SETTER - RNA SEcondary sTructure-based TERtiary Structure Similarity Algorithm

David Hoksza and Daniel Svozil

[1] Charles University in Prague, FMP, Department of software engineering,
Malostranské nám. 25, 118 00, Prague, Czech Republic
hoksza@ksi.mff.cuni.cz,
WWW home page: http://siret.ms.mff.cuni.cz/hoksza

[2] Institute of Chemical Technology Prague, Laboratory of Informatics and Chemistry, Technická 5, 166 28 Prague, Czech Republic
daniel.svozil@vscht.cz,
WWW home page: http://ich.vscht.cz/~svozil

Abstract. The recent interest in function of various RNA structures, reflected in the growth of solved RNA structures in PDB, calls for methods for effective and efficient similarity search in RNA structural databases. Here, we propose a method called **SETTER** (RNA **SE**condary s**T**ructure-based **TER**tiary structure similarity) based on partitioning of RNA structures into so-called generalized secondary structure units (GSSU). We introduce a fast similarity method exploiting RMSD-based algorithm allowing to assess distance to a pair of GSSU, and a method for aggregating these partial distances into a final distance corresponding to structural similarity of the examined RNA structures. Our algorithm yields not only the distance but also a superposition allowing to visualize the structural similarity. Comparative experiments show that our proposed method is competitive with the best existing solutions, both in terms of effectiveness and efficiency.

Keywords: RNA, RNA secondary structure, RNA tertiary structure, RNA structural similarity

1 Introduction

The primary components of living organisms - nucleic acids and proteins - are biopolymers, long linear molecules composed from the sequence of building blocks called monomers. While proteins are the active elements of cells, the instruction for their synthesis is stored in deoxyribonucleic acid (DNA). DNA is a biopolymer consisting of four types of units called nucleotides. Each nucleotide is composed from three parts: one of four possible bases (adenine (A), guanine (G), cytosine (C), thymine (T)), a sugar deoxyribose, and a phosphate group. Gene - the DNA sequence serving as a prescription for protein synthesis (expression) - determines which protein will be expressed in the organism in the given time at the given place. The bases within base-pairs are stabilized at their positions by the chemical interaction called hydrogen bond. In DNA the bases are complementary meaning that A pairs always with T, and C with G forming the so-called canonical (or Watson-Crick) base-pairs.

DNA is too valuable material to be used directly in the protein expression. Instead, the genetic information is first transcribed into another type of nucleic acid - ribonucleic acid RNA. The basic building blocks of RNA are similar to that of DNA with two important exceptions: thymine is substituted by uracil (U), and deoxyribose by ribose. Unlike DNA, most RNA molecules are single-stranded. However, the RNA chain is not stretched in biological conditions, instead it maintains a distinct 3D arrangement called conformation (or fold). The biological function of RNA is directly related to its conformation, and the study of 3D structure of biopolymers generally is very important for better understanding of the inner workings of living organisms. Resolved structures (i.e. xyz coordinates of all atoms in the molecule) are deposited into the PDB database [4] that is available free of charge to broad scientific community.

Single-stranded RNA molecules adopt very complex 3D structures, as the presence of ribose introduces additional hydrogen bonding site allowing for formation of various non-canonical base pairs [22]. RNA structure is hierarchical [5], and can be divided into primary (RNA sequence), secondary, tertiary and quaternary levels. RNA secondary structure motifs [17], that are stable independently of their 3D folds, can be defined as double helices combined with various types of loop structures, and they can be categorized based on the mutual positions of these simple elements. A single loop connecting the end of helix is a hairpin loop, two single strands linking a pair of double-helical segments comprise an internal loop (if one of these links is of zero length a bulge loop is formed), and three or more double-helical segments linked by a single-strand sequences form a junction loop. RNA motifs have been classified according to function, 3D structure or tertiary interaction in the SCOR (Structural classification of RNA) database [20, 27]. The SCOR classification system is based on the Directed Acyclic Graph (DAG) to reflect the fact, that RNA structural elements can have several distinct features and may belong to multiple classes. Characterization of secondary RNA motifs is important and it finds application in such areas as RNA design [18, 9], RNA structure prediction [25], RNA modeling [10] or RNA gene finding [7]. RNA plays a variety of essential roles in many cellular processes, including enzymatic activity [26], protein synthesis regulation [12], gene transcriptional regulation [2, 11] and chromosome replication [12, 16]. The knowledge of RNA 3D structure is indispensable for characterizing of such functions, and thus the ultimate goal remains the prediction of the tertiary structure.

Currently (January 2011), the PDB database stores 1980 RNA structures. Such a wealth of data allows the analysis and characterization of the RNA structural space, which may help to characterize RNA function. Since 3D structure is typically more evolutionary conserved than sequence, detecting structural similarities between RNA molecules can bring insights into their function that would not be detected by sequence information alone. The development of automatic tools capable of efficient and accurate RNA structural alignment and comparison has become an important part of structural bioinformatics of RNA. Detecting structural similarities between two RNA (or protein) molecules at the tertiary level is a difficult task that has been shown to be NP-hard [21]. Therefore currently available software tools for comparing two RNA 3D structures, such as ARTS [13, 14], DIAL [15], iPARTS [28], SARA [6] , SARSA [8] or LaJolla [3], are all based on some heuristic approaches.

The best existing approaches SARA and iPARTS to which SETTER is compared will now be briefly described. The SARA program represents distances among selected atoms as unit vectors existing in the unit spheres. All-to-all unit-vector RMS distances of consecutive unit spheres are computed and used as scoring matrix for subsequent dynamic programming based global alignment. Dynamic programming is also employed in iPARTS algorithm in which 3D RNA structures are represented as 1D sequences of 23 possible symbols, each of which corresponds to the distinct backbone conformational family.

In this paper, we propose a new pairwise RNA comparison method based on 3D similarity of the so-called general secondary structure units (GSSU) resembling secondary structure motifs. Each of the compared RNA structures is divided into non-intersecting set of GSSUs. For a pair of GSSUs, similarity measure is introduced based on executing multiple RMSD transformations on particular subsets from the GSSUs. The measure is then normalized to obtain the resulting distance/similarity (we will use the terms distance and similarity interchangeably throughout the text) of a pair of GSSUs. If the compared RNA structures contain more GSSUs, all-to-all distances are computed and aggregation takes place resulting in the pairwise RNA structure comparison. We show in the experimental section that our method outperforms SARA and iPARTS both in accuracy and runtime. Moreover, in SETTER there is essentially no limit on the size of aligned structures. This is in contrast with SARA and iPARTS which are (due to the use of dynamic programming) limited to structures having at most 1,000 and 1,900 nucleotides, respectively.

2 Method principles

For the purpose of our method, each nucleotide in an RNA structure is represented by its $C4'$ atom although any other backbone atom could be utilized. RNA structure is represented as a set of GSSUs that can be regarded as fundamental units of RNA structure. In contrast to the basic secondary structure motifs, GSSUs contain more information by comprising larger subsets of RNA. These subsets represent meaningful RNA partitioning being easy to work with.

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

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

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

Let i_{max} *and* j_{min} *be highest indices of the Watson-Crick paired bases. We define* ***loop*** *as* $\mathcal{L} = \{n_i\}_{i=i_{max}+1}^{j_{min}-1} \subset \mathcal{R}$ *and* ***stem*** *as* $\mathcal{R} \setminus \mathcal{L}$ *and* ***neck*** *as the pair* $\{n_{i_{max}}, n_{j_{min}}\}$.

Note that even a structure without a single Watson-Crick pair has a GSSU which is identical with the structure itself. Usually, a GSSU looks like a hairpin motif but compared to hairpin, GSSU can contain bulges and internal loops within its stem part (see e.g. Fig. 1).

Due to the limited space, we will only briefly describe the GSSU extraction algorithm instead of showing its exact version. In general, extraction processes an RNA structure in the order of its sequence generating GSSUs based on the presence/absence of Watson-Crick hydrogen bonding pattern of each nucleotide[3]. We differentiate two states — GSSU generation is proceeding and GSSU generation does not take place. If GSSU is not being generated, the nucleotides are pushed on the stack to be processed later. If a nucleotide hydrogen-bonded to the nucleotide in the stack is identified during the process of GSSU generation, all non hydrogen-bonded nucleotides lying between them and the boundary nucleotides are added to the GSSU. The process of GSSU $\mathcal{G}$ is finished when a pair $\{nt_1, nt_2\}$ is found where $nt_2 \notin \mathcal{G}$. An example of GSSUs found in the structure of glutamine tRNA (PDB code 1EXD) is shown in Fig. 1.

2.1 Single GSSU Pairwise Comparison

When SETTER compares structures consisting of multiple GSSUs, pairwise GSSU comparison is employed. Therefore, single GSSU comparison can be viewed as the principle component of SETTER.

Each GSSU is represented by the ordered set of 3D coordinates enhanced with bonding and nucleotide/atom type information. The common way how to assess similarity of two sets X and Y of points is to define pairing between them. The sets are then superposed by finding such translation and rotation that the mutual distances of individual paired points are minimized. Usually, the root mean square deviation (RMSD) is chosen as the distance measure, because there exists a polynomial time algorithm able to optimally superpose two structures given a pairing/alignment [19]. However, finding the optimal alignment is a hard problem. To evaluate the quality of alignments that can potentially be a part of global alignment SETTER uses Kabsch [19] RMSD algorithm. The search for the optimal superposition (including search for the optimal alignment) is NP-hard [21]. Because trying each possible alignment is not computationally feasible, suitable alignments with potential to participate in optimal alignment should be identified. That is the principle idea behind SETTER's structure comparison process.

[3] To obtain hydrogen bonding information from PDB files, we used the 3DNA utility [24, 23].

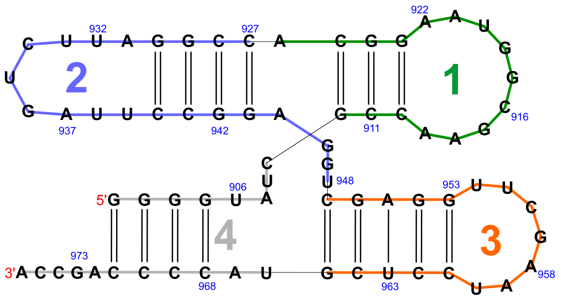


Fig. 1: 4 extracted GSSUs for RNA structure with PDB code 1EXD. The sequence starts at the 5' end and the colored numbers denote order of GSSU generation (number color corresponds with the respective GSSU's color). Note that, as GSSU 4 indicates, GSSU does not have to be comprised of a continuous chain of nucleotides but it has to correspond to the conditions of the Def. 1.

The nucleotides participating in necks of two GSSUs should not be missed in the optimal alignment. Otherwise stated, to superpose two GSSUs means to match their loops which implies also matching their necks. By matching necks, one can unambiguously superpose the structures in two dimensions but since in reality GSSUs exist in three dimensional space, at least three points are needed to define the superposition. We call these points *triplet*, and an *alignment* is formed by matching these points between two processed structures. Two matched points are further referred to as a "pair". Therefore, SETTER aligns necks and then tries to align each pair of loops' nucleotides one by one. The loop pair defines final pair in the triplet necessary to superpose the GSSUs. For example, if two GSSUs having loops consisting of n and m nucleotides to be aligned, $n \times m$ alignments are generated (see Figure 2)

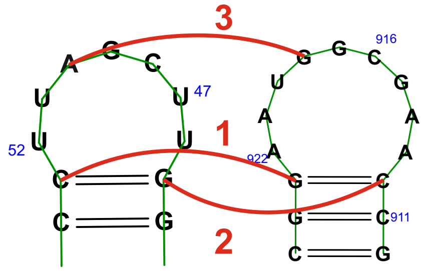


Fig. 2: Alignment of GSSU from tRNA domain of transfer-messenger RNA (PDB code 1P6V) with GSSU from glutamine tRNA (PDB code 1EXD). The final structural alignment is defined by three nucleotide pairs forming a triplet (the red lines 1, 2, and 3). To find an optimal superposition for the given neck pairs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

For each of the proposed alignments, a rotation matrix and a translation vector defining optimal superposition of the triplets is generated and subsequently used to superpose the whole GSSU. After that, nearest neighbor from the second GSSU in 3D space is identified for each nucleotide, and their distance is added to the overall distance of the two GSSUs. Finally the distance is normalized. The whole process can be formalized by equation 1.

$$
\begin{aligned}
NN_{\zeta}(x,\mathcal{G}) &= \begin{cases} \min_{1\le i\le|\mathcal{G}|}\{d_{nt}(x,\mathcal{G}_i)\}\times\zeta & \text{if } x=y \\ \min_{1\le i\le|\mathcal{G}|}\{d_{nt}(x,\mathcal{G}_i)\} & \text{otherwise} \end{cases} \\
\gamma(\mathcal{G}^{\mathcal{A}},\mathcal{G}^{\mathcal{B}}) &= \sum_{i=1}^{|\mathcal{G}^{\mathcal{A}}|} \begin{cases} 1 & \text{if } NN_1({\mathcal{G}^{\mathcal{A}}}_i,\mathcal{G}^{\mathcal{B}})\le\epsilon \\ 0 & \text{otherwise} \end{cases} \\
\delta(\mathcal{G}^{\mathcal{A}},\mathcal{G}^{\mathcal{B}}) &= \min_{t\in T}\Big\{\sum_{i=1}^{|\mathcal{G}^{\mathcal{A}}|} NN_{\alpha}({\mathcal{G}^{\mathcal{A}}}_i,\tau(\mathcal{G}^{\mathcal{B}},t))\}\Big\} \\
\mathit{\Delta}(\mathcal{R}^{\mathcal{A}},\mathcal{R}^{\mathcal{B}}) = \mathit{\Delta}(\mathcal{G}^{\mathcal{A}},\mathcal{G}^{\mathcal{B}}) &= \frac{\frac{\delta(\mathcal{G}^{\mathcal{A}},\mathcal{G}^{\mathcal{B}})}{\min\{|\mathcal{G}^{\mathcal{A}}|,|\mathcal{G}^{\mathcal{B}}|\}}\times(1+\frac{||\mathcal{G}^{\mathcal{A}}|-|\mathcal{G}^{\mathcal{A}}||}{\min\{|\mathcal{G}^{\mathcal{A}}|,|\mathcal{G}^{\mathcal{B}}|\}})}{\gamma(\mathcal{G}^{\mathcal{A}},\tau(\mathcal{G}^{\mathcal{B}},t_{opt}))}
\end{aligned}
\tag{1}
$$

In the formula, $\mathcal{G}^{\mathcal{A}}$ (identified with an RNA structure $\mathcal{R}^{\mathcal{A}}$) and $\mathcal{G}^{\mathcal{B}}$ (identified with an RNA structure $\mathcal{R}^{\mathcal{B}}$) represent the GSSUs to be compared, $\mathcal{G}_i$ stands for i-th nucleotide in the nucleotide sequence of $\mathcal{G}$ and $|\mathcal{G}|$ for its length. $NN(x,\mathcal{G})$ is the Euclidean distance from a nucleotide x to its nearest neighbor in $\mathcal{G}$. If x and its nearest neighbor have identical type, the distance is modified by factor ζ. δ computes the raw distance - T is a set of transpositions resulting from the candidate alignments and $\tau(\mathcal{G},t)$ transposes GSSU $\mathcal{G}$ using the transposition t. The normalized distance $\mathit{\Delta}$ then employs function γ counting number of nearest neighbors within the distance ϵ of the optimal transposition t_{opt}.

The whole process can be summarized in the following four steps:

1. Identify candidate set of alignments of triplet pairs (two nucleotides from neck, one from loop).
2. Compute superpositions (i.e. set of rotation matrices and translations vectors) for each of the alignments.
3. For each rotation matrix and translation vector superpose the structures.
4. For each superposition identify nearest neighbors, sum the distances to get δ and normalize it to obtain the final distance $\mathit{\Delta}$.

Sometimes identification of hydrogen bonds may not be correct and the real neck position within the GSSU is shifted. Therefore, SETTER also tries to simulate the neck shift by aligning the residues next to (under) the necks. Finally, when aligning the neck $\{n_1^{\mathcal{A}},n_2^{\mathcal{A}}\}$ of a GSSU $\mathcal{A}$ with then neck $\{n_1^{\mathcal{B}},n_2^{\mathcal{B}}\}$ of a GSSU $\mathcal{B}$ it is not clear in which direction the loops are oriented in 3D space (whether the correct alignment is $\{\{n_1^{\mathcal{A}},n_1^{\mathcal{B}}\},\{n_2^{\mathcal{A}},n_2^{\mathcal{B}}\}\}$ or $\{\{n_1^{\mathcal{A}},n_2^{\mathcal{B}}\},\{n_2^{\mathcal{A}},n_1^{\mathcal{B}}\}\}$)

and therefore both possibilities are investigated. These tweakings are necessary for accurate GSSU comparison, however they slightly increase the running time of SETTER.

Though in most cases the GSSU comprises of a stem and a loop, it is not a strict rule, as demonstrated in Fig. 1, GSSU number 4. Two particular situations can occur — GSSU has a zero-sized loop or the RNA does not have a single hydrogen bond (i.e., it does not have a secondary structure at all). In case of GSSU without the loop we select the third nucleotide for triplet alignment from the stem and we vary its position within the stem. When dealing with a GSSU having no secondary structure, several triplets covering whole structure are formed and used for the alignment.

2.2 Multiple GSSU Structure Comparison

For the comparison of RNA structures containing multiple GSSUs we utilized a straightforward solution. Consider the comparison of RNA structures $\mathcal{R}^{\mathcal{A}}$ and $\mathcal{R}^{\mathcal{B}}$ consisting of $n_{\mathcal{A}}$ and $n_{\mathcal{B}}$ GSSUs. We modify the Δ definition in the following way:

$$\Delta(\mathcal{R}^{\mathcal{A}}, \mathcal{R}^{\mathcal{B}}) = \min_{\substack{1 \leq i \leq n_A \\ 1 \leq j \leq n_B}} \left\{ \mathcal{G}_i^{\mathcal{A}}, \mathcal{G}_j^{\mathcal{B}} \right\} \times (1 + |n_{\mathcal{A}} - n_{\mathcal{B}}|) \times \beta \tag{2}$$

We compute all-to-all distances between the GSSUs and we choose the pair with minimal distance. Moreover, we multiply the distance by the difference in GSSUs count to favor structures with similar number of GSSU. The parameter β allows more distinct separation. Increasing value of β more noticeably favors similar-sized structures (in our experiment, we use $\beta = 2$).

SETTER uses only pairwise GSSU comparisons for matching RNA structures of any size including the largest ones such as ribosomal subunits. Since the mutual GSSU positions are rigid, the optimal superposition for a pair of GSSUs defines superposition for the whole structures which can be easily visualized (Fig. 3).

Though our solution follows the KISS principle (Keep It Simple and Stupid), it has several advantages over more elaborate approaches based for example on finding maximal common subgraphs in the network of interactions between individual GSSUs. Not only it is much faster, but it also allows to use effective early termination mechanism leading to additional speed improvements. This mechanism is introduced in the following section.

3 Speed up

The nearest neighbor search process needed for Δ computation is highly expensive since it has $O(n^2)$ time complexity with respect to the GSSU's length. Moreover, the process has to be done for each of the candidate alignment, noticeably decreasing efficiency of SETTER. Therefore we implemented simple early

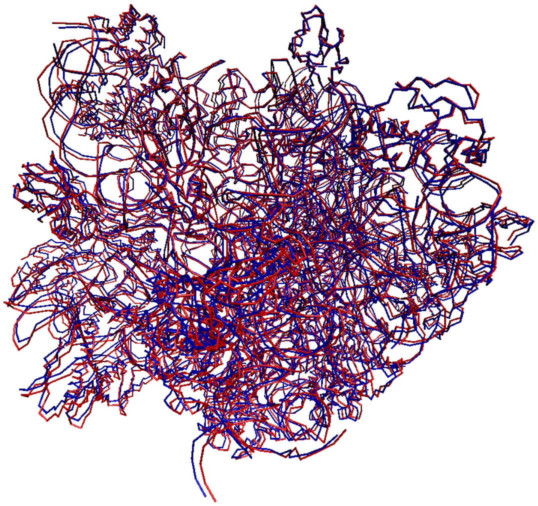


Fig. 3: The superposition of structures from 23S ribosomal RNA having PDB codes 1NWY:0 (2880 nucleotides, 84 GSSUs, blue) with 1SM1:0 (2880 nucleotides, 83 GSSUs, red) - RMSD = 2.43.

termination condition into the SETTER's GSSU comparison process. We identify alignments that are not likely part of the optimal superposition and for these alignments the nearest neighbor search is skipped. Because the superposition was optimized for the aligned triplets, their distance will be low compared to other nucleotides in the structure and they will very likely stay nearest neighbors also after the superposition of the whole GSSUs. Thus, Δ of the triplet-based GSSUs will be probably lower then Δ of the GSSU from which they come. If we align a triplet $\mathcal{T}^{\mathcal{A}} \subset \mathcal{G}^{\mathcal{A}}$ with a triplet $\mathcal{T}^{\mathcal{B}} \subset \mathcal{G}^{\mathcal{B}}$ with $\Delta(\mathcal{G}^{\mathcal{A}}, \mathcal{G}^{\mathcal{B}}) = \chi$ being the best result so far, the comparison computation can be terminated (i.e., we do not identify all the nearest neighbors) if $\Delta(\mathcal{T}^{\mathcal{A}}, \mathcal{T}^{\mathcal{B}}) \times \lambda > \chi$. Since the early termination is a heuristic ($\Delta(\mathcal{T}^{\mathcal{A}}, \mathcal{T}^{\mathcal{B}}) < \Delta(\mathcal{G}^{\mathcal{A}}, \mathcal{G}^{\mathcal{B}})$ does not have to be valid), we strengthen the early termination condition by introducing the parameter $\lambda \geq 1$. By varying the λ parameter, the trade-off between accuracy and speed can be set. In case of multiple GSSU comparison, the speed-up can be even more noticeable. The scope of χ variable can span multiple GSSU pairwise comparisons since we are searching for the minimum distance among all pairs of GSSUs. Such an approach can further emphasize the effect of the early termination condition.

4 Experimental Results

In order to evaluate SETTER and to compare it with other solutions we run test on datasets introduced in [6]. It contains three datasets — FSCOR, T-FSCOR, R-FSCOR based on functional classification obtained from the SCOR database [20], version 2.0.3. The FSCOR contains all RNA chains with more than three nucleotides with unique functional classification. The R-FSCOR is a structurally dissimilar subset of the FSCOR. The T-FSCOR set contains structures from the FSCOR set not present in the R-FSCOR set. Using these datasets we can evaluate quality of RNA similarity method in terms of functional assignment/classification ability. The task is to assign the functional (i.e. SCOR) classification to the query RNA structure by comparison with a database of classified RNA structure. Specifically, we performed two experiments — 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.

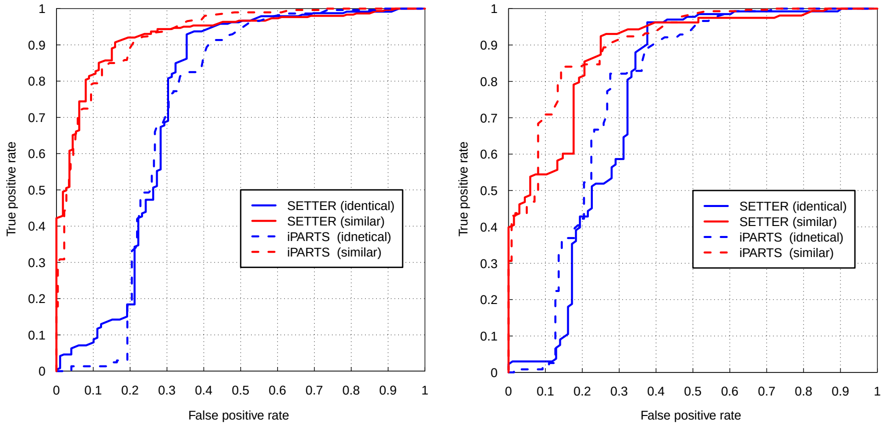
When comparing functions of two RNA structures, we differentiate between possessing identical and possessing similar function. Two structures have identical function if they share the deepest SCOR classification. If they do not agree at the deepest level but share classification at the parent level, they are said to have similar function.

In our experiments, we compute ROC curves and their AUC (area under the ROC curve) that is considered to be a robust indicator of quality of a classifier [1]. ROC is computed such that for each query we identify the most similar database structure (nearest neighbor) and the distance to it. The nearest neighbors for all queries are sorted according to their distances, and a distance threshold is varied from the most similar to the most dissimilar pair to generate points of the ROC curve. For a given threshold we identify number of structure pairs above the threshold with identical/similar function and denote them as true positives (TP). Rest of the above-threshold structures are denoted as false positives (FP). If P (positives) is the number of pairs with identical/similar function in the whole result set and N (negatives) the number of pairs with different functions, then $\frac{FP}{N}$ is called *false positive ratio* and $\frac{TP}{P}$ *true positive ratio*. The ROC curve consists of false positive ratio (x-axis) vs true positive ratio (y-axis) points.

Throughout the experimental section SETTER is used with following settings — $\alpha = 0.2$, $\beta = 2$, $\epsilon = 4\,\text{Å}$ and $\lambda = 1$ (see sections 2.1, 2.2 and 3 for details).

In Fig. 4, ROC curves of SETTER and iPARTS on FSCOR (Fig. 4a) and T-FSCOR (Fig. 4a) datasets are compared. We can see that on the FSCOR dataset, SETTER outperforms iPARTS with AUC equal to 0.74 (identical function) and 0.93 (similar function) in case of SETTER. iPARTS achieves AUC of 0.72 for identical function and of 0.92 for similar function. For SARA, only AUC values were presented in [6] being 0.61 and 0.83, respectively. When testing the T-FSCOR set against the R-FSOR set, SETTER is outperformed by iPARTS as is demonstrated by ROC curves in Fig. 4a. Specifically, AUC values are equal to 0.70 and 0.88 in case of SETTER, and to 0.77 and 0.90 in case of iPARTS. The results of SARA on the T-FSCOR set were again worse then the results of both SETTER and iPARTS — 0.58 for identical function and 0.85 for similar function.

We also carried out experiments measuring running time of SARA, SETTER and iPARTS. The runtime of SETTER was measured on Linux machine with 4 Intel(R) Xeon(R) CPUs E7540, 2GHz (the algorithm is not parallelized) and 132 GB of RAM (although the average memory size needed for an RNA structure



(a) FSCOR

(b) T-FSCOR vs R-FSCOR

Fig. 4: ROC curves of SETTER and iPARTS

from the FSCOR set is less then 3.3 MB) running Red Hat Linux. Runtime of SARA and iPARTS were taken from the output of their web interfaces. Thus, the comparison is only approximate. However, the variations between SETTER and SARA/iPARTS are substantial (Tab. 1) and can not be attributed to the hardware differences only.

Table 1: Runtime comparison of iPARTS, SARA and SETTER. The *tRNA* set contains structures 1EHZ:A, 1H3E:B, 1I9V:A, 2TRA:A and 1YFG:A structures (average length 76 nucleotides), *Ribozyme P4-P6 domain* contains 1GID:A, 1HR2:A and 1L8V:A (average length 157 nucleotides), *Domain V of 23S rRNA* contains 1FFZ:A and 1FG0:A (average length 496 nucleotides) and *16S rRNA* contains 1J5E:A and 2AVY:A (average length 1522 nucleotides).

data set	iPARTS	SARA	SETTER
tRNA	1.1 s	1.7 s	0.1 s
Ribozyme P4-P6 domain	2.6 s	9.2 s	1.8 s
Domain V of 23S rRNA	17.0 s	?	2.1 s
16S rRNA	2.8 min	?	8.1 s

In time of writing this paper, the SARA program was not able to handle sets *Domain V of 23S rRNA* and *16S rRNA*.

Table 1 shows times of all-to-all comparisons on four data sets. Note the difference between SETTER and iPARTS for growing structure size. It can be seen that SETTER's runtime grows more or less linearly with the size of the structure, in contrast to iPARTS where the growth is quadratical. That stems from iPARTS use of dynamic programming when searching for the alignment of 1D representations of the compared structures. In its original version, SETTER also uses $O(n^2)$ nearest neighbor identification procedure, but since it employs the speed optimization ($\lambda = 1$), the runtime of the algorithm, especially for large structures, is noticeably downsized.

5 Conclusion

In this paper, we have proposed a fast method for effective comparison of two RNA structures. The comparison is based on reasonably selected subsets of the nucleotide sequence resembling common secondary structure motifs. These subsets are then compared in three-dimensional space. Our method outperforms best existing solutions while maintaining high search speed.

In future, we would like to improve efficiency of our method by designing more sophisticated pruning method. We would also like to improve the effectivity by implementing multiple GSSU alignment and by introducing statistical methods, such as expectancy, into the classification process.

Acknowledgments. This work was supported by Czech Science Foundation project Nr. 201/09/0683 and by the Ministry of Education of the Czech Republic MSM6046137302.

References

1. P. Baldi, S. A. Brunak, Y. Chauvin, C. A. F. Andersen, and H. Nielsen. Assessing the accuracy of prediction algorithms for classification: an overview. *Bioinformatics*, 16(5):412–424, May 2000.
2. D. Bartel. MicroRNAs: Genomics, Biogenesis, Mechanism, and Function. *Cell*, 116(2):281–297, January 2004.
3. R. A. Bauer, K. Rother, P. Moor, K. Reinert, T. Steinke, J. M. Bujnicki, and R. Preissner. Fast structural alignment of biomolecules using a hash table, n-grams and string descriptors. *Algorithms*, 2(2):692–709, 2009.
4. H. M. Berman, J. D. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne. The protein data bank. *Nucleic Acids Res*, 28(1):235–242, 2000.
5. P. Brion and E. Westhof. Hierarchy and dynamics of rna folding. *Annual Review of Biophysics and Biomolecular Structure*, 26(1):113–137, 1997.
6. E. Capriotti and M. A. Marti-Renom. Rna structure alignment by a unit-vector approach. *Bioinformatics*, 24:i112–i118, August 2008.
7. R. J. Carter, I. Dubchak, and S. R. Holbrook. A computational approach to identify genes for functional RNAs in genomic sequences. *Nucl. Acids Res.*, 29(19):3928–3938, October 2001.
8. Y.-F. Chang, Y.-L. Huang, and C. L. Lu. Sarsa: a web tool for structural alignment of rna using a structural alphabet. *Nucleic Acids Res*, 36(Web-Server-Issue):19–24, 2008.

9. A. Chworos, I. Severcan, A. Y. Koyfman, P. Weinkam, E. Oroudjev, H. G. Hansma, and L. Jaeger. Building programmable jigsaw puzzles with RNA. *Science*, 306(5704):2068–2072, December 2004.
10. M. A. Ditzler, M. Otyepka, J. Šponer, and N. G. Walter. Molecular Dynamics and Quantum Mechanics of RNA: Conformational and Chemical Change We Can Believe In. *Accounts of Chemical Research*, 43(1):40–47, January 2010.
11. Y. Dorsett and T. Tuschl. siRNAs: applications in functional genomics and potential as therapeutics. *Nature Rev. Drug Discovery*, 3:318–329, 2004.
12. J. A. Doudna. Structural genomics of RNA. *Nat Struct Biol*, 7 Suppl:954–956, November 2000.
13. O. Dror, R. Nussinov, and H. Wolfson. ARTS: alignment of RNA tertiary structures. *Bioinformatics*, 21 Suppl 2, September 2005.
14. O. Dror, R. Nussinov, and H. J. Wolfson. The ARTS web server for aligning RNA tertiary structures. *Nucleic Acids Res*, 34(Web Server issue), July 2006.
15. F. Ferrè, Y. Ponty, W. A. Lorenz, and P. Clote. 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, 2007.
16. G. J. Hannon, F. V. Rivas, E. P. Murchison, and J. A. Steitz. The expanding universe of noncoding RNAs. *Cold Spring Harb Symp Quant Biol*, 71:551–564, 2006.
17. D. K. Hendrix, S. E. Brenner, and S. R. Holbrook. RNA structural motifs: building blocks of a modular biomolecule. *Q Rev Biophys*, 38(3):221–243, August 2005.
18. L. Jaeger, E. Westhof, and N. B. Leontis. TectoRNA: modular assembly units for the construction of RNA nano-objects. *Nucleic Acids Res*, 29(2):455–463, January 2001.
19. W. Kabsch. A solution for the best rotation to relate two sets of vectors. *Acta Crystallographica Section A*, 32(5):922–923, Sep 1976.
20. P. S. Klosterman, M. Tamura, S. R. Holbrook, and S. E. Brenner. SCOR: a Structural Classification of RNA database. *Nucleic Acids Res*, 30(1):392–394, January 2002.
21. R. Kolodny and N. Linial. Approximate protein structural alignment in polynomial time. *Proc Natl Acad Sci U S A*, 101(33):12201–12206, August 2004.
22. N. B. Leontis and E. Westhof. Geometric nomenclature and classification of RNA base pairs. *RNA*, 7(4):499–512, April 2001.
23. X.-J. Lu and W. K. Olson. 3DNA: a versatile, integrated software system for the analysis, rebuilding and visualization of three-dimensional nucleic-acid structures. *Nature Protocols*, 3(7):1213–1227, July 2008.
24. X.-J. J. Lu and W. K. Olson. 3DNA: a software package for the analysis, rebuilding and visualization of three-dimensional nucleic acid structures. *Nucleic Acids Res*, 31(17):5108–5121, September 2003.
25. B. Shapiro, Y. Yingling, W. Kasprzak, and E. Bindewald. Bridging the gap in RNA structure prediction. *Current Opinion in Structural Biology*, 17(2):157–165, April 2007.
26. D. W. Staple and S. E. Butcher. Pseudoknots: RNA structures with diverse functions. *PLoS biology*, 3(6), June 2005.
27. M. Tamura, D. K. Hendrix, P. S. Klosterman, N. R. Schimmelman, S. E. Brenner, and S. R. Holbrook. SCOR: Structural Classification of RNA, version 2.0. *Nucleic Acids Res*, 32(Database issue), January 2004.
28. C.-W. Wang, K.-T. Chen, and C. L. Lu. iPARTS: an improved tool of pairwise alignment of rna tertiary structures. *Nucleic Acids Res*, 38 Suppl:W340–7, 2010.