BIOINFORMATICS

Vol. 00 no. 00 2011
Pages 1–8

Efficient RNA pairwise structure comparison by SETTER method.

David Hoksza [1,2,*] and Daniel Svozil [2,*]

[1]SIRET Research Group, Department of Software Engineering, FMP, Charles University in Prague, Czech Republic
[2]Laboratory of Informatics and Chemistry, Institute of Chemical Technology Prague, Czech Republic

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Associate Editor: XXXXXXX

ABSTRACT

Motivation: Understanding the architecture and function of RNA molecules requires methods for comparing and analyzing their 3D structures. While a structural alignment of short RNAs is achievable in a reasonable amount of time, large structures represent much bigger challenge. However the growth of the number of large RNAs deposited in the PDB database calls for the development of fast and accurate methods for analyzing their structures, as well as for rapid similarity searches in databases.

Results: In this article a novel algorithm for an RNA structural comparison SETTER (SEcondary sTructure-based TERtiary Structure Similarity Algorithm) is introduced. SETTER utilizes a pairwise comparison method based on 3D similarity of the so-called generalized secondary structure units (GSSU). For each pair of structures, SETTER produces a distance score and an indication of its statistical significance. SETTER can be used both for the structural alignments of structures that are already known to be homologous, as well as for 3D structure similarity searches and functional annotation. The algorithm presented is both accurate and fast and does not impose limits on the size of aligned RNA structures.

Availability: The SETTER program, as well as all datasets, are freely available from http://siret.cz/hoksza/projects/setter/.

Contact: hoksza@ksi.mff.cuni.cz, svozild@vscht.cz

Supplementary information: Supplementary Information is available at Bioinformatics online.

1 INTRODUCTION

In addition to its role in the transfer of biological information, the evidence shows that RNA molecules also play key roles in a variety of cellular processes (Mattick and Makunin, 2006). RNA shows, among others, an enzymatic activity in ribozymes (Scott, 2007), it takes part in the transcription regulation (Bartel, 2004) and it is involved in the chromatin modeling (Kelley and Kuroda, 2000).

RNA 3D structure is hierarchical (Tinoco, 1999), and can be divided into primary, secondary, tertiary and quaternary levels. RNA secondary structure motifs (Hendrix *et al.*, 2005) can be defined as double helices interconnected by various types of loop structures and are stable independently of their 3D folds. Tertiary interactions (Holbrook, 2008) stabilize the overall arrangement and packing of double helices in large RNA structures. The first resolved RNA crystal structure was that for the yeast phenylalanine tRNA (Kim *et al.*, 1974). This achievement was followed by solving structures of other naturally occurring tRNAs (Arnez and Steitz, 1994), as well as of a variety of oligomeric RNA model structures (Holbrook *et al.*, 1991). With the improvements in molecular biological methods (Kim *et al.*, 1995) and in crystallographic techniques (Garman, 2003) the size of solved RNA structures has later increased dramatically. Large RNAs[1] allowed for the first detailed studies of RNA structure elements not found in smaller molecules, such as continuous interhelical base stacking, RNA domain structure, and helical packing.

The function of an RNA molecule is largely determined by the 3D structure that is typically more evolutionarily conserved than its sequence (Chursov *et al.*, 2012). Thus, methods for the comparative RNA function annotation based on structural similarity usually yield much better results than sequence based approaches. Though detecting optimal structural similarity between two biomolecules in 3 dimensions has been shown to be NP-hard (Kolodny and Linial, 2004), the development of automatic tools capable of an efficient and accurate RNA structural alignment has become an important part of structural bioinformatics. The study of RNA tertiary and quaternary structures must be facilitated by the software that is able to work both with small and large RNA molecules. To be computationally tractable, currently available software tools for comparing two RNA 3D structures, such as ARTS (Dror *et al.*, 2005, 2006), DIAL (Ferrè *et al.*, 2007), iPARTS (Wang *et al.*, 2010), SARA (Capriotti and Marti-Renom, 2008, 2009), SARSA (Chang *et al.*, 2008) or R3D Align (Rahrig *et al.*, 2010) are therefore based on heuristic approaches. ARTS (Dror *et al.*, 2005, 2006) detects a maximum common substructure between two RNA 3D structures using backbone phosphate atoms. Based on 3D similarity between 1333 solved RNA structures assessed by the ARTS algorithm a database of hierarchically classified structures DARTS was subsequently developed (Abraham *et al.*, 2008). ARTS is not practical for comparison of large RNA molecules due to its cubic time complexity. To overcome this problem the DIAL server

*to whom correspondence should be addressed

[1] An arbitrary limit of 100 residues is used to define large RNAs (Holbrook, 2008).

using a dynamic programming algorithm and running in a quadratic time was developed (Ferrè *et al.*, 2007). The DIAL alignment algorithm is based on torsion and/or pseudotorsion similarity, sequence similarity, and base pairing similarity, and it provides access to global, local and semi-global structural alignments. An improvement in the speed over the DIAL algorithm was later brought by PARTS (Chang *et al.*, 2008), an algorithm based on the so-called structural alphabet (SA). Structural alphabet is an emerging concept in the structural biology of proteins. A protein structure is represented as a limited series of "letters" each assigned to a well-characterized conformation (de Brevern *et al.*, 2000). PARTS uses the vector quantization approach to derive an RNA structural alphabet of 23 letters representing the most common backbone conformations. The RNA structures are represented as 1D sequences of SA letters, and these are aligned utilizing classical methods for pairwise and multiple sequence alignments. A new set of SA letters was derived for the improved version of PARTS called iPARTS (Wang *et al.*, 2010). iPARTS outperforms its previous version PARTS, as well as (in some aspects) another highly efficient algorithm SARA (Capriotti and Marti-Renom, 2008, 2009). In SARA, distances among selected atoms are represented as unit vectors existing on unit spheres (Chew *et al.*, 1999). All-to-all unit-vector RMSD distances of consecutive unit spheres are computed and used as scoring matrix for the dynamic programming based global alignment. Highly accurate alignments of homologous molecules are produced by the R3D Align approach (Rahrig *et al.*, 2010) which is based on local nucleotide by nucleotide superpositions that effectively accommodate the flexibility of RNA molecules. Local alignments are then merged to form a global alignment by employing a maximum clique algorithm on a specially defined 'local alignment' graph.

In the presented paper a novel pairwise RNA comparison method SETTER (SEcondary sTructure-based TERtiary Structure Similarity Algorithm) is proposed. The method divides the whole RNA structure into non-overlapping generalized secondary structure units (GSSUs). The structural alignment is then obtained by utilizing a distance measure based on RMSD transformations between all possible pairs of GSSUs. The algorithm scales as $O(n^2)$ with the size of GSSU and as $O(n)$ with the number of GSSUs in the structure. The segmentation to GSSUs offers the advantage of high speeds even for the largest structures. The algorithm can be used both for the 3D structural alignments (Fig. 1-12 in the Supplementary Information show several examples of pairwise structural alignments), as well as for 3D structural similarity searches and for functional annotation. High speed of the algorithm does not compromise its accuracy as is demonstrated by benchmarking of both structural alignment and functional annotation against other published methods.

2 METHODS

2.1 GSSU Identification

Three important elements are recognized in the GSSU: a loop, a neck, and a stem (see Fig. 1). A formal description of the GSSU is given by the following definition.

DEFINITION 1. *Let $\mathcal{R}$ be an RNA structure with a nucleotide sequence $\{nt_i\}_{i=1}^{n}$ and let $\mathcal{WC} \subseteq \mathcal{R}$ denote its subset participating in a Watson-Crick base pair. By a **generalized secondary structure unit (GSSU)** $\mathcal{G}$ we understand a pair of substrings of $\mathcal{R}$, $\{nt_i\}_{i=i_1}^{i_2}$ and $\{nt_j\}_{j=j_1}^{j_2}$ ($i_1 \leq i_2 < j_1 \leq j_2, i_2 = j_1 - 1$) of `maximum lengths` such that each nucleotide $nt_x \in \mathcal{G}$:*

- *$i_1 \leq x \leq i_2$: $nt_x \notin \mathcal{WC}$ or nt_x is paired with nt_y where $j_1 \leq y \leq j_2$*
- *$j_1 \leq x \leq j_2$: $nt_x \notin \mathcal{WC}$ or nt_x is paired with nt_y where $i_1 \leq y \leq i_2$*

*In case of ambiguity a maximum length is assigned to the substring occurring earlier in the sequence. Let i_{max} and j_{min} be the highest/lowest indexes of the Watson-Crick paired bases in $\mathcal{G}$. We define a **loop** as $\mathcal{L} = \{nt_i\}_{i=i_{max}+1}^{j_{min}-1} \subset \mathcal{G}$, a **stem** as $\mathcal{G} \setminus \mathcal{L}$ and a **neck** as the pair $\{nt_{i_{max}}, nt_{j_{min}}\}$.*

Nucleotides are represented by their P atoms. Watson-Crick (WC) hydrogen bonds are identified using 3DNA (Lu and Olson, 2008). Non-WC pairs are not used because they often mediate RNA tertiary contacts the presence of which does not allow an unambiguous distinction between GSSUs.

To identify all GSSUs the process iteratively applies two following steps. The RNA structure is processed in a sequence order, and in the first step nucleotides are stored on a stack. This process stops by encountering a nucleotide nt_i WC bonded with a nucleotide nt_j already in the stack. Then, in the second step, a new GSSU $\mathcal{G}$ starts to be formed from the pair $\{nt_i, nt_j\}$ (i.e., the neck) and from all nucleotides found between nt_i and nt_j (i.e., the loop). These residues are then removed from the stack. Finally, the stem is formed from all residues encountered either before the residue WC bonded to the residue not stored on the stack or before the residue WC bonded with the residue that was pushed on the stack before the previous GSSU was generated. By repeating these two steps, the algorithm iteratively searches for GSSUs, and it stops when the end of the sequence is reached. All residues remaining on the stack (if any) then form the last GSSU. Note that even a structure without a single Watson-Crick pair has a GSSU which is identical with the structure itself. A detailed description of the process of generating GSSUs from the Fig. 1, as well as the pseudocode algorithm are given in sections 2 and 3 in the Supplementary Information.

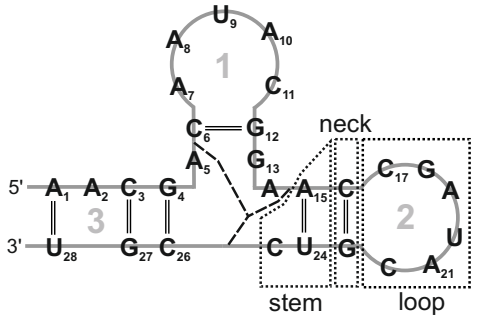


Fig. 1. Three GSSUs extracted from an RNA structure. The sequence starts at the 5' end. The borders between individual GSSUs are indicated by dashed lines and the numbers show the order of the GSSU generation.

2.2 Comparing Two GSSUs

GSSU pairwise comparison lies in the heart of the method. Each GSSU is represented by an ordered set of 3D coordinates of P atoms annotated with bonding and nucleotide/atom type information. A common way to assess similarity between two sets of points is to define a pairing between them. The sets are then superposed by finding translation and rotation of one of them over the other minimizing the mutual distances of the respective paired points. Usually, the root mean square deviation (RMSD)

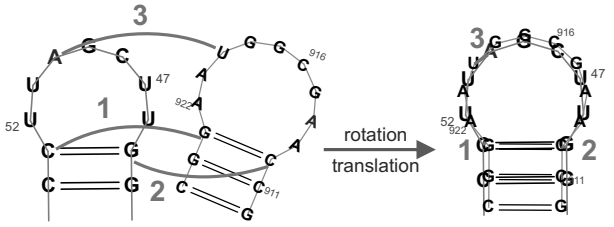


Fig. 2. The alignment of the GSSU from the tRNA domain of the transfer-messenger RNA (PDB code 1P6V) with the GSSU from the glutamine tRNA (PDB code 1EXD). The final structural alignment is defined by three nucleotide pairs forming a triplet (the lines 1, 2, and 3). To find the optimal superposition for the given neck pairs (lines 1 and 2), the position of the middle pair is varied (line 3).

is used as the distance measure, and two structures can be superposed given a pairing (alignment) in polynomial time (Kabsch, 1976). However, finding the optimal alignment is an NP-hard problem (Kolodny and Linial, 2004). The optimal solution can be found by exhaustive search, which is computationally not feasible. This problem can be resolved by identifying suboptimal alignments that will likely participate in the optimal alignment. This is the principle idea behind SETTER's structure comparison process. SETTER generates a set of short alignments which quality is evaluated by the Kabsch (Kabsch, 1976) RMSD algorithm. Working with relatively short alignments allows to superpose even the largest RNA structures in a reasonable amount of time.

To superpose two GSSUs means to match their loops which implies matching their necks (see Fig. 1). To define the superposition in 3D space unambiguously at least three pairs of points are needed. A set of these points is called a *triplet*, and the *alignment* is formed by matching triplets between two given structures. SETTER aligns necks first, and then the final pair in the triplet is identified by aligning each possible pair of loops' nucleotides. For example, if two GSSUs with loops consisting of n and m nucleotides are to be aligned, $n \times m$ alignments are generated (see Fig. 2).

For each alignment, the rotation matrix and the translation vector defining optimal superposition of two triplets is calculated. Though these are optimal for the given triplet pair only, they are used to superpose whole GSSUs. This possible inaccuracy is the trade-off for an efficiency.

After the superposition, for each nucleotide in GSSU A its nearest neighbor in GSSU B is found, and their distance is added to the distance of two GSSUs referred to as S-distance. Finally, the S-distance is normalized. The whole process is formalized by Eq. 1.

$$
\begin{aligned}
NN_\zeta(x,\mathcal{G}) &= \begin{cases} \min_{1\le i\le|\mathcal{G}|}\{d(x,\mathcal{G}_i)\}\times\zeta & \text{if } t(x)=t(\mathcal{G}_{i_{min}}) \\ \min_{1\le i\le|\mathcal{G}|}\{d(x,\mathcal{G}_i)\} & \text{otherwise} \end{cases} \\
\gamma(\mathcal{G}^{\mathcal{A}},\mathcal{G}^{\mathcal{B}}) &= \frac{1}{2}\Big(\sum_{i=1}^{|\mathcal{G}^{\mathcal{A}}|}\begin{cases} 1 & \text{if } NN_1(\mathcal{G}^{\mathcal{A}}{}_i,\mathcal{G}^{\mathcal{B}})\le\epsilon \\ 0 & \text{otherwise} \end{cases} \\
&\quad + \sum_{i=1}^{|\mathcal{G}^{\mathcal{B}}|}\begin{cases} 1 & \text{if } NN_1(\mathcal{G}^{\mathcal{A}},\mathcal{G}^{\mathcal{B}}{}_i)\le\epsilon \\ 0 & \text{otherwise} \end{cases}\Big) \\
\delta(\mathcal{G}^{\mathcal{A}},\mathcal{G}^{\mathcal{B}}) &= \min_{t\in T}\Big\{\frac{1}{2}\Big(\sum_{i=1}^{|\mathcal{G}^{\mathcal{A}}|}NN_\zeta(\mathcal{G}^{\mathcal{A}}{}_i,\tau(\mathcal{G}^{\mathcal{B}},t)) \\
&\quad + \sum_{i=1}^{|\mathcal{G}^{\mathcal{B}}|}NN_\zeta(\mathcal{G}^{\mathcal{A}},\tau(\mathcal{G}^{\mathcal{B}}{}_i,t))\Big)\Big\} \\
S(\mathcal{G}^{\mathcal{A}},\mathcal{G}^{\mathcal{B}}) &= \frac{\frac{\delta(\mathcal{G}^{\mathcal{A}},\mathcal{G}^{\mathcal{B}})}{\min\{|\mathcal{G}^{\mathcal{A}}|,|\mathcal{G}^{\mathcal{B}}|\}}\times(1+\frac{||\mathcal{G}^{\mathcal{A}}|-|\mathcal{G}^{\mathcal{B}}||}{\min\{|\mathcal{G}^{\mathcal{A}}|,|\mathcal{G}^{\mathcal{B}}|\}})}{\gamma(\mathcal{G}^{\mathcal{A}},\tau(\mathcal{G}^{\mathcal{B}},t_{opt}))}
\end{aligned}
\tag{1}
$$

where $\mathcal{G}^{\mathcal{A}}$ and $\mathcal{G}^{\mathcal{B}}$ represent two GSSUs to be compared, $\mathcal{G}_i$ stands for the i-th nucleotide in the sequence of $\mathcal{G}$, and $|\mathcal{G}|$ for its length. $NN(x,\mathcal{G})$ is the Euclidean distance from the nucleotide x to its nearest neighbor in $\mathcal{G}$. If x and its nearest neighbor share the same nucleotide type (the function $t(x)$ in the formula) the distance is modified by a factor ζ. It takes values from the interval $0 < \zeta \le 1$, the lower its value the more matching identical nucleotides are rewarded. δ computes the raw distance - T is the set of transpositions resulting from the candidate triplet alignments and $\tau(\mathcal{G},t)$ transposes GSSU $\mathcal{G}$ using the transposition t. The S-distance is then normalized by the function γ counting a number of nearest neighbors within the distance ϵ after the optimal transposition t_{opt}.

Since hydrogen bonds are identified using simple geometric criteria, their detection may sometimes be incorrect. This leads to the shift of the neck position within the GSSU. SETTER simulates the neck shift by aligning also the residues next to (under) the necks. These tweakings are necessary for accurate GSSU comparison, however they slightly increase the running time of SETTER.

Though in most cases the GSSU consists of a stem and a loop, it is not a strict rule. Two particular situations can occur — the GSSU has no loop or RNA does not have a single WC hydrogen bond at all. In the case of the GSSU without the loop the third nucleotide of the triplet is selected from the stem such that the S-distance is minimized. When dealing with the GSSU with no WC hydrogen bond, several triplets covering the whole structure are formed and used as a basis for the alignment. Such a simplified comparison may not lead to the best possible result, and SETTER was not developed for these cases. However, the probability of encountering such defective GSSUs in large RNA structures is very small (indeed, no such GSSU is found either in 16S rRNA or in 23S rRNA).

2.3 Comparing More Than Two GSSUs

If structures contain more than one GSSU the following three-step multiple GSSU comparison process is implemented (see Fig. 3):

1. All-to-all pairwise GSSU comparisons are performed.
2. Few best GSSU pairs (with low S-distances) are used as seeds for the alignment of the rest of the GSSUs.
3. The structures' GSSUs are aligned, their S-distances are aggregated into the $\overline{S}$-distance, and the alignment with the lowest $\overline{S}$-distance is identified.

If comparing structures $\mathcal{R}^{\mathcal{A}}$ and $\mathcal{R}^{\mathcal{B}}$, each GSSU from $\mathcal{R}_{\mathcal{A}}$ is compared to each GSSU from $\mathcal{R}^{\mathcal{B}}$, but only top κ pairs with the minimum distance

is processed further. For each of the κ selected pairs $\{\mathcal{G}_i^A, \mathcal{G}_j^B\}$ the value of $\overline{S}$ is set to $S(\mathcal{G}_i^A, \mathcal{G}_j^B)$. In the second step, S-distances of the neighboring GSSU pairs are iteratively added to the $\overline{S}$-distance. For the GSSU pair $\{\mathcal{G}_{i+1}^A, \mathcal{G}_{j+1}^B\}$ the value of $S(\mathcal{G}_{i+1}^A, \mathcal{G}_{j+1}^B)$ and the penalty for the rotation needed to transform the structures from the state corresponding to $S(\mathcal{G}_i^A, \mathcal{G}_j^B)$ to the state corresponding to $S(\mathcal{G}_{i+1}^A, \mathcal{G}_{j+1}^B)$ are added to $\overline{S}$. This process goes from $\{i+1, j+1\}$ until either $i+1$ or $j+1$ reaches the number of GSSUs in $\mathcal{R}^\mathcal{A}$ or $\mathcal{R}^\mathcal{B}$. Similarly, the other ends of the structures need to be aligned and so the process is repeated for $\{i-1, j-1\}$. However, the case when GSSUs in the RNA structure are oriented in the opposite direction must also be considered, and another κ alignments must be performed aligning $i-1$ residues with $j+1$ residues (not shown in Fig. 3).

The rotation between two GSSUs imposes a penalty to the $\overline{S}$. This penalty is calculated as a distance between the rotation matrices defining two consequent GSSU superpositions (see Fig. 3c). However, the penalty for translation is not included explicitly. The translation is limited only to the pair of GSSUs just being aligned, and such a translation is already implicitly present in the S-distance (see section 2.2 and Fig. 2).

Currently, there is no provision for a situation in which one structure is missing a GSSU that is present in the other structure. This potential limitation may have an undesirable effect on the alignment, however, it can not be improved without increase in computational demands.

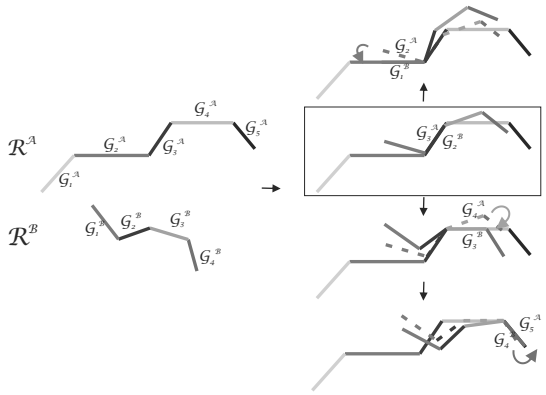
2.4 Early Termination

The nearest neighbor search, which is a part of the S-distance computation, has $O(n^2)$ time complexity with respect to the GSSU's length n. In addition, the search is performed for each of the candidate alignments decreasing the efficiency of SETTER. To increase the algorithm's speed a simple early termination condition is thus implemented. Alignments that are not likely to be the part of the optimal superposition are identified, and for these the nearest neighbor search is skipped. Such alignments will very likely have the triplet S-distance higher than the lowest GSSU distance obtained up to that time. Specifically, triplet-based S-distance will probably be lower than "real" S-distance. If the triplet $\mathcal{T}^\mathcal{A} \subset \mathcal{G}^\mathcal{A}$ is aligned with the triplet $\mathcal{T}^\mathcal{B} \subset \mathcal{G}^\mathcal{B}$ with $S(\mathcal{G}^\mathcal{A}, \mathcal{G}^\mathcal{B}) = \chi$ being the best result so far, the comparison computation can be terminated if $S(\mathcal{T}^\mathcal{A}, \mathcal{T}^\mathcal{B}) \times 1/\lambda > \chi$. Since the early termination is a heuristic ($S(\mathcal{T}^\mathcal{A}, \mathcal{T}^\mathcal{B}) < S(\mathcal{G}^\mathcal{A}, \mathcal{G}^\mathcal{B})$ does not have to be valid), the early termination condition is strengthened by introducing the parameter $\lambda \geq 1$. By varying the λ parameter, the trade-off between accuracy and speed can be set. The higher the λ is, the less often early termination occurs, and the more accurate and slower the algorithm is. The effect of the λ parameter on the quality of the functional annotation is demonstrated in section 4 of the Supplementary Information.

For each alignment SETTER outputs the list of residues forming individual GSSUs, $\overline{S}$-distance characterizing the overall quality of the alignment, p-value quantifying the statistical significance of the alignment, list of aligned GSSUs, rotation and translation matrices, 3D coordinates of each residue after the superposition, triplet pair of the best scoring GSSU pair, and list of residues with their respective nearest neighbors and the corresponding distances.

2.5 Structural Alignment Accuracy

The assessment of the quality of structural alignments is not an easy task because it is not possible to define what a perfect 3D-to-3D alignment is (Brown *et al.*, 2009). The commonly used measures such as e.g. the root-mean-square deviation (RMSD) require the knowledge of which residues are aligned against which ones. However, because SETTER is not based on a sequence alignment algorithm such information is missing. Therefore, the list of aligned residues was generated utilizing a simple geometric approach. Two residues A and B are considered to be aligned if A is the closest residue to B and, at the same time, B is the closest residue to A. Such a definition is, in our opinion, suitable for an evaluation of the quality of the superposition



$$\overline{S}(\mathcal{R}^\mathcal{A}, \mathcal{R}^\mathcal{B})) = (S(\mathcal{G}_3^A, \mathcal{G}_2^B) + \pi(\xi_2) S(\mathcal{G}_4^A, \mathcal{G}_3^B) +$$
$$\pi(\xi_3) S(\mathcal{G}_5^A, \mathcal{G}_4^B) + \pi(\xi_1) S(\mathcal{G}_2^A, \mathcal{G}_1^B))$$
$$\times \beta(|\mathcal{R}^\mathcal{A}|_{GSSU} - |\mathcal{R}^\mathcal{B}|_{GSSU})$$

b)

Fig. 3. A multiple GSSU structural comparison works in four steps. The figure represents a situation where only the optimal GSSU pair is considered ($\kappa = 1$), and the direction of the alignment is given by the order of GSSUs. **a)** A schematic representation of two RNA structures $\mathcal{R}^\mathcal{A}$ and $\mathcal{R}^\mathcal{B}$ with 5 and 4 GSSUs, respectively. The most similar GSSUs pair is $\{\mathcal{G}_3^A, \mathcal{G}_2^B\}$ (shown in rectangle). Structures $\mathcal{R}^\mathcal{A}$ and $\mathcal{R}^\mathcal{B}$ are aligned based on the rotation and translation of this pair. The superposition needed to optimally align $\mathcal{G}_4^A$ and $\mathcal{G}_3^B$ was obtained during the all-to-all GSSU pairwise comparison stage, and the rotation angle needed to get from the state $\{\mathcal{G}_3^A, \mathcal{G}_2^B\}$ to $\{\mathcal{G}_4^A, \mathcal{G}_3^B\}$ (first down arrow in the figure) is thus known. The current state is changed to $\{\mathcal{G}_4^A, \mathcal{G}_3^B\}$, and the process is repeated for the pair $\{\mathcal{G}_5^A, \mathcal{G}_4^B\}$ (second down arrow in the figure). Similarly, the algorithm must also process in the opposite direction from the position of $\mathcal{G}_2^A$ and $\mathcal{G}_1^B$ (up arrow in the figure). **b)** The rotation angles ξ_i from the previous step are used in the penalty function $\pi()$ which represents a weight function for the GSSU distances. To get the final $\overline{S}$-distance, the sum of weighted GSSU S-distances is normalized by a the ratio of non-aligned GSSUs over the maximum number of non-aligned GSSUs. The parameter β was empirically set to 6.

and can be used for an approximate comparison with others alignment-based methods.

In the present work the quality of the structural alignments was assessed by utilizing the following measures: the RMSD, the percentage of structural identity (PSI), the percentage of sequence identity (PID) (Capriotti and Marti-Renom, 2008, 2009) and the number of nucleotides aligned and the number of exact base matches (Rahrig *et al.*, 2010). RMSD captures the general 3D shape of RNA, but it can be misleading as the errors are spread over the whole molecule. PSI is defined as a percentage of superimposed residues within 4.0 Å with respect to the length of the shorter of the two structures. PID is the percentage of aligned nucleotides of the same type with respect to the length of the shorter of the two structures. Number of nucleotides aligned and number of exact base matches give similar information as PSI and PID. We note that these measures do not account for the specificity of RNA base-pairing and base-stacking interactions. Therefore, some new metrics particularly suitable to RNA structure comparison have been developed (Parisien *et al.*, 2009). However, these are not utilized in the present study as they would not allow to compare SETTER results with other approaches.

2.6 Statistical Significance of The Structural Alignment

The quality of the structural alignment can be assessed by means of statistical hypothesis testing. The key idea is to create a set of randomly generated structures, to align them and to fit the distribution of their $\overline{S}$-distances. For the given $\overline{S}$-distance its p-value can then be calculated. The alignment is a good one if its $\overline{S}$-distance is good compared to the distribution of $\overline{S}$-distances. This is reflected by its low p-value; the smaller the p-value, the more statistically significant the $\overline{S}$-distance is. To show how well data follow the fitted distribution a visual inspection of quantile-quantile plots (QQ-plots) can be used, or the fit can be tested by two-sample Kolmogorov-Smirnov test.

$\overline{S}$-distance follows the log-normal distribution (see section 5 in the Supplementary Information) which probability density function $\rho(x)$ is given as

$$\rho(x) = \frac{1}{x\sqrt{2\pi\sigma^2}} e^{-\frac{\ln x - \mu}{2\sigma^2}}$$

where parameters μ and σ are the mean and standard deviation, respectively, of the variable's natural logarithm that is, by definition, normally distributed. On a non-logarithmized scale μ is called a location parameter and σ a scale parameter. These parameters must be determined, and once they are known, they can be used to derive the statistical significance of the particular alignment given as its p-value. p-value corresponds to the probability $P(x \leq X)$ that the variable X takes a value lower or equal to x

$$P(X \leq x) = \frac{1}{2} + \frac{1}{2}\text{erf}\left(\frac{\ln(X) - \mu}{\sqrt{2\sigma^2}}\right)$$

where erf(x) is the error function defined as

$$\text{erf}(x) = \frac{2}{\pi} + \int_0^x e^{-t^2}\,\mathrm{d}t$$

For the determination of μ and σ parameters a set of reasonably unrelated structures was prepared. Such a set should cover the whole range of alignments starting from the exceptionally good ones going up to the very bad ones. The unrelatedness of the structures was based on their sequence similarity. The used threshold of 80% sequence similarity guarantees the uniform coverage of the alignments in terms of their quality. Because the μ and σ parameters depend on the length of the shorter structure N in the alignment they must be determined for different lengths separately. For each length a dataset containing 50,000 structure pairs was generated by randomly cutting the regions of the given length from structures longer than N. The data sets of lengths 5, 10, 15, 20, ...300 residues were prepared. The $\overline{S}$-distance was determined for each alignment in the given data set, and the parameters μ and σ of the log-normal distribution were found by a maximum likelihood fitting. All statistical calculations were performed using the R system version 2.13.1 (R Development Core Team, 2011) with the package MASS (version 7.3-14).

2.7 Functional Annotation Accuracy

The quality of the functional annotation was assessed by SETTER's ability to correctly assign the SCOR functional classification to the query RNA structure utilizing three datasets from the SCOR database (Tamura *et al.*, 2004). The FSCOR dataset contains all RNA chains with more than three nucleotides with a unique functional classification, the R-FSCOR dataset is a structurally dissimilar subset of the FSCOR, and the T-FSCOR dataset contains structures from the FSCOR set not present in the R-FSCOR set (Capriotti and Marti-Renom, 2008). Two RNA structures can be either functionally identical (referred to as the exact classification, they have the same deepest SCOR classification) or functionally similar (referred to as the similar classification, they do not agree at the deepest level but share classification at the parent level). Particularly, two experiments were performed — a leave-one-out test on the FSCOR dataset and a test assigning functions to structures from the T-FSCOR with the R-FSCOR serving as the database set. The accuracy of a functional annotation was assessed by utilizing two different measures: a classification accuracy (ACC), and an area under the ROC curve (AUC). ACC is calculated as a percentage of correctly classified structures. To obtain the ROC curve the alignments of all pairs of RNA structures were sorted by their p-values. A p-value threshold is the varied between minimum and maximum of the sorted p-values. For the fixed threshold, all pairs of aligned structures which p-values are above the threshold are assumed positive. Moreover, the pairs are counted as true positives (TP) if they belong to the same family (i.e. they are structurally similar) and false positives (FP) otherwise (i.e. they are structurally dissimilar). If P (positives) is the number of structurally similar pairs in the whole result set and N (negatives) is the number of structurally dissimilar pairs, then $\frac{FP}{N}$ is called a *false positive ratio* (FPR) and $\frac{TP}{P}$ a *true positive ratio* (TPR). The point on the ROC curve corresponding to the fixed threshold is produced by plotting its TPR (y-axis) against FPR (x-axis). The area under the ROC curve (AUC), a threshold independent measure, is considered a robust indicator of a classifier quality (Fawcett, 2006). An AUC of 1.0 indicates a perfect classifier and an AUC of 0.5 corresponds to a random classification. High AUC means that correct classifications are present mostly at the top of aligned structures sorted by their p-values. We also notice that it is difficult to obtain high values of both ACC and AUC simultaneously (see section 6 in the Supplementary Information), and both measures should thus be reported.

3 RESULTS AND DISCUSSION

3.1 Assessment of The Structural Alignment Quality

SETTER structural alignments were compared with SARA by calculating PSI values for the all-to-all comparisons using the FSCOR dataset. The results are summarized in the Fig. 4 showing that SETTER yields less alignments with very low PSI (up to 20%), and SARA returns slightly more alignments with PSI $> 90\%$. In terms of remaining PSI levels, both methods perform similarly.

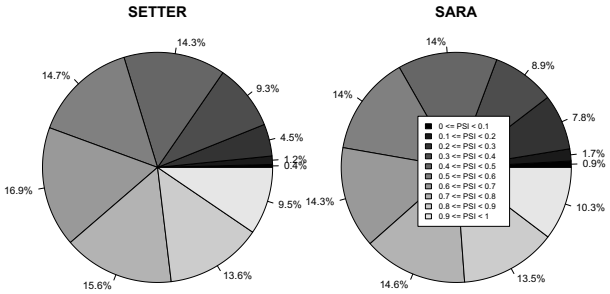


Fig. 4. PSI (percentage of structural identity) values produced by SARA and SETTER approaches for the all-to-all alignments using the FSCOR dataset.

In addition, SETTER was compared with R3D Align, ARTS, SARA and DIAL by calculating various measures reflecting the quality of the alignment of two 16S rRNA structures and of the alignment of the sarcin/ricin domain from 28S rRNA with the central part of the 5S rRNA. The results summarized in section 7 in the Supplementary Information also demonstrate that SETTER produces alignments of the quality comparable with other automated approaches considering its approximate nature in obtaining the list of aligned residues (see section 2.5). The ability of SETTER to produce good structural alignments is demonstrated on the visualizations of the superpositions of several 23S rRNA, 16S rRNA, 5S rRNA, tRNA and other RNA structures (see section 1 in the Supplementary Information).

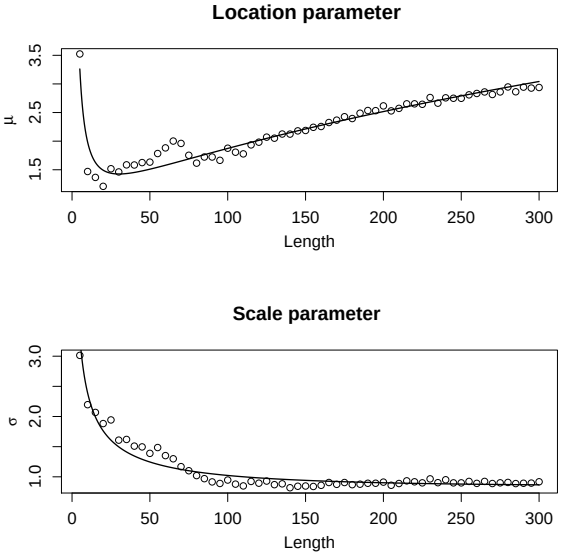


Fig. 5. Length dependence of log-normal parameters.

3.2 Assessment of The Functional Annotation Quality

3.2.1 Statistical Parameters Evaluation QQ-plots show that data follow the fitted log-normal distribution very closely except for the region of high $\overline{S}$-distances (see section 5 in the Supplementary Information). However, poor fitting in this region will not seriously influence the results of database searching, as we are generally not interested in highly dissimilar structures. The quality of the fit was further confirmed by Kolmogorov-Smirnov test that provided p-values close to zero for all sequence lengths. The location and scale parameters (μ and σ) of the log-normal distribution depend on the length N of the shorter structure in the alignment (Capriotti and Marti-Renom, 2008). The location parameter μ can be fitted by the following curve (see Fig. 5):

$$\mu = a \times \sqrt{N} + \frac{b}{N} \tag{2}$$

where $a = 0.1729$ and $b = 14.3753$. Similarly, the scale parameter was fitted (see Fig. 5) by

$$\sigma = a \times \frac{1}{\sqrt{N}} + b \times \ln(N)^2 \tag{3}$$

where $a = 7.3325$ and $b = 0.0136$. The relations (2) and (3) provide a simple way to calculate μ and σ parameters for any sequence length N.

3.2.2 Effectiveness Comparison The capabilities of SETTER for a functional annotation of new RNA structures were compared with SARA and iPARTS approaches using the published AUC (SARA, iPARTS) and ACC (SARA only) values on the FSCOR and T/R-FSCOR datasets. Following settings were used in SETTER: $\zeta = 0.2$, $\beta = 2$, $\epsilon = 6$ Å $\kappa = 3$, and $\lambda = 1$ (see sections 2.2, 2.3 and 2.4 for details).

The results are summarized in Table 1. The percentage of classified structures for the given p-value threshold is called a coverage, and two sets of results with different coverage are presented for SETTER. $STR_{pV=1}$ corresponds to the classification where the structures are sorted according to their p-values but no threshold is applied (coverage equals to 100%). At this coverage SETTER is compared with iPARTS which does not employ any filtering. Results in Table 1 show, that SETTER outperforms iPARTS in AUC for exact classification both in FSCOR and T/R-FSCOR datasets, and performs comparably for similar classification.

SARA was evaluated (Capriotti and Marti-Renom, 2009) at the coverage of 58.7%. To achieve this coverage in SETTER the p-value threshold of 0.013 was used. At this coverage SETTER performs better than SARA in terms of AUC both for the FSCOR and T/R-FSCOR datasets (see Table 1). In addition, SETTER can also be compared with SARA in terms of ACC. SETTER's ACC is comparable to that of SARA in the similar classification both for the FSCOR and T/R-FSCOR datasets, and in the exact classification for the FSCOR dataset. However, it is much higher (by 13.7%) for the exact classification for the T/R-FSCOR dataset.

Thus, it can be concluded that SETTER performs better than iPARTS and SARA in terms of AUC and is comparable with SARA in terms of ACC.

Table 1. ACC and AUC comparison of SETTER, iPARTS and SARA on the FSCOR and T/R-FSCOR datasets. The values are given in % and are reported for exact/similar classification. iPARTS should be compared to SETTER with the p-value threshold of 1.0 (i.e. no filtering applied), and SARA should be compared to SETTER with the p-value threshold of 0.013. For iPARTS, ACC was not reported and necessary tests can not be performed using the iPARTS web interface. ROC curves from which AUC value were calculated are shown in section 8 of the Supplementary Information.

	FSCOR		T/R-FSCOR	
	AUC	ACC	AUC	ACC
iPARTS	72/92	?	77/90	?
$STR_{pV=1.0}$	82/91	61.8/72.8	87/89	67.4/71.8
SARA	61/83	81.4/95.3	58/85	78.0/94.5
$STR_{pV=0.013}$	71/87	80.5/95.1	83/91	91.7/95.0

3.3 Speed Comparison

SETTER's nearest neighbor identification procedure scales as $O(n^2)$ with the size of the GSSU (not with the size of structure!), and employment of the heuristic speed optimization with $\lambda = 1$ further reduces the number of $O(n^2)$ computations. The speed of SETTER was compared to that of iPARTS and SARA measuring the runtimes of all-to-all comparisons on four datasets containing RNA structures of various sizes (Table 2). The runtime of SETTER was measured on Linux machine with 4 Intel(R) Xeon(R) CPUs E7540, 2GHz (the algorithm is not parallelized thus only one core per comparison was utilized) and 132 GB of RAM (however, the average memory size needed to store the representations of all RNA structures from the FSCOR set was less than 3.3 MB) running Red Hat Linux. The comparison was based on measuring the runtimes utilizing iPARTS and SARA server versions which limit the size of aligned structures (1,900 residues for iPARTS and 1,000 residues for SARA) from the performance reasons. Such a comparison can thus be only qualitative, however, the variations between SETTER and SARA/iPARTS are substantial (see Table 2) and can not be

attributed to a different hardware setup only. SETTER clearly outperforms both SARA and iPARTS methods and is much better suited for the alignment of even the largest RNA structures.

Table 2. Runtime comparison of iPARTS, SARA and SETTER for datasets of RNA structures of various size. The *D1* set contains tRNA structures 1EHZ:A, 1H3E:B, 1I9V:A, 2TRA:A and 1YFG:A (average length 76 nucleotides), *D2* set contains ribozyme P4-P6 domain 1GID:A, 1HR2:A and 1L8V:A (average length 157 nucleotides), *D3* contains domain V of 23S rRNA 1FFZ:A and 1FG0:A (average length 496 nucleotides), *D4* contains 16S rRNA 1J5E:A and 2AVY:A (average length 1522 nucleotides), and *D5* contains the currently largest RNA structures in PDB – yeast 25S rRNA 3O58:1 and 3O5H:1 (average length 3396 nucleotides). The runtimes of SARA and iPARTS were obtained from their server versions. A comparison with SARA is difficult, because for three of five data sets the server times out and returns no results.

data set	iPARTS	SARA	SETTER
D1	1.1 s	1.7 s	0.3 s
D2	2.6 s	9.2 s	2.4 s
D3	17.0 s	?	3.6 s
D4	2.8 min	?	21.3 s
D5	?	?	79.8 s

4 CONCLUSIONS

- The SETTER method divides the RNA structure into non-overlapping structural elements called generalized secondary structure units (GSSUs). The structural alignment is then based on the pairwise comparison utilizing 3D similarity of the GSSUs.
- SETTER was not developed for aligning RNA molecules not containing any secondary structure. However, such cases are very rare, no such a structure is present either in the FSCOR dataset or in large 16S or 23S rRNAs.
- The SETTER algorithm scales as $O(n^2)$ with the size of GSSU, and as $O(n)$ with the number of GSSUs in the structure. This scaling gives SETTER its unprecedented speed as the average size of GSSU remains constant irrespective of the size of the structure. However, it has been noted that due to the complexity of RNA 3D alignment the quadratic time algorithms (or better) can not be expected to be highly accurate (Ferrè *et al.*, 2007). Therefore, the main utility of the SETTER can be in identifying potential alignment regions which can further be processed by more accurate but computationally intensive methods such as R3D Align (Rahrig *et al.*, 2010).
- SETTER can be used both for pairwise structural alignment, as well as for functional annotation of a new RNA structure.
- The quality of the structural alignment was assessed by the comparison with R3D Align, ARTS, SARA and DIAL approaches. The results demonstrate that SETTER produces structural alignments of comparable quality.
- The functional assignment was benchmarked against iPARTS and SARA utilizing the classification accuracy (ACC) and the area under the ROC curve (AUC) measures for three datasets from the SCOR database. SETTER performs better than iPARTS and SARA in terms of AUC and is comparable with SARA in terms of ACC.
- SETTER method is capable of aligning even the largest RNA structures deposited in the PDB database in a reasonable amount of time (e.g., two structures of the 25S rRNA each having 3396 nucleotides and represented by 89 GSSUs are aligned in 1 minute and 20 seconds), and represents thus an important addition to the portfolio of automatic RNA structural analysis tools.

ACKNOWLEDGMENT

This work was supported by the Czech Science Foundation (GAČR) project Nr. P202/11/0968 and by the Ministry of Education of the Czech Republic MSM6046137302. We also thank anonymous referees whose stimulating critique and comments lead to the significant improvements of the manuscript.

REFERENCES

Abraham, M., Dror, O., Nussinov, R., and Wolfson, H. J. (2008). Analysis and classification of RNA tertiary structures. *RNA*, **14**(11), 2274–2289.

Arnez, J. G. and Steitz, T. A. (1994). Crystal structure of unmodified tRNA(Gln) complexed with glutaminyl-tRNA synthetase and ATP suggests a possible role for pseudo-uridines in stabilization of RNA structure. *Biochemistry*, **33**(24), 7560–7567. PMID: 8011621.

Bartel, D. P. (2004). MicroRNAs: genomics, biogenesis, mechanism, and function. *Cell*, **116**, 281–297.

Brown, J. W., Birmingham, A., Griffiths, P. E., Jossinet, F., Kachouri-Lafond, R., Knight, R., Lang, B. F., Leontis, N., Steger, G., Stombaugh, J., and Westhof, E. (2009). The RNA structure alignment ontology. *RNA (New York, N.Y.)*, **15**(9), 1623–1631.

Capriotti, E. and Marti-Renom, M. A. (2008). Rna structure alignment by a unit-vector approach. *Bioinformatics*, **24**, i112–i118.

Capriotti, E. and Marti-Renom, M. A. (2009). Sara: a server for function annotation of rna structures. *Nucl. Acids Res.*, page gkp433.

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**(Web-Server-Issue), 19–24.

Chew, L. P., Huttenlocher, D., Kedem, K., and Kleinberg, J. (1999). Fast detection of common geometric substructure in proteins. *Journal of computational biology : a journal of computational molecular cell biology*, **6**(3-4), 313–325.

Chursov, A., Walter, M. C., Schmidt, T., Mironov, A., Shneider, A., and Frishman, D. (2012). Sequence-structure relationships in yeast mRNAs. *Nucleic Acids Research*, **40**(3), 956–962.

de Brevern, A. G., Etchebest, C., and Hazout, S. (2000). Bayesian probabilistic approach for predicting backbone structures in terms of protein blocks. *Proteins*, **41**(3), 271–287.

Dror, O., Nussinov, R., and Wolfson, H. (2005). ARTS: alignment of RNA tertiary structures. *Bioinformatics*, **21 Suppl 2**.

Dror, O., Nussinov, R., and Wolfson, H. J. (2006). The ARTS web server for aligning RNA tertiary structures. *Nucleic Acids Res*, **34**(Web Server issue).

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recogn. Lett.*, **27**(8), 861–874.

Ferrè, 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**(Web-Server-Issue), 659–668.

Garman, E. (2003). 'Cool' crystals: macromolecular cryocrystallography and radiation damage. *Current Opinion in Structural Biology*, **13**(5), 545–551. PMID: 14568608.

Hendrix, D. K., Brenner, S. E., and Holbrook, S. R. (2005). RNA structural motifs: building blocks of a modular biomolecule. *Quarterly reviews of biophysics*, **38**(3), 221–243.

Holbrook, S. R. (2008). Structural principles from large RNAs. *Annual review of biophysics*, **37**(1), 445–464.

Holbrook, S. R., Cheong, C., Tinoco, I, J., and Kim, S. H. (1991). Crystal structure of an RNA double helix incorporating a track of non-Watson-Crick base pairs. *Nature*, **353**(6344), 579–581. PMID: 1922368.

Kabsch, W. (1976). A solution for the best rotation to relate two sets of vectors. *Acta Crystallographica Section A*, **32**(5), 922–923.

Kelley, R. L. and Kuroda, M. I. (2000). Noncoding RNA genes in dosage compensation and imprinting. *Cell*, **103**(1), 9–12.

Kim, R., Holbrook, E. L., Jancarik, J., and Kim, S. H. (1995). Synthesis and purification of milligram quantities of short RNA transcripts. *BioTechniques*, **18**(6), 992–994. PMID: 7546725.

Kim, S. H., Suddath, F. L., Quigley, G. J., McPherson, A., Sussman, J. L., Wang, A. H., Seeman, N. C., and Rich, A. (1974). Three-dimensional tertiary structure of yeast phenylalanine transfer RNA. *Science (New York, N.Y.)*, **185**(149), 435–440. PMID: 4601792.

Kolodny, R. and Linial, N. (2004). Approximate protein structural alignment in polynomial time. *Proc Natl Acad Sci U S A*, **101**(33), 12201–12206.

Lu, X.-J. and Olson, W. K. (2008). 3DNA: a versatile, integrated software system for the analysis, rebuilding and visualization of three-dimensional nucleic-acid structures. *Nature Protocols*, **3**(7), 1213–1227.

Mattick, J. S. and Makunin, I. V. (2006). Non-coding RNA. *Human Molecular Genetics*, **15**(suppl 1), R17–R29.

Parisien, M., Cruz, J. A. A., Westhof, E., and Major, F. (2009). New metrics for comparing and assessing discrepancies between RNA 3D structures and models. *RNA (New York, N.Y.)*, **15**(10), 1875–1885.

R Development Core Team (2011). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

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**(21), 2689–2697.

Scott, W. G. (2007). Ribozymes. *Current Opinion in Structural Biology*, **17**(3), 280–286.

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**(Database issue).

Tinoco, I. (1999). How RNA folds. *Journal of Molecular Biology*, **293**(2), 271–281.

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 Suppl**, W340–7.