loops into the RNA 3D structural alignment method. It splits the RNA sequences into GSSUs (generalized secondary structure units), each of which has a loop, a neck and a stem. The highly similar GSSU pairs are used as seeds to guide the alignment of other GSSUs. To simplify the computation, the exact mapping of nucleotides is ignored in this method.

It can be seen that the RNA secondary structural information, in particular the hierarchical topology of stacks, is not used in the reviewed methods. However, the enclosing and juxtaposing relations between stacks provide more detailed structural information than what has been used in the existing tools, such as 'paired' or 'unpaired' attributes, base pairing interactions and stack positions. The issue is the difficulty of integrating the conventional RNA secondary structure alignment algorithms into the RNA 3D structural comparison. Given the high time complexity of these algorithms [at least $O(n^3)$] (24–27), applying them directly to the relatively complicated atomic coordinates will increase the computational complexity significantly.

Here, we propose a novel RNA 3D structural alignment tool called STAR3D that explicitly makes use of the conservation of secondary structures with high efficiency. It aims at finding the consensus of stacks by using 2D topology and 3D geometry first, and then uses it to guide the alignments of the loop regions. To achieve this goal, first, the sub-stacks with similar 3D structures are detected and assembled into conserved stack pairs. Then, a compatible graph is constructed based on their secondary structural relations and spatial distances. In this graph, the maximum clique can be converted into a tree-like consensus structure of two RNAs. After that, the loop regions are ordered by the common tree. Each of them only needs to be compared with its partner by using 3D information. STAR3D has been implemented in Java. The benchmarking results show that STAR3D outperforms the state-of-the-art RNA 3D structural alignment tools with high efficiency.

MATERIALS AND METHODS

Preprocessing

The inputs of STAR3D are the atomic coordinates of two polymer RNA chains, which are presented in the corresponding PDB files. They are preprocessed to obtain the RNA secondary structures to guide the 3D structural alignment. All plausible pairing interactions are identified by using MC-Annotate (28,29). Among them, the Watson–Crick base pairs (A↔U, C↔G) and wobble base pairs (G↔U) are retrieved to form the RNA secondary structures (30). Other pairing interactions are considered during the loop alignment. In order to avoid excessive computation, we eliminate the crossing base pairs in the secondary structures by using the program RemovePseudoknots (31) in the RNAstructure package (32). The discarded stems are used as pairing interactions in the loops.

Stack decomposition

Helical structured stacks are formed by consecutively nested Watson–Crick base pairs and wobble base pairs. To detect the 3D structural conservation in the stacks efficiently, the double-stranded regions in the pseudoknot-free secondary structures are decomposed into consecutive sub-stacks of size k, namely k-stacks. A stack with l base pairs ($l \geq k$) can be divided into $l - k + 1$ overlapping k-stacks. All the possible k-stacks are collected for further processing.

Based on the definition of k-stack, we introduce some basic notations. Given an RNA A, the 3D coordinates of the i-th residue are denoted as $A[i]$. At the secondary structure level, the set of k-stacks in the pseudoknot-free structure is denoted as $\mathcal{P}^A$. For a specific k-stack $p^A \in \mathcal{P}^A$, the index of the leftmost base (5′ end) is represented as $b(p^A)$ and the index of the rightmost base (3′ end) is represented as $e(p^A)$. Thus the 3D coordinates of the double-stranded subsequences in p^A are $A[b(p^A)\ldots b(p^A) + k - 1]$ and $A[e(p^A) - k + 1\ldots e(p^A)]$, which are defined as $3D(p^A)$.

Detecting the conserved stack regions

STAR3D identifies the stack components conserved in 3D structures as anchors and uses them to constrain the global alignment. Similar approaches have been applied in numerous computation-efficient tools for genome alignment (33,34) and RNA secondary structure alignment (35,36). The difference is that STAR3D uses the 3D coordinates of atoms to detect the potential homologous regions. Given the fact that RNA stacks adopt an A-form helical conformation, a major issue needs to be addressed: whether the 3D structural similarity of conserved stack regions is significant enough to distinguish them from the random ones. To answer the question, we have conducted a statistical research on the matched stacks in a hand-crafted alignment (37). The survey results in Figure 1 indicate that the orthologous sub-stacks have highly similar 3D structures, and they can be

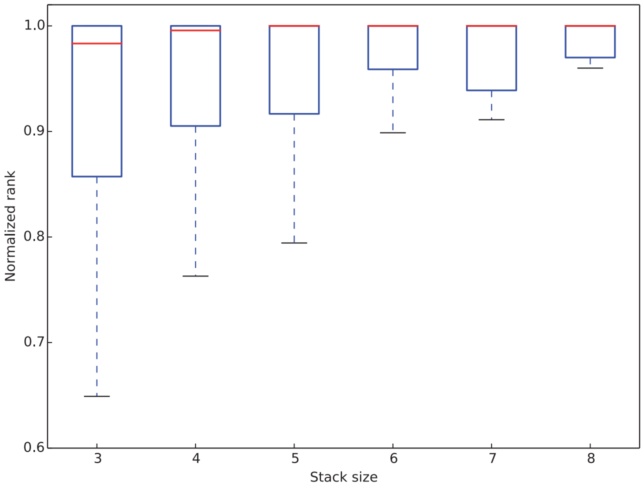

Figure 1. The normalized ranks of the matched stacks in two 23S rRNAs (PDB 2j01, chain A and PDB 2aw4, chain B). The structure similarity is measured with RMSD. For a stack of size k in 2j01, its RMSDs to all the size-k stacks in 2aw4 are computed.

detected by evaluating RMSD. In our method, the k-stacks (default value of k is 3) in the inputs are retrieved as the building blocks for the larger conserved regions. Shorter helices are not considered because of their low occurrence in the real RNAs.

Given two input RNAs A and B, the two sets of k-stacks $\mathcal{P}^A$ and $\mathcal{P}^B$ are sorted in ascending order by the leftmost bases. The three-dimensionally conserved k-stacks in A and B are determined by their RMSDs. $C_{i,j}$, the indicator of conservation for p_i^A and p_j^B, is computed using the following function:

$$C_{i,j} = \begin{cases} 1 & RMSD(3\mathrm{D}(p_i^A), 3\mathrm{D}(p_j^B)) < r_c \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

The function $RMSD$ measures the average spatial distance between the superimposed residues in $3\mathrm{D}(p_i^A)$ and $3\mathrm{D}(p_j^B)$, with r_c being the RMSD cutoff (default value is 4 Å) (21). In our implementation, the RMSD values are computed with the Kabsch method (38) by using the geometric center of six backbone atoms C3', C4', C5', O3', O5' and P (39). The indicators for all pairs of k-stacks ($\mathcal{P}^A \times \mathcal{P}^B$) are stored in a matrix. Then, we extend the consecutive matches of k-stacks to form larger ungapped alignments. For instance, p_i^A, p_{i+1}^A and p_j^B, p_{j+1}^B can be merged into two aligned stacks of size $k+1$, if $C_{i,j} = C_{i+1,j+1} = 1$, $b(p_i^A) = b(p_{i+1}^A) - 1$, $e(p_i^A) = e(p_{i+1}^A) + 1$, $b(p_j^B) = b(p_{j+1}^B) - 1$ and $e(p_j^B) = e(p_{j+1}^B) + 1$. This procedure continues through the diagonals of the matrix until all the constructed alignments can not be extended any further. The two stack components in an assembled alignment are called extended stacks, written shortly as e-stacks. Correspondingly, the aligned e-stacks form e-stack pairs. We define the sets of e-stacks in A and B as $\mathcal{Q}^A$ and $\mathcal{Q}^B$, and the set of e-stack pairs as $\mathcal{S}$. According to the definition of e-stack, the cardinalities of $\mathcal{Q}^A$, $\mathcal{Q}^B$ and $\mathcal{S}$ are identical. As a result, we denote the members of a specific e-stack pair $s_i (\in \mathcal{S})$ as $q_i^A (\in \mathcal{Q}^A)$ and $q_i^B (\in \mathcal{Q}^B)$ $(s_i = (q_i^A, q_i^B))$. Note that e-stacks may overlap with each other [see Figure 2A]. Unlike k-stacks, the sizes of e-stacks are not fixed. Therefore, we define a new notation $l(q^A)$ to represent the number of base pairs in q^A. Hence $3D(q^A)$ are $A[b(q^A) \ldots b(q^A) + l(q^A) - 1]$ and $A[e(q^A) - l(q^A) + 1 \ldots e(q^A)]$.

For some large RNAs, the numbers of e-stack pairs are too large for computation. To determine the highly significant ones, we consider two criteria: the RMSD between two e-stacks and their size. The significant scores of e-stack pairs are defined using the formula $RMSD(3D(q_i^A), 3D(q_i^B)) - 0.1 \times l(q_i^A)$. They are sorted in ascending order and only 200 top-ranked pairs are retained for further processing. Based on our study, 200 high-scoring e-stack pairs are sufficient to cover most of the conserved helical regions in 23S rRNAs, the largest RNAs in PDB. More e-stack pairs may be used by setting the parameter if more complex structures are presented.

Assembling the consensus of stacks

To generate a consensus of stacks, the positions of e-stack pairs in the secondary structures and 3D space are analyzed. In the pseudoknot-free secondary structure of A, two e-stacks q_i^A and q_j^A may have one of the three following relations: (i) q_i^A and q_j^A are overlapping (denoted by $q_i^A \otimes q_j^A$); (ii) q_i^A encloses q_j^A (denoted by $q_i^A \prec_E q_j^A$); (iii) q_i^A is before and juxtaposed to q_j^A (denoted by $q_i^A \prec_J q_j^A$). In our algorithm, q_i^A *directly* encloses q_j^A if $q_i^A \prec_E q_j^A$ and $\nexists k(q_i^A \prec_E q_k^A \prec_E q_j^A)$ (denoted by $q_i^A <_E q_j^A$). Similarly, we say q_i^A is *directly* before and juxtaposed to q_j^A if $q_i^A \prec_J q_j^A$ and $\nexists k(q_i^A \prec_J q_k^A \prec_J q_j^A)$ (denoted by $q_i^A <_J q_j^A$).

Notice that both $\prec_E$ and $\prec_J$ are strict partial orders, so the non-overlapping e-stacks in an RNA can form a directed acyclic graph. It is well-known that the RNA secondary structures have a tree-like topology (40–43). Thus we model the non-overlapping relations of e-stacks in A as a tree:

(i) Assign a pseudo stack $q_\bullet^A$ $(b(q_\bullet^A) = 0, e(q_\bullet^A) = |A| + 1, l(q_\bullet^A) = 0)$ to the root node.
(ii) Connect q_i^A to q_j^A if $q_i^A <_E q_j^A$.
(iii) Order the children nodes of q_i^A in ascending order based on $<_J$.

We also define the compatible e-stack pairs: s_i and s_j are compatible if $(q_i^A, q_j^A) \in R$, $(q_i^B, q_j^B) \in R$, and $R \in \{\prec_E, \succ_E, \prec_J, \succ_J\}$. The non-compatible e-stack pairs can not be in the consensus together because their members are disordered in the secondary structures.

Furthermore, we can prove the following lemma:

LEMMA 1 . *For a non-empty set $\mathcal{S}' \subseteq \mathcal{S}$, if any two of e-stack pairs $s_i = (q_i^A, q_i^B)$ and $s_j = (q_j^A, q_j^B) \in \mathcal{S}'$ are compatible, the corresponding two e-stack sets have the same tree structure.*

PROOF . Without loss of generality, we assume that q_i^A is a child of $q_\bullet^A$ and q_i^B is not a child of $q_\bullet^B$. Then q_i^B must be on

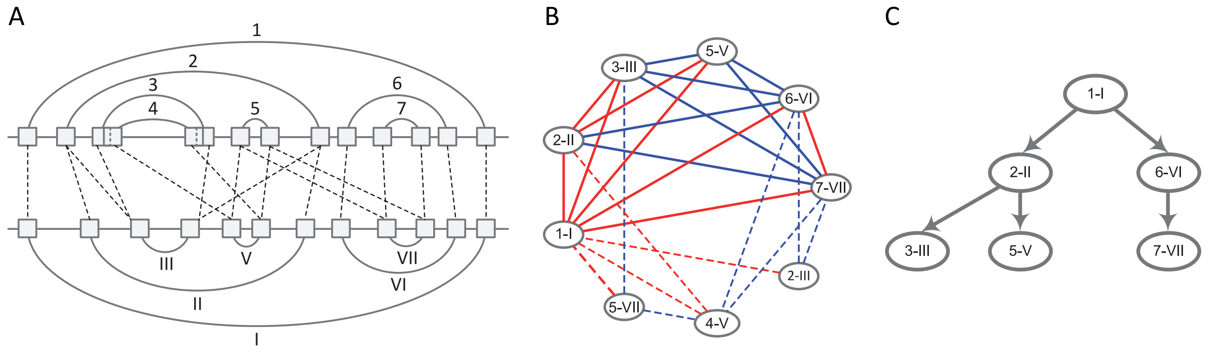

Figure 2. A description of the basic data structures used in STAR3D. (**A**) The *e*-stack pairs in two artificial RNAs. The gray boxes show the *e*-stacks and the dashed lines show the matching between them. *e*-stack 3 and *e*-stack 4 are overlapped with each other. (**B**) The compatible graph of *e*-stack pairs in (A). Red color marks the enclosing relations and blue color marks the juxtaposing relations. The solid lines show the edges in the maximum clique. To simplify the presentation, the 3D similarity requirement is not considered in the figure. (**C**) The tree-like consensus of *e*-stacks obtained from the clique in (B).

a subtree rooted at one child of $q_\bullet^B$. Thus at least two stacks, $q_\bullet^B$ and q_j^B, enclose q_i^B. However q_i^A has only one ancestor $q_\bullet^A$. It is a contradiction to the conditions, because s_i and s_j are not compatible. So the children of $q_\bullet^A$ and $q_\bullet^B$ are from the same set of *e*-stack pairs. Based on Step (iii) of the tree construction procedure, the orders of their children should be the same. Then it is proved that the lemma holds for the top two levels of the trees. Assume it is also true for the top *n* levels. Then for an *e*-stack $q_{i'}^A$ at *n*-th level, its partner $q_{i'}^B$ must also be at *n*-th level and their relative positions on the trees are the same. Assume an *e*-stack $q_{j'}^A$ is a child of $q_{i'}^A$ and $q_{j'}^B$ is not a child of $q_{i'}^B$. First $q_{j'}^B$ must be on a subtree rooted at $q_{i'}^B$. Otherwise $s_{i'}$ and $s_{j'}$ are not compatible. Second, $q_{j'}^B$ can not be at $(n+2)$-th or lower levels, otherwise the numbers of the ancestors of $q_{j'}^A$ and $q_{j'}^B$ are different, which is a contradiction to the conditions. So the children of $q_{i'}^A$ and $q_{i'}^B$ are also from the same set of *e*-stack pairs, and they can be sorted by the juxtaposing relation. By induction, we know the lemma is true.

Lemma 1 indicates how to find the 3D structural consensus of the stack regions in two input RNAs. Thus, to detect the *e*-stack configuration for the consensus, we construct a compatible graph. The vertices are the *e*-stack pairs. A vertex s_i is connected to another one s_j if they meet two requirements. First, s_i and s_j are compatible, which ensures the *e*-stacks in them are well-ordered in the secondary structures. Second, s_i and s_j must satisfy $RMSD(3D(q_i^A) \circ 3D(q_j^A), 3D(q_i^B) \circ 3D(q_j^B)) < r_c$, which implies that s_i and s_j share similar rigid transformation ('$\circ$' means the concatenation operation which joins the lists of 3D coordinates end-to-end). Based on the graph properties, the optimal stack configuration can be inferred from the maximum clique in the graph, which is detected by using the Bron-Kerbosch algorithm (44). After that, the 3D structural alignment in the double-stranded regions is determined by the topology of these vertices in the clique. Note that the *e*-stacks are the maximal 3D conversed regions in the helices (they can not be extended any more). Therefore, the 3D similarity requirement will filter most of the improper edges, and make the compatible graph very sparse. Although normally finding the maximum clique takes exponential time, it is solved very efficiently in our method. Figure 2B and C show a compatible graph and the corresponding consensus of stacks. The detected consensus is the 'core' of the 3D structural conservation and it will work as an anchor for the following loop alignment. The double-stranded regions not in the consensus, such as stack 4 in Figure 2A, are considered as loops in the following computation. The corresponding Watson–Crick base pairs and wobble base pairs are also used as interactions in the loop regions to assist the alignment.

Loop alignment using 3D information

With the tree-like consensus of stacks, all the other regions not in it can be divided into ordered loops. For one leaf node, two hairpin loops enclosed in two *e*-stacks can be identified. For the internal nodes, their enclosed regions are split by their children nodes into internal loops, bulges, or multi-loops. Hence, the numbers of loops in the inputs are the same, and we can find the mapping of them by traversing the tree. This approach has two benefits. First, the computational efficiency of loop alignment can be improved significantly for large RNAs, because only the matched loops need to be aligned together. Second, the superimposition of stack regions can be used to guide the 3D structural alignment of loop regions. For the functional RNAs, the stack regions are more conserved than the loop regions. Thus, any RMSD computation during the loop alignment uses the rotation and translation of the stack alignment.

A dynamic programming algorithm with quadratic-time complexity is applied to the 3D structural alignment of two loops. Assume the 3D structures of *k*-th pair of matched loops are $A[i_k \dots i_k + n_{k_1} - 1]$ and $B[j_k \dots j_k + n_{k_2} - 1]$. To simplify the description and computation, we denote them as $L_k^A[1 \dots n_{k_1}]$ and $L_k^B[1 \dots n_{k_2}]$, whose starting index is 1. Thus, the recursive function is ($1 \le i \le n_{k_1}$, $1 \le j \le n_{k_2}$):

$$\begin{aligned} I_{i,j} &= \max\{M_{i-1,j} + \epsilon_o + \epsilon_e,\ I_{i-1,j} + \epsilon_e,\ D_{i-1,j} + \epsilon_o + \epsilon_e\} \\ D_{i,j} &= \max\{M_{i,j-1} + \epsilon_o + \epsilon_e,\ I_{i,j-1} + \epsilon_o + \epsilon_e,\ D_{i,j-1} + \epsilon_e\} \qquad (2) \\ M_{i,j} &= \max\{I_{i-1,j-1},\ D_{i-1,j-1},\ M_{i-1,j-1}\} + \alpha(i,j) + \beta(i,j) \end{aligned}$$

Here, ε_o and ε_e are the gap open penalty and gap extension penalty. *I*, *D*, *M* denote the optimal alignment scores for insertions, deletions and substitutions, respectively. These functions are initialized with $M_{0,0} = I_{0,0} = D_{0,0} = 0$, $M_{i,0} = M_{0,j} = -\infty$, $I_{i,0} = \varepsilon_o + \varepsilon_e \times i$, $D_{0,j} = \varepsilon_o + \varepsilon_e \times j$, $I_{0,j} = D_{i,0} = -\infty$. The optimal score is $\max(I_{n_{k_1},n_{k_2}}, D_{n_{k_1},n_{k_2}}, M_{n_{k_1},n_{k_2}})$ and the exact 3D structural alignment for the two loops can be found by using traceback.

The scores for substitution contain two parts: $\alpha(i, j)$ and $\beta(i, j)$. The function $\alpha(i, j)$ is based on the 3D distance between two bases. The corresponding formula is:

$$\alpha(i, j) = \begin{cases} -\infty & d_{i,j} \geq 2 \cdot r_c \\ \text{mismatch\_score} & 2 \cdot r_c > d_{i,j} \geq r_c \\ 0.5 \times \text{match\_score} & r_c > d_{i,j} \geq 0.5 \cdot r_c \\ \text{match\_score} & 0.5 \cdot r_c > d_{i,j} \end{cases} \quad (3)$$

where $d_{i,j}$ denotes the RMSD between two nucleotides $L_k^A[i]$ and $L_k^B[j]$. Note that they are superimposed with the transformation of aligned stack regions. To capture the backbone conformation, STAR3D uses 3-nt regions, $L_k^A[i-1, i, i+1]$ for $L_k^A[i]$ and $L_k^B[j-1, j, j+1]$ for $L_k^B[j]$, in the computation of $d_{i,j}$. The possible values of $d_{i,j}$ can be categorized into three groups. The two nucleotides are not allowed to be aligned if the spatial distance is too large ($\geq 2 \cdot r_c$). Otherwise, they are defined to be 'matched' or 'mismatched' and the matched nucleotides may be assigned with two different scores.

The second function $\beta(i,j)$ calculates the bonus scores for the base pairs in loop regions. Pseudoknots, non-canonical base pairs and canonical base pairs in the unaligned stack regions are considered in the computation. Due to the potential crossing in pseudoknots and non-canonical base pairs, finding the optimal matching of these pairing interactions is an NP-hard problem. To reduce the running time, we propose a heuristic algorithm to solve the problem. Generally, each base has three possible pairs: Watson–Crick base pair, Hoogsteen base pair and Sugar base pair (45). All the predicted base pairs of two nucleotides $L_k^A[i]$ and $L_k^B[j]$ are compared in 3D space by using a similar approach of comparing nucleotides in $\alpha(i, j)$. The match of two base pairs is valid if the corresponding RMSD is less than r_c. The maximum number of matched pairs is returned as the result of $\beta(i,j)$. Thus the problem is converted into bipartite graph matching, which can be solved by dynamic programming.

RESULTS

Benchmarking tools

STAR3D is benchmarked with ARTS, LaJolla (v2.2), SARA (v1.0.7) and R3D Align in this section. Their batch programs are available and widely used for performance testing. In addition, they can output the exact one-to-one mapping of nucleotides, which is important for the analyzing of specific alignments of homologous and non-homologous RNAs. R3D Align is dedicated to homologous RNAs. To make the comparison fair, it is only used in the experiments for homologous rRNAs. An in-house modification of LaJolla is implemented to output not only the rigid transformation but also the exact alignments. All the experimental results were performed with default parameters. To evaluate the secondary structure similarity and optimize the superimposition, '-b' and '-s' options are specified for SARA. Both ARTS and LaJolla generate 'disordered alignments', e.g. a_i is aligned to b_j, a_k is aligned to b_l, while $i < k$ and $j > l$. For ARTS, the largest proper alignment is retrieved; for LaJolla, the improper alignment is discarded since only one result is returned from the modified implementation.

Alignment quality assessment with R-FSCOR dataset

The R-FSCOR dataset (46) contains 192 chains collected from the SCOR database (47). In SCOR, the chains with at least three base pairs and unique function annotations are clustered at 90% identity. The representative in each cluster is selected into the R-FSCOR dataset. The performance of four tools is compared by calculating PSI (percentage of structural identity) and PSS (percentage of aligned secondary structure) values of the all-to-all alignments for the R-FSCOR dataset. PSI is defined as the percentage of aligned nucleotides in 4 Å with respect to the length of the shorter sequence. PSS is defined as the percentage of aligned base pairs in 4 Å with respect to the smaller number of base pairs of two aligned RNA sequences. PSI and PSS have been used as replacement for RMSD to evaluate the quality of the 3D structural alignment (21,23). The base pairs in the tested chains, including both canonical base pairs and non-canonical base pairs, are predicted using MC-Annotate. All programs in this experiment were executed on a CentOS cluster with 100 nodes.

None of the tools can find alignments for all the inputs. ARTS outputs 11 385 proper alignments, LaJolla outputs 7771 proper alignment, SARA outputs 18 335 alignments and STAR3D outputs 17 455 alignments, respectively. For STAR3D, no alignment is generated if the sizes of all potential *e*-stacks in the inputs are less than $k(=3)$. However, the alignments for these inputs can be detected if a smaller k (e.g. 2) is specified. In addition, RNAMotifScanX (48), which is also designed by our lab for searching RNA 3D structural motifs in the single-stranded regions, can be applied since those RNAs are relatively short and are dominated by loops. STAR3D was compared with ARTS, SARA and LaJolla one by one. To make the comparison fair, the inputs are not considered if STAR3D or the corresponding benchmarking tool can not generate alignments for them. Table 1 summarizes the mean PSI and PSS values of four tools in the experiments. It can be seen that STAR3D outperforms the other three tools by a large margin: the PSIs are increased by 13 to 30% and the PSSs are increased by 10 to 190%. The low PSS values of LaJolla may be caused by ignoring of the secondary structural features. ARTS and SARA have relatively high PSS values because the base pairing information is integrated. For SARA, the optimization step after the backbone alignment may contribute to its better performance than ARTS. By considering the secondary structures of two input RNAs, STAR3D accurately predicts the matching of the stack regions, which is demonstrated by the high PSS values. And guided by the consensus of stacks, STAR3D provides best global alignments in all four tools without an optimization step, which is shown by the high PSI values. We also find that the running time of STAR3D for the whole procedure is much shorter (at least 1/10) than the other three tools. A detailed discussion about the computational efficiency will be shown in a later section.

Table 1. The comparison of mean PSI and PSS values between STAR3D and three other tools by using the R-FSCOR dataset

	# of overlapped alignments	ARTS		SARA		LaJolla		STAR3D	
		PSI	PSS	PSI	PSS	PSI	PSS	PSI	PSS
ARTS versus STAR3D	11 054	0.538	0.485					**0.682**	**0.632**
SARA versus STAR3D	17 454			0.601	0.580			**0.679**	**0.638**
LaJolla versus STAR3D	7397					0.580	0.251	**0.754**	**0.729**
Consensus	4451	0.600	0.549	0.683	0.668	0.627	0.318	**0.764**	**0.729**

The total number of inputs is 18 336. ARTS, SARA, LaJolla and STAR3D output 11 385, 18 335 , 7771 and 17455 alignments, respectively. Best performance is set to bold.

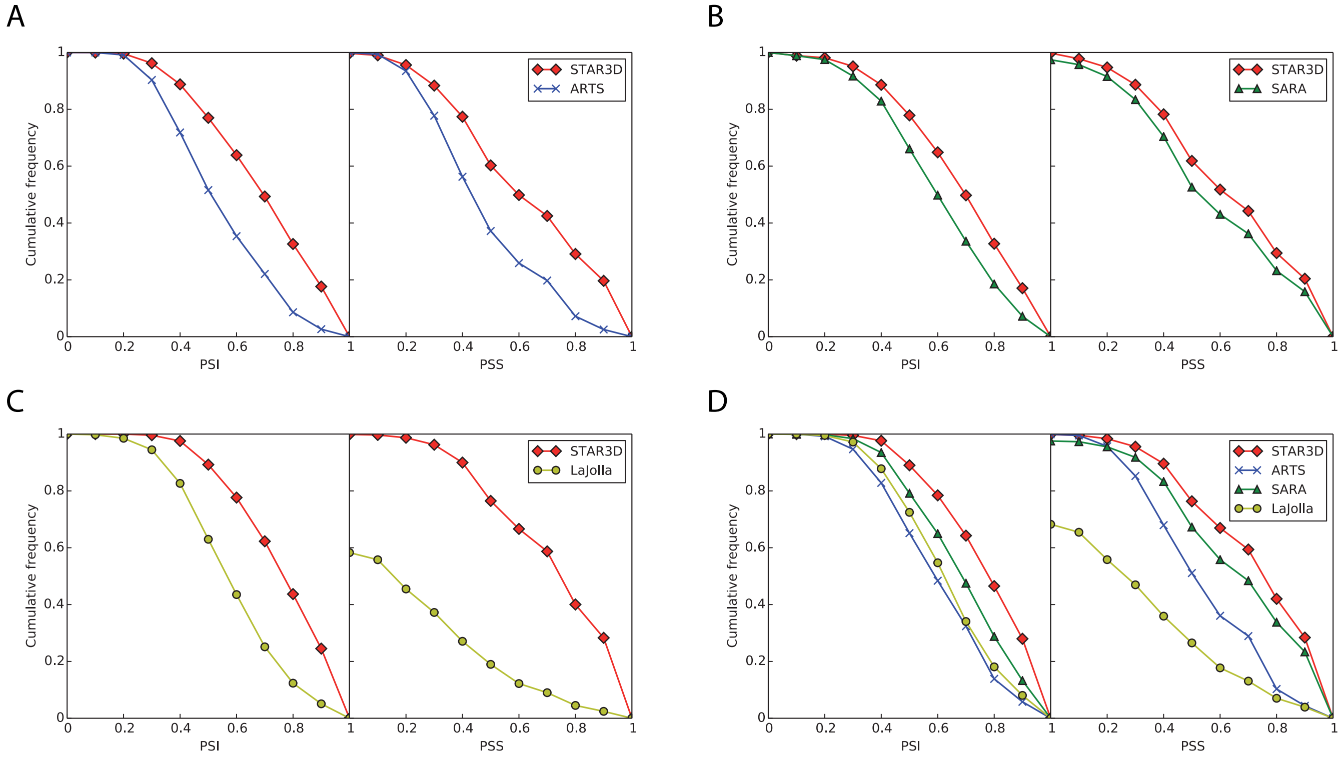

Figure 3. The cumulative frequencies of the PSI and PSS values of STAR3D, ARTS, SARA and LaJolla in different experiments. (**A**) STAR3D versus ARTS. (**B**) STAR3D versus SARA. (**C**) STAR3D versus LaJolla. (**D**) All four tools.

Figure 3 shows the cumulative frequencies of the PSI and PSS values in different comparisons. Figure 3D is based on the valid inputs for all tools. It can be seen that some alignments of SARA and LaJolla may not contain any base pair. On the other hand, the PSS values of ARTS and STAR3D are all greater than zero, because ARTS extends the base pair mapping and STAR3D relies on the stack mapping. What's more, from Figure 3D, we can see that PSI curves between 0.0 and 0.2 are very similar for all tools. The major performance difference between STAR3D and the other three tools is at the range from 0.4 to 0.7, which indicates STAR3D may be more sensitive to the local conservation of RNAs.

Structural alignments of non-homologous RNAs

Identifying the conserved regions in non-homologous RNAs is a major aim of the RNA 3D structural alignment tools. In this section, we will analyze the different strategies of STAR3D and three other tools by showing the alignments of non-homologous RNAs. The RNAs in the examples are obtained from the R-FSCOR dataset.

The first example is the alignment between a GNRA motif (PDB 1zih, chain A) and a *Deinococcus radiodurans* (*D. radiodurans*) 23S rRNA (PDB 1njo, chain 0). The aligned regions and the corresponding secondary structures are shown in Figure 4. Although it has a decent stack mapping (residue 2526 is paired with residue 2540 and residue 2527 is paired with residue 2539), the alignment produced by LaJolla is disordered: residue 2526–2527 should be at the 5′ side of residue 2539. For SARA, the aligned region of the rRNA is highly conserved with a segment of the motif. However, only one strand of the motif is aligned and the corresponding loop regions are very different. One possible reason is that the unit-vectors used by SARA only describe the conformation of the backbone. Furthermore, the alignments of base pairs and the whole 3D structures are computed separately. Thus it may overlook the pairing information if the partial structure alignment achieves the maximum score. ARTS finds the matching of base pairs

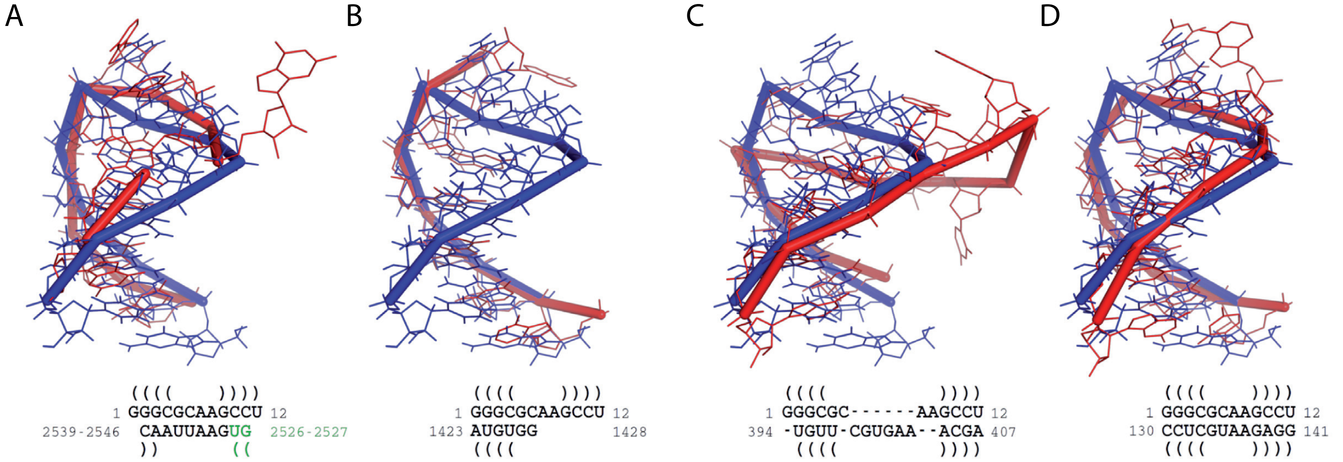

Figure 4. The alignment results for the GNRA motif (PDB: 1zih, chain A) and the 23S rRNA (PDB: 1njo, chain 0). (**A**) The result of LaJolla. (**B**) The result of SARA. (**C**) The result of ARTS. (**D**) The result of STAR3D. The blue ribbons show the 3D structure of the GNRA motif and the red ribbons show the 3D structures of the aligned regions in the 23S rRNA. The secondary structural alignments are listed below the 3D structure figures and the base pairs are predicted by MC-Annotate. The green letters in the LaJolla alignment mark the disordered nucleotides (2526–2527).

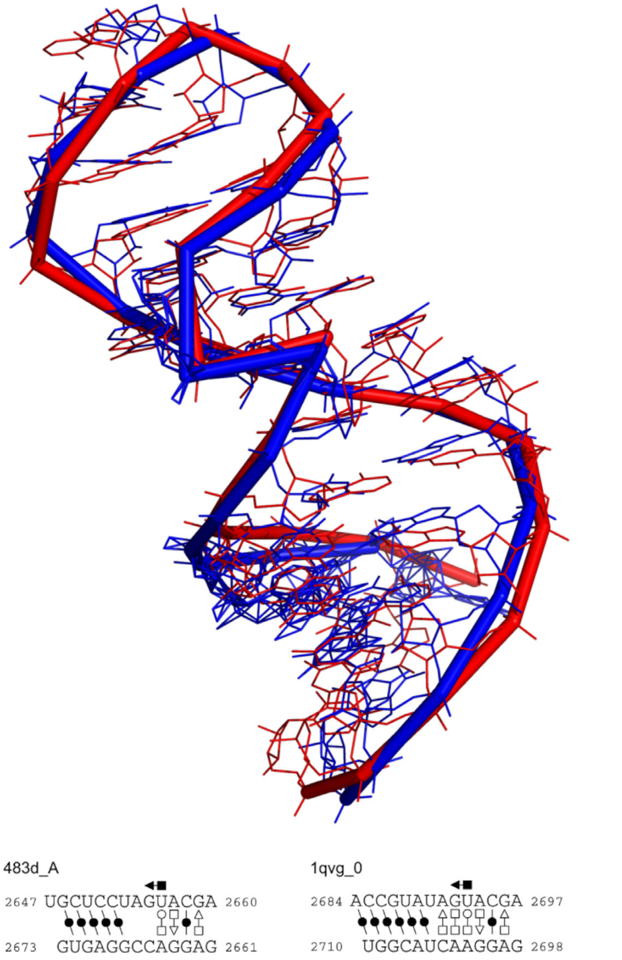

Figure 5. The alignment result of STAR3D for the sarcin-ricin motif (PDB: 483d, chain A) and the 23S rRNA (PDB: 1qvg, chain 0). The blue ribbon shows the 3D structure of the sarcin-ricin motif and the red ribbon shows the 3D structure of the 23S rRNA. The base pairs in the aligned regions are predicted by MC-Annotate.

first and then extends it to both 5′ and 3′ directions of the RNA strand. In Figure 4C, it can be seen that three base pairs are matched very well, while the 3D structures of the loop regions are distinct. This may be caused by the different treatment of stacks and the corresponding loops in the computation of ARTS. For STAR3D, the entire motif (residue 1–12) is aligned to the residues 130–141 in the rRNA. The tetraloop of the rRNA has the common structural characteristics of the GNRA motif: the four residues are ‘GUAA’ and the loop is closed by a ‘C↔G’ pair. The 3D structural alignment in Figure 4D also shows that the detected region in the 23S rRNA has a very high probability to be a GNRA motif. Similarly to the strategy of ARTS, the conserved stack regions are detected first in STAR3D. However, STAR3D ensures that the corresponding loops should have similar rigid transformation with the stacks, otherwise, the entire alignment will be assigned a low score.

The sarcin-ricin motif is an important structural motif involved in the interaction between rRNAs and the elongation factors (49). In the R-FSCOR dataset, there is one chain of 23S sarcin-ricin motif (PDB: 483d, chain A) and 22 23S rRNAs, 11 from *Haloarcula marismortui* (*H. marismortui*) and 11 from *D. radiodurans*. The 3D structural alignments of the motif and all the 23S rRNAs are analyzed. Compared with the GNRA motif, sarcin-ricin is more complex: it contains 27 residues, 6 canonical base pairs and 4 non-canonical base pairs. For ARTS and LaJolla, no highly conserved region is found in those rRNAs. SARA can detect potential motifs in all the *H. marismortui* 23S rRNAs, but none in the *D. radiodurans* 23S rRNAs. STAR3D not only finds the motif candidates in *H. marismortui* 23S rRNAs, but also in 6 of 11 *D. radiodurans* 23S rRNAs. To verify the detected motifs, the docking results of the alignments are analyzed. An example alignment of the sarcin-ricin motif and one *H. marismortui* 23S rRNA is shown in Figure 5. From the base pair profiles and the 3D structures, it can be seen that the hairpin loop (residues 2684–2710) has a high probability to be a sarcin-ricin motif. By checking the base pair annotation of all the *D. radiodu-*

Table 2. Running time (in seconds) of ARTS, LaJolla, SARA, R3D Align and STAR3D for the homologous alignments of 16S and 23S rRNAs

rRNAs	ARTS	LaJolla	SARA	R3D Align	STAR3D
H. marismortui and *E. coli* 23S	117.2	119 886.7	27 035.2	751.7	**1.7**
H. marismortui and *T. thermophilus* 23S	98.5	125 835.9	26 184.8	573.4	**2.0**
E. coli and *T. thermophilus* 23S	79.7	152 635.6	27 467.3	653.2	**1.4**
E. coli and *T. thermophilus* 16S	20.1	16 209.3	4714.4	308.9	**1.1**

Best performance is set to bold. The preprocessing time is not included for ARTS, SARA, R3D Align, and STAR3D.

Table 3. Summary of alignments between the *T. thermophilus* and *E. coli* 16S RNAs (PDB: 1j5e, chain A and PDB: 2avy, chain A)

	Manual		3D structural alignment				
	Crystallographer	Composite	ARTS	LaJolla	SARA	R3D Align	STAR3D
Number of aligned nucleotides	1488	1414	1116	1106	1343	1400	1466
Agreeing with Composite	1401	1414	1056	1101	1240	**1362**	**1362**
Agreeing with Crystallographer	1488	1401	1081	1030	1276	1354	**1414**

Best performance is set to bold.

rans 23S rRNA, we can find the structural variance in the motif regions. For the five rRNAs in which STAR3D can not detect the motifs, only four base pairs are annotated in the helix of the motif regions. The different annotation between these regions and the sarcin-ricin motif, which has five base pairs in the stack, disallows STAR3D to make the correct prediction.

Structural alignments of homologous rRNAs

We also tested the performance of STAR3D on aligning the 3D structures of homologous 16S and 23S rRNAs. The benchmarking dataset includes three 23S rRNA chains from three different species: *H. marismortui* (PDB 1s72, chain 0), *Escherichia coli* (*E. coli*; PDB 2aw4, chain B) and *Thermus thermophilus* (*T. thermophilus*; PDB 2j01, chain A); and two 16S rRNA chains from two species: *T. thermophilus* (PDB 2avy, chain A) and *E. coli* (PDB 1j5e, chain A). We also used the two manually generated alignments of these 16S rRNAs as references. The first one is the Crystallographer alignment, which is implied in the numbering system used by the crystallographers; the second one is the Composite alignment, which is hand-crafted and based on comparative analysis. They have been used as benchmarking dataset before (15,20) (note that no such alignments are available for 23S rRNAs). R3D Align is also included in this benchmarking. All the tools are installed locally on a DELL XPS Desktop with Intel i7-4770 CPU at 3.40 GHz with 16 GB of RAM. To make the comparison fair, only one thread is allowed in the experiments.

First, we compare the running time of five tools for the rRNA alignments (see Table 2). It can be seen that STAR3D improves the time efficiency of the other tools by ten to a thousand folds. The adoption of the MaxSub algorithm (50) to refine the original 3D structural alignments may cause the huge time consumption in SARA. For STAR3D, the major running time reduction comes from the computation of loop alignments. Assuming the total lengths of loop regions for two RNA sequences are n_1 and n_2, the time complexity of loop alignment is $O(n_1 \times n_2)$. In STAR3D, one loop region only needs to be compared with another one marked by the *e*-stacks. Thus the time complexity is $O((n_1 \times n_2)/m)$, where m denotes the number of *e*-stack pairs in the consensus. With the relatively large number of stacks in RNAs with complex structures, our method can significantly improve the efficiency of the loop alignments.

Based on manually generated alignments of two 16S rRNAs, the accuracy of five tools is also examined. The results are shown in Table 3. It can be seen that both R3D Align and STAR3D achieve the maximum true positive number if the background dataset is the Composite dataset. The accuracy of STAR3D is slightly lower than R3D Align because it detects more nucleotide matches in the two sequences. However, for the Crystallographer dataset, STAR3D outperforms the other three tools. So STAR3D is not only highly efficient but also an accurate algorithm when it is used to align large homologous RNA molecules.

DISCUSSION

In this article, we have proposed a novel tool, named STAR3D, for RNA 3D structural alignment. First it detects the conserved double-stranded regions in two input RNAs by joining the matches of small stack components. Then the consensus of stacks is assembled based on the 3D structural similarity and 2D compatible relationship. Its underlying tree-like topology leads to the ordering of loop regions. In addition, the rigid transformation of the aligned stacks can guide the 3D alignment of the loop regions. As a result, each loop only needs to be compared with its partner in the other sequence by using the superimposition of the conserved stacks. Finally, we combine the stack alignment and all the loop alignments as the final result. This ‘two-step’ strategy is derived on the basis of three observations. First, insertions and deletions are rarely seen in the conserved helical regions, which means that the ungapped extension is applicable to the stacks; second, the 3D structural similarity of conserved stacks is higher than that of random stacks; third, the stack regions are easier to annotate, even for low resolution PDB structures, so the stack alignment can be used as an anchor for the loop alignment. By integrating these properties into the design, STAR3D avoids the complex computation of secondary structure comparison. Furthermore, the one-to-one loop alignments, which replace the all-to-all base matching in entire single-stranded regions, reduce the running time of STAR3D for large RNAs significantly. The benchmark results show that the prediction accuracy of STAR3D outperforms the state-of-the-art tools, and does so with higher efficiency. What's more, STAR3D can be easily implemented with multi-thread support. The detection of *e*-stack pairs depends on ungapped alignment. The computation at each diagonal can be performed at an individual thread. For maximum clique finding, the Bron–Kerbosch algorithm can be implemented in parallel too. In the last step, the alignments of loop regions are independent, and can be deployed in different threads as well.

A potential expansion of STAR3D is to implement a local alignment version of the tool. From the experiment of aligning the GNRA motif and the 23S rRNA, it can be seen that STAR3D is sensitive to local similarities in RNA 3D structures. On the other hand, it is natural to convert STAR3D into finding local alignments. In the original method, only the maximum clique in the compatible graph is chosen to build the structural tree. To develop a local alignment approach, we can change STAR3D to deal with multiple cliques. For each one, a local alignment can be generated by only comparing the loops covered by the aligned stacks. With this new method, we anticipate that more structural motifs will be found in the functional ncRNAs.

Another direction for future study is to incorporate comparative methods into STAR3D. There are two approaches. The first one is to use the comparative methods to improve the alignments of the RNAs with low resolution coordinates. The excessive flexibility of atomic positions challenges the prediction of base pairing interactions, which may affect the performance of STAR3D. To solve the problem, we plan to design an iterative pipeline to find the base pairs in the low resolution RNA structures. First, we need a homology of the target that has high resolution 3D structural data. Hence a better annotation of base pairs for the target can be inferred by aligning two RNAs with STAR3D. In the following run, these predicted base pairs can be used as the secondary structural information in STAR3D to generate a more precise alignment. This procedure is continued until no new base pairs can be detected for the target. Considering the high efficiency of STAR3D, the time consumption of the pipeline should be practical. In addition, the low resolution RNAs can be aligned to other RNAs more accurately with the inferred base pairs. The second way is to find the 3D structural conservation among the RNAs in one family by using comparative methods. A hierarchical clustering based method, which is similar to CLUSTALW (51), is adopted. A 3D consensus structure of two RNAs can be constructed by connecting the centroids of the mapped nucleotides. Then, by merging the sub-clusters we can find the consensus for the whole family and its corresponding multiple sequence alignment.

AVAILABILITY

http://genome.ucf.edu/STAR3D.

ACKNOWLEDGEMENT

The authors thank Raphael André Bauer for helpful discussions about the LaJolla program.

FUNDING

National Institute of General Medical Sciences of the National Institutes of Health under Award Number [R01GM102515]. Funding for open access charge: National Institutes of Health [R01GM102515].
Conflict of interest statement. None declared.

REFERENCES

1. Eddy,S.R. (2001) Non-coding RNA genes and the modern RNA world. *Nat. Rev. Genet.*, **2**, 919–929.
2. Storz,G. (2002) An expanding universe of noncoding RNAs. *Science*, **296**, 1260–1263.
3. He,L. and Hannon,G.J. (2004) MicroRNAs: small RNAs with a big role in gene regulation. *Nat. Rev. Genet.*, **5**, 522–531.
4. Batista,P.J. and Chang,H.Y. (2013) Long noncoding RNAs: cellular address codes in development and disease. *Cell*, **152**, 1298–1307.
5. Jacquier,A. (2009) The complex eukaryotic transcriptome: unexpected pervasive transcription and novel small RNAs. *Nat. Rev. Genet.*, **10**, 833–844.
6. Woodson,S.A. (2010) Compact intermediates in RNA folding. *Annu. Rev. Biophys.*, **39**, 61–77.
7. Eidhammer,I., Jonassen,I. and Taylor,W.R. (2000) Structure comparison and structure patterns. *J. Comput. Biol.*, **7**, 685–716.
8. Holm,L. and Sander,C. (1993) Protein structure comparison by alignment of distance matrices. *J. Mol. Biol.*, **233**, 123–138.
9. Madej,T., Gibrat,J.F. and Bryant,S.H. (1995) Threading a database of protein cores. *Proteins*, **23**, 356–369.
10. Ye,Y. and Godzik,A. (2003) Flexible structure alignment by chaining aligned fragment pairs allowing twists. *Bioinformatics*, **19**(Suppl. 2), i246–i255.
11. Malod-Dognin,N. and Pržulj,N. (2014) GR-Align: fast and flexible alignment of protein 3D structures using graphlet degree similarity. *Bioinformatics*, **30**, 1259–1265.
12. Berman,H.M., Westbrook,J., Feng,Z., Gilliland,G., Bhat,T.N., Weissig,H., Shindyalov,I.N. and Bourne,P.E. (2000) The Protein Data Bank. *Nucleic Acids Res.*, **28**, 235–242.
13. 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.
14. Bauer,R.A., Rother,K., Moor,P., Reinert,K., Steinke,T., Bujnicki,J.M. and Preissner,R. (2009) Fast structural alignment of biomolecules using a hash table, N-grams and string descriptors. *Algorithms*, **2**, 692–709.
15. Wiegels,T., Bienert,S. and Torda,A.E. (2013) Fast alignment and comparison of RNA structures. *Bioinformatics*, **29**, 588–596.
16. Nguyen,M.N. and Verma,C. (2015) Rclick: a web server for comparison of RNA 3D structures. *Bioinformatics*, **31**, 966–968.
17. 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.
18. 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.
19. Dror,O., Nussinov,R. and Wolfson,H. (2005) ARTS: alignment of RNA tertiary structures. *Bioinformatics*, **21**(Suppl. 2), 47–53.
20. 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.
21. Capriotti,E. and Marti-Renom,M.A. (2008) RNA structure alignment by a unit-vector approach. *Bioinformatics*, **24**, i112–i118.
22. Ortiz,A.R., Strauss,C.E. and Olmea,O. (2002) MAMMOTH (matching molecular models obtained from theory): an automated method for model comparison. *Protein Sci.*, **11**, 2606–2621.

23. Hoksza,D. and Svozil,D. (2012) Efficient RNA pairwise structure comparison by SETTER method. *Bioinformatics*, **28**, 1858–1864.
24. Sankoff,D. (1985) Simultaneous solution of the RNA folding, alignment and protosequence problems. *SIAM J. Appl. Math.*, **45**, 810–825.
25. Havgaard,J.H., Torarinsson,E. and Gorodkin,J. (2007) Fast pairwise structural RNA alignments by pruning of the dynamical programming matrix. *PLoS Comput. Biol.*, **3**, 1896–1908.
26. Will,S., Reiche,K., Hofacker,I.L., Stadler,P.F. and Backofen,R. (2007) Inferring noncoding RNA families and classes by means of genome-scale structure-based clustering. *PLoS Comput. Biol.*, **3**, e65.
27. Zhong,C. and Zhang,S. (2013) Efficient alignment of RNA secondary structures using sparse dynamic programming. *BMC Bioinformatics*, **14**, 269.
28. Gendron,P., Lemieux,S. and Major,F. (2001) Quantitative analysis of nucleic acid three-dimensional structures. *J. Mol. Biol.*, **308**, 919–936.
29. 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.
30. Draper,D.E. (1992) The RNA-folding problem. *Acc. Chem. Res.*, **25**, 201–207.
31. 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.
32. Reuter,J.S. and Mathews,D.H. (2010) RNAstructure: software for RNA secondary structure prediction and analysis. *BMC Bioinformatics*, **11**, 129.
33. Brudno,M., Poliakov,A., Salamov,A., Cooper,G.M., Sidow,A., Rubin,E.M., Solovyev,V., Batzoglou,S. and Dubchak,I. (2004) Automated whole-genome multiple alignment of rat, mouse, and human. *Genome Res.*, **14**, 685–692.
34. Lippert,R.A., Zhao,X., Florea,L., Mobarry,C. and Istrail,S. (2005) Finding anchors for genomic sequence comparison. *J. Comput. Biol.*, **12**, 762–776.
35. Ji,Y., Xu,X. and Stormo,G.D. (2004) A graph theoretical approach for predicting common RNA secondary structure motifs including pseudoknots in unaligned sequences. *Bioinformatics*, **20**, 1591–1602.
36. Bafna,V., Tang,H. and Zhang,S. (2006) Consensus folding of unaligned RNA sequences revisited. *J. Comput. Biol.*, **13**, 283–295.
37. Stombaugh,J., Zirbel,C.L., Westhof,E. and Leontis,N.B. (2009) Frequency and isostericity of RNA base pairs. *Nucleic Acids Res.*, **37**, 2294–2312.
38. Kabsch,W. (1978) A discussion of the solution for the best rotation to relate two sets of vectors. *Acta Crystallogr. A*, **34**, 827–828.
39. Reijmers,T.H., Wehrens,R. and Buydens,L.M. (2001) The influence of different structure representations on the clustering of an RNA nucleotides data set. *J. Chem. Inf. Comput. Sci.*, **41**, 1388–1394.
40. Hochsmann,M., Toller,T., Giegerich,R. and Kurtz,S. (2003) Local similarity in RNA secondary structures. *Proc. IEEE Comput. Soc. Bioinform. Conf.*, **2**, 159–168.
41. Ma,B., Wang,L. and Zhang,K. (2002) Computing similarity between RNA structures. *Theor. Comput. Sci.*, **276**, 111–132.
42. Schirmer,S. and Giegerich,R. (2013) Forest alignment with affine gaps and anchors, applied in RNA structure comparison. *Theor. Comput. Sci.*, **483**, 51–67.
43. Zhang,S., Haas,B., Eskin,E. and Bafna,V. (2005) Searching genomes for noncoding RNA using FastR. *IEEE/ACM Trans. Comput. Biol. Bioinform.*, **2**, 366–379.
44. Bron,C. and Kerbosch,J. (1973) Algorithm 457: finding all cliques of an undirected graph. *Commun. ACM*, **16**, 575–577.
45. Leontis,N.B. and Westhof,E. (2001) Geometric nomenclature and classification of RNA base pairs. *RNA*, **7**, 499–512.
46. Capriotti,E. and Marti-Renom,M.A. (2009) SARA: a server for function annotation of RNA structures. *Nucleic Acids Res.*, **37**, W260–W265.
47. 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.
48. Zhong,C. and Zhang,S. (2015) RNAMotifScanX: a graph alignment approach for RNA structural motif identification. *RNA*, **21**, 333–346.
49. Szewczak,A.A., Moore,P.B., Chang,Y.L. and Wool,I.G. (1993) The conformation of the sarcin/ricin loop from 28S ribosomal RNA. *Proc. Natl. Acad. Sci. U.S.A.*, **90**, 9581–9585.
50. Siew,N., Elofsson,A., Rychlewski,L. and Fischer,D. (2000) MaxSub: an automated measure for the assessment of protein structure prediction quality. *Bioinformatics*, **16**, 776–785.
51. Thompson,J.D., Higgins,D.G. and Gibson,T.J. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.*, **22**, 4673–4680.