



POZNAN UNIVERSITY OF TECHNOLOGY

FACULTY OF COMPUTING AND TELECOMMUNICATIONS

Institute of Computing Science

DOCTORAL DISSERTATION

**Combinatorial Analysis
of RNA Tertiary Structures
with the Use of Angular
Representations**

Jakub Wiedemann, M.Sc.

Supervisor: **Maciej Antczak, Ph.D., Dr Habil.**

Co-Supervisor: **Maciej Miłostan, Ph.D.**

Poznań, 2022

Podziękowania

Chciałbym podziękować wszystkim, którzy przyczynili się do powstania niniejszej pracy.

Serdecznie dziękuję mojemu promotorowi dr hab. inż. Maciejowi Antczakowi, promotorowi pomocniczemu dr inż. Maciejowi Mitostanowi oraz prof. dr hab. inż. Marii Szachnick. Wasza wiedza, pomoc i wsparcie na każdym z etapów doktoratu oraz powstawania niniejszej pracy były nieocenione.

Dziękuję mojej rodzinie, bliskim i przyjaciołom za wsparcie, dobre rady i ciepłość.

Abstract

This dissertation summarizes the author's research in bioinformatics, a scientific field in which biology and computing science interlace each other. It describes selected issues in RNA structural biology and their solutions based on combinatorial optimization methodology.

The strong relationship between the structure and function of an RNA molecule is one of the paradigms underlying structural bioinformatics. It drives research on structure prediction, structural comparison, or quality assessment, leading to the exploration of molecule functions, drug design, and disease diagnosis. Many structural studies rely on atomic coordinates that form an algebraic representation of the molecular structure. Here, we focus mainly on an angular representation that illustrates chain folding and allows for significant optimization of the computational complexity of algorithms for structural data processing.

The presented doctoral study began with an exploration of the relationship between the sequence of the molecule and the tertiary structure. The work showed that high sequence identity was not always equivalent to global fold similarity in 3D space. Thus, a direct comparison of homologous tertiary structures is necessary. Moreover, local similarities should be considered, especially when looking for equivalent functional sites. This conclusion inspired the development of a new measure that assesses the angular similarity of 3D RNA structures. I developed the algorithm *LCS-TA* to identify

the Longest Continuous Segments in the Torsion Angle space within the context of the reference structure. The method compares 3D RNA structures, identifies fragments that expose similar folds, and returns their location and length (the latter acts as a measure of similarity). The algorithm *LCS-TA* was used to evaluate 3D RNA models submitted to Round IV of the RNA-Puzzles competition and is available in a toolkit provided by the RNA-Puzzles consortium. LCS-based analyses led my interest in RNA multiloops (also referred to as N-way junctions). These highly polymorphic structural motifs significantly affect the overall folding of RNA molecules. However, the lack of extensive analysis of their conformations in known, experimentally determined 3D structures of RNA makes their accurate in silico prediction extremely challenging. As a remedy for this, we developed the *RNAloops* database to collect multiloops identified in experimentally determined 3D RNA structures in a fully automated way and store them in a single repository. Their tertiary structures are described, i.a., by Euler and planar angles computed from atomic coordinates by own script provided by experimenters. The *RNAloops* enables multiparametric structural analysis and search for multiloops that meet user-defined criteria, for example, sequence, secondary structure, number of branches, etc. Such functionalities support, i.e., extracting structure motifs with specific features, their comparative analysis, or 3D structure modelling characterised by specific properties, e.g., when designing therapeutic solutions.

List of publications

Papers to form the basis of the dissertation:

- A1. **Wiedemann J**, Milostan M (2016) StructAnalyzer-a tool for sequence vs. structure similarity analysis. *Acta Biochimica Polonica* 63(4):753–757 (doi: 10.18388/abp.2016_1333).
- A2. **Wiedemann J**, Zok T, Milostan M, Szachniuk M (2017) LCS-TA to identify similar fragments in RNA 3D structures. *BMC Bioinformatics* 18(1): 456 (doi: 10.1186/s12859-017-1867-6).
- A3. Magnus M, Antczak M, Zok T, **Wiedemann J**, Lukasiak P, Cao Y, Bujnicki J, Westhof E, Szachniuk M, Miao Z (2020) RNA-Puzzles toolkit: a computational resource of RNA 3D structure benchmark datasets, structure manipulation, and evaluation tools. *Nucleic Acids Research* 48(2): 576–588 (doi: 10.1093/nar/gkz1108).
- A4. **Wiedemann J**, Kaczor J, Milostan M, Zok T, Blazewicz J, Szachniuk M, Antczak M (2022) RNAloops: a database of RNA multi-loops. *Bioinformatics* 38(17):4200–4205 (doi: 10.1093/bioinformatics/btac484).

Other papers:

- A5. Miao Z, Adamiak RW, Antczak M, Boniecki MJ, Bujnicki JM, Chen SJ, Cheng CY, Cheng Y, Chou FC, Das R, Dokholyan NV, Ding F,

Geniesse C, Jiang Y, Joshi A, Krokhotin A, Magnus M, Mailhot O, Major F, Mann TH, Piatkowski P, Pluta R, Popena M, Sarzynska J, Sun L, Szachniuk M, Tian S, Wang J, Watkins AM, **Wiedemann J**, Xu X, Yesselman JD, Zhang D, Zhang Z, Zhao C, Zhao P, Zhou Y, Zok T, Zyla A, Ren A, Batey RT, Golden BL, Huang L, Lilley DM, Liu Y, Patel DJ, Westhof E (2020) RNA-Puzzles Round IV: 3D structure predictions of four ribozymes and two aptamers. *RNA* 26(8):982–995 (doi: 10.1261/rna.075341.120).

Table 1: Bibliometric parameters.

Article ID	PY ¹	IF (PY ¹)	5-IF (2022)	MEiN ² (PY ¹)	MEiN ² (2021)	Quartile (WoS ³)	Rank (WoS ³)
A1	2016	1.159	2.175	15	70	Q4	246/295
A2	2017	2.213	3.629	35	100	Q2	16/58
A3	2020	16.971	15.542	200	200	Q1	8/295
A4	2022	6.937	7.197	200	200	Q1	3/58
A5	2020	4.942	4.941	140	140	Q2	93/295
Total		32.222	33.484	590	710	n/a	n/a

Table 2: Number of citations and H-index.

Article ID	Web of Science (all citations)	Web of Science (without self-citations)	Scopus	Google Scholar
A1	5	4	5	6
A2	9	8	10	9
A3	21	20	21	33
A4	0	0	0	0
A5	33	33	32	55
Total	68	65	68	103
H-index	4	4	4	4

All journals were qualified in the discipline of *Information and Communication Technology* by the MEiN². The rank of the journal (Table 2) is given for computational biology and bioinformatics, if possible, otherwise for the multidisciplinary area.

¹Publication Year

²The Ministry of National Education (Poland)

³Web of Science

Contribution to publications

I declare the following contribution to the publications underlying my doctoral dissertation:

- A1. I designed and implemented the *StructAnalyzer* tool to explore and analyze similarities between biological molecules and sequence-structure relationships. I developed new algorithms to compare tertiary structures following many-to-one or many-to-many scenarios. The first one performs a global comparison, and the second – a local comparison considering sequence alignment. I performed computational experiments and analyzed the results. I participated in the manuscript writing and prepared all the figures.
- A2. I designed and implemented the *LCS-TA* (Longest Continuous Segments in Torsion Angle space) algorithm to identify similar fragments of 3D RNA structures in torsion angle space. The proposed algorithm provides the sequence-dependent and sequence-independent comparison modes. I conducted computational experiments to evaluate the algorithm and interpreted the results. Following the latter, I adjusted the algorithm's default configuration. I participated in the writing of the manuscript by preparing the graphics and graphs included in the article. I did all the additional work during the revision process.
- A3. I adjusted the *LCS-TA* algorithm to fit the requirements of the RNA-Puzzles contest, where the comparison of multiple predictions was performed in the context of the reference structure. I performed computational experiments using *LCS-TA* for all the RNA-Puzzles challenges considered in this round. I was responsible for detecting and addressing data inconsistencies in the benchmark set. I aggregated the methods evaluation results as scoreboards prepared for all

considered challenges. I participated in the analysis of the pooled results, the writing of the manuscript, and the preparation of the figures.

- A4. I developed and implemented algorithms to compute planar and Euler angles between adjacent branches of a multiloop. I designed a prototype of *RNAloops*, a repository collecting data on multiloops identified in experimental 3D structures of RNA. I accumulated and analyzed statistics on both types of angles at N -way junctions. I aggregated these data and correlated them with information derived from the literature. I co-supervised the implementation of the *RNAloops* system, provided the preliminary database scheme, and evaluated the search function performance. I drafted the manuscript, performed all tests, and prepared the figures.

Jakub Wiedeman

Contents

Acknowledgements	i
Abstract	ii
List of publications	iii
Chapter 1 Introduction	1
1.1 RNA structure and function	1
1.2 Angle-based representation of RNA 3D structure	5
1.3 RNA 3D structure quality assessment	8
1.4 Computational basics	12
Chapter 2 Main results	15
2.1 Sequence vs. 3D structure analysis	15
2.2 Alignment of RNA 3D structures in torsion angle space . . .	19
2.3 RNA multiloops extraction and analysis	27
Bibliography	34
Publication reprints	42
Co-author declarations	86
Extended abstract in Polish	98
Appendices	101
A Participation in research projects	102
B Conference presentations	103
C Awards and distinctions	106
D Organizational activities	107

CHAPTER 1

Introduction

This chapter presents bioinformatics issues and concepts related to the research conducted in this dissertation.

1.1.

RNA structure and function

Ribonucleic acid (RNA) is a linear polymer composed of nucleotides linked by a phosphodiester bond. A single monomer (ribonucleotide) consists of a sugar (ribose) to which one of the four nitrogen bases via an N-glycosidic bond is attached. We can distinguish purine (adenine and guanine) and pyrimidine (cytosine and uracil) bases in RNA [Berg (2002)].

Four levels of RNA structural organisation can be distinguished: primary, secondary, tertiary, and quaternary structure (Figure 1.1). The primary structure of the RNA is represented by the sequence of nucleobases attached to the sugar-phosphate backbone, usually stored in FASTA-format files. The FASTA format consists of a header that includes the description of the molecule, followed by lines of sequence data [Lipman & Pearson (1985)].

The nucleobase pairing considering canonical [Halder & Bhattacharyya

1.1. RNA STRUCTURE AND FUNCTION

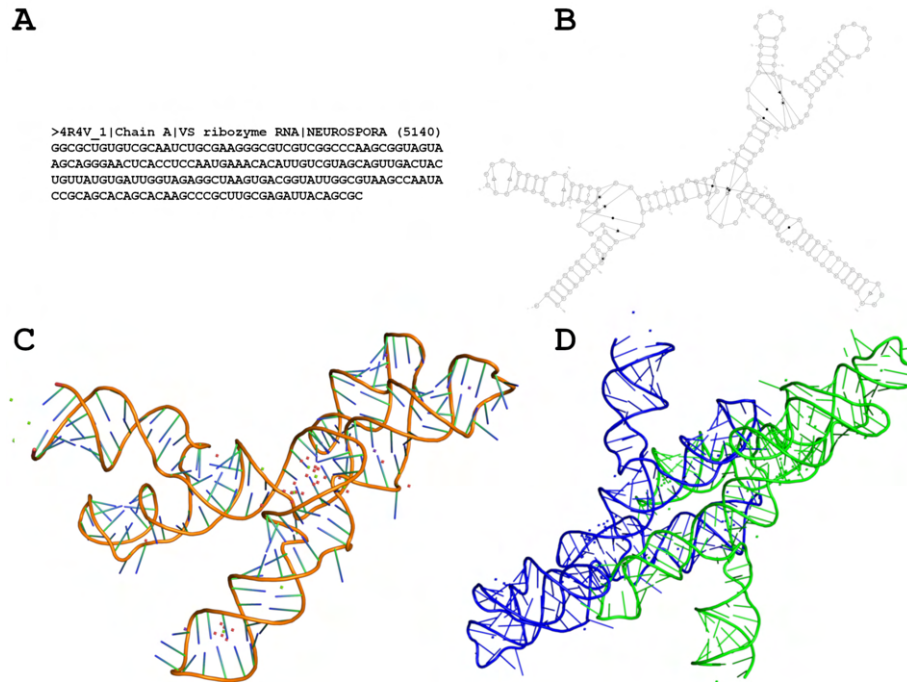


Figure 1.1: RNA organisation levels shown in the VS ribozyme structure (PDB id: 4V4R): (A) Primary structure (RNA sequence) presented in FASTA format, (B) Visualisation of RNA secondary structure prepared by VARNA [Darty et al. (2009)], (C) Tertiary and (D) Quaternary structures visualised by PyMOL [Schrödinger, LLC (2015)].

(2013); Šponer et al. (2005)] and noncanonical [Leontis & Westhof (2001); Hoehndorf et al. (2011)] base pairs describes the secondary structure of RNA. The former mainly forms a 3D-fold core of RNA. Nevertheless, non-canonical base pairs are usually crucial in ensuring proper function. Canonical pairings include base pairs A-U and G-C that form either two or three hydrogen bonds, respectively. [Watson & Crick (1974); Halder & Bhattacharyya (2013); Šponer et al. (2005)]. The wobble base pair G-U, formed by two hydrogen bonds, is also treated as canonical [Varani & McClain (2000)]. Information about secondary structures is usually stored using one of three formats: Connectivity Table (CT), BPSEQ, and Dot-Bracket [Ponty & Leclerc (2014)] (see Figure 1.2). The first (CT) consists of a header followed by separate records that describe each base. In the header, the number of bases is stored, and the name of the considered structure

1.1. RNA STRUCTURE AND FUNCTION

is given. The record describing each nucleobase considers six values: base number index (bni), one-letter nucleobase code (A, U, G, C), $bni - 1$, $bni + 1$, index of the paired nucleobase or 0 while it is unpaired, and repeated value of bni . The residue information in the BPSEQ format is described using three columns. The first represents the sequence position starting from one. The second and third columns represent the one-letter nucleobase code and the index of the paired nucleobase, or 0 while the residue is unpaired, respectively. The dot-bracket notation file consists of 3 lines: header, sequence, and secondary structure. The secondary structure in extended dot-bracket notation is represented as a linear string that includes dots and brackets. Its length corresponds to the length of the RNA sequence. In this format, dots represent unpaired residues and paired residues are denoted by various types of brackets representing pseudoknots.

The next level of RNA organisation is the tertiary structure. It describes a spatial arrangement of topological elements, building RNA structures stabilised by ions and hydrogen interactions. Many factors impact the 3D shape of RNA, e.g. sequence, environmental conditions, etc. [Eric & Pascal (2006); Hoehndorf et al. (2011); Zemora & Waldsich (2010)]. Prediction of the 3D RNA structure based on a given RNA sequence remains one of the unsolved challenges of structural bioinformatics [Leontis & Westhof (2012), Miao et al. (2020)], Townshend et al. (2021)]. Considering the gap between the number of both known RNA sequences and experimentally determined 3D structures and the fact that the overall fold is crucial to determine the function, it is essential to predict *in silico* the 3D shape of RNA molecules. Unfortunately, this is usually a nontrivial task because of the volatility and diversity of RNAs. According to the Anfinsen dogma, the RNA sequence with a high sequence identity with known experimentally determined structures should fold into a structurally similar 3D structure

1.1. RNA STRUCTURE AND FUNCTION

A

```
>strand_C
uGCCUGGCGGCCGUAGCGCGGUGGUCCC
ACCUGACCCCAUGCCGAACUCAGAAGUG
AAACGCCGUAGCGCCGAUGGUAGUGUGG
GGUCUCCCAUGCGAGAGUAGGGAACUG
CCAGGCAU
(((((((.(.....((.(.(.....
..(((.....)))).....
...)).)).).((.....((((
((...)))))).....))....)
)))))))).
```

B

```
120
1 u 0 2 119 1
2 G 1 3 118 2
3 C 2 4 117 3
4 C 3 5 116 4
5 U 4 6 115 5
6 G 5 7 114 6
7 G 6 8 113 7
8 C 7 9 112 8
9 G 8 10 0 9
10 G 9 11 110
```

C

```
1 u 119
2 G 118
3 C 117
4 C 116
5 U 115
6 G 114
7 G 113
8 C 112
9 G 0
10 G 110
```

Figure 1.2: Example secondary structure presented in: (A) dot-bracket notation, (B) CT format, and (C) BPSEQ format.

of RNA. However, in the literature, it can be found that even a point mutation in specific circumstances can significantly change the overall fold of the molecule [Wiedemann & Milostan (2017)]. However, there are also many cases where the 3D fold of the molecule is conservative and remains unchanged despite sequence changes [Hoehndorf et al. (2011)].

The last and highest level is the quaternary structure. It can be defined as interactions between RNA chains that form complexes such as dimers. Both tertiary and quaternary are mainly stored in mmCIF files [Bourne et al. (1997)] which is currently the successor to the PDB file format [(Bernstein et al., 1977)]. The mmCIF file format is built on dictionaries. The file contains entries describing 3D atom coordinates of a molecule extended by

1.2. ANGLE-BASED REPRESENTATION OF RNA 3D STRUCTURE

structure metadata such as determination experiment, authors, etc.

RNAs play a crucial role in various biological processes, such as regulation of gene expression or catalysis by chemical reaction [Berg (2002)]. For viruses, RNA acts as the primary genetic material. Many of them cause human diseases such as rabies [Albertini et al. (2007)], polio [Kitamura et al. (1981)], or COVID-19 [Hu et al. (2020)]. Furthermore, accumulation of noncoding RNA repeats can lead to diseases such as amyotrophic lateral sclerosis (ALS) or Huntington’s disease-like 2 (HDL-2) [Swinnen et al. (2019)]. Thus, predicting and understanding the 3D structure of RNA can drive the development of RNA therapies against many disorders, for example, neurodegenerative diseases.

1.2.

Angle-based representation of RNA 3D structure

RNA 3D structure representations include algebraic, geometric [Ryu et al. (2020); Gong & Fan (2019)], probabilistic [Frellsen et al. (2009)], and trigonometric models [Zok et al. (2013); Richardson et al. (2008)]. The most commonly used is the algebraic description of a structure. It describes the RNA model as a set of atoms together with their spatial coordinates using a Cartesian coordinate system. Another approach is a coarse-grained [Dawson et al. (2016)] in which the fully atomic representation is simplified to a model of beads. Geometric representation is usually based on distances between nucleotides [Gong & Fan (2019)] and the probabilistic approach is based on the distribution of atoms [Frellsen et al. (2009)]. The latter, the trigonometric model, uses an angle-based representation [Zok et al. (2013);

1.2. ANGLE-BASED REPRESENTATION OF RNA 3D STRUCTURE

Richardson et al. (2008)].

The use of an angular representation brings many benefits, e.g., in the case of comparing 3D structures of RNA. Measures operating on angular representation are superposition independent, thus allowing one to omit computationally demanding problem of 3D structures alignment. In some circumstances, the 3D structure superposition problem can be avoided by switching to the torsion angle space.

The 3D structure is often represented by the vector of torsion angles [Duarte & Pyle (1998); Zok et al. (2013); Richardson et al. (2008)]. The torsion angle (dihedral) is formed between two planes, intersecting in a 3D space. The planes are determined by a set of three atoms, so for the chain of four atoms A-B-C-D we can indicate a torsion angle between two planes defined by the triples of atoms A, B, C, and B, C, D, respectively (see Figure 1.3). To fully describe the shape of the RNA we need eight torsion angles α , β , γ ,

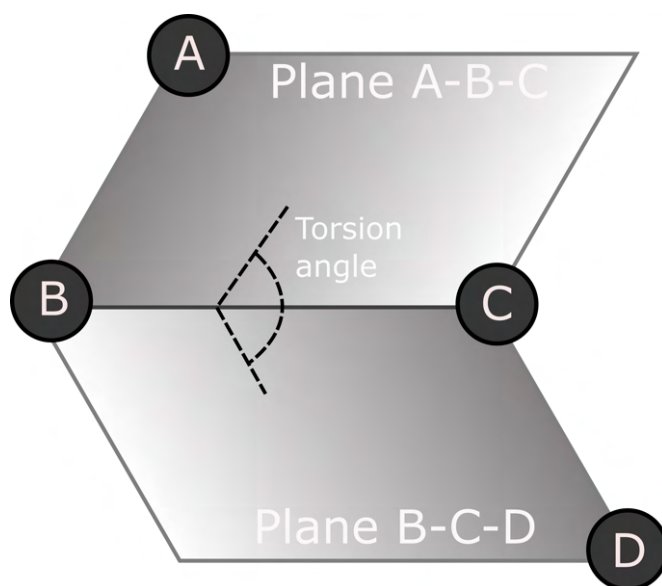


Figure 1.3: Torsion (dihedral) angle between planes.

δ , ϵ , ζ , P , and χ . Each of them is described by a chain of four consecutive atoms (see Figure 1.4, Table 1.1). P is a value representing the ribose ring that was defined as a pseudo-rotation of sugar pucker and calculated with the following formula (1.1) [Altona & Sundaralingam (1972)]:

1.2. ANGLE-BASED REPRESENTATION OF RNA 3D STRUCTURE

$$P = \arctan(\tau_4 + \tau_1 - \tau_3 - \tau_0, 2 * \tau_2 * (\sin 36^\circ + \sin 72^\circ)) \quad (1.1)$$

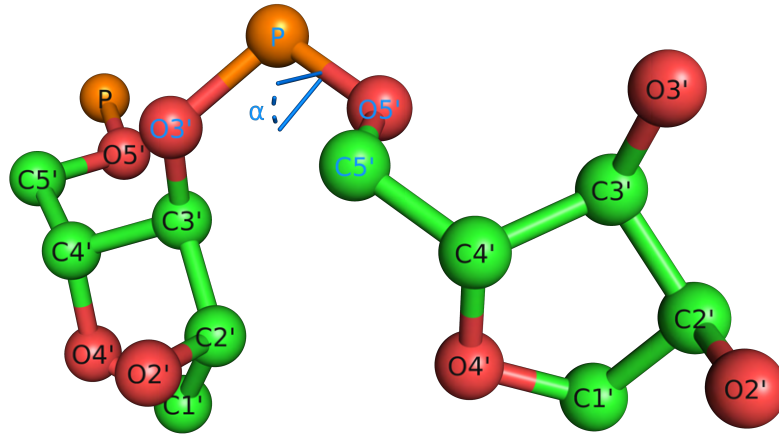


Figure 1.4: α torsion angle presented in RNA structure. Atoms defining the angle are denoted with blue labels.

The idea of using torsion angles was also used in the Ramachandran plot, where two pseudo-torsions angles (η , θ) are calculated for nucleotides and plotted against each other (see Figure 1.5) [Ramachandran et al. (1963)].

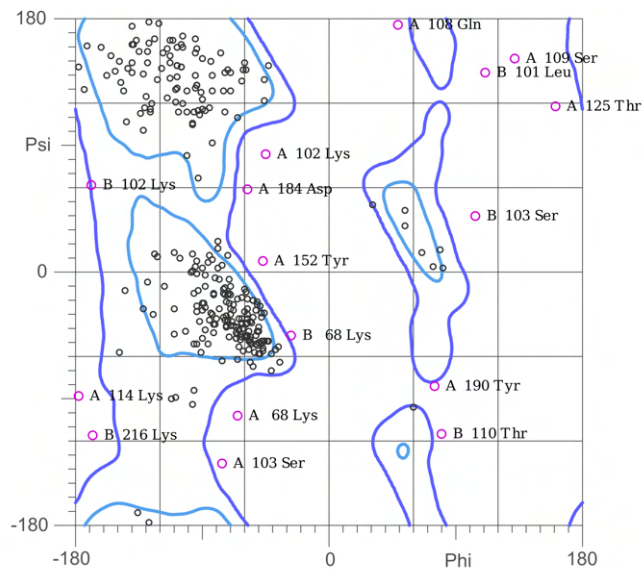


Figure 1.5: Example Ramachandran plot generated with MolProbity [Davis et al. (2007)] for structure 1HMP [Eads et al. (1994)].

Table 1.1: Torsion angles and atoms that define them [Saenger (2013)].

Torsion angle	Atoms
α	O3'(i-1)-P-O5'-C5'
β	P-O5'-C5'-C4'
γ	O5'-C5'-C4'-C3'
δ	C5'-C4'-C3'-O3'
ϵ	C4'-C3'-O3'-P(i+1)
ζ	C3'-O3'-P-O5'(i+1)
χ for pyrimidines (C and U)	O4'-C1'-N1-C2
χ for purines (G and A)	O4'-C1'-N9-C4
τ_0	C4'-O4'-C1'-C2'
τ_1	O4'-C1'-C2'-C3'
τ_2	C1'-C2'-C3'-C4'
τ_3	C2'-C3'-C4'-O4'
τ_4	C3'-C4'-O4'-C1'

1.3.

RNA 3D structure quality assessment

Comparison and assessment of structural models of biological molecules, predicted *in silico*, are among the crucial problems of structural bioinformatics. Comparison of 3D structures allows for the identification of structural similarities and the determination of the level of proximity of molecules in quantitative and qualitative manners, leading to a better understanding of their function. This is in line with a major paradigm of structural bioinformatics in which the sequence determines the structure and vice versa.

In 2011, RNA-Puzzle initiative [Cruz et al. (2012); Miao et al. (2020)] was established as a collective experiment to blindly predict the 3D structure of RNA. It aims to induce and encourage the RNA society to improve computational methods of 3D prediction of RNA structures. As more 3D structure prediction approaches for RNA emerge, the demand for reliable and efficient measures to assess the quality of predicted 3D models is of

1.3. RNA 3D STRUCTURE QUALITY ASSESSMENT

great interest.

A set of various measures that focus on the specific features of tertiary models is needed to compare and assess structures adequately in the overwhelming space of existing parameters. The RNA-Puzzle competition uses the following approaches: Root-Mean-Square Deviation (RMSD) [Kabsch (1976)], Interaction Network Fidelity (INF), Deformation Index (DI) [Parisien et al. (2009)], Clash score [Davis et al. (2007)], Mean of Circular Quantities (MCQ) [Zok et al. (2013)], and Longest Continuous Segments in Torsion Angle Space (LCS-TA) [Wiedemann et al. (2017)] (See Table 1.2). Root-mean-square deviation [Kabsch (1976)] is the most commonly known measure of 3D structure comparison. The calculation starts with the superposition of all the atoms considered between the compared structures. Next, the Euclidean distances are calculated for all the atom pairs considered. Finally, the result is calculated as the quadratic mean of these distances. It can be expressed as the following equation (1.2):

$$RMSD = \sqrt{\frac{1}{N} \sum_{i=1}^N \delta_i^2}, \quad (1.2)$$

where N is the number of atoms considered and δ is the distance between the corresponding atoms. Although it is a widely used and well-established measure, it could lead to misleading conclusions. The RMSD measure strongly depends on structural superposition. In addition, local differences, even small, can significantly influence the entire structure score. Unfortunately, different 3D folds or even little deviations in angles between outgoing stems of a multiloop can affect and disrupt the overall RMSD score. Finally, the RMSD value depends on the length of the RNA sequence.

The Interaction Network Fidelity measure (INF) [Parisien et al. (2009)] determines the correspondence between the interactions, considering Watson-Crick and non-canonical interactions, and the sequential stacking of bases,

1.3. RNA 3D STRUCTURE QUALITY ASSESSMENT

of the reference structure and that of the predicted structure. INF value is defined as the Matthews correlation coefficient (MCC) estimated using the following formula (1.3) [Parisien et al. (2009)]:

$$MCC = \sqrt{PPV \times STY}, \quad (1.3)$$

where specificity (PPV) (1.4) and sensitivity (STY) (1.5) are calculated based on sets of true positives (TP), false positives (FP), and false negatives (FN) using the following formulas:

$$PPV = \frac{|TP|}{|TP| + |FP|} \quad (1.4)$$

$$STY = \frac{|TP|}{|TP| + |FN|} \quad (1.5)$$

Compare two sets of interactions in two RNA structures. The first contains interactions within the structure of the target (S), and the second contains interactions within the structure of the model (S'). The interaction found at the intersection of both sets is a true positive, $TP = S \cap S'$. Interactions in S' that do not exist in S are false positives, $FP = S' \setminus S$. Interactions that are not present in S' but are present in S are false negatives, $FN = S \setminus S'$ [Parisien et al. (2009)].

Furthermore, the relationship between RMSD and INF is reflected in the Deformation Index measure (DI) defined as the ratio between RMSD and INF [Parisien et al. (2009)].

A Deformation Profile (DP) [Parisien et al. (2009)] is an average distance between the compared 3D structures of the RNA, presented using a two-dimensional matrix calculated in two steps. In the first step, for each aligned nucleotide pair, the superposition is calculated. The average distance was calculated for each pair of corresponding residues.

Clash score is used to measure the quality of the geometric parameters of

1.3. RNA 3D STRUCTURE QUALITY ASSESSMENT

the particular structure. It counts the number of overlapping atoms (for which the Euclidean distance is lower than 0.4\AA) per thousand atoms.

Mean of Circular Quantities (MCQ) [Zok et al. (2013)] compare 3D structures of RNA in torsion angle space. The MCQ value between two structures $\mathbf{St}, \mathbf{S't}$ is calculated with the following formula (1.6):

$$MCQ(\mathbf{St}, \mathbf{S't}) = \arctan\left(\frac{1}{r|\mathcal{T}|} \sum_{i=1}^r \sum_{j=1}^{|\mathcal{T}|} \sin \Delta(t_{ij}, t'_{ij}), \frac{1}{r|\mathcal{T}|} \sum_{i=1}^r \sum_{j=1}^{|\mathcal{T}|} \cos \Delta(t_{ij}, t'_{ij})\right) \quad (1.6)$$

where \mathbf{St} and $\mathbf{S't}$ are 3D structures given in the trigonometric representation, r is the number of residues in $\mathbf{S} \cap \mathbf{S'}$, \mathcal{T} is a set of torsion angles considered, and the distance between two corresponding angles t_{ij}, t'_{ij} is defined by (1.7):

$$\Delta(t, t') = \begin{cases} 0 & \text{if both } t \text{ and } t' \text{ are undefined} \\ \pi & \text{if either } t \text{ or } t' \text{ is undefined} \\ \min\{diff(t, t'), 2\pi - diff(t, t')\} & \text{otherwise} \end{cases} \quad (1.7)$$

where

$$diff(t, t') = |\text{mod}(t) - \text{mod}(t')| \quad (1.8)$$

and

$$\text{mod}(t) = (t + 2\pi) \text{ modulo } 2\pi \quad (1.9)$$

Longest Continuous Segments in Torsion Angle Space (*LCS-TA*) [Wiede-

1.4. COMPUTATIONAL BASICS

mann et al. (2017)] is a method to identify the longest continuous segments that are structurally similar. Two segments are considered similar if their *MCQ* value is below the predefined threshold.

Most of the comparison measures consider the 3D structures of the RNA as a set of atom coordinates. However, the 3D structure of RNA can be represented by a set of torsion angles that describe the course of its backbone and the arrangement of the bases. Such a trigonometric representation does not require one to superimpose 3D structures during comparison.

Table 1.2: Quality assessment measures for 3D RNA structures.

Method name	Structure representation	Assessment			
		Global	Local	Qualitative	Quantitative
RMSD	algebraic	✓		✓	
INF	algebraic	✓		✓	
DI	algebraic	✓		✓	
Clash score	algebraic	✓			✓
MCQ	trigonometric	✓		✓	
LCS-TA	trigonometric		✓		✓

1.4.

Computational basics

Solving biological problems usually involves performing complex calculations, analyses, or gathering and processing huge amounts of data. To answer these needs, computational biology emerged, together with the field of bioinformatics [Gauthier et al. (2018); Hagen (2000); Ouzounis & Valencia (2003)]. Its primary objective is to model and solve complex biologically inspired problems. This goal is usually achieved by developing and also applying widely known techniques originated from computer science and

1.4. COMPUTATIONAL BASICS

adjusting them to apply successfully in biological circumstances. One of the commonly used algorithms was proposed in 1970 by Saul Needleman and Christian Wunsch to align globally biological sequences using dynamic programming [Needleman & Wunsch (1970)].

An algorithm is usually defined as a set of specific instructions that allow for the achievement of a specific goal. It takes some value or a set of values and, by performing predefined steps, returns a result of these calculations [Rivest et al. (2009)]. They can be applied to solve various types of problems, and there are many ways to classify them. In general, we can identify two major groups, exact algorithms, and heuristics. Exact algorithms always return an optimal solution, while heuristics find an approximate solution that does not have to be optimal. Exact algorithms can be further split into a few major strategies, i.e.:

- Brute force search: In this naive strategy, the algorithm is enumerating every possible solution to find the optimal one [Rivest et al. (2009)].
- Divide-and-conquer technique – the problem is recursively divided into several subproblems of similar/related types until these become simple enough to solve directly [Rivest et al. (2009)].
- Branch-and-bound technique – this approach is based on searching the tree representing the problem’s solution space. Applied cut-offs reduce the number of search nodes, thus reducing the solution space to check. During branching, the set of solutions is divided by the particular node into subsets that include the successors of that node. The bounding procedure omits the branches of the tree whose paths will not always lead to the optimal solution [Rabiner (1984); Rivest et al. (2009)].

1.4. COMPUTATIONAL BASICS

- Dynamic programming – is a strategy of solving optimization problems (the so-called optimal substructure property problems) by dividing it into smaller subproblems and exploiting the fact that the partial solutions can be utilized to find the optimum to the main problem [Rabiner (1984); Rivest et al. (2009)].

With the continuously growing amount of biological data, there was a need to collect and fully automate the processing of these data, which resulted in the development of many biological databases, i.e., Protein Data Bank [Berman (2000)] or NCBI Reference Sequence Database (RefSeq) [Pruitt et al. (2011)]. Most biological databases use relational databases to store data. A relational database stores the data in a set of tables (entities) with columns and rows. Columns represent attributes collected in the table, while rows store data. Each row of a table is represented by a unique identifier called a primary key. The primary key can be used to refer to this record from other tables and to establish a relationship between two entities. All relations between considered entities define the database logical structure [Date (2006)].

An important part of database design is the selection of a fully functional and efficient database management system (DBMS), because data need to be easily modified or searched. Generally, a database management system is software for managing the database and data. It is an interface between entities stored in the database and end users or other applications [Date (2006)].

CHAPTER 2

Main results

The research described in this dissertation was mostly concerned with analysis of the tertiary structures of RNA. It was also related to the influence of various factors on the overall 3D fold of the RNA molecule. As a result of this analysis, new methods were proposed that allow the assessment of 3D RNA structures. This chapter briefly summarizes the research conducted and the results obtained. The full texts of the articles are included in the next section.

2.1.

Sequence vs. 3D structure analysis

The research described in this section concerns the exploration of the structural diversity of RNAs deposited in the Protein Data Bank [Burley et al. (2020)]. Conducted research resulted in introducing a new tool (*StructAnalyzer*) and identifying the importance of local comparison of structures.

2.1.1. Background

[A1] summarises the study of the projection between the sequence and the 3D structure of the RNA. Comparative analysis of biological sequences and

2.1. SEQUENCE VS. 3D STRUCTURE ANALYSIS

structures can lead to the determination of common, characteristic structural, and functional elements of biological compounds. Thorough exploration of sequence-structure relationships allows one to identify biological molecules with significantly different sequences but similar 3D structures and, on the other hand, molecules with quite similar sequences but significantly differing 3D folds. In this paper, we focus our attention on the latter case because it allows one to identify sequences prone to significant structural change due to the small number of point mutations. This kind of research is often crucial to a better understanding of the process of folding of molecules, which according to the Anfinsen dogma can lead to the discovery of their function. Thus, we decided to explore the structural diversity of similar RNA sequences (with sequence identity 90-100%) for 3D structures deposited in the Protein Data Bank [Burley et al. (2020)]. Analysis of multiple sequences and 3D structures is expensive in processing time. Therefore, the usage of concurrency processing is advisable in such cases. Therefore, we proposed a tool, called *StructAnalyzer*, that supports the analysis of the relationships between the RNA sequence and the 3D structure.

2.1.2. Results

StructAnalyzer was designed to compare RNA structures at two levels, sequential and structural. The proposed approach allows one to compare sets of 3D structures as well as to perform a pairwise comparison of them. The input requires 3D models in PDB format and multi-FASTA files (concatenating multiple single-sequence FASTA files). The comparison results can be exported to Comma-Separated Values files or visualised as heat maps, valuable during interpretation of the results.

StructAnalyzer workflow is shown on Figure 2.1. In the first stage, the

2.1. SEQUENCE VS. 3D STRUCTURE ANALYSIS

algorithm generates sequence alignment using MUSCLE software [Edgar (2004)]. The alignment is the basis for further analysis. The aligned fragments of the sequences are then selected. The selected fragments are structurally superimposed with their equivalents in the reference structure, and the algorithm calculates structural similarity using the Root-Mean-Square Deviation score (RMSD) [Kabsch (1976)]. In the case of pairwise comparison, *StructAnalyzer* allows for comparative analysis of all structurally similar fragments of a predetermined length (frame) identified between two molecules considered.

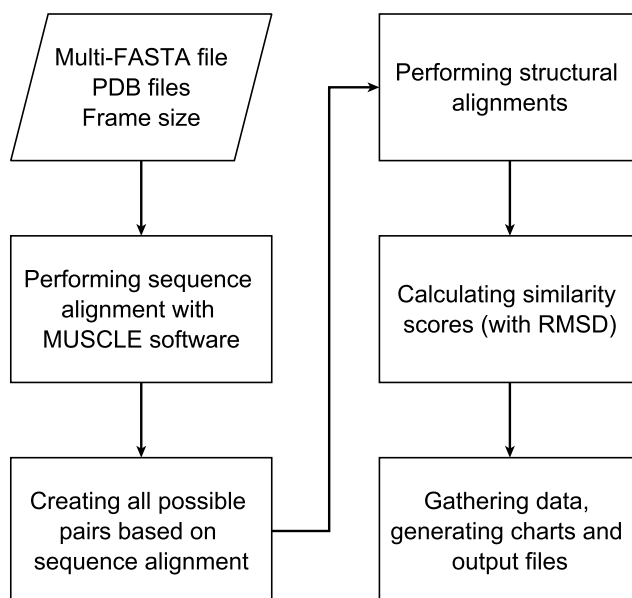


Figure 2.1: StructAnalyzer’s workflow.

The scope of the research was mainly to find structures with a high sequence identity that may differ significantly in the tertiary structure. To achieve this, we generated pairwise sequence alignments for all RNA structures stored in the PDB database. In the second step, we developed two sets of RNAs on the basis of sequence identities. The first set contained structures with 100% sequence identity. The second includes structures with sequence identity greater than or equal to 90% but less than 100%. For the sets considered, we created matrices of the global RMSD scores

2.1. SEQUENCE VS. 3D STRUCTURE ANALYSIS

(Fig 2.2). Reviewing this set allowed us to identify some structures with high sequential identity that differ significantly in the tertiary structure level. We further investigate those cases by performing a pairwise comparison of these structures to find the reasons for these differences (Fig. 2.3). The results obtained showed that when comparing the local motifs of the 3D structures of the RNA, we can observe significantly smaller differences than those obtained from the global perspective.

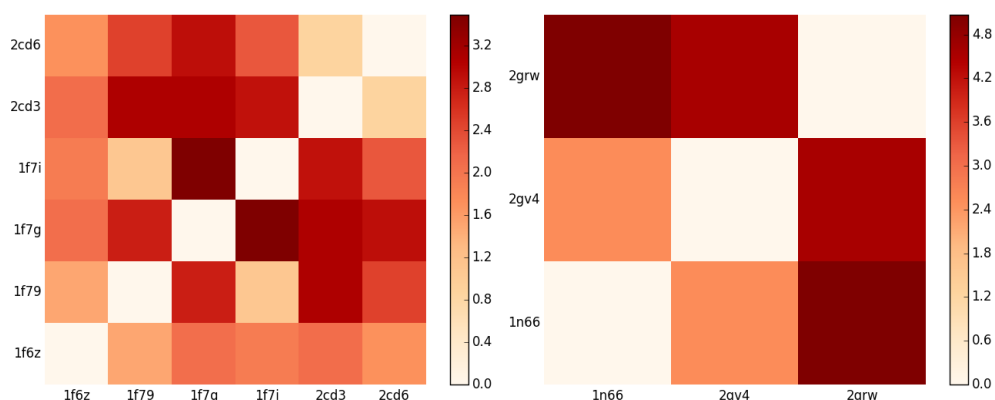


Figure 2.2: StructAnalyzer's many-to-many sequence vs. 3D structure global comparison. Adapted from Wiedemann & Miłostan (2017).

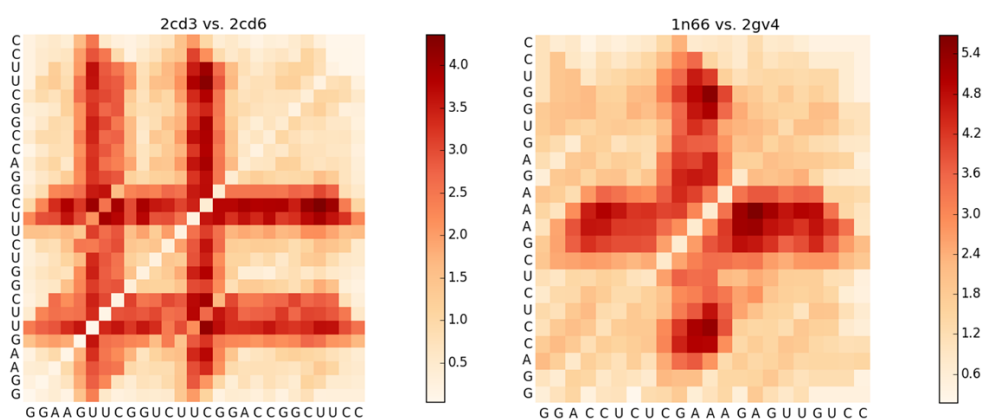


Figure 2.3: StructAnalyzer's many-to-many sequence vs. 3D structure local comparison (frame size is equal to 5). Adapted from Wiedemann & Miłostan (2017).

2.1.3. Conclusions

We found that, despite the high value of RMSD between molecules considered from a global perspective, when considering the local scope, we can find similar common motifs. However, even point mutations in the sequence can significantly affect the overall fold of the structure. This research highlights the need to assess structures both globally and locally. In the future, the tool will be extended with additional quality assessment measures, adjusted to utilize mmCIF file format, and equipped with an interactive graphical interface.

2.2.

Alignment of RNA 3D structures in torsion angle space

The research described in this section summarizes research on utilizing torsion angle representation for comparing and assessing the structural similarity of RNAs. Conducted research resulted in proposing a new measure for RNA structure assessment.

2.2.1. Background

Identification of common features and differences in 3D structures of biomolecules is a challenging task that requires the application of efficient computing methods. There is a necessity to develop new and tune existing similarity measures to reliably analyse and evaluate structures, especially those predicted in silico. The articles [A2], [A3], and [A5] refer

2.2. ALIGNMENT OF RNA 3D STRUCTURES IN TORSION ANGLE SPACE

to research focused on the local quality assessment of 3D structures of RNA. The research conducted in the article [A1] showed that even a point mutation in the structure can affect the overall 3D fold of RNAs. Thus, we identified the need to explore local similarities of the structures rather than analysing them globally. As a starting point, we reviewed the available and most commonly used quality assessment approaches that were used in the RNA-Puzzles competition [Miao et al. (2020)] to assess structures. During the review, we observed two major characteristics that were common for most of the analysed approaches: comparisons are performed from a global perspective, and 3D structures are described using algebraic representation.

Taking into account these results and observations from the article [A1], we concluded that there is a gap that needs to be filled. Therefore, we decided to propose a new method to locally align 3D RNA structures. At the time of RNA, an approach (*RNAnalyzer*) provided information on the quality of local alignment for the models considered. It uses the concept of spheres built with predefined radius that allowed the user to compare structures on different levels of structural detail. The method is available through the RNAssess web server. In the case of proteins, one of the most popular approaches, used in the CASP competition (Critical Assessment of methods of protein Structure Prediction) [Pereira et al. (2021)], for local evaluation is the Local-Global Alignment (LGA) [Zemla (2003)]. It combines two approaches, the Longest Continuous Segment (LCS) [Zemla et al. (1999)] and the Global Distance Test (GDT) [Zemla et al. (1999)], to assess proteins. The first locates the longest continuous segments that fit under the given RMSD cut-off threshold. The latter identifies the largest set of residues that fit a predefined RMSD cut-off, but the residues do not have to form a continuous segment.

Algebraic structure representation requires 3D structure superposition to

2.2. ALIGNMENT OF RNA 3D STRUCTURES IN TORSION ANGLE SPACE

compute the RMSD score during structure evaluation. However, this step can be omitted when one switches to the torsion angle representation, where the course of its backbone and the arrangement of the bases are described by the torsion angle vector. Such a representation allows for a comparison of structures independent of their superposition, simplifying the computation.

2.2.2. Results

Gathered observations resulted in the design of a new method, the so-called Longest Continuous Segments in the Torsion Angle space (*LCS-TA*) [A2]. The proposed solution identifies the longest continuous local alignment. *LCS-TA* operates in the torsion angle space, so it is superposition independent. Two segments are considered similar if their MCQ (Mean of Circular Quantities) value [Zok et al. (2013)] is below the predefined threshold. *LCS-TA* supports the following modes: sequence-dependent and sequence-independent. The first one searches for the longest continuous segment taking into account sequential alignments, so it aligns only fragments of the same sequence. The latter does not require 100% sequence identities between aligned fragments of given structures. The method has been incorporated into the *MCQ4Structures* software (<https://github.com/tzok/mcq4structures>) [Zok et al. (2013)].

2.2. ALIGNMENT OF RNA 3D STRUCTURES IN TORSION ANGLE SPACE

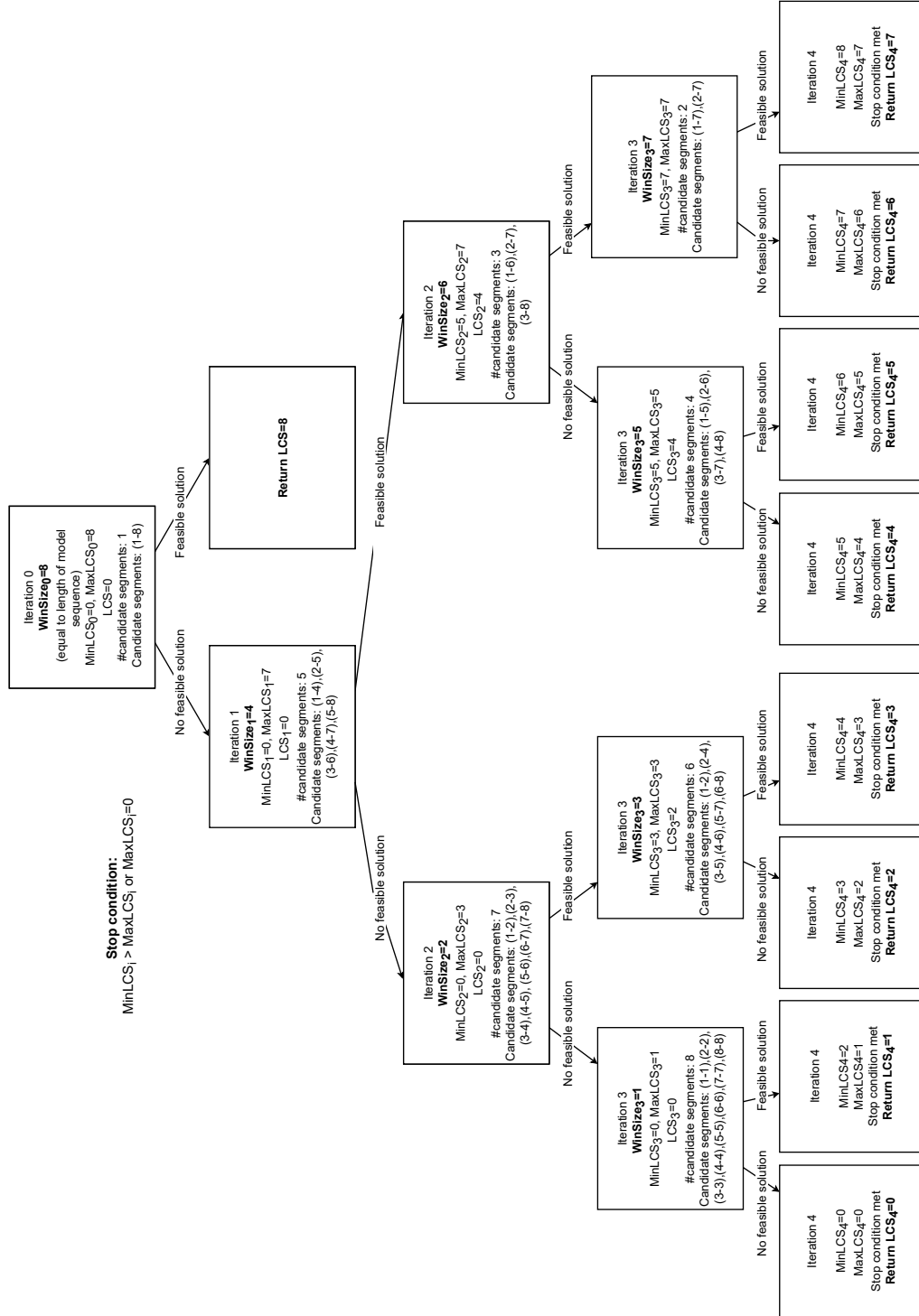


Figure 2.4: Recursion tree as an example visualisation of divide-and-conquer-driven computation applied in the LCS-TA algorithm. The LCS variable defines the length of the longest continuous segment. Reprinted from Wiedemann et al. (2017).

The algorithm *LCS-TA* is based on the divide-and-conquer approach. As input, it requires tertiary structures stored in the PDB format and the

2.2. ALIGNMENT OF RNA 3D STRUCTURES IN TORSION ANGLE SPACE

predefined MCQ cut-off value defined by the user. In the first step, for the molecule analysed, the algorithm (using the divide-and-conquer technique; see Fig. 2.4) chooses the promising length of the fragment to be analysed. At this stage, the algorithm verifies whether in the set of fragments (including all possible fragments with predetermined length) there exists a feasible solution (i.e., the solution that meets the requirements) by comparing them with the reference structure and looking for fragment with the highest value of the MCQ score within the predefined threshold. If the solution exists, then in the next step, the algorithm looks for a longer solution. In another case, if the set does not contain any feasible solution, the procedure searches the collections that contain shorter fragments. The procedure continues until the end of the division procedure.

To show the capabilities of the presented approach, we conduct two experiments. In the first, we run the *LCS-TA* for two models from the 18th challenge of the RNA Puzzle competition [Miao et al. (2020)]. The first model was predicted by the RNAComposer system [Antczak et al. (2016)] in a server category, and the second was submitted by the Chen group (in the human category). Both models were compared with the reference structure of exonuclease-resistant Zika virus RNA (PDB id: 5TPY) in the following modes: sequence-dependent (Fig 2.5) and sequence-independent (Fig 2.6).

In the second experiment, we have investigated models submitted for the 18th and 19th challenges of the RNA-Puzzles competition. The 18th and 19th challenges include 53 and 54 unique 3D models, respectively. From these sets, we have selected one model per participant (the first model submitted by each participant was selected) and compared them to the reference structure, i.e., exonuclease-resistant Zika virus RNA (PDB id: 5TPY) in the 18th challenge (Fig. 2.7), and twister sister ribozyme (TS) (PDB id: 5T5A) in the 19th challenge (Fig 2.8). For this analysis, we

2.2. ALIGNMENT OF RNA 3D STRUCTURES IN TORSION ANGLE SPACE

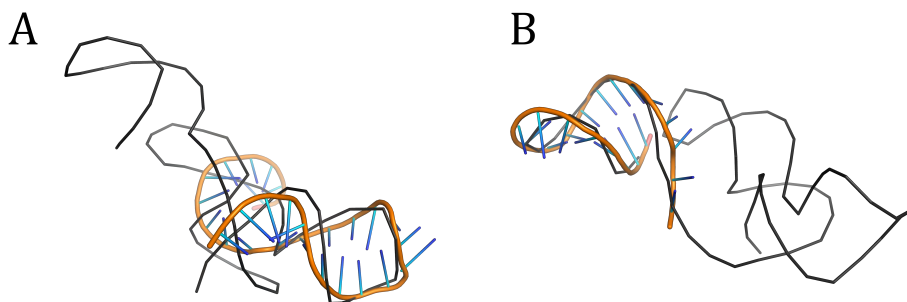


Figure 2.5: The longest continuous segments found by the LCS-TA algorithm in sequence-dependent mode (the MCQ threshold value set at 20°) for (A) Chen 1st and (B) RNAComposer 1st models, aligned with the reference structure (PDB id: 5TPY) (black backbone). Adapted from Wiedemann et al. (2017).

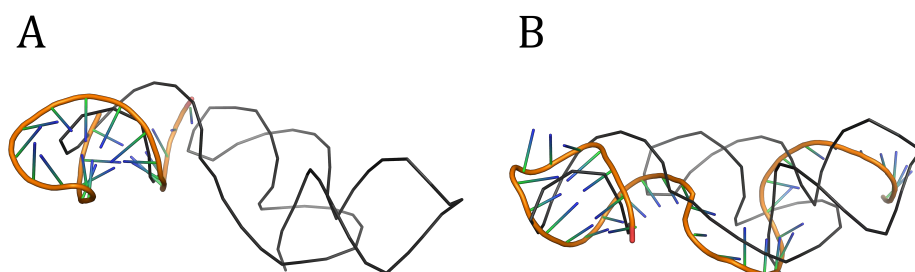


Figure 2.6: The longest continuous segments found by the LCS-TA algorithm in sequence-independent mode (the MCQ threshold value set at 20°) for (A) Chen 1st and (B) RNAComposer 1st models, aligned with the reference structure (PDB id: 5TPY) (black backbone). Adapted from Wiedemann et al. (2017).

wanted to investigate how different values of the MCQ threshold affect the length of the LCS. The results obtained showed us that when we consider models from the global perspective (MCQ threshold $> 20^\circ$) they seem to be similar even when local differences may be substantial and worth further investigation.

2.2. ALIGNMENT OF RNA 3D STRUCTURES IN TORSION ANGLE SPACE

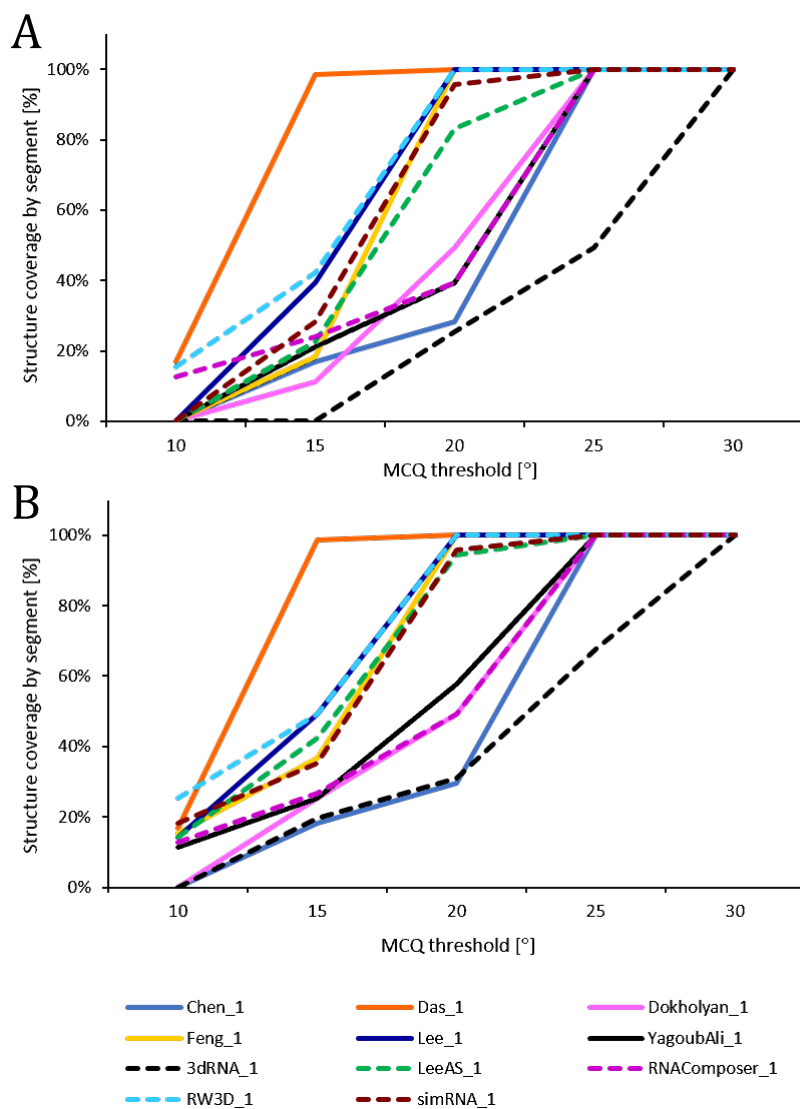


Figure 2.7: LCS-TA results for models submitted for the 18th challenge of RNA-Puzzles. The top graph (A) shows results for the sequence-dependent mode, and the bottom graph (B) shows results for the sequence-independent mode. Adapted from Wiedemann et al. (2017).

2.2. ALIGNMENT OF RNA 3D STRUCTURES IN TORSION ANGLE SPACE

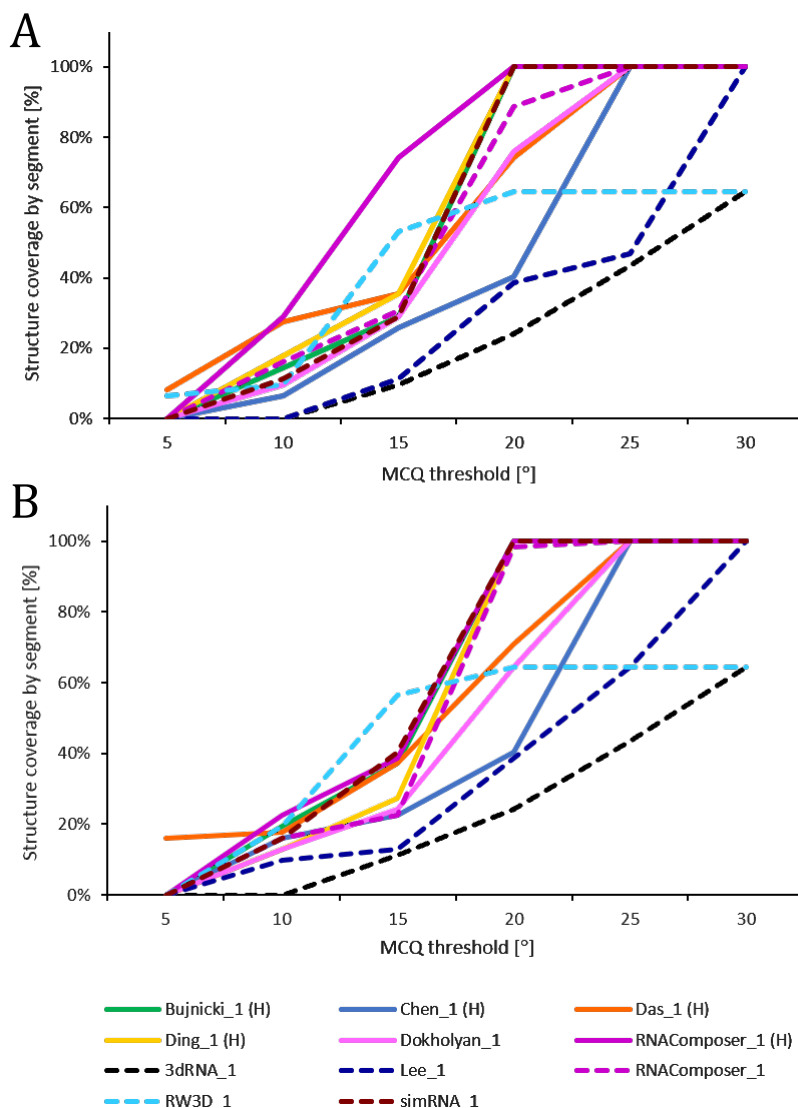


Figure 2.8: LCS-TA results for models submitted for the 19th challenge of RNA-Puzzles. The top graph (A) shows results for the sequence-dependent mode and the bottom graph (B) shows results for the sequence-independent mode. Adapted from Wiedemann et al. (2017).

2.2.3. Conclusions

As a result of this research, we have addressed the problem of aligning 3D RNA structures in the torsion angle space and evaluating the similarity of the tertiary structure from a local perspective. The *LCS-TA* method identifies similar fragments that show high similarity according to the MCQ measure. The method has been implemented in Java and was provided to

2.3. RNA MULTILoops EXTRACTION AND ANALYSIS

the community in the open source project named *MCQ4Structures*, freely available at <https://github.com/RNApolis/mcq4structures>. The *LCS-TA* method was included as part of the RNA-Puzzles toolkit [A3]. The RNA-Puzzles toolkit gathers well-established approaches to allow reliable evaluation of 3D RNA models within the context of the reference structure. This kind of approach aims to allow scientists interested in prediction and assessment of 3D structures of RNA to find relevant metrics and give a clear prediction assessment protocol for the RNA-Puzzles community. The RNA-Puzzles toolkit was also used later during the evaluation process in the RNA-Puzzles competition [A5]. Future work will provide the ability to analyse multiple structures with *LCS-TA* simultaneously or to automatically generate visualisations to support the researcher performing the evaluation.

2.3.

RNA multiloops extraction and analysis

This section contains results of analysing one of the RNA motifs – multiloops. The research concerns the identification, extraction, and analysis of multiloops, and resulted in the creation of the *RNAloops* database.

2.3.1. Background

[A4] summarises the study in which I focused on the analysis of RNA multiloops. Involvement in the summary of some challenges of the RNA puzzle [Miao et al. (2020)] [A3][A5] highlighted a few bottlenecks in the prediction process, i.e., particular RNA motifs, especially including noncanonical interactions, which usually decrease the accuracy of 3D predictions. One of

2.3. RNA MULTILoops EXTRACTION AND ANALYSIS

the structural motifs that most computational algorithms find difficult to reliably predict is the N -way junction. It, also known as a multiloop, can be defined as a set of single-stranded fragments that connect outgoing adjacent double-stranded regions (helices) and, therefore, significantly affect the spatial arrangement of the whole molecule.

2.3.2. Results

The research started with an overview of state-of-the-art resources that showed that there is no valid and up-to-date database that collects information about experimentally determined N -way junction 3D structures, along with their parameters and specialised visualisations. Thus, we developed *RNAloops* (<https://rnaloops.cs.put.poznan.pl>), a new repository that fills the gap. The *RNAloops* database collects information about N -way junctions, including, i.e., RNA sequence, secondary and tertiary structures, planar and Euler angles (Diebel, 2006). The Euler angles describe the relationship between the outgoing and adjacent helices. The proposed platform consists of the following layers: user-friendly interface, back-end providing RESTful API, database management layer, and fully automated repository update service. The user interface was developed using the React.js and Next.js frameworks. It presents the results retrieved through the RESTful API to the end user. The back-end layer is responsible for running all operations required during request handling. It also manages the database and the repository update service executions. The relational database used in *RNAloops* runs on PostgreSQL DBMS, which stores data describing all collected multiloops and allows users to query them comprehensively. An internal fully automated periodic update service is responsible for the weekly update of the repository of *RNAloops*.

2.3. RNA MULTILOOPS EXTRACTION AND ANALYSIS

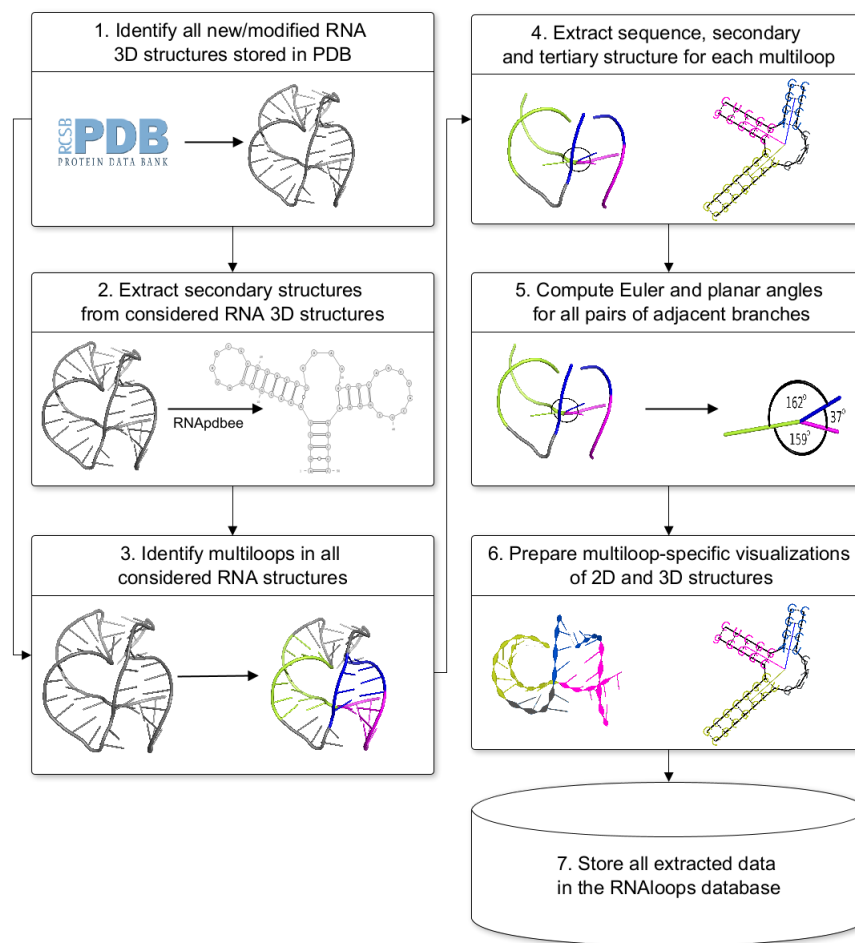


Figure 2.9: RNAloops’s workflow. Reprinted from Wiedemann et al. (2022).

The proposed repository is automatically updated once a week by a self-developed routine implemented in Python. The process starts by retrieving a list of entries from the PDB repository (Berman, 2000) that have been changed. In the next step, the system downloads all identified 3D structures of RNAs that are standardised prior to further processing. For every downloaded 3D RNA structure, its secondary structure is extracted in extended dot-bracket notation using the RNApdbee tool Zok et al. (2018). Next, all N -way junctions ($N > 2$) are identified for each RNA structure by a specialised scan of its secondary structure presented in (Procedure 1). Each extracted multiloop is saved into the *RNAloops* database presenting the features extracted from both the secondary and

2.3. RNA MULTILoops EXTRACTION AND ANALYSIS

the tertiary structures. Next, the planar and Euler angle values are computed for every pair of adjacent helices and provided within the description of the single-stranded region connecting them in the loop. The pipeline provided within the *RNAloops* platform is shown in Figure 2.9.

Procedure 1 A procedure to extract all N -way junctions from the RNA secondary structure.

Input: structure2D - RNA secondary structure in extended dot-bracket notation

Output: listOfJunctions[0...n-1] - list of identified multiloops

```

1: procedure IDENTIFYJUNCTIONS(structure2D)
2:   bpOpen = ['(', '[', '{', '<', 'A', 'B', 'C', 'D', 'E', 'F']
3:   bpClose = [')', ']', '}', '>', 'a', 'b', 'c', 'd', 'e', 'f']
4:   bps, bpsDict  $\leftarrow$  identifyBasePairs(structure2D, bpOpen, bpClose)
5:   outerBps  $\leftarrow$  identifyOuterBP(bps)
6:   for j in (0, len(outerBps)-1) do
7:     bp  $\leftarrow$  outerBps[j]
8:     bpo  $\leftarrow$  outerBps[j][0]
9:     bpc  $\leftarrow$  outerBps[j][1]
10:    complexity  $\leftarrow$  0 ▶ represents the loop topological complexity
11:    while bpc+1  $\leq$  outerBps[j][1] do
12:      openID  $\leftarrow$  -1
13:      closeID  $\leftarrow$  -1
14:      for k in (bpo, len(structure2D)-1) do
15:        for z in (0, len(bpOpen)-1) do
16:          if structure2D[k] = bpOpen[z] then
17:            openID  $\leftarrow$  z
18:            bpc  $\leftarrow$  k
19:          for z in (0, len(bpClose)-1) do
20:            if structure2D[k] = bpClose[z] then
21:              closeID  $\leftarrow$  z
22:              bpc  $\leftarrow$  k
23:          if (openID  $\geq$  0 and openID  $\leq$  bp[2])
24:             or (closeID  $\geq$  0 and closeID  $\leq$  bp[2]) then
25:             break
26:          if ((openID  $\geq$  0) or (closeID  $\geq$  0)) then
27:            if (bpc in bpsDict) and ((bpc+1)  $\leq$  outerBps[j][1]) then
28:              bpo  $\leftarrow$  bpc+1
29:              bpc  $\leftarrow$  bpsDict[bpc]+1
30:              if bpo < bpc then
31:                pair  $\leftarrow$  [bpo, bpc]
32:                bpo  $\leftarrow$  bpc
33:              else
34:                pair  $\leftarrow$  [bpc, bpo]
35:                bpc  $\leftarrow$  bpo
36:            pairs.append(pair)
37:    complexity  $\leftarrow$  complexity + 1

```


2.3. RNA MULTILoops EXTRACTION AND ANALYSIS

```

37:         else
38:             break
39:             bpo = bpsDict[bpc]+1
40:         if complexity  $\geq$  3 then
41:             listOfJunctions.append([complexity, pairs])
42:     return listOfJunctions
Input: structure2D - RNA secondary structure in extended dot-bracket notation
        bpOpen - list of opening chars in bp in extended dot-bracket notation
        bpClose - list of closing chars in bp in extended dot-bracket notation
Output: bpsDict - dictionary of identified base pairs
        bps - list of identified base pairs
43: procedure IDENTIFYBASEPAIRS(structure2D, bpOpen, bpClose)
44:     stacks  $\leftarrow$  []
45:     for i in (0, len(bpOpen) do
46:         stacks[i]  $\leftarrow$  []
47:     for i in (0, len(structure2D)-1) do
48:         openID  $\leftarrow$  -1
49:         for j in (0, len(bpOpen)-1) do
50:             if structure2D[i] = bpOpen[j] then
51:                 openID  $\leftarrow$  j
52:         closeID  $\leftarrow$  -1
53:         for j in (0, len(bpClose)-1) do
54:             if structure2D[i] = bpClose[j] then
55:                 closeID  $\leftarrow$  j
56:         if openID  $\geq$  0 then
57:             stacks[openID].append(i)
58:         else if closeID  $\geq$  0 then
59:             j  $\leftarrow$  stacks[closeID].pop()
60:             bps.append([j+1,i+1,closeID])
61:             bpsDict[key $\leftarrow$  j, value $\leftarrow$  i]
62:             bpsDict[key $\leftarrow$  i, value $\leftarrow$  j]
63:     bps.sort()  $\triangleright$  sorting bps according to first elements in sublists
64:     return bpsDict, bps
Input: bps - list of identified base pairs
Output: outerBps - list of outer base pairs
65: procedure IDENTIFYOUTERBP(bps)
66:     for i in (0, len(bps)-1) do
67:         if i = len(bps)-1 then
68:             outerBps.append(bps[i])
69:         else if i < len(bps)-1 then
70:             i1, j1, l1  $\leftarrow$  bps[i]
71:              $\triangleright$  i1 = bps[i][0], j1 = bps[i][1], l1 = bps[i][2]
72:             i2, j2, l2  $\leftarrow$  bps[i+1]
73:             if (not(i2=i1+1) or not(j2=j1-1)) and (l2=l1) then
74:                 outerBps.append(bps[i])
75:     return outerBps

```

Euler and planar angles allow one to describe the N -way junction topol-

2.3. RNA MULTILOOPS EXTRACTION AND ANALYSIS

ogy and mutual relations between adjacent helices. Calculation of angular features of multiloops begins with the construction of a *simplified junction model* determined within two stages:

1. *Determination of a geometric centre of the multiloop.* In the first step, the algorithm identifies all base pairs (i,j) directly adjacent to every single-stranded fragment that connects outgoing helices. The geometric centre of the particular loop is computed as a centroid of all nonhydrogen residue atoms included within these base pairs. Later, the centroid is used as a shared point for identifying the directional lines for each outgoing helix of the particular loop.
2. *Determination of directional line for every outgoing helix within the particular loop.* Helices are represented as directional lines constructed between geometric centres of both the loop and each considered helix, respectively. The geometric centre of the particular helix is calculated based on all non-hydrogen residue atoms for the first or third base pair for short (less than 2 bps) or longer helices, respectively.

Based on the aforementioned simplified multiloop representation, where the particular junction is described by directional lines of all included helices, the algorithm computes the following angular features:

1. *Planar angle value* is computed between two rays, sharing an initial point in the plane, with the following equation (Eqn. 2.1)

$$\alpha = \arccos[(\vec{a}\vec{b})/(|\vec{a}| * |\vec{b}|)] \quad (2.1)$$

where \vec{a} and \vec{b} are normal vectors to the two planes and α is an angle between these planes.

2.3. RNA MULTILOOPS EXTRACTION AND ANALYSIS

2. *Euler angle values.* A set of three angles that describe the orientation of an object in Euclidean space. The angles α , β , γ represent a rotation around the axes X, Y, Z, respectively (Heyde & Wood, 2020). The Euler angles are calculated first by projecting vectors (\vec{a}, \vec{b}) , sharing starting point on the planes perpendicular to all axes of the coordinate system and next computing the corresponding angle value between those vectors using the equation (Eqn. 2.1). The projection is based on finding the shortest path between the particular point and the plane.

2.3.3. Conclusions

RNAloops is a fully automated web-accessible repository that allows users to find information on N -way junctions observed in experimentally determined 3D RNA structures ($N > 2$) deposited in the Protein Data Bank [Burley et al. (2020)]. The data collected include the RNA sequence, the secondary and tertiary structure, and the planar and Euler angles. The latter describes the relationship between every pair of adjacent double-stranded regions (helices) of the multiloop integrated by the particular single-stranded region. A new representation of the relationship between outgoing helices allows for a comprehensive investigation of adjacent RNA domains. These types of data can be used to design 3D RNA structures that characterise the expected structural features. Furthermore, the tertiary structures of N -way junctions retrieved from the *RNAloops* database can be directly used as structural elements applied in semi-automated, expert modelling of RNA 3D models using, e.g., the RNAComposer system. Currently *RNAloops* stores information about circa 85k N -way junctions that were identified in nearly 1,900 experimentally determined 3D structures of RNA. This means that more than 30% of all RNA structures

2.3. RNA MULTILoops EXTRACTION AND ANALYSIS

deposited in the Protein Data Bank [Burley et al. (2020)] include at least one 3-way junction or more, which confirms their importance.

Future work for *RNAloops* should focus on further integration with other tools from the RNAPolis platform [Szachniuk (2019)]. Models and information about N -way junctions can be used directly in the RNAComposer system [Antczak et al. (2016)] to improve generated models.

Bibliography

Albertini, A. A. V., Schoehn, G., Weissenhorn, W., & Ruigrok, R. W. H. (2007). Structural aspects of rabies virus replication. *Cellular and Molecular Life Sciences*, 65(2), 282–294.

Altona, C. & Sundaralingam, M. (1972). Conformational analysis of the sugar ring in nucleosides and nucleotides. new description using the concept of pseudorotation. *Journal of the American Chemical Society*, 94(23), 8205–8212.

Antczak, M., Popena, M., Zok, T., Sarzynska, J., Ratajczak, T., Tomczyk, K., Adamiak, R. W., & Szachniuk, M. (2016). New functionality of RNAComposer: application to shape the axis of miR160 precursor structure. *Acta Biochim Pol*, 63(4), 737–744.

Berg, J. (2002). *Biochemistry*. New York: W.H. Freeman.

Berman, H. M. (2000). The protein data bank. *Nucleic Acids Research*, 28(1), 235–242.

Bernstein, F. C., Koetzle, T. F., Williams, G. J., Meyer, E. F., Brice, M. D., Rodgers, J. R., Kennard, O., Shimanouchi, T., & Tasumi, M. (1977). The protein data bank: A computer-based archival file for macromolecular structures. *Journal of Molecular Biology*, 112(3), 535–542.

BIBLIOGRAPHY

- Bourne, P. E., Berman, H. M., McMahon, B., Watenpaugh, K. D., Westbrook, J. D., & Fitzgerald, P. M. (1997). Macromolecular crystallographic information file. In *Methods Enzymol* (pp. 571–590). Elsevier.
- Burley, S. K., Bhikadiya, C., Bi, C., Bittrich, S., Chen, L., Crichlow, G. V., Christie, C. H., Dalenberg, K., Costanzo, L. D., Duarte, J. M., Dutta, S., Feng, Z., Ganesan, S., Goodsell, D. S., Ghosh, S., Green, R. K., Guranović, V., Guzenko, D., Hudson, B. P., Lawson, C. L., Liang, Y., Lowe, R., Namkoong, H., Peisach, E., Persikova, I., Randle, C., Rose, A., Rose, Y., Sali, A., Segura, J., Sekharan, M., Shao, C., Tao, Y.-P., Voigt, M., Westbrook, J. D., Young, J. Y., Zardecki, C., & Zhuravleva, M. (2020). RCSB Protein Data Bank: powerful new tools for exploring 3D structures of biological macromolecules for basic and applied research and education in fundamental biology, biomedicine, biotechnology, bioengineering and energy sciences. *Nucleic Acids Research*, 49(D1), D437–D451.
- Cruz, J. A., Blanchet, M.-F., Boniecki, M., Bujnicki, J. M., Chen, S.-J., Cao, S., Das, R., Ding, F., Dokholyan, N. V., Flores, S. C., Huang, L., Lavender, C. A., Lisi, V., Major, F., Mikolajczak, K., Patel, D. J., Philips, A., Puton, T., Santalucia, J., Sijenyi, F., Hermann, T., Rother, K., Rother, M., Serganov, A., Skorupski, M., Soltysinski, T., Sripakdeevong, P., Tuszynska, I., Weeks, K. M., Waldsich, C., Wildauer, M., Leontis, N. B., & Westhof, E. (2012). RNA-puzzles: A CASP-like evaluation of RNA three-dimensional structure prediction. *RNA*, 18(4), 610–625.
- Darty, K., Denise, A., & Ponty, Y. (2009). VARNA: Interactive drawing and editing of the RNA secondary structure. *Bioinformatics*, 25(15), 1974–1975.

BIBLIOGRAPHY

- Date, C. (2006). *An Introduction to Database Systems*. Pearson Education India.
- Davis, I. W., Leaver-Fay, A., Chen, V. B., Block, J. N., Kapral, G. J., Wang, X., Murray, L. W., Arendall, W. B., Snoeyink, J., Richardson, J. S., & Richardson, D. C. (2007). MolProbity: all-atom contacts and structure validation for proteins and nucleic acids. *Nucleic Acids Research*, 35(Web Server), W375–W383.
- Dawson, W. K., Maciejczyk, M., Jankowska, E. J., & Bujnicki, J. M. (2016). Coarse-grained modeling of RNA 3d structure. *Methods*, 103, 138–156.
- Duarte, C. M. & Pyle, A. M. (1998). Stepping through an RNA structure: a novel approach to conformational analysis. *Journal of Molecular Biology*, 284(5), 1465–1478.
- Eads, J. C., Scapin, G., Xu, Y., Grubmeyer, C., & Sacchettini, J. C. (1994). The crystal structure of human hypoxanthine-guanine phosphoribosyl-transferase with bound GMP. *Cell*, 78(2), 325–334.
- Edgar, R. C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research*, 32(5), 1792–1797.
- Eric, W. & Pascal, A. (2006). *RNA Tertiary Structure*, chapter Nucleic Acids Structure and Mapping. John Wiley & Sons, Ltd.
- Frellsen, J., Moltke, I., Thiim, M., Mardia, K. V., Ferkinghoff-Borg, J., & Hamelryck, T. (2009). A probabilistic model of RNA conformational space. *PLoS Computational Biology*, 5(6), e1000406.
- Gauthier, J., Vincent, A. T., Charette, S. J., & Derome, N. (2018). A brief history of bioinformatics. *Briefings in Bioinformatics*, 20(6), 1981–1996.

BIBLIOGRAPHY

- Gong, W. & Fan, X.-Q. (2019). A geometric characterization of DNA sequence. *Physica A: Statistical Mechanics and its Applications*, 527, 121429.
- Hagen, J. B. (2000). The origins of bioinformatics. *Nature Reviews Genetics*, 1(3), 231–236.
- Halder, S. & Bhattacharyya, D. (2013). RNA structure and dynamics: A base pairing perspective. *Progress in Biophysics and Molecular Biology*, 113(2), 264–283.
- Heyde, K. & Wood, J. L. (2020). Representation of rotations, angular momentum and spin. In *Quantum Mechanics for Nuclear Structure, Volume 2*, 2053-2563 (pp. 1–1 to 1–46). IOP Publishing.
- Hoehndorf, R., Batchelor, C., Bittner, T., Dumontier, M., Eilbeck, K., Knight, R., Mungall, C. J., Richardson, J. S., Stombaugh, J., Westhof, E., & et al. (2011). The RNA Ontology (RNAO): An ontology for integrating RNA sequence and structure data. *Applied Ontology*, 6(1), 53–89.
- Hu, B., Guo, H., Zhou, P., & Shi, Z.-L. (2020). Characteristics of SARS-CoV-2 and COVID-19. *Nature Reviews Microbiology*, 19(3), 141–154.
- Kabsch, W. (1976). A solution for the best rotation to relate two sets of vectors. *Acta Crystallographica Section A*, 32(5), 922–923.
- Kitamura, N., Semler, B. L., Rothberg, P. G., Larsen, G. R., Adler, C. J., Dorner, A. J., Emini, E. A., Hanecak, R., Lee, J. J., van der Werf, S., Anderson, C. W., & Wimmer, E. (1981). Primary structure, gene organization and polypeptide expression of poliovirus RNA. *Nature*, 291(5816), 547–553.

BIBLIOGRAPHY

- Leontis, N. & Westhof, E. (2012). Modeling RNA Molecules. In *Nucleic Acids and Molecular Biology* (pp. 5–17). Springer Berlin Heidelberg.
- Leontis, N. B. & Westhof, E. (2001). Geometric nomenclature and classification of RNA base pairs. *RNA*, 7(4), 499–512.
- Lipman, D. J. & Pearson, W. R. (1985). Rapid and sensitive protein similarity searches. *Science*, 227(4693), 1435–1441.
- Miao, Z., Adamiak, R. W., Antczak, M., Boniecki, M. J., Bujnicki, J., Chen, S.-J., Cheng, C. Y., Cheng, Y., Chou, F.-C., Das, R., Dokholyan, N. V., Ding, F., Geniesse, C., Jiang, Y., Joshi, A., Krokhotin, A., Magnus, M., Mailhot, O., Major, F., Mann, T. H., Piątkowski, P., Pluta, R., Popenda, M., Sarzynska, J., Sun, L., Szachniuk, M., Tian, S., Wang, J., Wang, J., Watkins, A. M., Wiedemann, J., Xiao, Y., Xu, X., Yesselman, J. D., Zhang, D., Zhang, Y., Zhang, Z., Zhao, C., Zhao, P., Zhou, Y., Zok, T., Żyła, A., Ren, A., Batey, R. T., Golden, B. L., Huang, L., Lilley, D. M., Liu, Y., Patel, D. J., & Westhof, E. (2020). RNA-Puzzles Round IV: 3D structure predictions of four ribozymes and two aptamers. *RNA*, 26(8), 982–995.
- Needleman, S. B. & Wunsch, C. D. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology*, 48(3), 443–453.
- Ouzounis, C. A. & Valencia, A. (2003). Early bioinformatics: the birth of a discipline—a personal view. *Bioinformatics*, 19(17), 2176–2190.
- Parisien, M., Cruz, J. A., Westhof, É., & Major, F. (2009). New metrics for comparing and assessing discrepancies between RNA 3d structures and models. *RNA*, 15(10), 1875–1885.
- Pereira, J., Simpkin, A. J., Hartmann, M. D., Rigden, D. J., Keegan, R. M.,

BIBLIOGRAPHY

- & Lupas, A. N. (2021). High-accuracy protein structure prediction in casp14. *Proteins: Structure, Function, and Bioinformatics*, 89(12), 1687–1699.
- Ponty, Y. & Leclerc, F. (2014). Drawing and Editing the Secondary Structure(s) of RNA. In *Methods in Molecular Biology* (pp. 63–100). Springer New York.
- Pruitt, K. D., Tatusova, T., Brown, G. R., & Maglott, D. R. (2011). NCBI reference sequences (RefSeq): current status, new features and genome annotation policy. *Nucleic Acids Research*, 40(D1), D130–D135.
- Rabiner, L. (1984). Combinatorial optimization: algorithms and complexity. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 32(6), 1258–1259.
- Ramachandran, G., Ramakrishnan, C., & Sasisekharan, V. (1963). Stereochemistry of polypeptide chain configurations. *Journal of Molecular Biology*, 7(1), 95–99.
- Richardson, J. S., Schneider, B., Murray, L. W., Kapral, G. J., Immormino, R. M., Headd, J. J., Richardson, D. C., Ham, D., Hershkovits, E., Williams, L. D., Keating, K. S., Pyle, A. M., Micallef, D., Westbrook, J., & Berman, H. M. (2008). RNA backbone: Consensus all-angle conformers and modular string nomenclature (an RNA ontology consortium contribution). *RNA*, 14(3), 465–481.
- Rivest, R. L., Cormen, T. H. D. C., Leiserson, C. E. M., & Stein, C. C. U. (2009). *Introduction to Algorithms*. MIT Press Ltd.
- Ryu, M. W., Oh, S. M., Kim, M. J., Cho, H. H., Son, C. B., & Kim, T. H. (2020). Algorithm for generating 3D geometric representation based on indoor point cloud data. *Applied Sciences*, 10(22), 8073.

BIBLIOGRAPHY

- Saenger, W. (2013). *Principles of Nucleic Acid Structure*. Springer US.
- Schrödinger, LLC (2015). The PyMOL molecular graphics system, version 1.8.
- Šponer, J. E., Špačková, N., Leszczynski, J., & Šponer, J. (2005). Principles of RNA Base Pairing: Structures and Energies of the Trans Watson-Crick/Sugar Edge Base Pairs. *The Journal of Physical Chemistry B*, 109(22), 11399–11410.
- Swinnen, B., Robberecht, W., & Bosch, L. V. D. (2019). RNA toxicity in non-coding repeat expansion disorders. *The EMBO Journal*, 39(1).
- Szachniuk, M. (2019). RNAPolis: Computational platform for RNA structure analysis. *Foundations of Computing and Decision Sciences*, 44(2), 241–257.
- Townshend, R. J. L., Eismann, S., Watkins, A. M., Rangan, R., Karelina, M., Das, R., & Dror, R. O. (2021). Geometric deep learning of RNA structure. *Science*, 373(6558), 1047–1051.
- Varani, G. & McClain, W. H. (2000). The g•u wobble base pair. *EMBO reports*, 1(1), 18–23.
- Watson, J. D. & Crick, F. H. C. (1974). Molecular structure of nucleic acids: a structure for Deoxyribose Nucleic Acid. *Nature*, 248(5451), 765–765.
- Wiedemann, J., Kaczor, J., Miłostan, M., Zok, T., Blazewicz, J., Szachniuk, M., & Antczak, M. (2022). RNALoops: a database of RNA multiloops. *Bioinformatics*.
- Wiedemann, J. & Miłostan, M. (2017). StructAnalyzer - a tool for sequence vs. structure similarity analysis. *Acta Biochimica Polonica*, 63(4), 753—757.

- Wiedemann, J., Zok, T., Milostan, M., & Szachniuk, M. (2017). LCS-TA to identify similar fragments in RNA 3d structures. *BMC Bioinformatics*, 18(1).
- Zemla, A. (2003). LGA: a method for finding 3d similarities in protein structures. *Nucleic Acids Research*, 31(13), 3370–3374.
- Zemla, A. T., Česlovas Venclovas, Moult, J., & Fidelis, K. (1999). Processing and analysis of casp3 protein structure predictions. *Proteins: Structure*, 37.
- Zemora, G. & Waldsich, C. (2010). RNA folding in living cells. *RNA Biology*, 7(6), 634–641.
- Zok, T., Antczak, M., Zurkowski, M., Popena, M., Blazewicz, J., Adamiak, R. W., & Szachniuk, M. (2018). RNApdbee 2.0: multifunctional tool for RNA structure annotation. *Nucleic Acids Res*, 46(W1), W30–W35.
- Zok, T., Popena, M., & Szachniuk, M. (2013). MCQ4Structures to compute similarity of molecule structures. *Central European Journal of Operations Research*, 22(3), 457–473.

StructAnalyzer – a tool for sequence *versus* structure similarity analysis

Jakub Wiedemann^{1*} and Maciej Miłostan^{1,2*✉}

¹Institute of Computing Science, Poznan University of Technology, Poznań, Poland; ²Institute of Bioorganic Chemistry, Polish Academy of Sciences, Poznań, Poland

In the world of RNAs and proteins, similarities at the level of primary structures of two comparable molecules usually correspond to structural similarities at the tertiary level. In other words, measures of sequence and structure similarities are in general correlated – a high value of sequence similarity imposes a high value of structural similarity. However, important exceptions that stay in contrast to this general rule can be identified. It is possible to find similar structures with very different sequences, as well as similar sequences with very different structures. In this paper, we focus our attention on the latter case and propose a tool, called StructAnalyzer, supporting analysis of relations between the sequence and structure similarities. Recognition of tertiary structure diversity of molecules with very similar primary structures may be the key for better understanding of mechanisms influencing folding of RNAs or proteins, and as a result for better understanding of their function. StructAnalyzer allows exploration and visualization of structural diversity in relation to sequence similarity. We show how this tool can be used to screen RNA structures in Protein Data Bank (PDB) for sequences with structural variants.

Key words: sequence similarity, structural similarity, RNA

Received: 30 May, 2016; **revised:** 28 June, 2016; **accepted:** 19 July, 2016; **available on-line:** 02 November, 2016

INTRODUCTION

Despite technological progress in laboratory pipelines, computing methods and computational facilities, determination of three dimensional structures of RNAs and proteins in-situ or in-silico is not a trivial task (cf. Lukasiak *et al.*, 2010). Comparison of deposition statistics between Protein Data Bank (PDB) (Berman *et al.*, 2000) and NCBI's RefSeq (Pruitt *et al.*, 2012), shows how large is the gap between the known sequences and structures. In-silico methods attempt to reduce this gap but as the RNA-Puzzles (Miao *et al.*, 2015) competition has shown, they are still far from being perfect.

Nowadays, the most successful structure prediction methods are often somehow based on correlations between the sequence and structure similarities, for example they are transformed in the form of libraries of fragments like in the RNA Composer (Popenda *et al.*, 2012) and FARNA (Cheng *et al.*, 2015; Das & Baker, 2007). It is a known fact that the similarity in structure (cf. Zok *et al.*, 2014) for information about structural similarities) of molecules, like proteins or RNAs, highly correlates with sequence similarities, under assumption that all of the

structures compared were obtained under similar conditions. Similar conditions are important from the perspective of thermodynamics – changes in conditions are the driving force of folding and unfolding. However, in practice it is not feasible to impose the same conditions for all molecules in the process of structure determination because of various factors, e.g. physiological conditions of molecular activity and stability. Let us stress that the RNA structures, in comparison to proteins, are more flexible and less thermodynamically stable due to a larger number of degrees of freedom (Rother *et al.*, 2011) (e.g. torsional angles in the backbone). Thus, we can assume that even small changes in the environment may cause a substantial change in the RNA conformation. The intriguing question is: *how structurally diverse are similar RNA sequences whose structures are deposited in PDB?*

Thus, the primary aim of our work is to provide a tool, called StructAnalyzer, that allows us to explore and visualize structural diversity in relation to sequence similarity for RNAs and proteins. In contrast to other similar tools, like RNAnalyzer (Lukasiak *et al.*, 2013) or RNAAssess (Lukasiak *et al.*, 2015), our aim is not to assess quality of the model versus the reference structure, but rather the analysis of structural diversity of the real structures determined by biochemical experiments (e.g. crystallography or NMR). This exploration should allow to identify twilight zones where the high sequence similarity does not impose structural identity. It is worth to note that, purposely, we would like to analyze only sets of highly similar sequences (90–100% of pairwise similarity). We do not want to construct a minimal library of structural fragments that covers as large area of the sequence space as possible (in such a case, it is common to keep the sequence similarity to below some level). We would like to support identification and visualization of structural variants of almost identical sequences. We assume that within clusters obtained by grouping molecules by sequence similarity, it should be possible to find diverse structures. Moreover, within these structures it should be possible to identify fragments with a relatively high and low stability. It is worth to note that some of the structures stored in PDB were obtained as complexes or in the presence of metal ions or with ligands and immersed in different chemical solutions. Interactions between proteins, RNAs, ions and ligands may lead to substantial structural changes. Experiments with proteins (Alexander *et al.*, 2009) showed that sometimes even a point mutation, or a small set of point mutations, in the sequence

✉ e-mail: Maciej.Milostan@cs.put.poznan.pl

*Contributed equally to this work

Abbreviations: PDB, Protein Data Bank

Table 1. Molecules' PDB IDs and general description

PDB ID	Description from PDB database
2O3X	Crystal structure of the prokaryotic ribosomal decoding site complexed with paromamine derivative NB30
3BNT	Crystal structure of the homo sapiens mitochondrial ribosomal decoding site in the presence of [CO(NH ₃) ₆]CL3 (A1555G mutant, BR-derivative)
1FYO	Eukaryotic decoding region A-site RNA

can switch the structure into a totally different structural fold. We believe that such cases also exist in RNAs and our tool may help to identify them.

MATERIALS AND METHODS

Data sources. We show features and test performance of StructAnalyzer on 3 datasets. First set consists of two proteins differing in single amino acid and originating from the paper by Alexander and coworkers (2009). The last two sets contain only RNAs and were generated by the following approach.

In the first step, we generated pairwise sequence alignments for all possible pairs of RNA structures deposited in PDB and computed a matrix of relevant similarity scores. For that purpose, we used the MUSCLE software (Edgar, 2004; <http://www.drive5.com/muscle/>) which accepts FASTA files as input. Be aware of the fact that the FASTA sequences stored in the PDB database sometimes differ from the sequences contained in the structure files (in particular if we consider a specific chain). Thus, we extracted sequences of the RNA molecules directly from the files containing structures (*.pdb) by means of a self-written Python script. This script generates one FASTA sequence file for each chain of molecules stored in a particular pdb file.

In the second step, based on the above mentioned matrix of sequence similarities, we constructed two sets of molecules. The first one contains all the pairs of structures having 100% sequence similarity and the second one consists of all the pairs of structures with sequence similarity over or equal to 90%, but less than 100%. Relations between pairs of molecules from each set had been depicted in the form of a graph (see Supplementary Data for details). Molecules are denoted by vertices which are labelled using relevant PDB IDs. Edges connect the molecules (denoted by vertices) with a similarity score over the defined cut off. It is worth noting that in case of both datasets we obtained graphs containing disjoint subgraphs.

The results of procedures described above for both sets are presented in the Supplementary Data (at www.actabp.pl). From the first set, containing pairs of sequences with 100% sequence similarity, the algorithm created 383 subgraphs. For the second set (sequences with similarity above 90%, but less than 100%) we obtained 93 subgraphs. From both sets we chose one subgraph to show features of the presented tool.

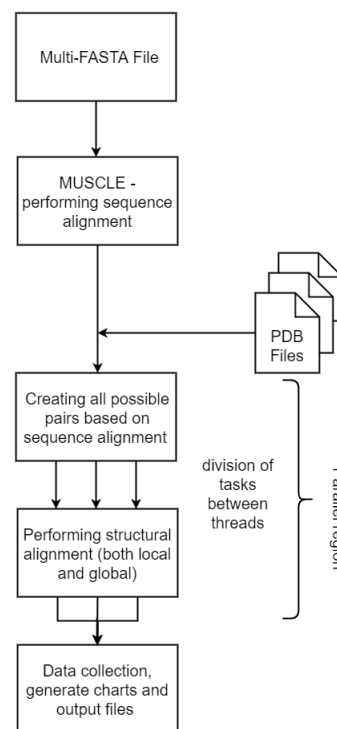
Algorithm description. The tool presented here allows to perform both, one-to-many and many-to-many sequence and structure comparisons. Our program uses PDB and Multi-FASTA files as input. On the basis of the data obtained and computational analysis, StructAnalyzer generates graphical interpretation of the results. The general workflow of StructAnalyzer is shown in Fig. 1.

In the first stage, our algorithm generates sequence alignment using the MUSCLE software. This alignment

is the basis for further analysis. We can distinguish two general modes of comparisons: many-to-one (one sequence is treated as the reference one) and many-to-many. Results of each mode are visualized in a different manner.

In both cases (many-to-one and many-to-many) the algorithm selects corresponding fragments of sequences based on the sequence alignment. Selected fragments are aligned with corresponding fragments of the reference structure and the algorithm calculates their structural similarity. RMSD is used as a measure of structural similarity. The program also allows merging of spatially neighbouring fragments into larger entities to increase the number of atoms used to perform the structural comparisons. To do this, the algorithm searches the spatial neighbourhood of each of the atoms of the previously obtained fragments. The scope of the spatial neighbourhood is restricted by the user defined radius (in Angstroms). The identified neighbours are added to the base fragment. Fragments extended by the added atoms are aligned and similarity of their structures is calculated.

In case of pairwise comparison, besides the previously described function, StructAnalyzer allows to perform a comparison of all fragments with a predetermined length of one molecule, to fragments (with the same length) of another molecule. The predetermined length is further referenced as the frame.

**Figure 1. StructAnalyzer workflow.**

The scheme presents the most important steps of the analysis performed by StructAnalyzer.

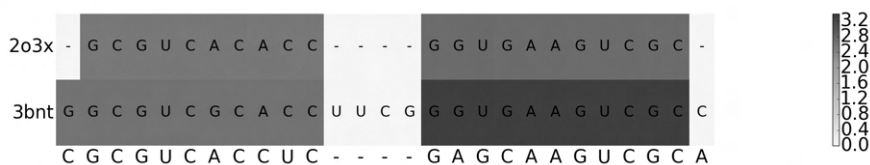


Figure 2. Example of a linear alignment for 1fy0 molecule against other structures (Table 1) in the set of molecules with sequence similarity 70%.

RESULTS AND THEIR REPRESENTATION

StructAnalyzer allows the user to save their results to a .csv file but its undeniable advantage is an ability to visualize them. For global alignment of structures, our tool presents the results using heat maps and a linear diagram (see Fig. 2). The linear diagram is particularly useful for comparing a sequence with low similarity or discontinuous fragments. In local alignment, we need to consider two cases. The first one is when dealing with large gaps in the alignment is necessary. The results can be visualized as a linear alignment with scores for each fragment determined individually. The second considered case is a situation when the local alignment is determined using a previously described frame. In this case, the results are shown on a heat map.

DISCUSSION

In order to show the capability of our tool, we conducted analysis for the three previously described sets. For each set, the StructAnalyzer determined the RMSD value for all molecules by performing both, a local and global alignment.

The first set consists of two protein structures differing in a single amino acid: 2KDM and 2KDL (Fig. 3). For molecules with such high sequence similarity, the results are surprising. The heat map (Fig. 4) for the global comparison shows the RMSD value is above 12 Angstroms. If we consider local comparison of this structures (Fig. 5), we can see some resemblance at the diagonal (or regions close to the diagonal) of the heat map. As we can easily deduce, high similarity scores at the diagonal indicate the identity of local structures for alignment under consideration, while similarities at the regions surrounding the diagonal can signal potential mismatches in the proposed alignment. The case under consideration shows that even a point mutation can influence the structure and function to a large extent. From the perspective of function, it is worth to stress that both

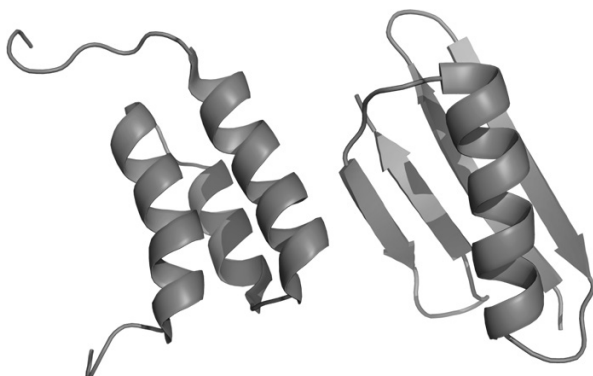


Figure 3. 2KDL (left) and 2KDM (right) spatial structures.

proteins have affinity to bind different molecules (see Table 2).

The second set consists of six RNA structures. As shown in the heat map (Fig. 6), the RMSD values within the set range from 0.5 to about 4 Angstroms. The results are quite unexpected considering the sequence similarity in the presented collection (equal to 100%). It is easy to spot, in that case, how large influence on the structure of the RNA the environmental conditions have. In order to demonstrate factors affecting the development of the analysed molecules, a brief description (extracted from PDB) of the structures has been gathered in Table 3. Despite large differences at the level of global alignment (see Fig. 6), it is worth to take a look at the differences at the local level. The results of the local alignment are presented in heat maps (see Figs. 7 and 8), generated for the structure pairs 2CD3 and 2CD6 (with the frame sizes equal to 5; Fig. 7; and 7; Fig. 8). As we can see, at the local alignment level there are many fragments with either good or very bad RMSD values. Based on these results we can deduce that despite big differences between the molecules observed from a global perspective, when we consider the local perspective, e.g. smaller fragments of structures, we can find many similarities. These similar fragments in globally different structures can stand for conservative regions which are characterized by low volatility and may determine similar functions of the considered molecules. On the other hand, sequence fragments which in many structures are characterized by a significant diversity, may designate potentially disordered regions. Another example of local comparison is presented in a heat map (Fig. 9) for the 1F7G and 1F7I structures (Fig. 10). In

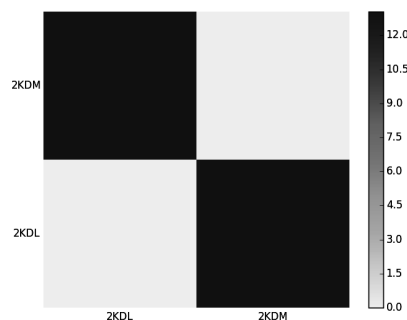


Figure 4. Heat map for global comparison of 2KDL and 2KDM structures. The value of RMSD determines the colour.

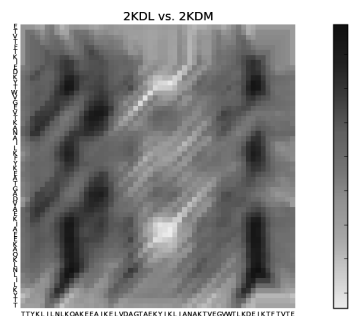


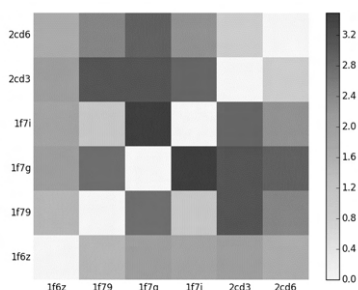
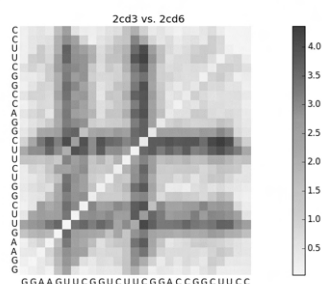
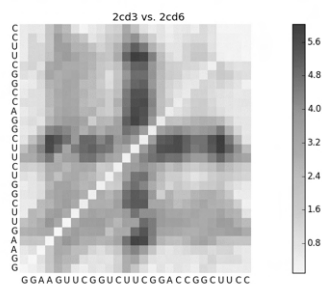
Figure 5. Heat map for structures (PDB IDs = 2KDL, 2KDM) differing by only one amino acid. The heat map was generated by using a frame size equal to 15.

Table 2. Molecules' PDB IDs, general description and classification

PDB ID	Description from PDB database	Classification
2KDM	NMR structures of GA95 AND GB95, two designed proteins with 95% sequence identity but different folds and functions	IGG binding protein
2KDL	NMR structures of GA95 AND GB95, two designed proteins with 95% sequence identity but different folds and functions	Human serum albumin binding protein

Table 3. Molecules' PDB IDs and general description

PDB ID	Description from PDB database
1F6Z	Solution structure of the RNase P RNA (M1 RNA) P4 stem C70U mutant oligoribonucleotide
1F7I	Solution structure of the RNase P RNA (M1 RNA) P4 stem C70U mutant oligoribonucleotide complexed with cobalt (III) hexamine, NMR, ensemble of 12 structures
1F7G	Solution structure of the RNase P RNA (M1 RNA) P4 stem C70U mutant oligoribonucleotide, ensemble of 17 structures
1F79	Solution structure of RNase P RNA (M1 RNA) P4 stem C70U mutant oligoribonucleotide complexed with cobalt (III) hexamine, NMR, minimized average structure
2CD3	Refinement of RNase P P4 stemloop structure using residual dipolar coupling data – C70U mutant
2CD6	Refinement of RNase P P4 stemloop structure using residual dipolar coupling data, C70U mutant cobalt (III) hexamine complex

**Figure 6. Heat map for all molecules against each other. The value of RMSD determines the colour.****Figure 7. Heat map for structures (PDB IDs = 2CD3, 2CD6) with the same sequence. The heat map was generated by using a frame size equal to 5.****Figure 8. Heat map for structures (PDB IDs = 2CD3, 2CD6) with the same sequence. The heat map was generated by using a frame size equal to 7.**

this case, at the diagonal (and near the diagonal) of the heat map we can see similar and dissimilar fragments. Dissimilarities can be the result of a metal (cobalt) presence during the structure determination process of the 1F7G molecule. This case shows how the environment can influence folding of the structure and also how even small structural differences at the local level can change overall fold of the analysed molecule.

The third set contains 3 RNA structures. Sequence similarity between molecules that were at the edges of the sub-graph containing the analysed structures are shown in Table 4. As in the previous example, despite high sequence similarity we can observe significant structural differences between all molecules (see Fig. 11). From the analysis of local comparison we can see a huge range of RMSD, from 0 to almost 6 Angstroms. It is worth noting the obvious fact that in the case of the analysed molecules, the best RMSD values for the fragments compared are most frequently located at the diagonal of the presented heat map (see Fig. 12). Consideration of values outside of the diagonal may be useful in detection of misalignments and when we look for reoccurring local spatial motifs between the analysed molecules – e.g. larger or smaller affinity for the sequence to adopt some spatial structure.

CONCLUSIONS

StructAnalyzer is a new, promising tool for structural analysis of RNA and proteins. This tool is still under active development, and thus new features will be incorporated shortly; this tool will be available for the general

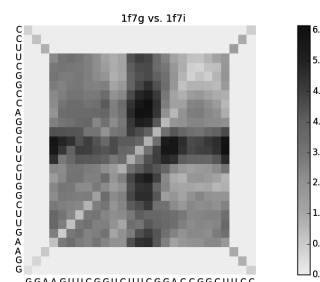
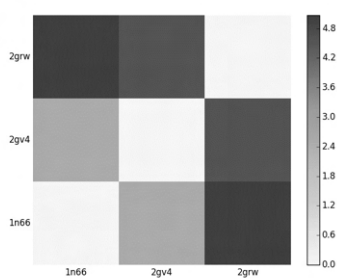
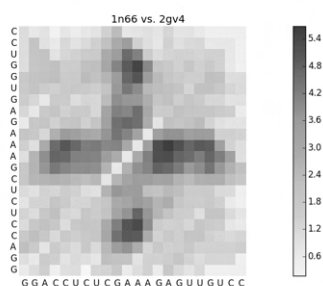
**Figure 9. Heat map for structures (PDB IDs = 1F7G, 1F7I) with the same sequence. The heat map was generated by using a frame size equal to 7.**

Table 4. Sequence similarity between molecules in the analyzed set

PDB ID:CHAIN ID	PDB ID 2:CHAIN ID	Sequence similarity
1N66:A	2GRW:A	0.95
1N66:A	2GV4:A	0.95
2GRW:A	2GV4:A	0.95
PDB ID:CHAIN ID	Description from PDB database	
1N66:A	Structure of the pyrimidine-rich internal loop in the Y-domain of poliovirus 3'-UTR	
2GRW:A	Solution structure of the poliovirus 3'-UTR Y-stem	
2GV4:A	Solution structure of the poliovirus 3'-UTR Y-stem	

**Figure 10. Superposition of spatial structures of 1F7G (black) and 1F7I (grey).****Figure 11. Heat map for all molecules against each other. The value of RMSD determines the colour.****Figure 12. Heat map for structures (PDB IDs = 2F88, 2LPT) with sequence similarity equal to 95%. The heat map was generated by using a frame size equal to 5.**

public (structanalyzer.cs.put.poznan.pl). In the current release it can perform both, global and local structure comparisons on the basis of sequence alignment and visualize the obtained results in an attractive manner. The presented approach enables to examine, for example, how different conditions or sequence differences af-

fect development of the structures. Moreover, a visual representation of the results makes them much easier to interpretate. Global comparison of structures shows us, in general, if there are any differences. In large sets of structures it allows us to screen through the whole set, so there is no need to examine the structures one by one, and it immediately indicates where and how big these structural differences are. After global comparison, we can decide for which structures we want to run the comparison locally or we can terminate the job. Results of local comparison provide us information about influence of the local differences, like point mutations or deletions, on the global shape of the molecule. Those results also allow identification of potential conservative or disordered regions. Another important feature of the tool presented here, is support for parallel processing which significantly reduces the duration of analysis.

Acknowledgements

Supported by the Polish Ministry of Science and Higher Education, under the KNOW program.

REFERENCES

- Alexander PA, He Y, Chen Y, Orban J, Bryan PN (2009) A minimal sequence code for switching protein structure and function. *Proc Natl Acad Sci USA* **106**: 21149–21154. doi: 10.1073/pnas.0906408106
- Berman HM, Westbrook J, Feng Z., Gilliland G., Bhat TN, Weissig H, Shindyalov IN, Bourne PE (2000) The Protein Data Bank. *Nucleic Acids Res* **28**: 235–242. doi: 10.1093/nar/28.1.235
- Cheng CY, Chou F-C, Das R, (2015) Chapter Two – modeling complex RNA tertiary folds with Rosetta. In *Methods in Enzymology*, Chen S-J, Burke-Aguero DH, eds, **55**: 35–64. Academic Press. doi:10.1016/bs.mie.2014.10.051
- Das R, Baker D (2007) Automated de novo prediction of native-like RNA tertiary structures. *Proc Natl Acad Sci USA* **104**: 14664–14669. doi:10.1073/pnas.0703836104
- Edgar RC (2004a) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* **32**: 1792–1797. Print 2004. PubMed PMID: 15034147. doi: 10.1093/nar/gkh340
- Edgar RC (2004b) MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics* **5**: 113. PubMed PMID: 15318951. doi: 10.1186/1471-2105-5-113
- Lukasiak P, Blazewicz J, Milostan M (2010) Some operations research methods for analyzing protein sequences and structures. *Annals Operations Res* **175**: 9–35
- Lukasiak P, Antczak M, Ratajczak T, Szachniuk M, Popenda M, Adamiak RW, Blazewicz J (2015) RNAssess – a webserver for quality assessment of RNA 3D structures. *Nucleic Acids Res* **43**: W502–W506. doi:10.1093/nar/gkv557
- Lukasiak P, Antczak M, Ratajczak T, Bujnicki JM, Szachniuk M, Popenda M, Adamiak RW, Blazewicz J (2013) RNAnalyzer – novel approach for quality analysis of RNA structural models. *Nucleic Acids Res* **41**: 5978–5990. doi:10.1093/nar/gkt318
- Miao Z, Adamiak RW, Blanchet M-F, Boniecki M, Bujnicki JM, Chen S-J, Cheng C, Chojnowski G, Chou F-C, Cordero P, Cruz JA, Ferre-D'Amare A, Das R, Ding F, Dokholyan NV, Dunin-Horkawicz S, Kladwang W, Krokhotin A, Lach G, Magnus M, Major F, Mann TH, Masquida B, Matelska D, Meyer M, Peselis A, Popenda M, Purzycka KJ, Serganov A, Stasiewicz J, Szachniuk M, Tandon A, Tian S, Wang J, Xiao Y, Xu X, Zhang J, Zhao P, Zok T, Westhof E (2015) RNA-puzzles round II: Assessment of RNA structure prediction programs applied to three large RNA structures. *RNA* **21**: 1–19. doi:10.1261/rna.049502.114
- Popenda M, Szachniuk M, Antczak M, Purzycka KJ, Lukasiak P, Bartol N, Blazewicz J, Adamiak RW, (2012) Automated 3D structure composition for large RNAs. *Nucleic Acids Res* **40**: e112. doi:10.1093/nar/gks339
- Pruitt KD, Tatusova T, Brown GR, Maglott DR (2012) NCBI Reference Sequences (RefSeq): current status, new features and genome annotation policy. *Nucleic Acids Res* **40**: D130–D135. PMID: 22121212. PMCID: PMC3245008. doi:10.1093/nar/gkr1079
- Rother K, Rother M, Boniecki M, Puton T, Bujnicki JM (2011) RNA and protein 3D structure modeling: similarities and differences. *J Mol Model* **17**: 2325–2336. doi:10.1007/s00894-010-0951-x
- Zok T, Popenda M, Szachniuk M (2014) MCQ4Structures to compute similarity of molecule structures. *Central Eur J Operations Res* **22**: 457–474. doi:10.1007/s10100-013-0296-5

RESEARCH ARTICLE

Open Access

LCS-TA to identify similar fragments in RNA 3D structures



Jakub Wiedemann¹, Tomasz Zok^{1,2}, Maciej Milostan^{1,2} and Marta Szachniuk^{1,3*}

Abstract

Background: In modern structural bioinformatics, comparison of molecular structures aimed to identify and assess similarities and differences between them is one of the most commonly performed procedures. It gives the basis for evaluation of in silico predicted models. It constitutes the preliminary step in searching for structural motifs. In particular, it supports tracing the molecular evolution. Faced with an ever-increasing amount of available structural data, researchers need a range of methods enabling comparative analysis of the structures from either global or local perspective.

Results: Herein, we present a new, superposition-independent method which processes pairs of RNA 3D structures to identify their local similarities. The similarity is considered in the context of structure bending and bonds' rotation which are described by torsion angles. In the analyzed RNA structures, the method finds the longest continuous segments that show similar torsion within a user-defined threshold. The length of the segment is provided as local similarity measure. The method has been implemented as LCS-TA algorithm (Longest Continuous Segments in Torsion Angle space) and is incorporated into our MCQ4Structures application, freely available for download from <http://www.cs.put.poznan.pl/tzok/mcq/>.

Conclusions: The presented approach ties torsion-angle-based method of structure analysis with the idea of local similarity identification by handling continuous 3D structure segments. The first method, implemented in MCQ4Structures, has been successfully utilized in RNA-Puzzles initiative. The second one, originally applied in Euclidean space, is a component of LGA (Local-Global Alignment) algorithm commonly used in assessing protein models submitted to CASP. This unique combination of concepts implemented in LCS-TA provides a new perspective on structure quality assessment in local and quantitative aspect. A series of computational experiments show the first results of applying our method to comparison of RNA 3D models. LCS-TA can be used for identifying strengths and weaknesses in the prediction of RNA tertiary structures.

Keywords: RNA 3D structure, Structure comparison, Local similarity, Torsion angles

Background

A comparison of contents stored in NCBI Reference Sequence Database (RefSeq) [1] and Protein Data Bank (PDB) [2] brings to a conclusion that there is a large, ever-widening gap between the numbers of known sequences and structures of biomolecules. Today, this gap is being filled with the use of computational methods that address the problem of RNA and protein 3D

structure prediction. Following that, a necessity to estimate the quality of computational models and fidelity of predictors arises. Since the 1990s, CASP (Critical Assessment of protein Structure Prediction) experiment has taken the challenge of assessing protein structure prediction [3]. RNA-Puzzles initiative launched in 2011 and drawing on the solutions implemented in CASP, followed to support the RNA community [4, 5]. Both experiments have significantly contributed to a development of measures and methods for validation and assessment of 3D structure models predicted in silico [6]. The resulting algorithms have been applied not only in the evaluation of predicted proteins and RNAs. They are also used for validation and analysis of experimentally

* Correspondence: marta.szachniuk@cs.put.poznan.pl

¹Institute of Computing Science & European Centre for Bioinformatics and Genomics, Poznan University of Technology, Piotrowo 2, 60-965 Poznan, Poland

³Institute of Bioorganic Chemistry, Polish Academy of Sciences, Noskowskiego 12/14, 61-704 Poznan, Poland

Full list of author information is available at the end of the article



© The Author(s). 2017 **Open Access** This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The Creative Commons Public Domain Dedication waiver (<http://creativecommons.org/publicdomain/zero/1.0/>) applies to the data made available in this article, unless otherwise stated.

solved structures, clustering 3D models, identification of structure motifs, tracking conformational changes, exploring the sequence-structure relationship, etc. [6–14].

RNA-Puzzles, a collective experiment for blind RNA structure prediction, uses the following approaches to assess submitted RNA 3D models: (i) Root Mean Square Deviation (RMSD), (ii) Interaction Network Fidelity (INF) [15], (iii) Deformation Index (DI), (iv) Clash score by MolProbity [16], and (v) Mean of Circular Quantities (MCQ) [17]. Except that, a few other RNA evaluation methods have been developed and applied in various projects [8, 18]. All of them relate to various attributes of the considered RNA 3D structures, but their common feature is that the structures are mainly evaluated globally. Similarly, most structure assessment methods in CASP treat protein models globally, and only a few touch an aspect of local similarity. Such approach is fully understood and seems sufficient when we deal with the evaluation and ranking of many models submitted to the competition. However, when analyzing individual structures, finding their strengths and weaknesses, comparing substructures, or identifying motifs, a local assessment is necessary. In such cases, local evaluation of the 3D model complements global analysis and significantly enhances our knowledge of the structure.

So far, one approach has been proposed to enable a local view on predicted RNA 3D model compared to the target structure. It is based on a concept of spheres built along RNA backbone and providing the scene for preview and RMSD-based evaluation of sphere-enclosed atom subsets. It has been first implemented as a standalone application named RNAnalyzer [8], and later released as RNAssess webserver [19]. In the case of proteins, Local-Global Alignment (LGA) is one of the most common approaches enabling local analysis [20]. LGA comprises two methods, Longest Continuous Segments (LCS) and Global Distance Test (GDT). The first one identifies the longest continual fragment within predicted protein structure which – compared to the target – has the RMSD below a given threshold. The second method computes the percentage of residues fitting below predefined distance cut-off. LGA is the reference method used to evaluate protein structures in CASP.

The methods mentioned in the previous paragraph operate in Euclidean space where each structure is represented as a set of atoms with coordinates in the Cartesian system. As all other approaches which consider molecule structures in Euclidean space and apply RMSD-based evaluation, they deal with the computationally demanding problem of optimum 3D structure alignment. This problem can be omitted when switching to the space of torsion angles. The 3D structure of RNA can be represented by a set of eight torsion angles that describe the course of its backbone and arrangement of

the bases. Such representation makes a comparison of structures independent of their alignment in space and simplifies the computation. This concept has been followed in MCQ4Structures method [17] that expresses structure similarity as Mean of Circular Quantities (MCQ).

Here, we propose a new method that integrates a concept of RNA 3D structure comparison in the space of torsion angles [17] with the idea of identifying longest continuous segments displaying local similarity [20]. Two segments are considered similar if their MCQ value is below the predefined threshold. The method has been implemented as LCS-TA algorithm (Longest Continuous Segments in Torsion Angle space) and incorporated into MCQ4Structures software. It is freely available at <http://www.cs.put.poznan.pl/tzok/mcq/>.

Methods

LCS-TA has been designed as the local similarity measure. It aims to compare two RNA 3D structures, S (structure of the target) and S' (structure of the model), and identify similar fragments within them. It runs either in sequence-independent or sequence-dependent mode. In the first mode, the compared structures can have different lengths, and the relationship between their residues can be unknown. Thus, no preliminary analysis of the sequences of S and S' is required here. In the second mode, the method processes structures of the same length. LCS-TA operates in the space of torsion angles, so it is superposition-independent and does not involve finding the optimum alignment of structures. The method scans both structures stepwise along their backbones and uses a moving search window to select segments for a comparison. In this routine, a divide and conquer formula is followed to determine the window size in each step. For a pair of window-highlighted segments, LCS-TA computes MCQ value over a set of torsion angles related to the segments. Next, it checks whether the MCQ value is below the threshold. At the output, LCS-TA provides the length of the longest continuous segment satisfying similarity condition (i.e., fitting below the threshold) and segment location (its first and last residue numbers). The resulting segment's length (referred to as LCS) is the measure of local similarity. Both components of the method, that is divide and conquer procedure and MCQ-based measure, are described in the following paragraphs.

Divide and conquer procedure

Divide and conquer (D&C) is a technique used to optimize the process of solving the problem by recursively splitting it into smaller subproblems and using their solutions to build the solution of the input problem. In our method, we apply D&C approach to

determine lengths of the search window in consecutive steps of the algorithm. The example recursion tree visualizing divide-and-conquer-driven computation in LCS-TA algorithm is presented in Fig. 1.

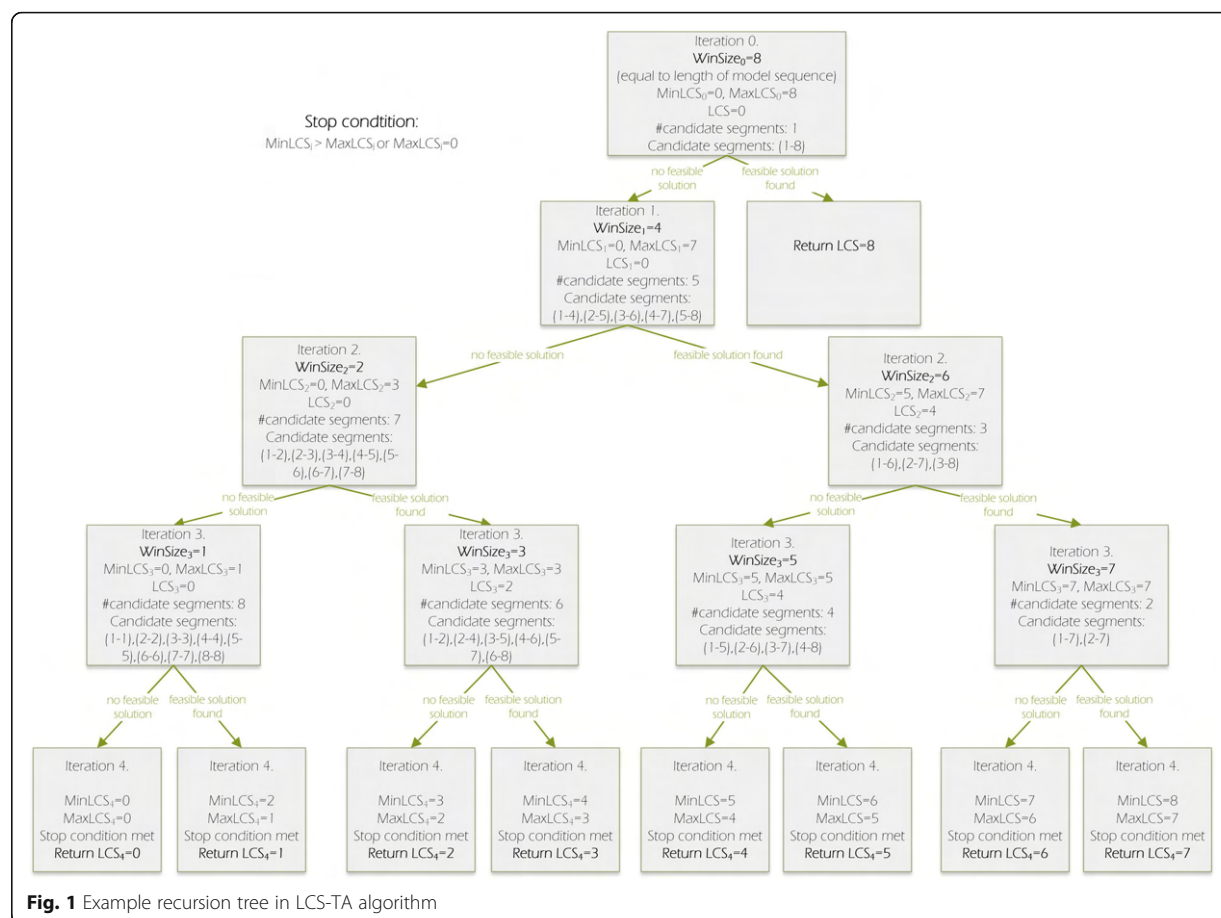
The initial window size in LCS-TA is equal to the number n of residues in the predicted model ($WinSize = n$). In each iteration, the algorithm checks whether a feasible solution (namely continuous segment with MCQ below the threshold) exists for current window size. In the case of a negative result, $WinSize$ is divided by 2 (and rounded up to the least succeeding integer). Otherwise, it is incremented to a value halfway between current size and $WinSize$ of grandparent iteration (i.e., iteration $i-2$, where i is the order number of current iteration) except the first iteration where $n-1$ is taken as an upper bound of $WinSize$. Next, the computation runs recursively for both sizes of the search window, thus branching into two subproblems. The algorithm stops if further reduction of the window size is impossible ($WinSize = 1$) and all possible solutions for that $WinSize$ value have been checked, or if the optimum solution is found. Such computation pattern, known as binary tree recursion, is one of the most commonly used

in the implementation of the D&C method. Its time complexity is $O(\log_2 n)$, where n is the instance size (in our problem n is the number of residues in S' – structure of predicted model).

MCQ-based measure

The MCQ-based distance measure has been developed for trigonometric representation of the molecule 3D structure [17]. In this representation, a shape of every RNA residue is described by eight torsion angles from the set $T = \{\alpha, \beta, \gamma, \delta, \epsilon, \zeta, \rho, \chi\}$. Each torsion angle in RNA molecule is defined by atom quadruple (the details can be found in [17, 21]) and determines rotation around particular chemical bond. It is computed as a dihedral angle between two planes defined by a pair of overlapping atom triples. Having a chain A-B-C-D of four atoms, we can easily determine the torsion angle between the plane passing through A, B, C, and the plane passing through B, C, D.

When the RNA structure is composed of n residues, then its trigonometric representation is a matrix containing $8n$ values of torsion angles t_{ij} , where $i = 1, \dots, n$, $j = 1, \dots, |T|$, and T is a set of torsion angles defined for



RNA (t_{ij} is torsion angle of type j within residue i). To measure the distance between two structures, S and S' , of equal length (n residues), given in trigonometric representations, we apply formula (1) for computing mean of circular quantities [17]:

$$\text{MCQ}(S, S') = \arctan\left(\frac{\sum_{i=1}^n \sum_{j=1}^{|T|} \sin\Delta(t_{ij}, t'_{ij})}{\sum_{i=1}^n \sum_{j=1}^{|T|} \cos\Delta(t_{ij}, t'_{ij})}\right) \tag{1}$$

The two-argument $\arctan(y, x)$ is used to distinguish results from the whole range $[-\pi; \pi)$. This is possible, because the function calculates angle value from the positive X half-axis to the vector between points $(0, 0)$ and (x, y) in a Cartesian coordinate system. In particular, this means that, unlike one-argument $\arctan(y/x)$ the two-argument variant is well-defined for $x = 0$ and in general $\arctan(y, x) \neq \arctan(-y, -x)$ which is not true for one-argument function.

In formula (1), the following function is used to obtain the distance between two angles:

$$\Delta(t, t') = \begin{cases} 0 & \text{If } t \text{ and } t' \text{ are undefined} \\ \pi & \text{if either } t \text{ or } t' \text{ is undefined} \\ \min\{\text{diff}(t, t'), 2\pi - \text{diff}(t, t')\} & \text{otherwise} \end{cases} \tag{2}$$

Where

$$\text{diff}(t, t') = |\text{mod}(t) - \text{mod}(t')| \tag{3}$$

and

$$\text{mod}(t) = (t + 2\pi) \text{ modulo } 2\pi \tag{4}$$

MCQ has been defined as a distance measure, and it shows the dissimilarity of two three-dimensional structures of the same length. Thus, the greater is its value, the more the two structures differ. And accordingly, the smaller the MCQ value, the greater is the similarity of compared structures.

It should be noted, that set T of torsion angles defined for RNA originally contained eight types of angles. However, MCQ is flexible, and any subset of T can be used to measure it. For example, if the user is interested to consider ribose ring only, then MCQ can be computed involving pseudotorion angle P (or, alternatively, $\tau_0, \tau_1, \tau_2, \tau_3, \tau_4$ angles). In the presented version of the algorithm we use original set $T = \{\alpha, \beta, \gamma, \delta, \epsilon, \zeta, P, \chi\}$.

Finally, let us add that originally MCQ value is computed in radians. In our application, it is next converted into degrees and so presented to the user.

LCS-TA algorithm

The LCS-TA algorithm compares two RNA 3D structures (hereby referred to as the target and the model) provided in PDB or mmCIF file formats. At the input, the user should also specify the MCQ threshold value in degrees and select the mode (sequence-independent or sequence-dependent). At the output, the algorithm provides the longest continuous segment (its location within both structures), its length and actual MCQ value. If more than one solution exists, all of them are shown to the user.

LCS-TA applies divide and conquer approach (Fig. 1) to find the optimum solution, i.e., the longest continuous segment in the model whose MCQ-based similarity to the target fragment is below the specified MCQ threshold. The computation proceeds as follows. First, the algorithm computes MCQ between entire structures. If its value does not exceed the threshold, the whole model structure is returned as the optimum solution. Otherwise, the size of the current search window is determined according to the D&C procedure described in the previous sections. Next, a set of candidate segments is constructed based on the model structure: the search window moves along the model from its 5' to 3'-end, and all window-highlighted fragments are put into the candidate set. Thus, the current candidate set contains all segments with length equal to the current window size. After that, for every segment from the candidate set the algorithm checks if it is a feasible solution. This part of the algorithm differs between the modes. In the sequence-independent mode, the check is done by positioning the candidate segment stepwise along the target structure, i.e., the candidate segment moves along the target structure every single residue. In the sequence-dependent mode, the candidate segment is compared to the corresponding fragment of the target structure. Two sets of torsion angles, one describing the candidate and the other describing the target segment, are computed. Based on that, the MCQ value between the positioned segments is determined. If the MCQ is below the user-defined threshold, the candidate segment is a feasible solution. If the feasible solution exists in the candidate set, the algorithm tries to find the longer segment (window size is enlarged for the next iteration). Otherwise, shorter segments are considered (window size is reduced for the next iteration). The procedure iterates until the stopping condition is satisfied.

Below, we show the pseudocode of LCS-TA focusing on the general steps of the algorithm running in the sequence-independent mode. In the sequence-dependent mode, the comparison of corresponding segments is done within one *FOR EACH* loop, instead of two nested loops.

Algorithm LCS-TA

Input: *Target* - 3D structure of the target in PDB or mmCIF format
Model - 3D structure of the model in PDB or mmCIF format
MCQthreshold - MCQ threshold in degrees

Output: *BestSolutions* - set of longest continuous segments
LCS - length of the longest continuous segment

```

1: FUNCTION LCS_TA(Target, Model, MCQthreshold)
2:   WinSize = Model.size
3:   MinLCS = 0
4:   MaxLCS = Model.size
5:   LCS = 0
6:   BestSolutions = []
7:   IF MCQ(Target, Model) <= MCQthreshold
8:     BestSolutions.push(Model)
9:     LCS=Model.size
10:  ELSE
11:    MaxLCS = MaxLCS - 1
12:    WHILE ((MinLCS <= MaxLCS) and (MaxLCS > 0))
13:      WinSize = [MinLCS + MaxLCS] / 2
14:      Found = false
15:      TargetSegs = CreateSetOfSegments(Target, WinSize)
16:      ModelSegs = CreateSetOfSegments(Model, WinSize)
17:      FOR EACH msegment in ModelSegs
18:        FOR EACH tsegment in TargetSegs #sequence-independent mode
19:          IF MCQ(tsegment, msegment) <= MCQthreshold
20:            IF not Found
21:              BestSolutions=[]
22:              Found = True
23:              BestSolutions.push(msegment)
24:              LCS = msegment.size
25:            IF Found
26:              MinLCS = WinSize + 1
27:            ELSE
28:              MaxLCS = WinSize - 1
29:      RETURN (BestSolutions, LCS)
30:  END FUNCTION

```

The LCS-TA algorithm in sequence-independent mode runs with the worst-case computational complexity of $O(n^2 \log_2 n)$. In the sequence-dependent mode the complexity is $O(n \log_2 n)$, where n denotes the number of residues in the predicted model. This computational complexity is due to the complexity of D&C being $O(\log_2 n)$, and the number of comparisons performed for every candidate segment in a single iteration.

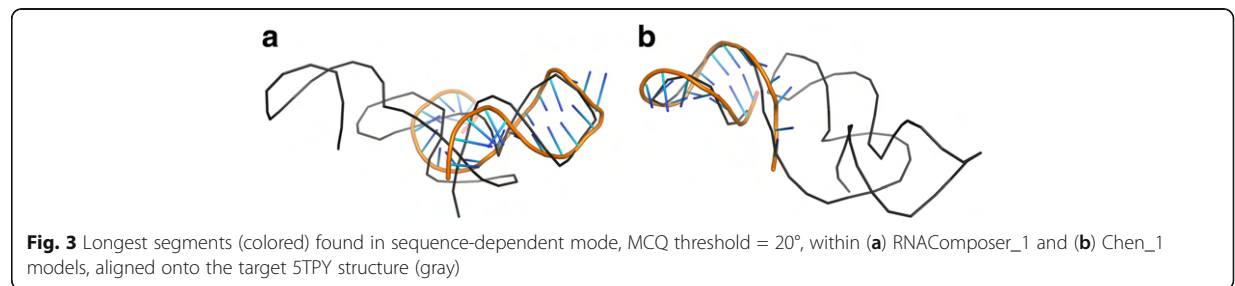
Accessibility and usage

LCS-TA algorithm has been implemented as a new functionality of MCQ4Structures [17], running as standalone Java Web start application. It is freely available for download at <http://www.cs.put.poznan.pl/tzok/mcq/>.

Results and discussion

In this section, we present the results of LCS-TA experimental runs over selected RNA 3D structures. We analyze the algorithm's output in the case of structure processing in sequence-independent and sequence-dependent mode, and we observe the impact of MCQ threshold value on local and global similarity assessment.

For a pair of compared RNA structures, LCA-TA algorithm provides the following output data: (i) LCS - a length of optimum solution (the longest continuous segment) measured as the number of residues in the segment, (ii) target structure coverage by the resulting segment, that is the ratio of segment to structure length (in percentages), (iii) actual MCQ value of the segment,



threshold), while global RMSD of the model was only 3.144 Å.

In the second experiment, we have investigated multiple models predicted in RNA-Puzzles challenge 18 and challenge 19. Altogether, 53 models were submitted in challenge 18, and 54 in challenge 19. From these sets, we have selected one model per each participant (namely, model 1) and we compared it to the target structure, i.e., exonuclease resistant RNA from Zika virus (PDB id: 5TPY) [22] in challenge 18, and twister sister (TS) ribozyme (PDB id: 5T5A) [26] in challenge 19. Experimental results concerning the selected models are presented in Tables 3–4 and Fig. 5 for challenge 18, and Tables 5–6 and Fig. 6 for challenge 19. In the tables, one can see LCS value, i.e., the length of the resulting segment found within each model for different MCQ thresholds, and actual MCQ of this segment. The best solution (LCS of the longest continuous segment found among all models) in human and server category is printed in bold. If more

models include a segment with the biggest LCS, the one with the smallest actual MCQ is considered the winner. The figures complement tabular data by showing, for each model and MCQ threshold, the percentage of target structure covered by the optimum solution.

Eleven participants submitted their predictions for challenge 18. Thus, 11 RNA 3D models were selected for the analysis with LCS-TA (Tables 3–4, Fig. 5). This number includes six human predictions (Fig. 5, solid lines) and five server-predicted ones (Fig. 5, dotted lines). In the human category, the Das_1 model has appeared to win for all MCQ thresholds. Among server predictions, RW3D_1 model, generated by Das server (unpublished), has been the best. This is true for both modes of LCS-TA. In the case of sequence-independent analysis and MCQ threshold set to 10°, RW3D_1 dominates Das_1 (Table 3). However, this relationship is not the same in the sequence-dependent mode (Table 4). A comparison of the results for Das_1 and RW3D_1 with

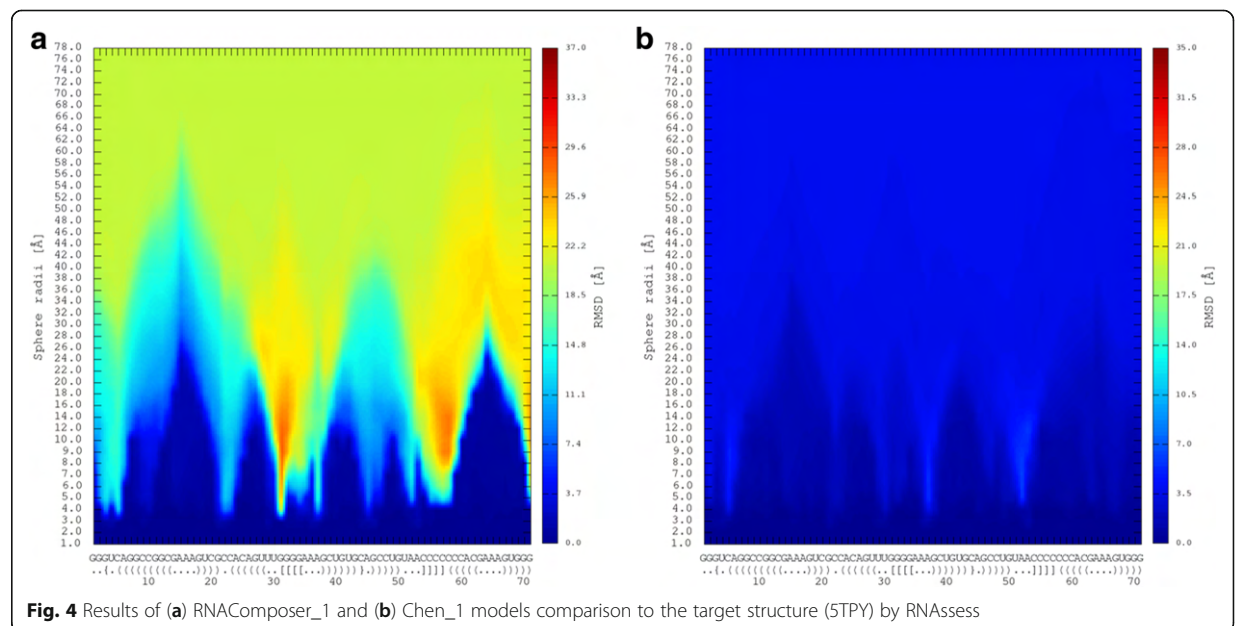


Table 3 LCS-TA results for predicted models of 5TPY structure in the sequence-independent mode

Model	MCQ threshold	10°		15°		20°		25°		≥30°	
		LCS	MCQ	LCS	MCQ	LCS	MCQ	LCS	MCQ	LCS	MCQ
(a) Human category											
Chen_1		0	n/a	13	14.80°	21	19.67°	71	23.81°	71	23.81°
Das_1		12	8.78°	70	14.98°	71	15.33°	71	15.33°	71	15.33°
Dokholyan_1		0	n/a	18	14.52°	35	19.40°	71	23.21°	71	23.21°
Feng_1		11	9.67°	26	14.90°	71	19.41°	71	19.41°	71	19.41°
Lee_1		10	9.83°	35	14.87°	71	18.57°	71	18.57°	71	18.57°
YagoubAli_1		8	9.70°	18	14.66°	41	19.69°	71	23.79°	71	23.79°
(b) Server category											
3dRNA_1		0	n/a	14	14.20°	22	18.58°	48	24.98°	71	26.37°
LeeAS_1		10	9.74°	30	14.99°	67	19.77°	71	20.71°	71	20.71°
RNAComposer_1		9	9.24°	19	14.91°	35	19.93°	71	23.48°	71	23.48°
RW3D_1		18	9.88°	35	14.77°	71	17.20°	71	17.20°	71	17.20°
simRNA_1		13	9.78°	25	14.5°	68	19.81°	71	20.61°	71	20.61°

MCQ threshold = 10° in both modes shows that there is one, accurately predicted 12 nt-long segment in Das_1 which is identified by LCS-TA in both modes. However, for RW3D_1 the longest segment below 10° threshold (with LCS = 18) corresponds very well to the other part of the target structure. This influences the overall quality of RW3D_1 prediction and makes it globally a little worse than that of Das_1. Nevertheless, the accuracy and quality of both models are very high. MCQ computed for each of these models in total, does not exceed 20 degrees. Thus, starting from threshold set to 20°, the optimum solution in both cases covers 100% of the structure (Fig. 5).

Challenge 19 has also attracted 11 participants, including six in the human category (Fig. 6, solid lines) and five in the group of servers (Fig. 6, dotted lines). Thus, 11 predicted models were processed with LCS-TA (Tables 5–6 and Fig. 6). This experiment's results show a greater diversity in the relationship between the models than in the case of challenge 18. In the human category, the situation is similar for both LCS-TA modes. Das_1 proves the best for MCQ threshold = 5°, however, when the threshold value increases by accepting values 10, 15, 20, 25 and 30 degrees, RNAComposerH_1 dominates all other models as far as LCS and actual MCQ are concerned. In the server category, the longest segments have

Table 4 LCS-TA results for predicted models of 5TPY structure in the sequence-dependent mode

Model	MCQ threshold	10°		15°		20°		25°		≥30°	
		LCS	MCQ	LCS	MCQ	LCS	MCQ	LCS	MCQ	LCS	MCQ
(a) Human category											
Chen_1		0	n/a	12	14.44°	20	19.62°	71	23.81°	71	23.81°
Das_1		12	8.78°	70	14.98°	71	15.33°	71	15.33°	71	15.33°
Dokholyan_1		0	n/a	8	13.14°	35	19.40°	71	23.21°	71	23.21°
Feng_1		0	n/a	13	14.25°	71	19.41°	71	19.41°	71	19.41°
Lee_1		0	n/a	28	15.0°	71	18.57°	71	18.57°	71	18.57°
YagoubAli_1		0	n/a	15	14.45°	28	19.68°	71	23.79°	71	23.79°
(b) Server category											
3dRNA_1		0	n/a	0	n/a	18	19.39°	35	23.81°	71	26.37°
LeeAS_1		0	n/a	16	14.87°	59	19.89°	71	20.71°	71	20.71°
RNAComposer_1		9	9.24°	17	13.69°	28	19.63°	71	23.48°	71	23.48°
RW3D_1		11	9.98°	30	14.56°	71	17.20°	71	17.20°	71	17.20°
simRNA_1		0	n/a	20	14.93°	68	19.95°	71	20.61°	71	20.61°

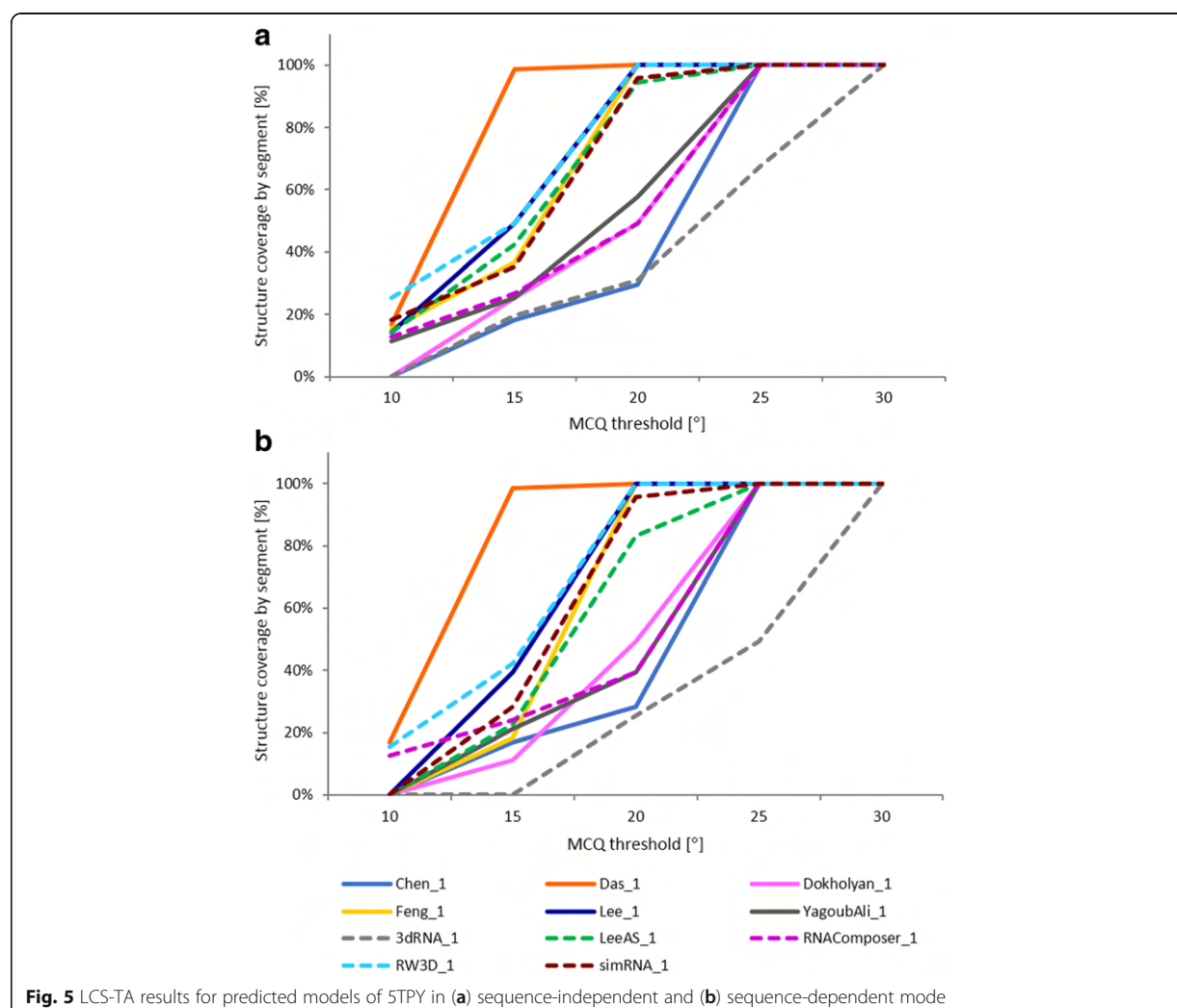
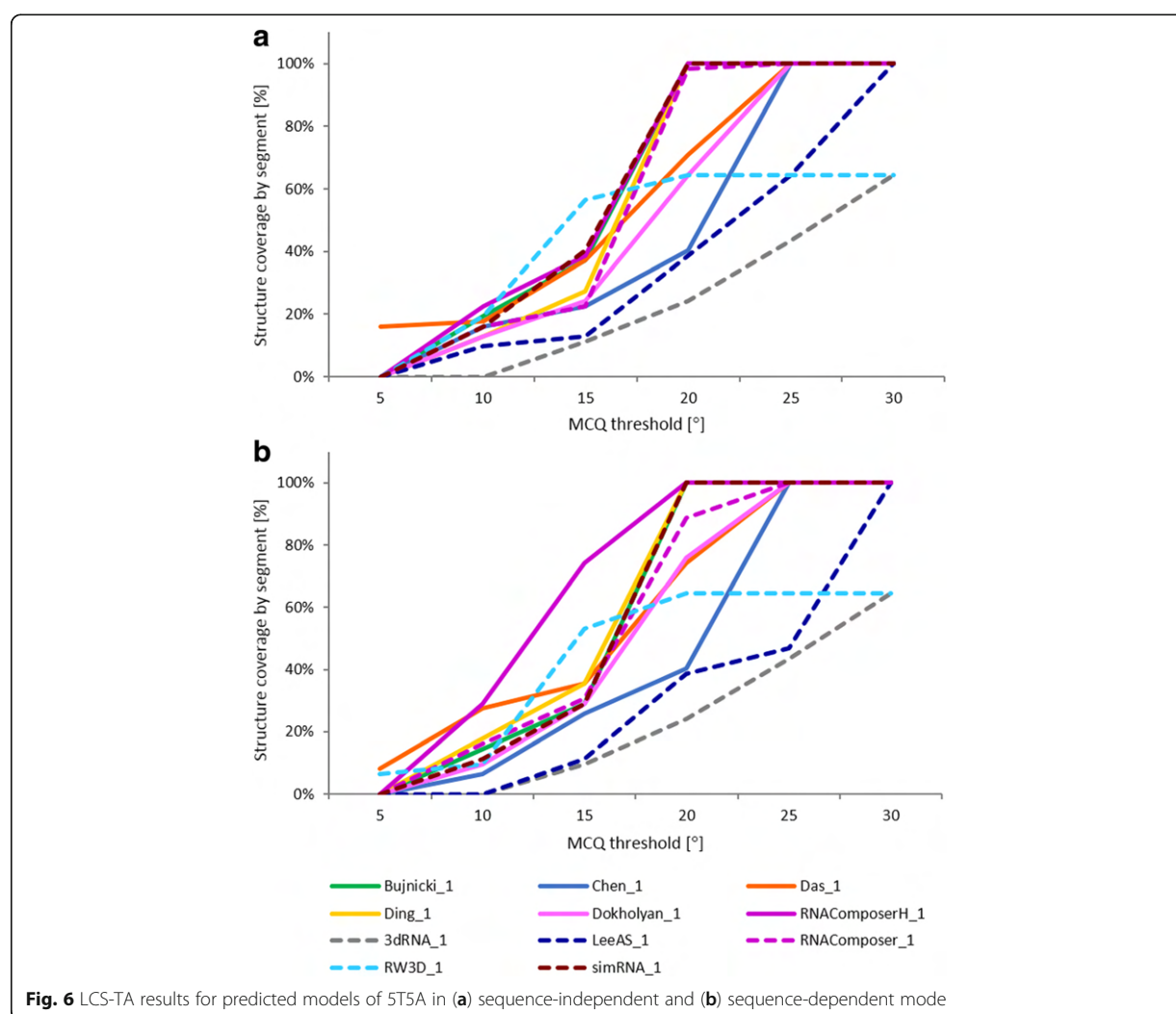


Table 5 LCS-TA results for predicted models of 5T5A structure in the sequence-independent mode

Model	MCQ threshold	5°		10°		15°		20°		25°		≥30°	
		LCS	MCQ	LCS	MCQ	LCS	MCQ	LCS	MCQ	LCS	MCQ	LCS	MCQ
(c) Human category													
Bujnicki_1	0	n/a	12	8.70°	23	14.60°	62	18.92°	62	18.92°	62	18.92°	
Chen_1	0	n/a	10	9.05°	14	13.53°	25	18.63°	62	22.88°	62	22.88°	
Das_1	10	4.61°	11	8.95°	23	13.20°	44	19.72°	62	21.41°	62	21.41°	
Ding_1	0	n/a	8	9.67°	17	14.44°	62	18.10°	62	18.10°	62	18.10°	
Dokholyan_1	0	n/a	8	9.67°	15	14.84°	40	19.36°	62	21.42°	62	21.42°	
RNAComposerH_1	0	n/a	14	9.56°	24	14.35°	62	18.04°	62	18.04°	62	18.04°	
(d) Server category													
3dRNA_1	0	n/a	0	n/a	7	14.71°	15	19.38°	27	24.21°	40	28.16°	
Lee_1	0	n/a	6	9.41°	8	14.89°	24	19.33°	40	23.97°	62	25.30°	
RNAComposer_1	0	n/a	10	6.79°	14	13.00°	61	19.70°	62	20.50°	62	20.50°	
RW3D_1	0	n/a	12	9.00°	35	14.66°	40	15.64°	40	15.64°	40	15.64°	
simRNA_1	0	n/a	10	9.18°	25	14.64°	62	19.36°	62	19.36°	62	19.36°	

Table 6 LCS-TA results for predicted models of 5T5A structure in the sequence-dependent mode

Model	MCQ threshold	5°		10°		15°		20°		25°		≥30°	
		LCS	MCQ	LCS	MCQ	LCS	MCQ	LCS	MCQ	LCS	MCQ	LCS	MCQ
(a) Human category													
Bujnicki_1		0	n/a	9	9.94°	18	14.11°	62	18.92°	62	18.92°	62	18.92°
Chen_1		0	n/a	4	9.49°	16	14.62°	25	19.85°	62	22.88°	62	22.88°
Das_1	5	4.91°	17	9.26°	22	14.24°	46	19.87°	62	21.41°	62	21.41°	
Ding_1		0	n/a	11	9.29°	22	13.86°	62	18.10°	62	18.10°	62	18.10°
Dokholyan_1		0	n/a	6	9.61°	18	14.65°	47	19.45°	62	21.42°	62	21.42°
RNAComposerH_1		0	n/a	18	9.91°	46	14.98°	62	18.04°	62	18.04°	62	18.04°
(b) Server category													
3dRNA_1		0	n/a	0	n/a	6	14.63°	15	19.38°	27	24.21°	40	28.16°
Lee_1		0	n/a	0	n/a	7	12.89°	24	19.96°	29	24.48°	62	25.30°
RNAComposer_1		0	n/a	10	8.84°	19	14.90°	55	19.98°	62	20.50°	62	20.50°
RW3D_1	4	4.08°	6	8.48°	33	14.94°	40	15.64°	40	15.64°	40	15.64°	
simRNA_1		0	n/a	7	9.24°	18	14.95°	62	19.36°	62	19.36°	62	19.36°



been found in RNAComposer_1 [23, 24], RW3D_1 and simRNA_1 [27] models, depending on the MCQ threshold and LCS-TA mode. This shows that although globally the considered models seem quite similar, the differences on a local level can be significant. Thus, local analysis of the model can indicate the direction for further development and improvement of the prediction approach. From these results, we can also see that global ranking of models based on LCS-TA value highly depends on the MCQ threshold.

Molecules selected for the above analysis are medium-size RNA structures. Their processing by both alignment-based and alignment-free algorithms is possible, although it is more time-consuming in the case of the first group of methods. The difference between computing times by both groups increases significantly with the increase in molecule size. The length of RNA chain can also influence the quality of results generated by alignment-based algorithms which provide a suboptimum solution. However, this is not the case of alignment-free approach, including LCS-TA. To show that our algorithm also works for longer RNAs, we have applied it to process RNA 3D models submitted to RNA-Puzzles challenge 7 and challenge 8. In the first case, we have chosen one model per each participant (namely, model 1) and we compared it to the target structure of Varkud satellite ribozyme (PDB id: 4R4V) [28]. Similarly, the first model submitted by each participant in challenge 8 was selected and analyzed with reference to the target structure of SAM I/IV-riboswitch (PDB id: 4 L81) [29]. Altogether, we have processed seven models from challenge 7 and 6 models from challenge 8. For all cases LCS-TA algorithm provided the results, finding similar fragments positioned along the entire structure. These experiments' results are presented in Additional file 1.

Conclusions

In the paper, we have addressed the problem of identifying similar fragments within RNA 3D structures and tertiary structure similarity assessment on the local level. We have introduced LCS-TA method that finds fragments displaying high similarity in torsion angle space. The method has been implemented in Java and added to MCQ4Structures standalone application, freely available at <http://www.cs.put.poznan.pl/tzok/mcq/>. We have shown an example application of the method in processing and analysis of RNA 3D structures predicted within RNA-Puzzles challenge 18 and 19.

Our algorithm is computationally non-demanding and user-friendly. At the input, it requires PDB or mmCIF files with RNA 3D structures and MCQ threshold value. The results are easy to compare and interpret. Thus, we hope it will be of wide interest in the RNA community.

LCS-TA has the potential to open new avenues in the RNA structural bioinformatics, particularly in the field of evaluating predicted RNA 3D models, local similarity assessment, as well as in structure motif/module identification and examination. Our future works will follow in this direction. We are going to perform large-scale tests of the method to define reliable MCQ thresholds. We plan to analyze the relationship between LCS-TA results and the secondary structure motifs of the analyzed RNA structures. This kind of analysis can indicate RNA motifs or fragments which are particularly hard (or easy) to predict. Finally, we plan to supplement the algorithm with the graphical output.

Additional file

Additional file 1: Table S1. LCS-TA results for predicted models of 4R4V structure in the sequence-independent mode. **Table S2.** LCS-TA results for predicted models of 4R4V structure in the sequence-dependent mode. **Table S3.** LCS-TA results for predicted models of 4 L81 structure in the sequence-independent mode. **Table S4.** LCS-TA results for predicted models of 4 L81 structure in the sequence-dependent mode. **Figure S1.** LCS-TA results for predicted models of 4R4V in (a) sequence-independent and (b) sequence-dependent mode. **Figure S2.** LCS-TA results for predicted models of 4 L81 in (a) sequence-independent and (b) sequence-dependent mode. **Table S5.** Longest segments found within example models of 4 L81 structure in the sequence-dependent mode. **Figure S3.** Results of (a) Bujnicki_1, (b) Das_1, and (c) Dokholyan_1 model comparison to the target structure (4 L81) by RNAssess. (PDF 465 kb)

Abbreviations

CASP: Critical Assessment of protein Structure Prediction; CSV: Comma-Separated Values; D&C: Divide and conquer; GDT: Global Distance Test; INF: Interaction Network Fidelity; LCS: Longest Continuous Segments; LCS-TA: Longest Continuous Segments in Torsion Angle space; LGA: Local-Global Alignment; MCQ: Mean of Circular Quantities; RMSD: Root Mean Square Deviation

Acknowledgements

This research was carried in the European Centre for Bioinformatics and Genomics, Poznan University of Technology (Poznan, Poland) and supported by the Leading National Research Centre Program (KNOW) granted by the Polish Ministry of Science and Higher Education.

Funding

This work has been supported by the Polish Ministry of Science and Higher Education and the Institute of Bioorganic Chemistry, PAS within intramural financing program. The authors acknowledge partial support by the National Science Center, Poland [2016/23/B/ST6/03931, 2016/23/N/ST6/03779].

Availability of data and materials

All predicted RNA 3D models used in our computational experiments are available at RNA-Puzzles website: <http://ahsoka.u-strasbg.fr/rnapuzzlesv2/results/>. The target structures can also be accessed via this webpage.

Authors' contributions

JW, TZ, and MS conceived the study. MM and MS prepared a specification of the project. JW and MM designed the LCS-TA algorithm. JW made an implementation, supported by TZ who authored the basic method for MCQ computation. JW carried computational tests further analyzed with the aid of MM and MS. MS coordinated the project. JW, MM, and MS drafted the manuscript, JW and MM prepared the figures. All authors were involved in discussions, as well as reading and approving the final manuscript.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Author details

¹Institute of Computing Science & European Centre for Bioinformatics and Genomics, Poznan University of Technology, Piotrowo 2, 60-965 Poznan, Poland. ²Poznan Supercomputing and Networking Center, Jana Pawla II 10, 61-139 Poznan, Poland. ³Institute of Bioorganic Chemistry, Polish Academy of Sciences, Noskowskiego 12/14, 61-704 Poznan, Poland.

Received: 9 June 2017 Accepted: 9 October 2017

Published online: 23 October 2017

References

- Pruitt KD, Tatusova T, Brown GR, Maglott DRNCBI. Reference sequences (RefSeq): current status, new features and genome annotation policy. *Nucleic Acids Res.* 2012;40:D130–5.
- Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. The Protein Data Bank. *Nucleic Acids Res.* 2000;28:235–42.
- Moult J, Pedersen JT, Judson R, Fidelis KA. Large-scale experiment to assess protein structure prediction methods. *Proteins.* 1995;23:ii–v.
- Cruz JA, Blanchet MF, Boniecki M, Bujnicki JM, Chen SJ, Cao S, et al. RNA-puzzles: a CASP-like evaluation of RNA three-dimensional structure prediction. *RNA.* 2012;18:610–25.
- Miao Z, Adamiak RW, Antczak M, Batey RT, Becka A, Biesiada M, et al. RNA-puzzles round III: 3D RNA structure prediction of five riboswitches and one ribozyme. *RNA.* 2017;23:655–72.
- Miao Z, Westhof E. RNA structure: advances and assessment of 3D structure prediction. *Annu Rev Biophys.* 2017;46:483–503.
- Blazewicz J, Szachniuk M, Wojtowicz ARNA. Tertiary structure determination: NOE pathway construction by tabu search. *Bioinformatics.* 2005;21:2356–61.
- Lukasiak P, Antczak M, Ratajczak T, Bujnicki JM, Szachniuk M, Popenda M, Adamiak RW, Blazewicz J. RNALyzer - novel approach for quality analysis of RNA structural models. *Nucleic Acids Res.* 2013;41:5978–90.
- Szostak N, Royo F, Rybarczyk A, Szachniuk M, Blazewicz J, del Sol A, Falcon-Perez JM. Sorting signal targeting mRNA into hepatic extracellular vesicles. *RNA Biol.* 2014;11:836–44.
- Zok T, Antczak M, Riedel M, Nebel D, Villmann T, Lukasiak P, Blazewicz J, Szachniuk M. Building the library of RNA 3D nucleotide conformations using clustering approach. *Int J Appl Math Comp.* 2015;25:689–700.
- Rybarczyk A, Szostak N, Antczak M, Zok T, Popenda M, Adamiak RW, Blazewicz J, Szachniuk M. New in silico approach to assessing RNA secondary structures with non-canonical base pairs. *BMC Bioinformatics.* 2015;16:276.
- Gudanis D, Popenda L, Szpotkowski K, Kierzek R, Gdaniec Z. Structural characterization of a dimer of RNA duplexes composed of 8-bromoguanosine modified CGG trinucleotide repeats: a novel architecture of RNA quadruplexes. *Nucleic Acids Res.* 2016;44:2409–16.
- Wiedemann J, Milostan M. StructAnalyzer - a tool for sequence versus structure similarity analysis. *Acta Biochim Pol.* 2016;63:753–7.
- Miskiewicz J, Tomczyk K, Mickiewicz A, Sarzynska J, Szachniuk M. Bioinformatics study of structural patterns in plant microRNA precursors. *Biomed Res Int.* 2017; doi: 10.1155/2017/6783010.
- Parisien M, Cruz JA, Westhof E, Major F. New metrics for comparing and assessing discrepancies between RNA 3D structures and models. *RNA.* 2009; 15:1875–85.
- Chen VB, Arendall WB 3rd, Headd JJ, Keedy DA, Immormino RM, Kapral GJ, Murray LW, Richardson JS, Richardson DC. MolProbity: all-atom structure validation for macromolecular crystallography. *Acta Crystallogr D Biol Crystallogr.* 2010;66:12–21.
- Zok T, Popenda M, Szachniuk M. MCQ4Structures to compute similarity of molecule structures. *Cent Eur J Oper Res.* 2014;22:457–74.
- Wang J, Zhao Y, Zhu C, Xiao Y. 3dRNAscore: a distance and torsion angle dependent evaluation function of 3D RNA structures. *Nucleic Acids Res.* 2015;43:e63.
- Lukasiak P, Antczak M, Ratajczak T, Szachniuk M, Popenda M, Adamiak RW, Blazewicz J. RNAssess - a webserver for quality assessment of RNA 3D structures. *Nucleic Acids Res.* 2015;43:W502–6.
- Zemla A. LGA: a method for finding 3D similarities in protein structures. *Nucleic Acids Res.* 2003;31:3370–4.
- Richardson JS, Schneider B, Murray LW, Kapral GJ, Immormino RM, Headd JJ, et al. RNA backbone: consensus all-angle conformers and modular string nomenclature (an RNA ontology consortium contribution). *RNA.* 2008;14:465–81.
- Akiyama BM, Laurence HM, Massey AR, Costantino DA, Xie X, Yang Y, Shi PY, Nix JC, Beckham JD, Kieft JS. Zika virus produces noncoding RNAs using a multi-pseudoknot structure that confounds a cellular exonuclease. *Science.* 2016;354:1148–52.
- Popenda M, Szachniuk M, Antczak M, Purzycka KJ, Lukasiak P, Bartol N, et al. Automated 3D structure composition for large RNAs. *Nucleic Acids Res.* 2012;e112:40.
- Antczak M, Popenda M, Zok T, Sarzynska J, Ratajczak T, Tomczyk K, Adamiak RW, Szachniuk M. New functionality of RNAComposer: an application to shape the axis of miR160 precursor structure. *Acta Biochim Pol.* 2016;63:737–44.
- Xu X, Zhao P, Chen SJ. Vfold: a webserver for RNA structure and folding thermodynamics prediction. *PLoS One.* 2014;9:e107504.
- Liu Y, Wilson TJ, Lilley DMJ. The structure of a nucleolytic ribozyme that employs a catalytic metal ion. *Nat Chem Biol.* 2017;13:508–13.
- Boniecki MJ, Lach G, Dawson WK, Tomala K, Lukasz P, Soltysinski T, Rother KM, Bujnicki JM. SimRNA: a coarse-grained method for RNA folding simulations and 3D structure prediction. *Nucleic Acids Res.* 2016;44:e63.
- Suslov NB, DasGupta S, Huang H, Fuller JR, Lilley DMJ, Rice PA, Piccirilli JA. Crystal structure of the Varkud satellite ribozyme. *Nat Chem Biol.* 2015;11:840–6.
- Trausch JJ, Xu Z, Edwards AL, Reyes FE, Ross PE, Knight R, Batey RT. Structural basis for diversity in the SAM clan of riboswitches. *PNAS.* 2014;111:6624–9.

Submit your next manuscript to BioMed Central and we will help you at every step:

- We accept pre-submission inquiries
- Our selector tool helps you to find the most relevant journal
- We provide round the clock customer support
- Convenient online submission
- Thorough peer review
- Inclusion in PubMed and all major indexing services
- Maximum visibility for your research

Submit your manuscript at
www.biomedcentral.com/submit



Additional file 1 (supplementary data)

LCS-TA to identify similar fragments in RNA 3D structures

Jakub Wiedemann¹, Tomasz Zok^{1,2}, Maciej Milostan^{1,2}, Marta Szachniuk^{1,3*}

¹Institute of Computing Science & European Centre for Bioinformatics and Genomics, Poznan University of Technology, Piotrowo 2, 60-965 Poznan, Poland.

²Poznan Supercomputing and Networking Center, Jana Pawla II 10, 61-139 Poznan, Poland.

³Institute of Bioorganic Chemistry, Polish Academy of Sciences, Noskowskiego 12/14, 61-704 Poznan, Poland.

*corresponding author: marta.szachniuk@cs.put.poznan.pl

Table S1: LCS-TA results for predicted models of 4R4V structure in the sequence-independent mode.

Model \ MCQ threshold	5°		10°		15		20°		25°		≥30°	
	LCS	MCQ	LCS	MCQ	LCS	MCQ	LCS	MCQ	LCS	MCQ	LCS	MCQ
Adamiak_1	4	4.61°	10	8.58°	25	14.79°	25	19.52°	138	24.89°	185	26.80°
Bujnicki_1	4	4.43°	13	9.21°	39	14.52°	176	19.96°	185	20.33°	185	20.33°
Chen_1	0	n/a	6	8.64°	14	14.59°	24	19.69°	41	24.77°	75	29.95°
Das_1	5	4.45°	14	9.79°	25	14.01°	64	19.88°	185	23.09°	185	23.09°
Ding_1	0	n/a	12	9.49°	34	14.83°	91	18.98°	185	22.00°	185	22.00°
Dokholyan_1	0	n/a	10	9.63°	26	14.90°	55	19.90°	181	24.91°	185	25.43°
Major_1	0	n/a	6	9.41°	15	14.42°	23	19.80°	42	24.98°	63	29.84°

Table S2: LCS-TA results for predicted models of 4R4V structure in the sequence-dependent mode.

Model \ MCQ threshold	5°		10°		15°		20°		25°		≥30°	
	LCS	MCQ	LCS	MCQ	LCS	MCQ	LCS	MCQ	LCS	MCQ	LCS	MCQ
Adamiak_1	0	n/a	7	9.72°	13	14.38°	30	19.96°	63	24.88°	185	26.80°
Bujnicki_1	0	n/a	11	9.59°	28	14.62°	97	20.00°	185	20.33°	185	20.33°
Chen_1	0	n/a	5	9.88°	11	14.97°	18	19.99°	29	24.99°	62	33.33°
Das_1	0	n/a	7	9.83°	14	14.85°	33	19.96°	185	23.09°	185	23.09°
Ding_1	0	n/a	12	9.49°	34	14.83°	65	19.99°	185	22.00°	185	22.00°
Dokholyan_1	0	n/a	4	9.44°	14	14.41°	23	19.58°	118	24.96°	185	25.43°
Major_1	0	n/a	0	n/a	11	14.90°	16	18.45°	28	24.53°	63	29.89°

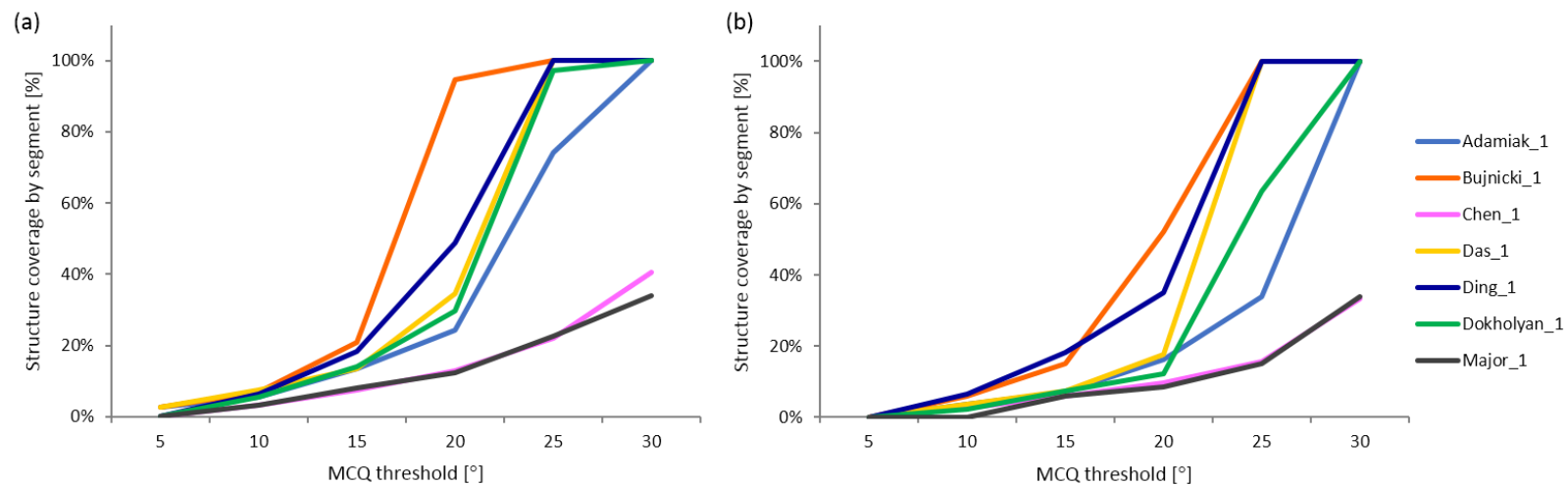
Table S3: LCS-TA results for predicted models of 4L81 structure in the sequence-independent mode.

Model \ MCQ threshold	5°		10°		15°		20°		≥25°	
	LCS	MCQ	LCS	MCQ	LCS	MCQ	LCS	MCQ	LCS	MCQ
Adamiak_1	0	n/a	13	9.47°	35	14.88°	86	19.91°	96	20.89°
Bujnicki_1	8	4.97°	22	9.41°	43	14.16°	96	17.04°	96	17.04°
Chen_1	0	n/a	8	9.65°	21	14.82°	45	19.89°	96	23.07°
Das_1	6	4.98°	18	9.87°	87	14.88°	96	15.79°	96	15.79°
Ding_1	0	n/a	13	9.85°	35	14.93°	95	19.53°	96	20.87°
Dokholyan_1	0	n/a	9	9.33°	18	14.17°	59	19.94°	96	22.42°

Table S4: LCS-TA results for predicted models of 4L81 structure in the sequence-dependent mode.

Model \ MCQ threshold	5°		10°		15°		20°		≥25°	
	LCS	MCQ	LCS	MCQ	LCS	MCQ	LCS	MCQ	LCS	MCQ
Adamiak_1	0	n/a	8	9.49°	27	14.84°	85	19.75°	96	20.89°
Bujnicki_1	8	4.97°	22	9.41°	43	14.16°	96	17.04°	96	17.04°
Chen_1	0	n/a	0	n/a	14	14.70°	45	19.89°	96	23.07°
Das_1	0	n/a	18	9.87°	87	14.88°	96	15.79°	96	15.79°
Ding_1	0	n/a	6	9.96°	23	14.76°	81	19.83°	96	20.87°
Dokholyan_1	0	n/a	5	9.32°	8	14.93°	31	19.73°	96	22.42°

Figure S1: LCS-TA results for predicted models of 4R4V in (a) sequence-independent and (b) sequence-dependent mode.



69

Figure S2: LCS-TA results for predicted models of 4L81 in (a) sequence-independent and (b) sequence-dependent mode.

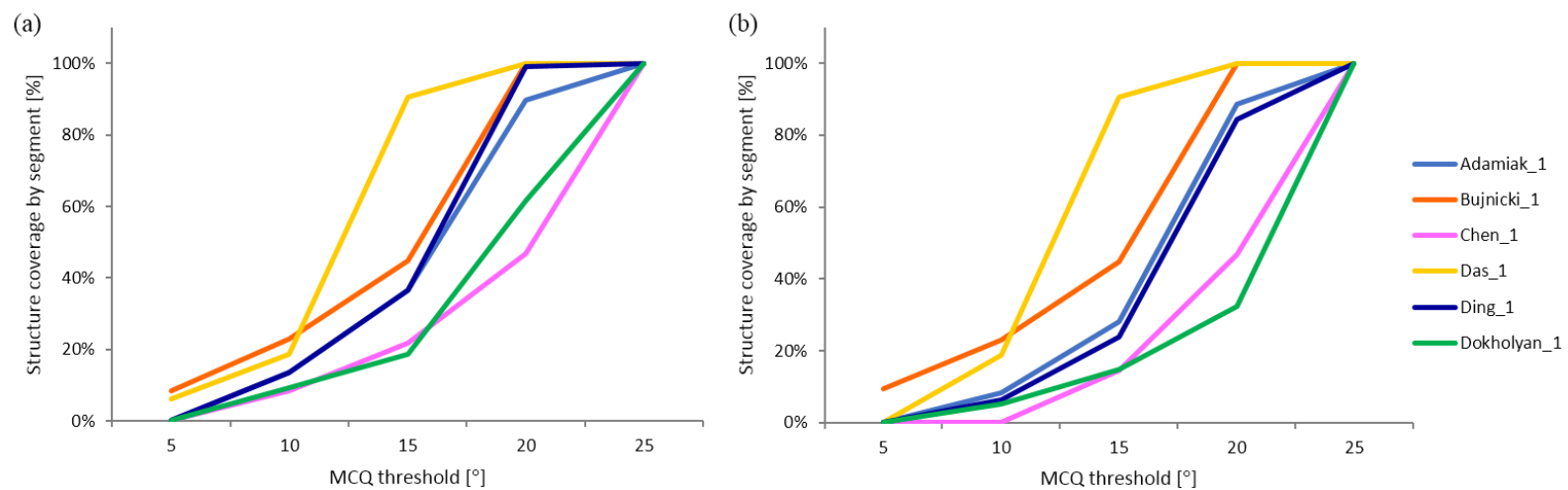
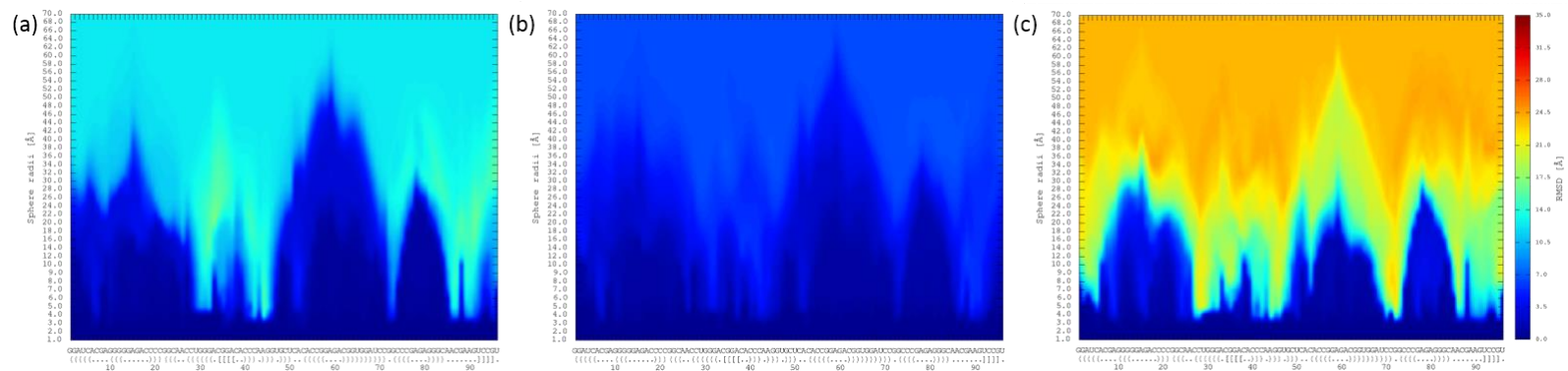


Figure S3: Results of (a) Bujnicki_1, (b) Das_1, and (c) Dokholyan_1 model comparison to the target structure (4L81) by RNAAssess.



RNA-Puzzles toolkit: a computational resource of RNA 3D structure benchmark datasets, structure manipulation, and evaluation tools

Marcin Magnus^{1,2}, Maciej Antczak^{3,4}, Tomasz Zok³, Jakub Wiedemann^{3,4},
Piotr Lukasiak^{3,4}, Yang Cao⁵, Janusz M. Bujnicki^{1,6}, Eric Westhof⁷,
Marta Szachniuk^{3,4,*} and Zhichao Miao^{8,9,10,*}

¹International Institute of Molecular and Cell Biology in Warsaw, 02-109 Warsaw, Poland, ²ReMedy-International Research Agenda Unit, Centre of New Technologies, University of Warsaw, 02-097 Warsaw, Poland, ³Institute of Computing Science & European Centre for Bioinformatics and Genomics, Poznan University of Technology, 60-965 Poznan, Poland, ⁴Institute of Bioorganic Chemistry, Polish Academy of Sciences, 61-704 Poznan, Poland, ⁵Center of Growth, Metabolism and Aging, Key Laboratory of Bio-Resource and Eco-Environment of Ministry of Education, College of Life Sciences, Sichuan University, Chengdu 610065, PR China, ⁶Institute of Molecular Biology and Biotechnology, Faculty of Biology, Adam Mickiewicz University, Poznan, Poland, ⁷Architecture et Réactivité de l'ARN, Université de Strasbourg, Institut de biologie moléculaire et cellulaire du CNRS, 12 allée Konrad Roentgen, 67084 Strasbourg, France, ⁸Translational Research Institute of Brain and Brain-Like Intelligence and Department of Anesthesiology, Shanghai Fourth People's Hospital Affiliated to Tongji University School of Medicine, Shanghai 200081, China, ⁹European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Wellcome Genome Campus, Cambridge CB10 1SD, UK and ¹⁰Newcastle Fibrosis Research Group, Institute of Cellular Medicine, Faculty of Medical Sciences, Newcastle University, Newcastle upon Tyne, UK

Received August 19, 2019; Revised November 06, 2019; Editorial Decision November 07, 2019; Accepted November 15, 2019

ABSTRACT

Significant improvements have been made in the efficiency and accuracy of RNA 3D structure prediction methods during the succeeding challenges of RNA-Puzzles, a community-wide effort on the assessment of blind prediction of RNA tertiary structures. The RNA-Puzzles contest has shown, among others, that the development and validation of computational methods for RNA fold prediction strongly depend on the benchmark datasets and the structure comparison algorithms. Yet, there has been no systematic benchmark set or decoy structures available for the 3D structure prediction of RNA, hindering the standardization of comparative tests in the modeling of RNA structure. Furthermore, there has not been a unified set of tools that allows deep and complete RNA structure analysis, and at the same time, that is easy to use. Here, we present RNA-Puzzles toolkit, a computational resource including (i) decoy sets generated by different RNA 3D structure pre-

diction methods (raw, for-evaluation and standardized datasets), (ii) 3D structure normalization, analysis, manipulation, visualization tools (RNA.format, RNA_normalizer, rna-tools) and (iii) 3D structure comparison metric tools (RNAQUA, MCQ4Structures). This resource provides a full list of computational tools as well as a standard RNA 3D structure prediction assessment protocol for the community.

INTRODUCTION

RNA 3D structure prediction, which dates back to the late 1960s (1), is nowadays being widely studied with the help of computer science. An increasing number of programs with different prediction approaches are being designed and continuously improved (2,3). Like in protein 3D structure prediction, it is important to benchmark the prediction programs to assess the capabilities of the prediction and the bottleneck in the field. CASP (Critical Assessment of Protein Structure Prediction) (4) is the largest worldwide event of protein structure prediction. And RNA-Puzzles (5–7) is a CASP-like assessment of RNA 3D structure prediction,

*To whom correspondence should be addressed. Tel: +44 1223 49 4554; Fax: +44 1223 49 4554; Email: zmiao@ebi.ac.uk
Correspondence may also be addressed to Marta Szachniuk. Email: mszachniuk@cs.put.poznan.pl

which is supported by dozens of research groups around the world.

RNA has its own structural and evolutionary features. Most importantly, the RNA secondary structure, determined by the set of *cis*-Watson-Crick base pairs, can be generally determined using sequence comparisons (8,9). However, the formation of a 3D structure requires, in addition, non-Watson-Crick base pairs (10), structural modules (11), and sometimes pseudoknots (12). Thus, the secondary structure description of RNA structure is insufficient. Precise sequence and covariation analysis (13), and/or chemical/enzymatic probing (14,15) are therefore necessary to predict relevant 3D structures. In RNA-Puzzles, we highlight the fact that 3D structure models can severely deviate from the reference structures even if the model retains perfect secondary structure (100% correct in terms of *cis*-Watson-Crick base pairing) (6) (see Supplementary Figure S1). In this context, RNA 3D structure prediction needs independent benchmarking systems that include both datasets and assessment metrics.

With the progress in protein structure prediction, many benchmark datasets and assessment metrics have been curated and developed (16). One available dataset for RNA structure benchmarking is the non-redundant dataset maintained by Leontis and Zirbel (17). Alternatively, the Rfam database, which links RNA sequence families with crystallographic structures when available, can also be used in prediction benchmarking (18). However, only 99 Rfam families have their 3D structures available. Such benchmarks are not blind and are biased towards RNAs with many homologous sequences. This is not always the case in prediction: some rare RNA structures do not necessarily have homologous sequence available, e.g. Varkud satellite ribozyme (19), in which case sequence alignment-dependent prediction methods may not be helpful. The RNA-Puzzles benchmark sets have been successfully used in developing RNA quality assessment methods (20) to identify the models similar to experimental structures without reference. Potentially, they will also serve as decoy sets for proposing structure-based force field or scoring functions, RNA design and other utilities.

Reliable evaluation of dozens of RNA 3D models cannot be performed manually and is usually preceded by normalization to comply with a common 3D structure representation. Since the start of RNA-Puzzles, a good number of RNA structure manipulation tools and structure comparison metrics, some of which are being used by the RNA-Puzzles community, have been conceived and designed. They are helpful in various ways, including structure analysis, comparison, and function inference. Here, we gather and summarize a computational resource ‘RNA-Puzzles toolkit’ that includes a set of datasets and various computational tools accumulated in the practice of RNA-Puzzles, which cover important aspects to understand RNA structure. RNA-Puzzles toolkit includes tools for structure formatting, analysis, manipulation, visualization, mutagenesis study and structure comparison. This computational resource will benefit biologists working with RNA structure and RNA structure prediction. All the datasets and codes are available as open-source on GitHub (<https://github.com/RNA-Puzzles>).

MATERIALS AND METHODS

Datasets

We provide three datasets derived from RNA-Puzzles: (i) *raw_dataset* - a dataset of raw submissions, which were generated by various prediction methods, (ii) *for-evaluation_dataset* - dataset used for official evaluation of the prediction methods in RNA-Puzzles, which does not change the coordinates of the predicted structures or add missing atoms, and (iii) *standardized_dataset* - a standardized dataset optimized with rna-tools, which not only unified the residue and atom names but also completed the missing atoms in incomplete RNA structures to standardize all the structures to the same format. All the datasets follow the same rules to name the structural files, which is a combination of the RNA-Puzzles identifier, prediction group name, and the structure model number, e.g. 19_RNAComposer_3.pdb means the third model predicted by RNAComposer (21) for Puzzle 19 in RNA-Puzzles. The reference structures were obtained from the crystallographers, renamed according to the puzzle name and marked as ‘solution’, e.g. 19_solution_0 means the first reference model of Puzzle 19. If one sequence has multiple solved structures or multiple chains in the asymmetric biological unit, all of them are used as reference structures. And the one with the lowest root mean square deviation (RMSD) to a given model is used as the reference structure to report the scores for that model.

RNA_format, RNA_normalizer and RNA_assessment

RNA_format, RNA_normalizer and RNA_assessment constitute a set of computational tools for the data formatting, processing and evaluation in RNA-Puzzles. They are implemented as Python packages making use of the BioPython (22) structure I/O library. The algorithms to compute RMSD, *P*-value (23), Deformation Profile, and Interaction Network Fidelity (24) are implemented in the Python package RNA_assessment, which makes use of BioPython, MC-Annotate (25) and NumPy (26). Deformation Profile was also implemented as an independent Python package.

rna-tools

rna-tools is a core library written in Python and a set of command-line programs execute various functions to process structural files in the PDB format but also to process RNA sequences, folding simulations, sequence alignments. Some tools in rna-tools are dependent on other programs or libraries such as ModeRNA (27), ClaRNA (28), BioPython (22).

RNAQUA

RNAQUA (RNA QUality Assessment tool) is a RESTful web service client developed in Java using Jersey (<https://jersey.github.io/>). It provides services for RNA 3D structure normalization and comparison, including the metrics of RMSD, *P*-value (23), Deformation Profile, Interaction Network Fidelity (24) and clash score (29). It uses selected functions from RNAnalyzer (30) and RNAssess (31), both of which are in the RNApolis platform (32).

MCQ4Structures

MCQ4Structures is a set of computational tools for RNA 3D structure comparison in the torsion angle space. It includes algorithms to compute *Mean of Circular Quantities* (MCQ) (33) and *Longest Continuous Segments in Torsion Angle space* (LCS-TA) (34) that compare structures, compute structure similarity, cluster and visualize the results, identify similar structural fragments, and rank the structural models. The package is implemented in Java, while functional modules of structure I/O and geometric statistics, on which both MCQ and LCS-TA depend, are implemented as separate packages of BioCommons (<https://github.com/tzok/BioCommons>) and Circular (<https://github.com/tzok/Circular>).

RESULTS

The overview of the resource

Our computational resource includes (i) the benchmark datasets from RNA-Puzzles, (ii) structure analysis, manipulation, visualization, clustering and normalization tools, (iii) and 3D structure comparison metrics (Figure 1). Considering an RNA structure comparison workflow given both a list of predicted structures and several reference structures, it is first necessary to standardize the predicted and reference structures to the same length and the same format. Structural features, such as clash score, which is based on the structure model, can be calculated and compared with the scores derived from the reference structures. Furthermore, our resource provides a set of tools for RNA structure manipulation and visualization, which can greatly facilitate manual inspection of the structures. Finally, our structure comparison metrics demonstrate the similarity/dissimilarity between the prediction and the reference structures in various aspects. The tools can be accessed via command-line, Jupyter Notebook, Docker image or web service. The user-friendly interfaces enable different usage scenarios throughout the community. Supplementary Table S1 gives a list of the datasets and computational tools in this resource, which are described in detail in the next sections.

Benchmark datasets of RNA 3D structure

In a structure prediction scenario, a good predictor should be robust in predicting structures of different types accounting for the characteristics of each prediction target. Therefore, a good benchmark must cover diverse structures (Figure 2A). The datasets from RNA-Puzzles, as listed in Supplementary Table S2, cover crucial aspects for the selection of puzzles, such as symmetry (35), ion binding (36), ligand binding (37,38), protein binding (39), the conformational change (40), and structural modules (7). Our datasets include 972 decoy RNA structures for 20 RNAs. They can be used as: (i) a standard dataset to compare with existing prediction methods, e.g. (41); (ii) a decoy dataset to develop effective structure scoring function, e.g. (20). The theoretical models were generated by the best existing RNA 3D structure prediction programs (21,42–46). The similarities of these theoretical models to crystal structures range

from low quality to the near-native (*cf.* Figure 2 and Supplementary Table S2), which provides a wide range of decoy structures that exist during structure modeling. The presented benchmark dataset can benefit the development of energy function or scoring function to discriminate the near-native structures from those far away decoys. This is an important step to identify high-quality prediction when the reference structure is unknown. In RNA-Puzzles, each group (or each prediction method) provides five candidate models (in the first 17 challenges, up to 10 models were allowed) and ranks these models according to its own prediction reliability index. However, some of the near-native structures are not ranked as the top models. The detection of such instances would improve prediction accuracy. In RNA-Puzzles, the scores for ‘quality prediction’ were obtained in Puzzles 4, 7, 8, 12, 13 and 14. The structure data from this resource is a good starting point for developing and benchmarking model ranking methods (20). According to the RMSD distribution (Figure 2C), longer structures are more difficult to predict unless homologous templates are available. Although this is consistent with the previous report (47), RNA-Puzzles includes the best RNA structure prediction approaches and demonstrates better performance in *de novo* prediction. Further, the Interaction Network Fidelity distribution highlights the insufficient prediction of non-Watson–Crick interactions. Other available datasets of the same kind are: (i) RASP (48) dataset, which includes 85 RNAs with 500 decoys for each structure and (ii) the KB (49) dataset, which includes 23 950 decoys for 20 RNAs. However, the decoy structures in these datasets were generated using only a couple of prediction methods, while our dataset covers a much wider variability in RNA structure prediction.

Standardizing the structure format considering all types of variations is the first step of a fair structure comparison. Different prediction methods result in a wide range of variations in the format of the predicted structures, ranging from nomenclature (chain names, residue names, atom names and their ordering) to structural variations (i.e. the structure at the 5′ and 3′ ends). For example, some prediction methods may use the molecular dynamics force field to minimize the energy of the predicted structure at their final steps, thus the output format depends on the force field used. Besides, the predicted structures need to be normalized according to the reference structure allowing unsolved fragments.

The RNA-Puzzles dataset can be used as (i) a standard dataset to benchmark with existing prediction methods; (ii) a decoy dataset to develop and test effective structure scoring function. To fulfill these two tasks, we provide *standardized_dataset* including structural data standardized and missing atoms completed using *rna-tools*. *rna-tools* was used to (i) add the missing atoms, especially at the 5′ and 3′ ends; (ii) mutate variant nucleotides in the predictions to make them consistent with the sequence of the reference structure. All the steps of processing and the detailed analysis of the differences between predicted models and the references, such as gaps, mismatches, etc. are described in the README files provided with the structures. The *standardized_dataset* is under active maintenance. The advanced users can also use *rna-tools* to process their own datasets.

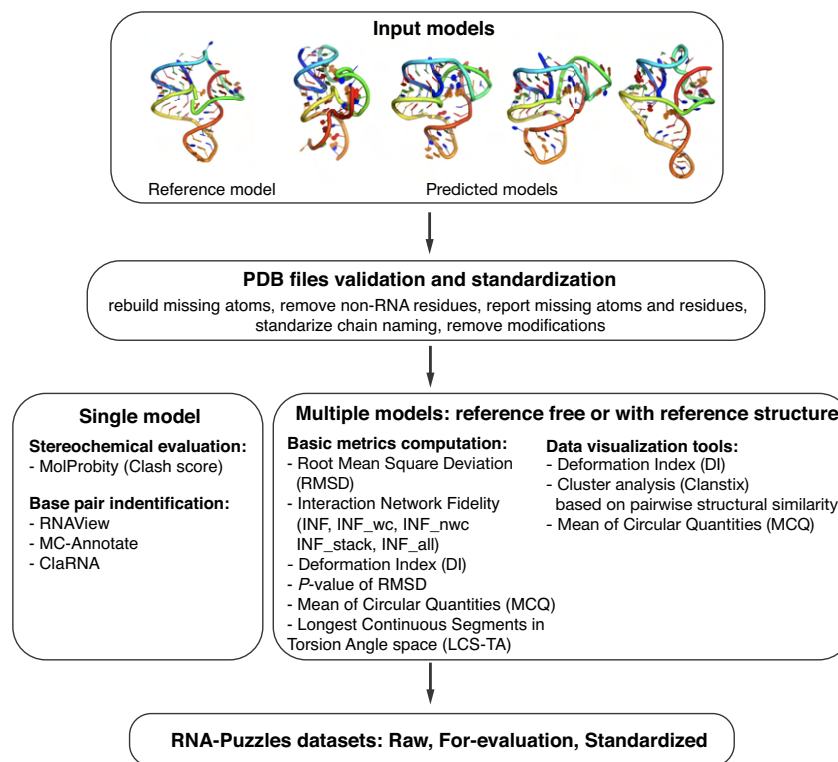


Figure 1. Scheme of the RNA-Puzzles toolkit. The toolkit is composed of three parts: tools for validation and standardization of PDB structure files, tools for analyzing the models, and the dataset of standardized submissions to the RNA-Puzzle. The user can start an analysis with single or multiple models. The first step is to standardize the formatting of analyzed structural models. Then, the user can run an analysis for a single model, such as Clash Score evaluation or base pair identification using various methods; or, for multiple models, various comparison methods are implemented. The tools can be accessed via command-line or Jupyter. The toolkit can be also executed as a Docker image that can be easily used.

RNA 3D structure formatting, manipulation, analysis and visualization tools

RNA_normalizer and rna-tools are two RNA oriented structure format tools providing semi-automatic RNA structure processing workflows.

RNA_normalizer

RNA_normalizer is an RNA structure formatting tool used in RNA-Puzzles evaluation workflow. It can: (i) normalize the residue names and atom names; (ii) order residues and atoms; (iii) extract pre-defined regions of an RNA structure. RNA_normalizer uses mapping dictionaries to normalize the non-canonical residue and atom names to the standard nomenclature. The idea of RNA_normalizer is to keep the maximum number of fragments that can be compared while keeping the prediction structures untouched. In a couple of cases, the sequence used in prediction slightly differ from the sequence of the crystal structure: e.g., single nucleotides variants or chain break because of the unsolved dynamic region in the reference structure. RNA_normalizer focuses on the consensus structure regions between the crystal sequence and the sequence in prediction. However, the skipped nucleotide makes the structure incomplete. Considering the need of complete structures for scoring function testing or molecular dynamics simulation, we provide

rna-tools to add the missing atoms in the structures. After normalizing the structure formats, we suggest to use 'RNA_format' or 'diffpdb' from rna-tools (Figure 3E) to check the consistency between the results and the standard format.

rna-tools

rna-tools includes a set of tools dedicated to (i) RNA structural handling and manipulating, i.e. rebuilding missing atoms, (ii) structure clustering, (iii) standardization of RNA structures, (iv) visualization of secondary RNA structures, i.e. drawing RNA arc diagrams of secondary structure, (v) visualization of RNA sequence alignments, and more.

The core library shared with the tools. The core part of the rna-tools package is the 'rna_pdb_toolsx.py' program that was used to prepare the *standardized_dataset*. The program facilitates many tedious operations on structural files. For example, one tool is the 'get-rmapuzzle-ready', which is used to get a standardized naming of atoms, residues, chains to be compatible with the format required by RNA-Puzzles. All structures from the *standardized_dataset* are compatible with this format, which makes it easy to compare them and use for further analysis. Another example of structure manipulation is introducing mutations. The rna-tools package

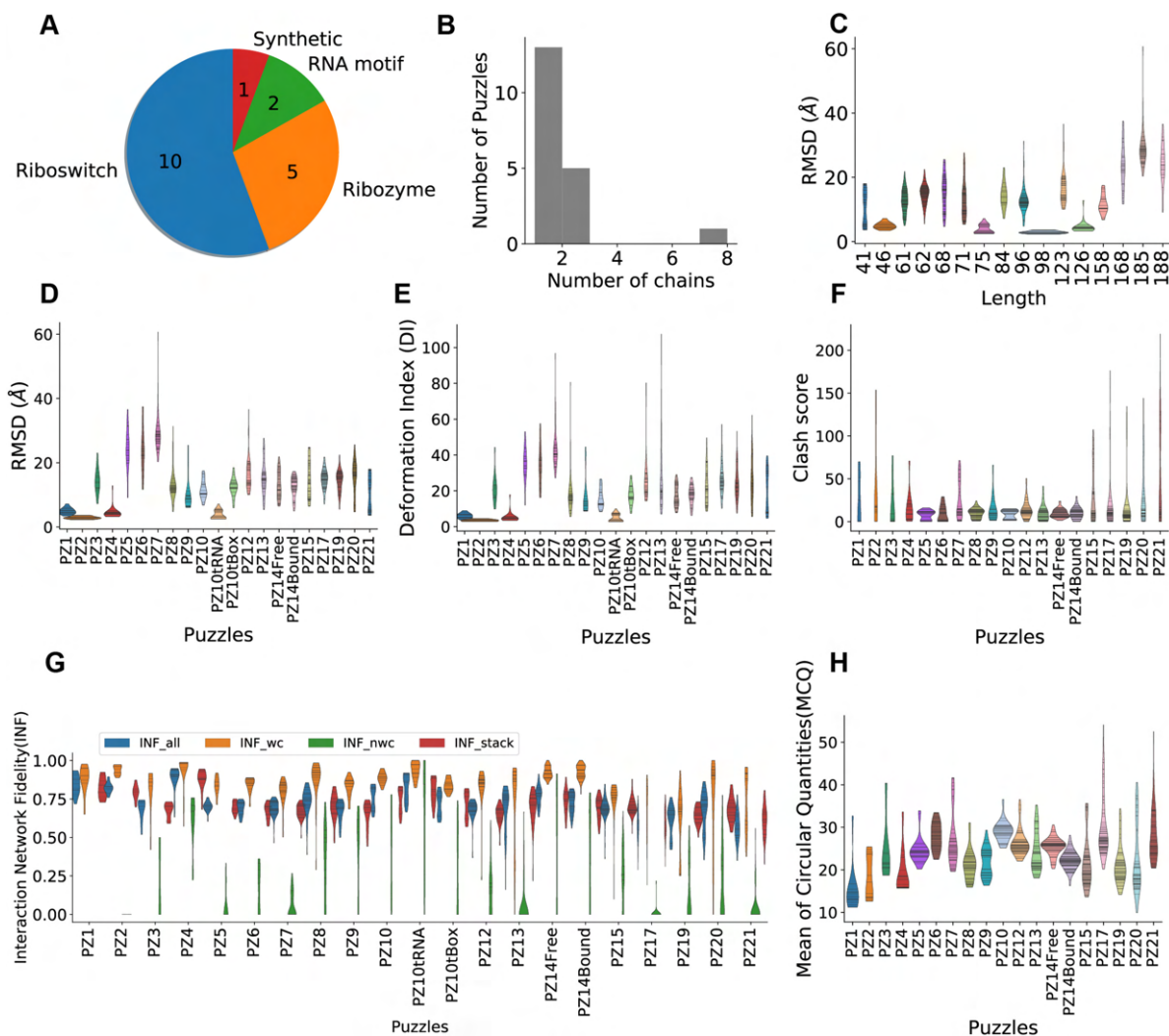


Figure 2. The structure diversity and comparison of the dataset. (A) The dataset is composed of 18 Puzzles of different types of RNA. (B) Most of them are one-chain or two-chain structures, except Puzzle 2 is of eight chains. (C) Correlation plot between the lengths of RNAs and the RMSD distributions, shown as violin plots, indicates that shorter RNA structures tend to be easier to predict. The RMSD, deformation index and clash score are shown in (D–F). The distributions of Interaction Network Fidelities are shown in (G), including stacking interactions (INF stacking), Watson-Crick interactions (canonical) (INF wc), non-Watson-Crick interactions (non-canonical) (INF nwc) and all interactions (INF all). MCQ assesses structure similarity based on torsion angles (H).

uses ModeRNA (25) to introduce single or double mutations in structures. But it overcomes ModeRNA's limitation in processing only one chain at the time (Figure 3A). Multiple mutations in multiple chains can be introduced.

Furthermore, rna-tools includes tools operating on various levels of RNA data: sequences, secondary structures, alignments, and 3D structures. rna-tools includes a collection of almost one hundred functionalities that facilitate common operations in RNA structural bioinformatics. It can be easily imported into 3rd party programs or pipelines. The full list of functionalities can be found in Supplementary Table S3.

RNA sequence tools. The first group of tools deals with RNA sequences. The tools help to perform searches us-

ing both Blast (50) on the PDB database and Infernal (51) on the Rfam database (52). Furthermore, multiple wrapper tools of RNA secondary structure prediction are implemented (Figure 3f), including RNAsubopt, RNAeval, RNAfold from ViennaRNA (45), CentroidFold (46), ContextFold (47), MC-Fold (53) and IPknot (54). All tools are compatible with Jupyter Notebook.

RNA secondary structure tools. The second group of tools aims to facilitate operations on RNA secondary structure that can be executed from Jupyter Notebooks (Figure 3F). The functionalities include visualization of a sequence and a structure with VARNA (55), evaluation of free energy, parsing secondary structure into a list of pairs, and various tools for secondary structure format conversions, etc.

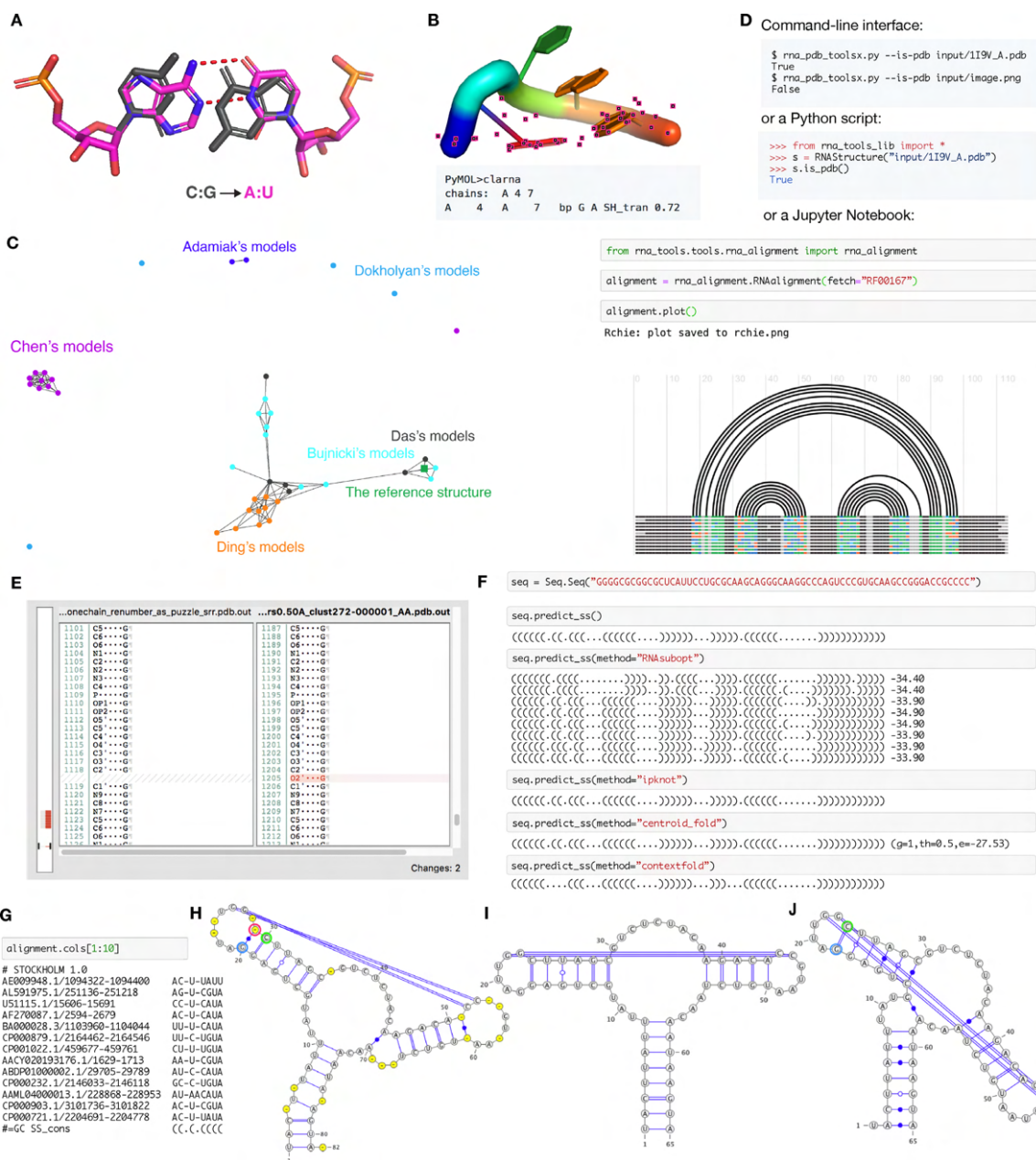


Figure 3. rna-tools is a set of tools dedicated to RNA structural file manipulation and analysis. (A) Mutate functionality allows for exchanging bases, in this case, C:G pair was replaced with A:U pair from two chains. (B) Contact classification for selected residues can be performed directly in PyMOL. In this case, trans Sugar-Hogsteen interaction was detected for closing residues of a tetraloop. (C) One of the tools implemented in rna-tools, clanstix. Clanstix can be used for visualizing RNA 3D structures based on pairwise structural similarity (as RMSD) with CLANS. The tool can be used for interactive clustering analysis when various RMSD thresholds can be tested. Here, the clustering of submission for RNA-Puzzle 8 was visualized. Dokholyan submitted four models very different from each other. Models of Chen and of Adamiak were similar respectively and made separate clusters. Models of Ding were similar to each other, and additionally, clustered with models of Das and Bujnicki. When the reference structure was released, it could be added to the visualization. Interestingly, the reference model clustered with two structures of Bujnicki and Das. (D) The functions of the package can be accessed from command-line, from Python scripts, and from Jupyter Notebooks, giving multiple ways to access the functionality. (E) diffpdb checks the consistency between annotations of two structural files. The tool ignores 3D coordinates of atoms and compares only text-content of two files in the PDB to identify the difference in the annotation of atoms, missing atoms (missing the O2' atom) and missing fragment (shown on the left side with the gray-red bar). (F) Multiple wrappers are implemented allowing for secondary structure prediction performed directly in Jupyter Notebooks, with methods such as RNAsubopt, IPknot, Centroidfold and Contextfold. (G) For RNA alignments it is possible to select only a subset of columns and work on them as a new alignment (in this case on the 1st to the 9th column). Sequences from RNA alignments and their secondary structures can be visualized with VARNA including gaps (H) and without gaps (I). The algorithm checks if residues are 'paired' with a gap position ('-') (position in red circle) for proper extraction of secondary structure. In this case, after wrong gap removal (J), G (in blue circle) is incorrectly paired with C (in green circle) and all other pairings are shifted by one.

RNA alignment tools. The third group includes tools that process RNA sequence alignments. The analysis of RNA sequence alignment is a crucial part of the structure prediction process used in RNA-Puzzle. To process and analyze RNA sequence alignments, rna-tools includes a collection of tools to load alignments, subset columns (Figure 3G) or sequences (rows), save a subset to a new file, plot an RNA arc diagrams (Figure 3D) (56), obtain a secondary structure in the dot-bracket notation, and visualize the data using VARNA of each of sequences in the alignment. Sequences and their secondary structures can be visualized with gaps (Figure 3H) and without gaps (Figure 3I). The algorithm checks if residues are ‘paired’ with a gap position (‘-’) to avoid the common problem with other tools with the wrong secondary structure after gap removal (Figure 3J).

RNA 3D structure tools. The last group of tools operates on RNA 3D structure. This group includes (i) tools for the analysis of 3D models (such as contact classifications) and (ii) tools for RNA 3D structure prediction, including the whole pipeline of structure prediction. First, to perform contact classifications, we provide two wrappers, that are ClaRNA (28) and 3DNA/DSSR (57). Using the wrappers together with the PyMOL4RNA tool in rna-tools, it is possible to perform contact classifications in PyMOL for a selected set of residues (Figure 3B). Second, the package contains scripts to help the RNA 3D structure prediction processes, both for SimRNA (42) (including SimRNAweb (58)), and Rosetta (59). Tools for SimRNA and Rosetta help to prepare input files, run modeling, cluster results, and extract models from trajectory files. Moreover, the program for SimRNAweb allows the users to download SimRNAweb prediction models and trajectory files. For processing trajectories of SimRNA, a Python interface is provided to parse trajectories into atoms, residues, simulation frames to prepare for further analysis. At the final step of a structure modeling process, a user can run the RNA refinement procedure implemented in a wrapper of QRNAS (60).

Auxiliary tools. In the package, there is a set of auxiliary tools of novel functions. One of them is diffpdb. It is a simple tool to compare two files of PDB format to identify the difference in the annotation of atoms, missing atoms, missing fragments (Figure 3e). Another standalone tool implemented in rna-tools is Clanstix. Clanstix can be used to interactively visualize the clustering results from CLANS (49). CLANS uses the Fruchterman–Reingold graph layout algorithm to visualize pairwise sequence similarities in either two-dimensional or three-dimensional space. The program was initially designed to calculate pairwise attraction values to compare protein sequences. However, it is possible to load a matrix of precomputed attraction values and thereby display any type of data based on pairwise interactions. Therefore, the Clanstix program from the rna-tools package can convert the all-vs-all distance (e.g., Root Mean Square Deviation) matrix into an input file for CLANS. An example of Clanstix is shown in Figure 3C, which is the result for RNA-Puzzle Puzzle 8. Models with a pairwise distance of RMSD lower than 8 Å are connected. The reference structure was added to this clustering. Interestingly, the reference structure was mapped to the small clus-

ter with two models from Das’s group and two models from Bujnicki’s group. The visualization can provide useful insights into a set of analyzed models or models obtained from a simulation trajectory. Another example of the usage of Clanstix can be found in the publication of EvoClustRNA (61), which shows how 3D models of various homologous sequences are clustered with respect to each other and the reference models.

The documentation with step-by-step tutorials. The description in this publication only briefly reports functionalities implemented in rna-tools. To facilitate the finding of the right tool, the package is well documented in both online documentation and tutorials that will walk the users through various use cases. The step-by-step tutorial that explains how to prepare files for the submission to RNA-Puzzles is also included.

Extensibility by design. The rna-tools package was developed with the goal in mind of providing a framework for various tools specifically to support extensibility. A new script can be easily drafted just by copying-pasting to a new folder in ‘rna_tools/tools/<new tool>’. Many core functionalities are coded in the ‘rna_tools.lib.py’ file that is shared between scripts; hence, the functions can be imported to new scripts. This design speeds up the development of new programs since many of them need some low-level common functionalities, e.g., Python engine for parsing selection of residues, atoms, parsing/converting various types of data.

Example of a complete analysis of the blind prediction of the RNA-Puzzle Puzzle 19. The functionality implemented in rna-tools can be accessed via command-line, imported in Python scripts or in Jupyter Notebooks (Figure 3D). One such notebook is released together with rna-tools and illustrates the steps performed by the Bujnicki group to collect information about the RNA-Puzzles Puzzle 19, the Twister Sister ribozyme (62) (<https://github.com/mmagnus/rna-tools/blob/master/rp19.ipynb>). The analysis started with the secondary structure prediction using multiple wrappers implemented in rna-tools followed by the Rfam search for an RNA family that the sequence belongs to. At the time of this analysis, no RNA family for the sequence of the puzzle was presented in the Rfam database. A useful piece of information was provided by a successful hit in the PDB database, to the structure in the PDB database, Xrn1-resistant RNA from the 3’ untranslated region of a flavivirus (PDB: 4PQV) (63). This structure was considered as a homolog of the Puzzle and was used for comparative modeling.

Metrics in RNA 3D structure comparison

Root mean square deviation (RMSD). Root Mean Square Deviation (RMSD) is a widely used metric for 3D structure comparison. The RMSD calculation aligns all the atoms that are found both in the predicted structure and the reference structure. A superimposition is performed based on these aligned atoms, and the result is calculated as the Root Mean Square Deviation based on the Euclidean distances of the aligned atoms.

Although RMSD is a well-established metric in structure comparison, it generalizes the errors over the whole structure. Thus, the final result can be misleading. When a linker region takes a different path or a hairpin loop has a different angle with respect to the core region, the overall RMSD may be large even if the core region is properly folded. In addition, RNA structure has more degrees of freedom in the backbone than proteins do and the accuracy of the base-pairing interactions requires inspection. To overcome the limitations of the RMSD metric, the concepts of Interaction Network Fidelity (INF) and Deformation Profile (DP) were introduced (24). These metrics, RMSD, INF, DP and *P*-value (23) are included in the packages of RNA_assessment and RNAQUA.

Interaction Network Fidelity (INF). The whole RNA structure can be considered as a large interaction network composed of Watson-Crick interactions, non-Watson-Crick interactions and base stackings. The correct prediction of all these interactions determines the success of the prediction. The interactions of an RNA structure can be extracted by programs such as MC-Annotate (25) and 3DNA (64). The Interaction Network Fidelity (INF) is defined as the Matthews correlation coefficient (MCC) between the interactions of the reference structure and that of the predicted structure. A higher INF score indicates higher consistency between the prediction and the reference structure in terms of interactions. The Interaction Network Fidelity can also assess a specific type of interaction. Thus, INF_wc, INF_nwc, INF_stack and INF_all, which define the Interaction Network Fidelity of Watson-Crick interactions, non-Watson-Crick interactions, stackings, and overall interactions, are used in the evaluation of RNA-Puzzles. Further, to account for the relationship between RMSD and INF, Deformation Index (DI) is defined as the ratio between RMSD and INF.

Deformation profile (DP). To complement single value evaluation metrics, Deformation Profile is a 2D distance matrix representing the average distance between a prediction and the reference structure (Figure 4). The deformation profile matrix calculation includes two steps: (i) computing 1-nt superimposition of predicted model over reference structure for each aligned nucleotide; (ii) computing the average distance between each base in the reference structure and the corresponding base in a predicted structure for each superimposition. The Deformation Profile displays the regions that depart most from the rest of the structure.

The deformation profile is effective in detecting the 'poorly predicted' regions. All comparisons between the model and the reference structure determined experimentally (e.g., by X-ray crystallography) rely on an assumption that the reference structure is 100% accurate, which may not always be true. Figure 4 shows that a poorly predicted region in the deformation profile (in red) corresponds to a region with a high B factor and insufficient electron density. One cannot exclude an error in the native structure during the modeling and fitting of the native structure.

***P*-value.** *P*-value represents the confidence that a prediction is significantly different from a randomly generated

RNA 3D structure (23). It was designed as a quality measure for RNA 3D structure prediction resulting from empirical relations for RMSD distribution as a function of RNA length. Therefore, it is independent of the molecule size. *P*-value is capable to differentiate *de novo* algorithms predicting all interactions from those who require to input base-pairing information. Normally, *P*-value lower than 0.01 indicates a successful prediction.

Clash score. Clash score (29) reports serious steric clashes identified in the RNA 3D structure. The score is computed as the number of disallowed (<0.4 Å) overlaps of atom pairs per thousand atoms. All-atom contacts are computed by PROBE (65) that uses van der Waals atom radii and identifies probes intersecting any not-covalently-bonded atom. In general, the existence of interatomic clashes indicates that a local conformation is not stereochemically accurate and should be refined. A high clash score indicates more severe steric clashes. However, clashes can exist also in high-resolution structures. Moreover, even if the global 3D fold of a modeled structure is close to the native one, the clash score value can be quite high when base-base interactions are not accurately reconstructed. Clash score is computed by MolProbity (29) incorporated into RNAQUA.

Mean of circular quantities (MCQ). In the practice of RNA structure modeling, several approaches try to represent the RNA structure with simplified models, such as a network model (66), and reconstruct the RNA 3D structure with standard bond lengths and bond angles. Assuming the standard bond lengths and bond angles are constant values, it is important to understand the accuracy of the torsion angles, which are the only degrees of freedom in the modeling in this context. Therefore, the Mean of Circular Quantities (MCQ) is a metric to compare RNA 3D structures in the torsion angle space. A nucleotide can be described by six torsion angles from the backbone, while the δ dihedral is constrained by the sugar ring (Figure 5A). The residue-wise comparison in the torsion angle space highlights the dissimilarity in local structure. We divide the torsion angle difference into four bins: <15°, 15–30°, 30–60° and >60°. MCQ value <15° means the best similarity, while >60° implies severe structural change. Dissimilar regions can be highlighted on the secondary structure plot by coloring the four bins in gradient color (Figure 5B). MCQ can measure the similarity between whole structures or selected fragments. It also allows multiple models comparison with the reference structure (Figure 5C).

When the reference structure is unknown, clustering the structures to identify consensus structural cores may give biological insights to the folding and function of the RNA structure. MCQ enables structure clustering in the torsion angle space. Pairwise MCQ comparison scores are used as similarity distance and structures can be clustered using the resulted distance matrix (Figure 5D).

Longest continuous segments in torsion angle space (LCS-TA). In the comparison of two RNA 3D structures, LCS-TA (34) identifies the longest continuous segments that display local similarity in the torsion angle space (Figure 5F). Two segments from different structures are considered sim-

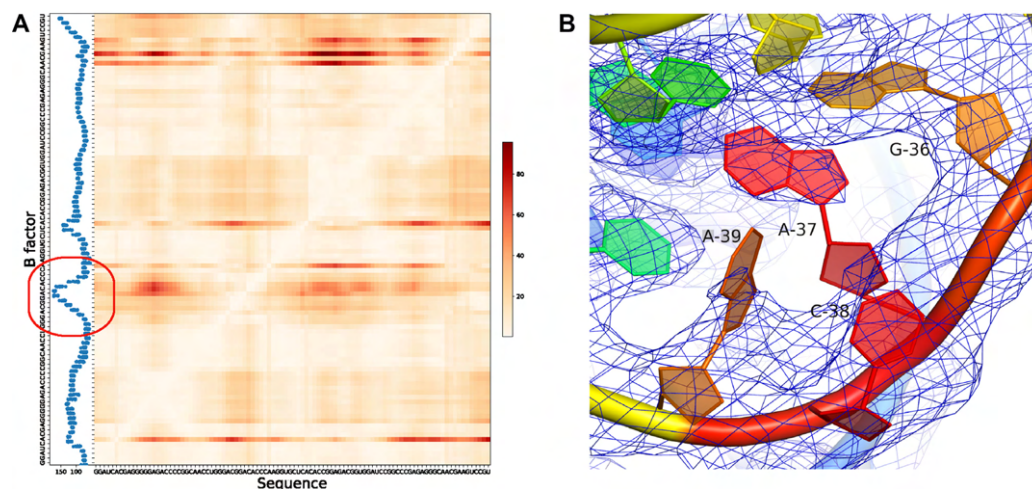


Figure 4. Deformation Profile comparison between predicted structure and reference structure. (A) Deformation Profile heatmap aligned with average B factor histogram, showing the Puzzle 8 (6) solution structure (PDB ID: 4L81) compared to the model 3 predicted by Das lab. (B) Electron density map of the high B factor region, G36–A39, shown in the red circle region in (A). This region is highly mobile, while A37 and A39 do not have a full density in the $2f_o - f_c$ electron density map to support the coordinates proposed by the crystal structure.

ilar if their angular distance (MCQ) does not exceed a predefined MCQ threshold which ranges between 10° and 20° . LCS-TA performs an iterative search using a slide-window approach until the longest continuous segment is found.

The structure comparison performed by LCS-TA can be either independent or dependent on the sequence. Sequence-dependent comparison assumes the same sequence in both the prediction and the reference structure and it finds similar segments with the same sequence. Sequence-independent comparison attempts to perform a structural alignment to identify the longest continuous segments which are similar in torsion angle space ignoring the sequence. In this mode, LCS-TA finds similar fragments with different sequences. When more than one segment is found to be similar in the sequence-independent comparison, all possible segments are listed. LCS-TA is also capable of performing a global comparison: with a fixed MCQ threshold, the prediction model with a longer identified segment has a higher similarity to the reference structure (Figure 5E).

DISCUSSION

The ability to predict RNA 3D structure attracts lots of attention because it opens great opportunities for new developments in biotechnology and basic science. The establishment of RNA-Puzzles boosted the improvement in RNA 3D structure prediction methods, as reported. Furthermore, through active and dynamic collaborations among research groups in RNA-Puzzles (5–7), new ideas were generated, validated and valuable tools were developed and implemented in the past eight years. These tools cover various functions that may be useful for RNA structure formatting, analysis, manipulation, visualization and comparison, which can be used in new exploratory studies.

Although biophysical rules are being learned from the experimentally determined RNA structures, the prediction

of RNA structure is a data-driven problem. Unbiased assessment of a prediction is the key to understand its performance and usability. It is beneficial to have a standard dataset, which can be used to benchmark the performance of a new method against all other prediction approaches. The RNA-Puzzles toolkit directly provides such a benchmark and has been used to demonstrate the accuracy of a novel prediction (46). Although it is possible to run RNA structure prediction programs on other public datasets, such as Rfam and non-redundant dataset (17), RNA-Puzzles prediction stands for the best state-of-the-art blind prediction performance and includes structural diversity. In addition, selecting the top-quality model from a set of models generated by different prediction methods is another important step for an accurate prediction. Our benchmark set has also proved its usability in developing such a scoring model (20). Our datasets can be used as a standard test allowing for methods development and comparison.

Moreover, we provide a unified kit of tools used already by our groups in previous research projects. RNA_format, RNA_normalizer and RNA_assessment were used before to support all calculations in the RNA-Puzzle experiment. The rna-tools package was used in various scientific projects, to calculate stability of various U6 RNAs of the spliceosome (67), to process input files for SimRNAweb (RNA 3D structure prediction method) (58) and NPdock (RNA/DNA-protein docking method) (68), and to analyze data for RNArchitecture database (a classification system of RNA families with a focus on structural information) (69) and EvoClustRNA (RNA 3D structure prediction using multiple sequence alignment information) (61). MCQ-based methods were used *i.a.* to evaluate models in the second (7) and third (6) round of RNA-Puzzles, to identify structural patterns in plant pre-miRNAs (70), to build a database of conformers within the RNAfitme system (71,72). For the first time, we describe these tools and show how they can be

integrated into one robust pipeline giving the users a way to provide a broad perspective on an RNA structure.

The installation of computational tools is non-trivial and can sometimes cost much time even for computational experts. A user-friendly implementation will greatly help the use of a computational tool. Considering that users may have diverse preferences, our resource tools provide both command-line executives and Jupyter Notebook (73) based tutorials, while all the tools are documented. Furthermore, we installed all the tools on a Docker image that can be easily downloaded and launched by the user, in particular, a biologist without programming skills. The Docker image saves the complicated actions required for installing all the tools. Finally, we release all of our datasets and computational tools at GitHub, which can be continuously updated if any bugs are detected. The ‘fork’ function of Github also facilitates novel computational methods or datasets being developed based on our resource, i.e. RNA-ligand interaction prediction.

The Jupyter Notebook (74) workflow in the resource provides a standard example for RNA structure prediction evaluation. Jupyter Notebook is an open-source web application that allows users to create and share documents that contain live code, equations, visualizations, and explanatory text. The tools implemented in the toolkit can be imported to such notebooks to create reproducible analyses that can be uploaded online and shared with the RNA structural bioinformatics community. One example of such analysis was described in the Result section for rna-tools. This approach of describing RNA bioinformatic analyses should help scientists to share their pipelines, e.g., protocols used for modeling in RNA-Puzzles, that can be later reproduced and/or improved by others. And since the Jupyter Notebook has support for over 40 programming languages, including those popular in Data Science such as Python, R, Julia and Scala, this is a great approach to incorporate the toolkit into pipelines written in other languages. In this way, all the RNA structure analysis work can be efficiently shared and reproduced. In addition, RNAQUA provides all the RNA structure comparison tools as a web service, which can alleviate the burden of software installation for non-computationally oriented users.

RNA structure comparison metrics have been developed since a decade ago (24). The availability of these metrics as computational tools is limited and not systematic, which highlights the importance of our toolkit. We also share every detail in a standard workflow accepted by the RNA-Puzzles community, i.e., when multiple structures have been solved for the same sequence, it is fair to consider all of them as native structures and use the nearest one to the prediction as the reference. Secondary structure analysis and visualization are useful aspects in understanding RNA 3D structure: rna-tools implements the easy transformation from 3D structure visualization in PyMOL(75) to 2D structure contacts annotation, thus enabling the intuitive comprehension from the biophysics aspects.

Our resource brings various tools and datasets into one unified resource that can be easily downloaded and used by biologists interested in RNA 3D structure prediction and analysis. We think that the toolkit with its open code should be considered as a library of functions and tools rather than

a complete package with a fixed set of functionalities. The toolkit is a framework of various functions. The users are invited to extend it with their scripts on the top of the existing tools. In this way, it is possible to adapt our tools for future cases. For example, to have a particular wrapper or variant of tools that can be used for a very specific application saving time and brainpower of the user to write the code from scratch. We believe that the RNA-Puzzle Toolkit will prompt new advances in the applications of the RNA 3D structure prediction and in method development.

DATA AVAILABILITY

All the datasets, computational tools, and related documentation are available as open-source at <https://github.com/RNA-Puzzles>.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

Authors contributions: M.M. developed rna-tools, combined the workflow, created the *standardized_dataset* and wrote parts of the manuscript. M.A. developed and tested RNAQUA, and wrote parts of the manuscript, P.L. supervised the development of RNAQUA, T.Z. implemented MCQ algorithm, J.W. and T.Z. implemented LCS-TA. Y.C. provided technical support to web development. J.M.B. supervised the development of rna-tools. M.S. developed a concept for MCQ and LCS-TA algorithms, supervised their implementation and wrote parts of the manuscript. E.W. supervised the entire project and drafted the manuscript. Z.M. conceived the project, cleaned up the datasets, implemented the RNA structure format tool, format check tool and evaluation metrics (INF, DP), designed the website and wrote parts of the manuscript. All authors contributed to manuscript preparation.

FUNDING

Polish National Science Centre [NCN, 2015/17/N/NZ2/03360 to M.M., 2016/23/B/ST6/03931 to M.S., 2016/23/N/ST6/03779 to T.Ż., 2017/26/A/NZ 1/01083 to J.M.B.]; Foundation for Polish Science co-financed by the European Union under the European Regional Development Fund [MAB/20172 carried out within the International Research Agendas Program]; Institute of Computing Science, Poznan University of Technology [09/91/SBAD/0681 to M.A.]; Single Cell Gene Expression Atlas grant from the Wellcome Trust [108437/Z/15/Z]. Funding for open access charge: Shanghai Fourth People’s Hospital.

Conflict of interest statement. None declared.

REFERENCES

1. Levitt, M. (1969) Detailed molecular model for transfer ribonucleic acid. *Nature*, **224**, 759–763.
2. Miao, Z. and Westhof, E. (2017) RNA structure: advances and assessment of 3D structure prediction. *Annu. Rev. Biophys.*, **46**, 483–503.

3. Dawson,W.K. and Bujnicki,J.M. (2016) Computational modeling of RNA 3D structures and interactions. *Curr. Opin. Struct. Biol.*, **37**, 22–28.
4. Moulton,J., Fidelis,K., Kryshchuk,A., Schwede,T. and Tramontano,A. (2018) Critical assessment of methods of protein structure prediction (CASP)-Round XII. *Proteins*, **86**, 7–15.
5. Cruz,J.A., Blanchet,M.-F., Boniecki,M., Bujnicki,J.M., Chen,S.-J., Cao,S., Das,R., Ding,F., Dokholyan,N.V., Flores,S.C. *et al.* (2012) RNA-Puzzles: a CASP-like evaluation of RNA three-dimensional structure prediction. *RNA*, **18**, 610–625.
6. Miao,Z., Adamiak,R.W., Antczak,M., Batey,R.T., Becka,A.J., Biesiada,M., Boniecki,M.J., Bujnicki,J.M., Chen,S.-J., Cheng,C.Y. *et al.* (2017) RNA-Puzzles Round III: 3D RNA structure prediction of five riboswitches and one ribozyme. *RNA*, **23**, 655–672.
7. Miao,Z., Adamiak,R.W., Blanchet,M.-F., Boniecki,M., Bujnicki,J.M., Chen,S.-J., Cheng,C., Chojnowski,G., Chou,F.-C., Cordero,P. *et al.* (2015) RNA-Puzzles Round II: assessment of RNA structure prediction programs applied to three large RNA structures. *RNA*, **21**, 1066–1084.
8. Noller,H.F. and Woese,C.R. (1981) Secondary structure of 16S ribosomal RNA. *Science*, **212**, 403–411.
9. Haas,E., Morse,D., Brown,J., Schmidt,F. and Pace,N. (1991) Long-range structure in ribonuclease P RNA. *Science*, **254**, 853–856.
10. Leontis,N.B. and Westhof,E. (2001) Geometric nomenclature and classification of RNA base pairs. *RNA*, **7**, 499–512.
11. Cruz,J.A. and Westhof,E. (2011) Sequence-based identification of 3D structural modules in RNA with RMDetect. *Nat. Methods*, **8**, 513–521.
12. Kucharik,M., Hofacker,I.L., Stadler,P.F. and Qin,J. (2016) Pseudoknots in RNA folding landscapes. *Bioinformatics*, **32**, 187–194.
13. Michel,F. and Westhof,E. (1990) Modelling of the three-dimensional architecture of group I catalytic introns based on comparative sequence analysis. *J. Mol. Biol.*, **216**, 585–610.
14. Brunel,C., Romby,P., Westhof,E., Ehresmann,C. and Ehresmann,B. (1991) Three-dimensional model of Escherichia coli ribosomal 5 S RNA as deduced from structure probing in solution and computer modeling. *J. Mol. Biol.*, **221**, 293–308.
15. Westhof,E., Romby,P., Romaniuk,P.J., Ebel,J.P., Ehresmann,C. and Ehresmann,B. (1989) Computer modeling from solution data of spinach chloroplast and of Xenopus laevis somatic and oocyte 5 S rRNAs. *J. Mol. Biol.*, **207**, 417–431.
16. Rychlewski,L. and Fischer,D. (2005) LiveBench-8: the large-scale, continuous assessment of automated protein structure prediction. *Protein Sci.*, **14**, 240–245.
17. Leontis,N.B. and Zirbel,C.L. (2012) Nonredundant 3D structure datasets for RNA knowledge extraction and benchmarking. *Nucleic Acids Mol. Biol.*, **27**, 281–298.
18. Weinreb,C., Riesselman,A.J., Ingraham,J.B., Gross,T., Sander,C. and Marks,D.S. (2016) 3D RNA and functional interactions from evolutionary couplings. *Cell*, **165**, 963–975.
19. Suslov,N.B., DasGupta,S., Huang,H., Fuller,J.R., Lilley,D.M.J., Rice,P.A. and Piccirilli,J.A. (2015) Crystal structure of the Varkud satellite ribozyme. *Nat. Chem. Biol.*, **11**, 840–846.
20. Li,J., Zhu,W., Wang,J., Li,W., Gong,S., Zhang,J. and Wang,W. (2018) RNA3DCNN: Local and global quality assessments of RNA 3D structures using 3D deep convolutional neural networks. *PLoS Comput. Biol.*, **14**, e1006514.
21. Antczak,M., Popena,M., Zok,T., Sarzynska,J., Ratajczak,T., Tomczyk,K., Adamiak,R.W. and Szachniuk,M. (2016) New functionality of RNAComposer: an application to shape the axis of miR160 precursor structure. *Acta Biochim. Pol.*, **63**, 737–744.
22. Cock,P.J.A., Antao,T., Chang,J.T., Chapman,B.A., Cox,C.J., Dalke,A., Friedberg,I., Hamelryck,T., Kauff,F., Wilczynski,B. *et al.* (2009) Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics*, **25**, 1422–1423.
23. Hajdin,C.E., Ding,F., Dokholyan,N.V. and Weeks,K.M. (2010) On the significance of an RNA tertiary structure prediction. *RNA*, **16**, 1340–1349.
24. Parisien,M., Cruz,J.A., Westhof,E. and Major,F. (2009) New metrics for comparing and assessing discrepancies between RNA 3D structures and models. *RNA*, **15**, 1875–1885.
25. Gendron,P., Lemieux,S. and Major,F. (2001) Quantitative analysis of nucleic acid three-dimensional structures. *J. Mol. Biol.*, **308**, 919–936.
26. Oliphant,T.E. (2006) *A Guide to NumPy. USA: Trelgol Publishing.* <https://www.scipy.org/citing.html>.
27. Rother,M., Rother,K., Puton,T. and Bujnicki,J.M. (2011) ModeRNA: a tool for comparative modeling of RNA 3D structure. *Nucleic Acids Res.*, **39**, 4007–4022.
28. Waleń,T., Chojnowski,G., Gierski,P. and Bujnicki,J.M. (2014) ClaRNA: a classifier of contacts in RNA 3D structures based on a comparative analysis of various classification schemes. *Nucleic Acids Res.*, **42**, e151.
29. Davis,I.W., Leaver-Fay,A., Chen,V.B., Block,J.N., Kapral,G.J., Wang,X., Murray,L.W., Arendall,W.B. 3rd, Snoeyink,J., Richardson,J.S. *et al.* (2007) MolProbity: all-atom contacts and structure validation for proteins and nucleic acids. *Nucleic Acids Res.*, **35**, W375–W383.
30. Lukasiak,P., Antczak,M., Ratajczak,T., Bujnicki,J.M., Szachniuk,M., Adamiak,R.W., Popena,M. and Blazewicz,J. (2013) RNAnalyzer—novel approach for quality analysis of RNA structural models. *Nucleic Acids Res.*, **41**, 5978–5990.
31. Lukasiak,P., Antczak,M., Ratajczak,T., Szachniuk,M., Popena,M., Adamiak,R.W. and Blazewicz,J. (2015) RNAssess—a web server for quality assessment of RNA 3D structures. *Nucleic Acids Res.*, **43**, W502–W506.
32. Szachniuk,M. (2019) RNAPolis: computational platform for RNA structure analysis. *Found. Comput. Decision Sci.*, **44**, 241–257.
33. Zok,T., Popena,M. and Szachniuk,M. (2014) MCQ4Structures to compute similarity of molecule structures. *Central Eur. J. Oper. Res.*, **22**, 457–473.
34. Wiedemann,J., Zok,T., Milostan,M. and Szachniuk,M. (2017) LCS-TA to identify similar fragments in RNA 3D structures. *BMC Bioinformatics*, **18**, 456.
35. Dibrov,S.M., McLean,J., Parsons,J. and Hermann,T. (2011) Self-assembling RNA square. *Proc. Natl. Acad. Sci. U.S.A.*, **108**, 6405–6408.
36. Ren,A., Vušurović,N., Gebetsberger,J., Gao,P., Juen,M., Kreutz,C., Patel,D.J. (2016) Pistol ribozyme adopts a pseudoknot fold facilitating site-specific in-line cleavage. *Nat. Chem. Biol.*, **12**, 702–708.
37. Baird,N.J., Zhang,J., Hamma,T. and Ferre-D'Amare,A.R. (2012) YbxF and YlxQ are bacterial homologs of L7Ae and bind K-turns but not K-loops. *RNA*, **18**, 759–770.
38. Peselis,A. and Serganov,A. (2012) Structural insights into ligand binding and gene expression control by an adenosylcobalamin riboswitch. *Nat. Struct. Mol. Biol.*, **19**, 1182–1184.
39. Zhang,J. and Ferré-D'Amare,A.R. (2013) Co-crystal structure of a T-box riboswitch stem I domain in complex with its cognate tRNA. *Nature*, **500**, 363–366.
40. Ren,A., Xue,Y., Peselis,A., Serganov,A., Al-Hashimi,H.M. and Patel,D.J. (2015) Structural and dynamic basis for low-affinity, high-selectivity binding of L-glutamine by the glutamine riboswitch. *Cell Rep.*, **13**, 1800–1813.
41. Watkins,A.M. and Das,R. (2019) FARFAR2: Improved de novo Rosetta prediction of complex global RNA folds. bioRxiv doi: <https://doi.org/10.1101/764449>, 10 September 2019, preprint: not peer reviewed.
42. Boniecki,M.J., Lach,G., Dawson,W.K., Tomala,K., Lukasz,P., Soltysinski,T., Rother,K.M. and Bujnicki,J.M. (2016) SimRNA: a coarse-grained method for RNA folding simulations and 3D structure prediction. *Nucleic Acids Res.*, **44**, e63.
43. Cheng,C.Y., Chou,F.-C. and Das,R. (2015) Modeling complex RNA tertiary folds with Rosetta. *Methods Enzymol.*, **553**, 35–64.
44. Sharma,S., Ding,F. and Dokholyan,N.V. (2008) iFoldRNA: three-dimensional RNA structure prediction and folding. *Bioinformatics*, **24**, 1951–1952.
45. Zhao,C., Xu,X. and Chen,S.-J. (2017) Predicting RNA Structure with Vfold. *Methods Mol. Biol.*, **1654**, 3–15.
46. Watkins,A.M., Geniesse,C., Kladwang,W., Zakrevsky,P., Jaeger,L. and Das,R. (2018) Blind prediction of noncanonical RNA structure at atomic accuracy. *Sci Adv.*, **4**, eaar5316.
47. Kerpedjiev,P., Siederdisen,Höner Zu and Hofacker,I.L. (2015) Predicting RNA 3D structure using a coarse-grain helix-centered model. *RNA*, **21**, 1110–1121.
48. Capriotti,E., Norambuena,T., Marti-Renom,M.A. and Melo,F. (2011) All-atom knowledge-based potential for RNA structure prediction and assessment. *Bioinformatics*, **27**, 1086–1093.

49. Bernauer, J., Huang, X., Sim, A.Y.L. and Levitt, M. (2011) Fully differentiable coarse-grained and all-atom knowledge-based potentials for RNA structure evaluation. *RNA*, **17**, 1066–1075.
50. Altschul, S.F., Gish, W., Miller, W., Myers, E.W. and Lipman, D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.
51. Nawrocki, E.P. and Eddy, S.R. (2013) Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics*, **29**, 2933–2935.
52. Kalvari, I., Argasinska, J., Quinones-Olvera, N., Nawrocki, E.P., Rivas, E., Eddy, S.R., Bateman, A., Finn, R.D. and Petrov, A.I. (2018) Rfam 13.0: shifting to a genome-centric resource for non-coding RNA families. *Nucleic Acids Res.*, **46**, D335–D342.
53. Parisien, M. and Major, F. (2008) The MC-Fold and MC-Sym pipeline infers RNA structure from sequence data. *Nature*, **452**, 51–55.
54. Sato, K., Kato, Y., Hamada, M., Akutsu, T. and Asai, K. (2011) IPknot: fast and accurate prediction of RNA secondary structures with pseudoknots using integer programming. *Bioinformatics*, **27**, i85–93.
55. Darty, K., Denise, A. and Ponty, Y. (2009) VARNA: Interactive drawing and editing of the RNA secondary structure. *Bioinformatics*, **25**, 1974–1975.
56. Lai, D., Proctor, J.R., Zhu, J.Y.A. and Meyer, I.M. (2012) R-CHIE: a web server and R package for visualizing RNA secondary structures. *Nucleic Acids Res.*, **40**, e95.
57. Hanson, R.M. and Lu, X.-J. (2017) DSSR-enhanced visualization of nucleic acid structures in Jmol. *Nucleic Acids Res.*, **45**, W528–W533.
58. Magnus, M., Boniecki, M.J., Dawson, W. and Bujnicki, J.M. (2016) SimRNAweb: a web server for RNA 3D structure modeling with optional restraints. *Nucleic Acids Res.*, **44**, W315–W319.
59. Das, R. and Baker, D. (2007) Automated de novo prediction of native-like RNA tertiary structures. *Proc. Natl. Acad. Sci. U.S.A.*, **104**, 14664–14669.
60. Stasiewicz, J., Mukherjee, S., Nithin, C. and Bujnicki, J.M. (2019) QRNAS: software tool for refinement of nucleic acid structures. *BMC Struct. Biol.*, **19**, 5.
61. Magnus, M., Kappel, K., Das, R. and Bujnicki, J.M. (2019) RNA 3D structure prediction guided by independent folding of homologous sequences. *BMC Bioinformatics*, **20**, 512.
62. Liu, Y., Wilson, T.J. and Lilley, D.M.J. (2017) The structure of a nucleolytic ribozyme that employs a catalytic metal ion. *Nat. Chem. Biol.*, **13**, 508–513.
63. Chapman, E.G., Costantino, D.A., Rabe, J.L., Moon, S.L., Wilusz, J., Nix, J.C. and Kieft, J.S. (2014) The structural basis of pathogenic subgenomic flavivirus RNA (sfRNA) production. *Science*, **344**, 307–310.
64. Lu, X.-J. and Olson, W.K. (2003) 3DNA: a software package for the analysis, rebuilding and visualization of three-dimensional nucleic acid structures. *Nucleic Acids Res.*, **31**, 5108–5121.
65. Word, J.M., Lovell, S.C., LaBean, T.H., Taylor, H.C., Zalis, M.E., Presley, B.K., Richardson, J.S. and Richardson, D.C. (1999) Visualizing and quantifying molecular goodness-of-fit: small-probe contact dots with explicit hydrogen atoms. *J. Mol. Biol.*, **285**, 1711–1733.
66. Kim, N., Petingi, L. and Schlick, T. (2013) Network theory tools for RNA Modeling. *WSEAS Trans. Math.*, **9**, 941–955.
67. Eysmont, K., Matylla-Kulińska, K., Jaskulska, A., Magnus, M. and Konarska, M.M. (2019) Rearrangements within the U6 snRNA core during the transition between the two catalytic steps of splicing. *Mol. Cell*, **75**, 538–548.
68. Tuszyńska, I., Magnus, M. and Jonak, K. (2015) NPdock: a web server for protein–nucleic acid docking. *Nucleic Acids Res.*, **43**, W425–W430.
69. Boccaletto, P., Magnus, M., Almeida, C., Zyla, A., Astha, A., Pluta, R., Baginski, B., Jankowska, E., Dunin-Horkawicz, S., Wirecki, T.K. et al. (2018) RNArchitecture: a database and a classification system of RNA families, with a focus on structural information. *Nucleic Acids Res.*, **46**, D202–D205.
70. Miskiewicz, J., Tomczyk, K., Mickiewicz, A., Sarzynska, J. and Szachniuk, M. (2017) Bioinformatics study of structural patterns in plant MicroRNA precursors. *Biomed. Res. Int.*, **2017**, 6783010.
71. Zok, T., Antczak, M., Riedel, M., Nebel, D., Villmann, T., Lukasiak, P., Blazewicz, J. and Szachniuk, M. (2015) Building the library of RNA 3D nucleotide conformations using the clustering approach. *Int. J. Appl. Math. Comput. Sci.*, **25**, 689–700.
72. Antczak, M., Zok, T., Osowiecki, M., Popenda, M., Adamiak, R.W. and Szachniuk, M. (2018) RNAfitme: a webserver for modeling nucleobase and nucleoside residue conformation in fixed-backbone RNA structures. *BMC Bioinformatics*, **19**, 304.
73. Yakimchik, A.I. (2019) Jupyter Notebook: a system for interactive scientific computing. *Geofizicheskiy Zhurnal*, **41**, 121.
74. Basu, A. Reproducible research with jupyter notebooks. *Authorea*, doi:10.22541/au.151460905.57485984.
75. Rigsby, R.E. and Parker, A.B. (2016) Using the PyMOL application to reinforce visual understanding of protein structure. *Biochem. Mol. Biol. Educ.*, **44**, 433–437.

Databases and ontologies

RNAloops: a database of RNA multiloops

Jakub Wiedemann¹, Jacek Kaczor¹, Maciej Milostan^{1,2}, Tomasz Zok^{1,2},
Jacek Blazewicz^{1,3}, Marta Szachniuk^{1,3,*} and Maciej Antczak^{1,3,*}

¹Institute of Computing Science, Poznan University of Technology, 60-965 Poznan, Poland, ²Poznan Supercomputing and Networking Center, 61-131 Poznan, Poland and ³Institute of Bioorganic Chemistry, Polish Academy of Sciences, 61-704 Poznan, Poland

*To whom correspondence should be addressed.

Associate Editor: Peter Robinson

Received on April 5, 2022; revised on June 26, 2022; editorial decision on July 5, 2022; accepted on July 6, 2022

Abstract

Motivation: Knowledge of the 3D structure of RNA supports discovering its functions and is crucial for designing drugs and modern therapeutic solutions. Thus, much attention is devoted to experimental determination and computational prediction targeting the global fold of RNA and its local substructures. The latter include multi-branched loops—functionally significant elements that highly affect the spatial shape of the entire molecule. Unfortunately, their computational modeling constitutes a weak point of structural bioinformatics. A remedy for this is in collecting these motifs and analyzing their features.

Results: RNAloops is a self-updating database that stores multi-branched loops identified in the PDB-deposited RNA structures. A description of each loop includes angular data—planar and Euler angles computed between pairs of adjacent helices to allow studying their mutual arrangement in space. The system enables search and analysis of multiloops, presents their structure details numerically and visually, and computes data statistics.

Availability and implementation: RNAloops is freely accessible at <https://rnaloops.cs.put.poznan.pl>.

Contact: mszachniuk@cs.put.poznan.pl or mantczak@cs.put.poznan.pl

Supplementary information: [Supplementary data](#) are available at *Bioinformatics* online.

1 Introduction

RNA molecules play a significant role in the functioning of living organisms and viruses. They carry out a broad range of functions—from translating genetic information through regulating the activity of genes to catalyzing biochemical reactions. Their participation in diverse processes has made them the center of researchers' interest for many years (Berg *et al.*, 2002; Miskiewicz *et al.*, 2017). In particular, studies focus on the structure of RNA molecules, trying to bridge the gap between knowledge of the sequences (Kudla *et al.*, 2021; O'Leary *et al.*, 2015; Zok *et al.*, 2022) and how they fold in space (Berman *et al.*, 2000; Blazewicz *et al.*, 2005; Wiedemann and Milostan, 2017; Zemora and Waldsich, 2010). In recent years, *in silico* methods for RNA 3D structure prediction have increasingly supported this research by generating spatial prototypes of various RNA molecules (Li *et al.*, 2020). Still, many computationally generated models are far from their native counterparts, as can be observed in subsequent RNA-Puzzles challenges (Cruz *et al.*, 2012; Miao *et al.*, 2015). A detailed analysis of their results allows the identification of weaknesses of the prediction methods (Carrasco *et al.*, 2022; Lukasiak *et al.*, 2015; Miao *et al.*, 2015; Popena *et al.*, 2021). They include modeling non-canonical base-pairs, long-range interactions, or selected structure motifs, like *n*-way junctions, also known as multiloops (Laing and Schlick, 2010; Parlea *et al.*, 2016; Rybarczyk *et al.*, 2015; Zuker and Sankoff, 1984).

n-Way junction in the RNA structure is an internal loop with *n* outgoing helices, where $n \geq 3$. The size of this motif, its 3D shape, and directions of outgoing stems determine the spatial arrangement of various structural elements in the molecule and significantly affect its general fold (Bailor *et al.*, 2011; Hao and Kieft, 2016; Lamiable *et al.*, 2012; Leontis and Westhof, 1998; Lescaute and Westhof, 2006; Parlea *et al.*, 2016; Westhof *et al.*, 1996; Zhao *et al.*, 2012). Our knowledge of multiloops comes primarily from experimentally determined structures deposited in the Protein Data Bank (Berman *et al.*, 2000). Partly, it is also available through databases dedicated to RNA fragments and motifs, such as RNA FRABASE (Popena *et al.*, 2008), RNA Bricks (Chojnowski *et al.*, 2014), RNA 3D motif atlas (Parlea *et al.*, 2016; Petrov *et al.*, 2013), or RAG-3D (Zahran *et al.*, 2015). These computational resources catalog a wide range of structural elements described with the details common to all motifs' primary, secondary, and tertiary structures. Multiloops themselves are collected in the RNAJunction database (Bindewald *et al.*, 2008). It stores over 12 000 junctions and kissing loops with annotations covering PDB ID, sequence, tertiary structure, and inter-helix angles and allows searching by PDB ID or RNA sequence. Unfortunately, the database was not updated after 2008 and therefore contains multiloops derived from <30% of the RNA structures currently deposited in PDB. As a result, no complete repository of *n*-way junctions or their efficient and precise search engine exists. Available bioinformatics systems do not collect multiloop-specific up-to-date

data from experimental and computational studies (Bailor *et al.*, 2011; Byron *et al.*, 2013; Hao and Kieft, 2016; Hohng *et al.*, 2004; Hua *et al.*, 2016; Laing *et al.*, 2012; Lescoute and Westhof, 2006). It makes comparative analysis and accurate modeling of these structural motifs difficult or nearly impossible.

Here, we present RNAloops, a database of multi-branched loops identified in the experimental RNA structures from the Protein Data Bank (Berman *et al.*, 2000). The data collected include, i.a. RNA sequence, secondary and tertiary structures, planar and Euler angles (Diebel, 2006) to describe the relationship between outgoing helices. The repository self-updates automatically every week. RNAloops comes with a handy mechanism to query the database contents based on several criteria, for example RNA sequence, secondary structure, the number of branches, and ranges of angle values. It automatically collects statistics about the data in the database and presents them in a user-friendly way. The output is available in text, numeric and graphical form. Retrieved multiloop structures are ready to apply in modeling topologically complex RNAs by the template- and fragment assembly-based prediction methods. They can be used to create learning sets for machine learning-oriented predictors (Townshend *et al.*, 2021), thus, complementing data from the other resources developed for this task (Adamczyk *et al.*, 2022; Becquey *et al.*, 2021). Finally, angular values can control the energy minimization process by constraining the arrangement of branching helices. We believe that due to the systematic collecting of all multiloop-specific data, RNAloops will contribute to improving RNA structure study and modeling.

2 Materials and methods

2.1 Data acquisition into the database

Every week, the RNAloops repository updates with new data taken from the Protein Data Bank (Berman *et al.*, 2000) and supplemented with additional information. The process (see Fig. 1) starts by retrieving PDB IDs of newly deposited, removed, or updated RNA 3D structures. The system downloads the corresponding PDB files in mmCIF format (Bourne *et al.*, 1997). Their contents are standardized using the Biopython functions (Cock *et al.*, 2009): the first model is taken in the case of multimodel files, non-RNA chains and incomplete residues are filtered-out, modified residues are transformed into their non-modified equivalents.

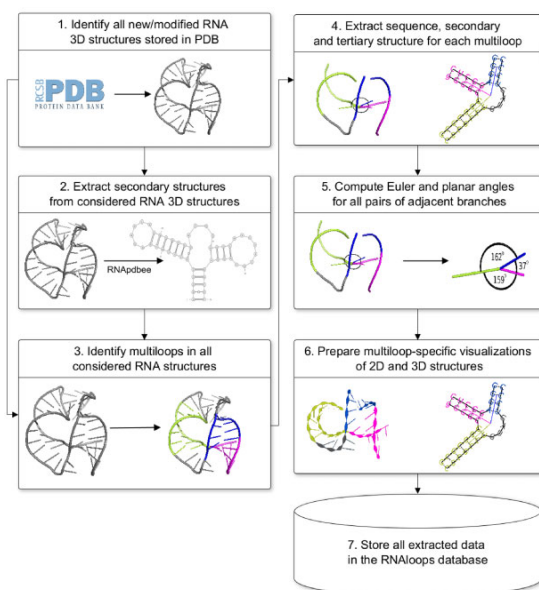


Fig. 1. Data flow in the RNAloops database

The secondary structure is derived for each RNA and encoded in extended dot-bracket notation using the RNApdbee algorithms (Antczak *et al.*, 2018a; Popenda *et al.*, 2008). This structure representation is scanned for n -way junctions ($n \geq 3$), taking pseudoknot interactions into account. 2D and 3D structures of each identified motif are extracted and uploaded into the database. In the RNAloops system, the structure of a multiloop is described by the loop and the outgoing full-length helices.

The mutual positions of all pairs of adjacent helices protruding from the loop are designated for each multiloop. For this purpose, planar and Euler angles are computed between directional vectors of these helices. The beginning and the end of each vector are in the geometric centers of the multiloop and helix, respectively. The first point is the centroid in the set of all non-hydrogen atoms that belong to the first base pairs of outgoing helices. The second point is based on all non-hydrogens from the third base pair in the helix or the first pair if the helix has < 3 bp. Planar angle ϕ (Fig. 2) is computed according to Equation (1) between two directional vectors, \vec{a} and \vec{b} , projected onto the plane. Euler angles, α , β , and γ (Fig. 2), reflect the orientation of a directional vector to the other. They define rotations to be made about the three coordinate axes to superimpose two helices (Heyde and Wood, 2020). The helix-representing vectors, \vec{a} and \vec{b} , are projected onto the planes perpendicular to all axes of the coordinate system. An angle between the vectors computes from Equation (1) separately for each dimension.

$$\phi = \arccos[(\vec{a} \cdot \vec{b}) / (|\vec{a}| \cdot |\vec{b}|)] \quad (1)$$

2.2 The RNAloops system implementation

The RNAloops system consists of a frontend layer providing a user interface, a backend with RESTful API, and the database management and update service. The interface uses React.js and Next.js frameworks and retrieves the searched data via RESTful API.

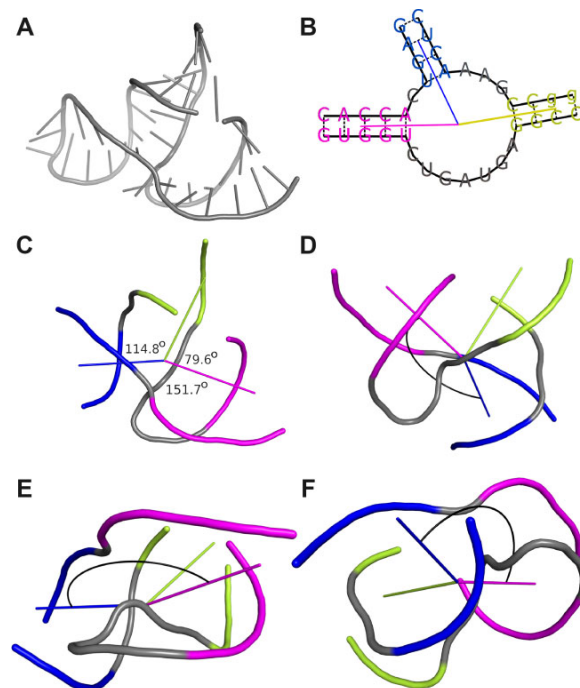


Fig. 2. (A) The 3D structure of hammerhead ribozyme (PDB ID: 1NYI; Dunham *et al.*, 2003). (B) The 2D diagram of the three-way junction identified in this structure with the directional vectors plotted. (C) The 3D model of this junction with planar angle values displayed. Euler angles between the blue and the magenta helix shown from the perspective of the (D) X, (E) Y, and (F) Z axes of the coordinate system, respectively

The backend layer executes all user-initiated operations and communicates with the database management and update services. The relational database of RNAloops, operating on PostgreSQL DBMS, is automatically updated weekly. The system is hosted and maintained by the Institute of Computing Science, Poznan University of Technology, Poland.

3 Results

3.1 Database content

Currently (as of March 31, 2022), RNAloops stores entries for 84 256 multiloops identified in 1831 RNAs from the Protein Data Bank (Berman et al., 2000). We obtained these by processing 5729 RNA-containing structures, that is stand-alone RNAs, RNAs derived from protein–RNA complexes, and DNA–RNA hybrids. Sixty-eight percent of all RNAs examined had no multiloops, whereas 32% contained at least one—these populated the database. They came from structures determined by X-ray (42.6%), fiber diffraction (0.2%), and electron microscopy methods (57.2%). For each of these molecules, we counted how many multiloops it included. The highest proportion (16.37%) is RNAs having exactly two n -way junctions. Structures containing four multiloops constitute 7.14% of the dataset, with 22, 51, and 76 multiloops—3.57%, 1.95%, and 2.23%, respectively. The percentage of structures with other numbers of loops oscillates (for each count) around 0–1%. The collection also includes some RNAs having >100 multiloops. An analysis of n -way junction multiplicity shows the highest number of three-way junctions—they account for nearly 40% of the total collection. The branching multiplicity per multiloop ranges between 3 and 14, but only 0.44% of the motifs have 14 branches. Figure 3 shows the distribution of branching multiplicities in the database.

3.2 User interface

RNAloops operates via a web application in any modern web browser. To run it, users open the address <https://rnaloops.cs.put.poznan.pl>.

The interface consists of five pages: *Home*, *Search result*, *Help*, *Statistics*, and *Cite us*. Four of them are visible by default. The *Home* page enables defining the query to search for data on RNA multiloops. *Help* explains all the options of web application. In *Statistics*, users can see stats of current database contents with charts showing data growth and distribution—the total number of RNA structures and multiloops, the number of multiloops by topology, and multiloops by topologies grouped by experimental method or PDB IDs. Statistical data are recomputed automatically after each database update. The *Cite us* page informs about the RNAloops-related publications. The *Search result* page displays output data and gets visible when the search completes.

3.2.1 Search modes

The search engine works in two modes: basic and auxiliary. By default, the tab with basic mode opens when entering the RNAloops homepage. It allows defining several search criteria: PDB ID(s),

number of branches in a multiloop (specified as a range or exact number), sequence pattern, secondary structure in dot-bracket notation, the range of planar angle values, and ranges of Euler angle values. When searching with an angle criterion, the system returns any multiloop in which at least one angle satisfies the criterion. If several search criteria are defined, the system combines them into a single query and looks for motifs meeting their conjunction. If the users do not enter any criteria and click the *Search* button, RNAloops outputs the list of all records in the database. The auxiliary mode allows scanning the database for RNAs containing the specified number of n -way junctions. In both modes, users can search hierarchically—the subset resulting from one search can be searched further using the basic mode criteria.

3.2.2 Search results

Basic search outputs the number of items found and their collection divided into pages. By default, each result page displays 10 item tiles. Page capacity is user-adjustable. Each tile contains a thumbnail diagram of the multiloop secondary structure, the type of the loop, and the PDB identifier of the source structure. A detailed description presents by clicking *Show details*, in four sections. The *General information* panel contains the multiloop type, a clickable identifier of the source structure linked to the PDB, sequence, and secondary structure in the dot-bracket notation. The file icon in the top right corner allows downloading the PDF file with general information and the details of all components of the multiloop. The right-positioned panel includes sections with the information on multiloop units—branching helices (sequence, length as the number of base pairs, source structure-derived residue numbering), planar and Euler angles between them, and single-stranded connectors (sequence, length as the number of residues, source structure-derived residue numbering). The secondary structure is displayed in two views—with and without directional vectors. The first one (default) is generated using RNAplot (Lorenz et al., 2011) and colored to ease distinguishing multiloop components. To keep clarity, RNAplot does not draw directional vectors if pseudoknotted base pairs are part of the loop. The second view is prepared with VARNA (Darty et al., 2009). Both diagrams are available for download in the Scalable Vector Graphics (SVG) format. The interactive view of the 3D structure is generated using the LiteMol library (Sehnal et al., 2017). The top three buttons allow switching between the structure of the multiloop itself, the source PDB structure, and the source structure with the multiloop highlighted in a different color. Clicking the gear in the 3D window displays the settings panel to manipulate the display parameters. Users can download the 3D structure in mmCIF format (Bourne et al., 1997).

The auxiliary search outputs the number of RNAs found and their list divided into pages. Each structure is described by PDB ID, PDB record title, resolution, experimental method, number of multiloops, and their types. By clicking the item, users get the result page as in the basic search. However, it displays only multiloops included in the selected structure. These results can be processed just like the output from the basic search.

3.3 RNAloops applications

RNAloops enables multi-parametric structural analysis and a search for multiloops meeting user-defined criteria. Such functionalities support, i.a. extracting significant features of structure motifs, their comparative analysis, or 3D structure modeling. Below we present sample applications of the system for three problems—RNA design, determining the spatial shape of an RNA motif based on the similarity of its secondary structure to experimental structures, and homology modeling.

In the first example, we tackled the RNA design problem. It aims to identify RNA sequence(s) that fold to a predefined secondary structure. Here, we targeted the four-way junction discussed in (Ivry et al., 2009). Given the dot-bracket representation of this multiloop secondary structure— $(\cdot(-)\cdot(-)\cdot(-))$ —we used the RNAloops search facility to find RNA sequences that could form a loop like this. We ran the search applying strand shifts to include various orientations

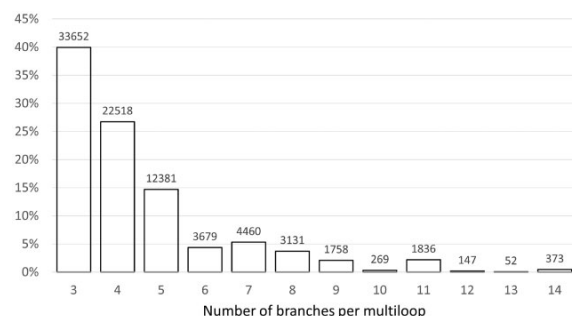


Fig. 3. Coverage of the RNAloops dataset by n -way junctions for $n \in \{3, 4, \dots, 14\}$

Table 1. Results of the RNAloops search aimed to find RNA sequences that fold to the target structure

Target secondary structure: $(-)(-)(-)(-)$				
Query	Matching four-way junctions found in RNAloops	Secondary structure	Sequence	Score
	PDB ID(s)			
$(-)(-)$	5wlc	$((((-((-)))((-)(-)))$	UGCAGCUC-GAGAGG-CCAC-GGGCA	23
$(-)(-)($	4v3p	$((((-)(-)(-)((-)))$	UCGGG-CUU-AGCAC-GUGUCCGG	21
$(-)(-)$	6spf	$((..((-)(-))..((-)(-))$	ACGAAGGU-GCUUGACU-AGGU-AGGU	18
$(-)(-)($	6q98	$(((((((-)))((-)(-)))$	UUGGCCGG-CCGGUUU-GAC-GCAA	18
$(-)(-)($	4u27, 4u1v, 4wf1, 4u25, 4u26, 4u24	$(((((((-)))((-)(-)))$	UGUUGGCCGGG-CCCGUUU-GAC-GCAACA	18
$(-)(-)($	6otr	$(-)(-)((((((-))))(-).....$	UGC-GAGCCGGC-GCCGGCA- UGAACUGGCCGUGAAGA	7

Note: Hits are sorted by a global alignment score computed for the four-way junction secondary structures regardless of the context (i.e. for bolded fragments only). The best one is in the first row of the table.

of the multiloop in the whole molecule context. The system output six four-way junctions—each structure composed of a loop and four branching arms of various lengths (Table 1). The results were ranked based on a global alignment score computed for the loop alone aligned with the target secondary structure. Here, we applied the Needleman–Wunsch algorithm assuming two points for a match, -2 for a mismatch, -1 for a gap (Needleman and Wunsch, 1970). A four-way junction derived from the small-subunit processome (PDB ID: 5WLC) (Barandun *et al.*, 2017) scored the best. Its secondary structure displays the highest similarity to the target with one insertion only. Thus, we obtained the following sequence of a multiloop as the input problem solution: CAGC-GAG-CAC-GGG.

The second example involves searching for the 3D shape of the RNA motif having specified secondary structure and unknown atomic coordinates. As a target, we selected purine riboswitch (Rfam ID: RF00167) with the secondary structure topology encoded as $(((((..(((((-))))))..(((((-))))))..))))$. The key motif shaping this molecule's 3D structure is a three-way junction with a 19-nucleotide internal loop (Barash and Gabdank, 2010). There is no experimental data for such a 2D structure in the Protein Data Bank (Berman *et al.*, 2000). Therefore, querying the RNAloops database for the corresponding dot-bracket input yielded zero results. Following this, we looked for molecules with secondary structures similar (but not the same) to the target. Multiple queries ran separately for each single-stranded component of the three-way junction with and without deletions at unpaired positions. This multi-step procedure resulted in 143 different three-way junctions, which we sorted by the global alignment score (Supplementary Table S1). The best match was a multiloop found in 10 PDB structures. Sorting them by resolution allowed us to select the best solution, a multiloop 3D structure from the *Thermus thermophilus* 70S ribosome (PDB ID: 4Y4O, 2.30 Å) (Polikanov *et al.*, 2015) (Supplementary Fig. S1).

In the third experiment, we evaluated the utility of RNAloops in modeling the RNA 3D structure. We used the RNAComposer system (Antczak *et al.*, 2016) to predict the core of the Alu domain of mammalian SRP RNA (PDB ID: 1E8O) (Weichenrieder *et al.*, 2000). We first ran the system for sequence and secondary structure data in the default fully automatic mode. The resulting prediction aligned at the reference structure had an RMSD of 4.23 Å. Next, we applied RNAComposer in a semi-automated mode. It allows users to insert particular structural elements into the predicted model. Knowing that multiloops significantly affect the shape of the whole molecule, we decided to model the latter using own three-way junction. To find the best fitting 3D structure for this multiloop, we searched the RNAloops database by giving the secondary structure and loop type as search criteria. We obtained 44 results, and we further reduced this set by excluding motifs originating from the target and those with planar angles outside the range 120° – 135° . The remaining three-way junctions were ranked according to the global alignment score computed for their secondary structures (Supplementary Table S2). The best rated multiloop originated from

the signal recognition particle interacting with the elongation-arrested ribosome (PDB ID: 1RY1) (Halic *et al.*, 2004). Its 3D structure was extracted from that molecule and used as a structural element in the RNAComposer modeling. The final prediction obtained this way has an RMSD of 2.40 Å. Thus, we improved the modeling accuracy by 56% and confirmed the RNAloops usefulness in RNA 3D structure prediction.

3.4 RNAloops versus other databases

Several existing databases catalog motifs found in experimental RNA 3D structures and make them searchable. They include RNA FRABASE (Antczak *et al.*, 2018b; Popenda *et al.*, 2008), RNA 3D motif atlas (Parlea *et al.*, 2016; Petrov *et al.*, 2013), RNA Bricks (Chojnowski *et al.*, 2014), and RNAJunction (Bindewald *et al.*, 2008) collecting only multiloops. In Table 2, we present the essential features of these resources to assist users in choosing the right fit. They fall into the following groups: supported motif types, database contents, filtering criteria, download options, and other facilities. Four of the six tools cover arbitrary RNA structural motifs. RNAJunction and RNAloops focus on multiloops, facilitating their exploration at the sequence (both), secondary (RNAloops), and tertiary (both) structure levels. The uniqueness of RNAloops includes calculating and sharing angular parameters for neighboring helices protruding out of a loop. RAG-3D, on the other hand, is the only one to show graph-based structure representation. All databases have associated data search engines allowing queries of varying syntax and complexity. Some of the resulting data (e.g. 3D structure in PDB or mmCIF format, graphical 2D and 3D models, structure parameter values) is ready for download with a single click. In the last category, we have included features important for many users—data visualization, self-updating statistics, systematic populating of the database with new data, and secure communication protocol.

4 Conclusions

So far, only one database has collected data on RNA multi-branched loops (Bindewald *et al.*, 2008). Unfortunately, it has not been updated since 2008, storing a constant number of ~ 12 000 multiloops extracted from RNA structures available at that time. Over the following 14 years, the number of RNAs in the Protein Data Bank has tripled. However, multiloops from newly determined RNA structures were not collected anywhere. The lack of fast and easy access to up-to-date data on these motifs and the need to study them in connection with RNA 3D structure modeling made us design a multiloop-dedicated bioinformatics system. The result of our work is RNAloops, a self-updating database that collects information about multi-branched loops identified in experimental RNA structures. The advantage of the presented tool is that the structural data, which come directly from the PDB, is supplemented with extra

Table 2. Selected features of databases collecting RNA structure motifs

	RNA FRABASE	RNA 3D motif atlas	RNA bricks	RAG-3D	RNAJunction	RNAloops
I Supported RNA motifs	Any	Any	Any	Any	Multiloops	Multiloops
II Database content						
Sequence	✓	✓	✓	✓	✓	✓
Secondary structure	✓		✓	✓		✓
Tertiary structure	✓	✓	✓	✓	✓	✓
Graph-based features				✓		
Angular data	✓					✓
III Search criteria						
PDB ID	✓	✓	✓	✓	✓	✓
Sequence	✓	✓	✓	✓	✓	✓
Secondary structure	✓		✓	✓		✓
Motif topology search	✓		✓			✓
Angular data	✓					✓
IV Download options						
Tertiary structure	✓	✓	✓	✓	✓	✓
Other motif-specific data	✓					✓
Table with search results	✓	✓	✓			✓
Visualizations				✓		✓
V Other facilities						
Output data visualization			✓			✓
Stats of database contents			✓			✓
Regular data updates	✓	✓	✓	✓		✓
Secure communication (HTTPS)						✓

information—i.e. Euler angles, planar angles, or branching multiplicities—and visualized in a user-friendly way. Each database update automatically launches a statistical module to provide users with information on data distribution due to various structural parameters. Currently (March 31, 2022), RNAloops contains >84 000 multiloops extracted from 1832 RNAs. The system supports accurate modeling of RNA 3D structures and studying their properties. It complements the collection of RNApolis tools (Szachniuk, 2019) that address various problems of RNA structural studies.

Acknowledgement

This work was carried in the European Centre for Bioinformatics and Genomics, Poznan University of Technology.

Funding

The financial support was provided by the Młoda Kadra grant for young researchers from Poznan University of Technology [0311/SBAD/0705 to J.W. and T.Z.]; the National Science Center, Poland [2019/35/B/ST6/03074 to M.S.]; and statutory funds of the Institute of Bioorganic Chemistry, PAS.

Conflict of Interest: none declared.

Data availability

RNAloops is accessible freely at <https://rnaloops.cs.put.poznan.pl/> with no login requirements. It can be operated via all modern web browsers.

References

Adamczyk, B. et al. (2022) RNAsolo: a repository of clean, experimentally determined RNA 3D structures. *Bioinformatics*, **38**, 3668–3670.

Antczak, M. et al. (2016) New functionality of RNAComposer: application to shape the axis of miR160 precursor structure. *Acta Biochim. Pol.*, **63**, 737–744.

Antczak, M. et al. (2018a) New algorithms to represent complex pseudoknotted RNA structures in dot-bracket notation. *Bioinformatics*, **34**, 1304–1312.

Antczak, M. et al. (2018b) RNAfitme: a webservice for modeling nucleobase and nucleoside residue conformation in fixed-backbone RNA structures. *BMC Bioinformatics*, **19**, 304.

Bailor, M.H. et al. (2011) 3D maps of RNA interhelical junctions. *Nat. Protoc.*, **6**, 1536–1545.

Barandun, J. et al. (2017) The complete structure of the small-subunit processome. *Nat. Struct. Mol. Biol.*, **24**, 944–953.

Barash, D. and Gabdank, I. (2010) Energy minimization methods applied to riboswitches: a perspective and challenges. *RNA Biol.*, **7**, 90–97.

Becquey, L. et al. (2021) RNA.Net: an automatically built dual-source dataset integrating homologous sequences and RNA structures. *Bioinformatics*, **37**, 1218–1224.

Berg, J.M. et al. (2002). *Biochemistry*. W.H. Freeman, New York.

Berman, H.M. et al. (2000) The protein data bank. *Nucleic Acids Res.*, **28**, 235–242.

Bindewald, E. et al. (2008) RNAJunction: a database of RNA junctions and kissing loops for three-dimensional structural analysis and nanodesign. *Nucleic Acids Res.*, **36**, D392–D397.

Blazewicz, J. et al. (2005) RNA tertiary structure determination: NOE pathways construction by Tabu search. *Bioinformatics*, **21**, 2356–2361.

Bourne, P.E. et al. (1997). Macromolecular crystallographic information file. *Methods Enzymol.*, **277**, 571–590.

Byron, K. et al. (2013) A computational approach to finding RNA tertiary motifs in genomic sequences: a case study. *Recent Pat. DNA Gene Seq.*, **7**, 115–122.

Carrascoza, F. et al. (2022) Evaluation of the stereochemical quality of predicted RNA 3D models in the RNA-Puzzles submissions. *RNA*, **28**, 250–262.

Chojnowski, G. et al. (2014) RNA bricks – a database of RNA 3D motifs and their interactions. *Nucleic Acids Res.*, **42**, D123–D131.

Cock, P. et al. (2009) Biopython: freely available python tools for computational molecular biology and bioinformatics. *Bioinformatics*, **25**, 1422–1423.

Cruz, J.A. et al. (2012) RNA-puzzles: a CASP-like evaluation of RNA three-dimensional structure prediction. *RNA*, **18**, 610–625.

Darty, K. et al. (2009) Varna: interactive drawing and editing of the RNA secondary structure. *Bioinformatics*, **25**, 1974–1975.

Diebel, J. (2006) Representing attitude: Euler angles, unit quaternions, and rotation vectors. *Matrix*, **58**, 1–35.

- Dunham, C. *et al.* (2003) A helical twist-induced conformational switch activates cleavage in the hammerhead ribozyme. *J. Mol. Biol.*, **332**, 327–336.
- Halic, M. *et al.* (2004) Structure of the signal recognition particle interacting with the elongation-arrested ribosome. *Nature*, **427**, 808–814.
- Hao, Y. and Kieft, J.S. (2016) Three-way junction conformation dictates self-association of phage packaging RNAs. *RNA Biol.*, **13**, 635–645.
- Heyde, K. and Wood, J.L. (2020). Representation of rotations, angular momentum and spin. In: Heyde, K. and Wood, J.L. (eds) *Quantum Mechanics for Nuclear Structure*. Vol. 2. IOP Publishing, pp. 1–46.
- Hohng, S. *et al.* (2004) Conformational flexibility of four-way junctions in RNA. *J. Mol. Biol.*, **336**, 69–79.
- Hua, L. *et al.* (2016) CHSalign: a web server that builds upon Junction-Explorer and RNAJAG for pairwise alignment of RNA secondary structures with coaxial helical stacking. *PLoS One*, **11**, e0147097.
- Ivry, T. *et al.* (2009) An image processing approach to computing distances between RNA secondary structures dot plots. *Algorithms Mol. Biol.*, **4**, 4.
- Kudla, M. *et al.* (2021) Virxicon: a lexicon of viral sequences. *Bioinformatics*, **36**, 5507–5513.
- Laing, C. and Schlick, T. (2010) Computational approaches to 3D modeling of RNA. *J. Phys. Condens. Matter*, **22**, 283101.
- Laing, C. *et al.* (2012) Predicting coaxial helical stacking in RNA junctions. *Nucleic Acids Res.*, **40**, 487–498.
- Lamiable, A. *et al.* (2012) Automated prediction of three-way junction topological families in RNA secondary structures. *Comput. Biol. Chem.*, **37**, 1–5.
- Leontis, N.B. and Westhof, E. (1998) A common motif organizes the structure of multi-helix loops in 16 S and 23 S ribosomal RNAs. *J. Mol. Biol.*, **283**, 571–583.
- Lescoute, A. and Westhof, E. (2006) Topology of three-way junctions in folded RNAs. *RNA*, **12**, 83–93.
- Li, B. *et al.* (2020) Advances in RNA 3D structure modeling using experimental data. *Front. Genet.*, **11**, 574485.
- Lorenz, R. *et al.* (2011) ViennaRNA package 2.0. *Algorithms Mol. Biol.*, **6**, 26.
- Lukasiak, P. *et al.* (2015) RNAssess – a web server for quality assessment of RNA 3D structures. *Nucleic Acids Res.*, **43**, W502–W506.
- Miao, Z. *et al.* (2015) RNA-Puzzles round II: assessment of RNA structure prediction programs applied to three large RNA structures. *RNA*, **21**, 1066–1084.
- Miskiewicz, J. *et al.* (2017) Bioinformatics study of structural patterns in plant microRNA precursors. *Biomed. Res. Int.*, **2017**, 6783010–6783018.
- Needleman, S.B. and Wunsch, C.D. (1970) A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.*, **48**, 443–453.
- O’Leary, N.A. *et al.* (2015) Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res.*, **44**, D733–D745.
- Parlea, L.G. *et al.* (2016) The RNA 3D motif atlas: computational methods for extraction, organization and evaluation of RNA motifs. *Methods*, **103**, 99–119.
- Petrov, A.I. *et al.* (2013) Automated classification of RNA 3D motifs and the RNA 3D motif atlas. *RNA*, **19**, 1327–1340.
- Polikanov, Y.S. *et al.* (2015) Structural insights into the role of rRNA modifications in protein synthesis and ribosome assembly. *Nat. Struct. Mol. Biol.*, **22**, 342–344.
- Popenda, M. *et al.* (2008) RNA FRABASE version 1.0: an engine with a database to search for the three-dimensional fragments within RNA structures. *Nucleic Acids Res.*, **36**, D386–D391.
- Popenda, M. *et al.* (2021) Entanglements of structure elements revealed in RNA 3D models. *Nucleic Acids Res.*, **49**, 9625–9632.
- Rybarczyk, A. *et al.* (2015) New in silico approach to assessing RNA secondary structures with non-canonical base pairs. *BMC Bioinformatics*, **16**, 276.
- Sehnal, D. *et al.* (2017) LiteMol suite: interactive web-based visualization of large-scale macromolecular structure data. *Nat. Methods*, **14**, 1121–1122.
- Szachniuk, M. (2019) RNApolis: computational platform for RNA structure analysis. *Found. Comput. Decis. Sci.*, **44**, 241–257.
- Townshend, R.J.L. *et al.* (2021) Geometric deep learning of RNA structure. *Science*, **373**, 1047–1051.
- Weichenrieder, O. *et al.* (2000) Structure and assembly of the Alu domain of the mammalian signal recognition particle. *Nature*, **408**, 167–173.
- Westhof, E. *et al.* (1996) RNA tectonics: towards RNA design. *Fold. Des.*, **1**, R78–R88.
- Wiedemann, J. and Milostan, M. (2017) StructAnalyzer – a tool for sequence vs. structure similarity analysis. *Acta Biochim. Pol.*, **63**, 753–757.
- Zahran, M. *et al.* (2015) RAG-3D: a search tool for RNA 3D substructures. *Nucleic Acids Res.*, **43**, 9474–9488.
- Zemora, G. and Waldsich, C. (2010) RNA folding in living cells. *RNA Biol.*, **7**, 634–6418.
- Zhao, W. *et al.* (2012) A Three-Helix junction is the interface between two functional domains of prohead RNA in 29 DNA packaging. *J. Virol.*, **86**, 11625–11632.
- Zok, T. *et al.* (2022) ONQUADRO: a database of experimentally determined quadruplex structure. *Nucleic Acids Res.*, **50**, D253–D258.
- Zuker, M. and Sankoff, D. (1984) RNA secondary structures and their prediction. *Bull. Math. Biol.*, **46**, 591–621.

Co-author declarations




Poznań, 12.04.2022

Oświadczenie

W związku z postępowaniem dotyczącym nadania stopnia naukowego doktora panu mgr Jakubowi Wiedemanowi oświadczam, że mój udział w badaniach przedstawionych w publikacji

Jakub Wiedemann, Jacek Kaczor, Maciej Milostan, Tomasz Zok, Jacek Błazewicz, Marta Szachniuk, Maciej Antczak, RNAloops: a database of RNA multiloops, zgłoszona do druku w czasopiśmie Bioinformatics (Oxford Academic)

polegał na udziale w dyskusjach dotyczących wyboru platformy sprzętowej i systemowej dla narzędzia bioinformatycznego RNAloops oraz konsultowaniu warstwy optymalizacyjnej algorytmu wyszukiwania multipętli.


Prof. dr hab. inż. Jacek Błazewicz



Warszawa, 20.10.2020

Oświadczam, że jestem współautorem publikacji:

Magnus M, Antczak M, Zok T, Wiedemann J, Lukasiak P, Cao Y, Bujnicki JM, Westhof E, Szachniuk M, Miao Z (2020) Nucleic Acids Research 48(2):576-588

Mój udział w przygotowaniu ww. publikacji polegał na pracy koncepcyjnej, współuczestnictwie w kreowaniu projektów rozwiązań, nadzorowaniu pracy realizowanej przez mojego doktoranta Marcina Magnusa związanej z tworzeniem narzędzi pakietu rna-tools, krytycznej analizie wyników oraz udziale w przygotowaniu manuskryptu.

Janusz M. Bujnicki.

Prof. dr hab. Janusz M. Bujnicki, MAE
Laboratorium Bioinformatyki i Inżynierii Białka,
Międzynarodowy Instytut Biologii Molekularnej i Komórkowej w Warszawie
e-mail: iamb@genesilico.pl



Prof Yang Cao
Center of Growth, Metabolism
and Aging, Key Laboratory of Bio-
Resource and Eco-Environment of
Ministry of Education, College of
Life Sciences, Sichuan University,
Chengdu 610065, PR China.
cao@scu.edu.cn

To Whom It May Concern

Chengdu, 20.10.2020

Due to my co-authorship in the following publication:

Magnus M, Antczak M, Zok T, Wiedemann J, Lukasiak P, Cao Y, Bujnicki JM,
Westhof E, Szachniuk M, Miao Z (2020) *Nucleic Acids Research* 48(2):576-
588

I am willing to state that within the scope of the research project
presented in the above paper, I contributed to manuscript preparation.

Yang Cao

Poznań, 18.04.2022

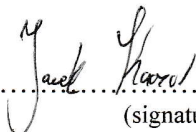
Jacek Kaczor
ul. Dzieci Wrzesińskich 12
61-066 Poznań

Declaration

Hereby, I declare that as a co-author of the paper

Jakub Wiedemann, Jacek Kaczor, Maciej Milostan, Tomasz Zok, Jacek Blazewicz, Marta Szachniuk, Maciej Antczak, *RNAloops: a database of RNA multiloops* (submitted to Bioinformatics)

I participated in the research carried under the supervision of prof. Marta Szachniuk and dr Maciej Antczak described in this paper. My task was to design and implement RNAloops – the bioinformatics database system collecting structures of RNA n-way junctions.


.....
(signature)



Dr hab. inż. Piotr Łukasiak, prof. PP

Zakład Teorii Algorytmów i Systemów Programowania

Piotr.Lukasiak@put.poznan.pl

Poznań, 12.04.2022

Oświadczenie

W związku z postępowaniem o nadanie stopnia doktora panu mgr Jakubowi Wiedemannowi oświadczam, że mój udział w badaniach przedstawionych w poniższej publikacji:

M. Magnus, M. Antczak, T. Zok, J. Wiedemann, P. Łukasiak, Y. Cao, J.M. Bujnicki, E. Westhof, M. Szachniuk, Z. Miao. RNA-Puzzles toolkit: A computational resource of RNA 3D structure benchmark datasets, structure manipulation, and evaluation tools, *Nucleic Acids Res* 48, 2020, pp. 576-588.

polegał na nadzorowaniu rozwoju programu RNAQUA oraz udziale w przygotowaniu w/w publikacji.



Dr Marcin Magnus

Międzynarodowy Instytut Mechanizmów i Maszyn Molekularnych

Polskiej Akademii Nauk

e-mail: m.magnus@imol.institute

telefon: 601714835

Warszawa, 11.04.2022

Oświadczenie

Niniejsze oświadczenie składam w związku z postępowaniem o nadanie stopnia doktora panu mgr Jakubowi Wiedemannowi. Oświadczam, że publikacja naukowa Magnus M, Antczak M, Zok T, Wiedemann J, Lukasiak P, Cao Y, Bujnicki JM, Westhof E, Szachniuk M, Miao Z. *Nucleic Acids Res* 2020, 48(2):576-588 (PDF: <https://academic.oup.com/nar/article/48/2/576/5651330>) powstała w wyniku pracy zespołowej, w ramach której opracowałem oraz zaimplementowałem pakiet *rna-tools*, zintegrowałem zestaw narzędzi stworzonych przez współautorów projektu w potoku wykonawczym, stworzyłem zbiór danych znormalizowanych oraz uczestniczyłem w pisaniu manuskryptu.



Dr Marcin Magnus



Dr Zhichao Miao
European Molecular Biology
Laboratory, European
Bioinformatics Institute
(EMBL-EBI),
Wellcome Genome Campus,
UK
zmiao@ebi.ac.uk

To Whom It May Concern

Cambridge, 20.10.2020

Due to my co-authorship in the following publication:

Magnus M, Antczak M, Zok T, Wiedemann J, Lukasiak P, Cao Y, Bujnicki JM, Westhof E, Szachniuk M, Miao Z (2020) *Nucleic Acids Research* 48(2):576-588

I am willing to state that I participated in the research presented in the above paper. In particular, I conceived the project, cleaned up the datasets, implemented the RNA structure format tool, format check tool and evaluation metrics (INF, DP), designed the website and wrote parts of the manuscript.



WYDZIAŁ INFORMATYKI I TELEKOMUNIKACJI

Instytut Informatyki

ul. Piotrowo 2, 60-965 Poznań, tel. +48 61 665 2997, fax +48 61 877 1525

e-mail: office_cs@put.poznan.pl, www.cs.put.poznan.pl

Poznań, 12.04.2022r.

Oświadczenie współautora

W związku z postępowaniem dotyczącym nadania stopnia naukowego doktora panu mgr Jakubowi Wiedemanowi oświadczam, że mój udział w badaniach przedstawionych w publikacjach doktoranta przedstawia się następująco:

[1] Wiedemann J, Milostan M (2016) StructAnalyzer-a tool for sequence vs. structure similarity analysis. *Acta Biochimica Polonica* 63(4):753-757 (doi:10.18388/abp.2016_1333).

- Określenie tematyki i zakresu badań
- Konsultacje prowadzonych badań
- Wsparcie na etapie tworzenia manuskryptu
- Przygotowanie i udostępnienie środowiska deweloperskiego

[2] Wiedemann J, Zok T, Milostan M, Szachniuk M (2017) LCS-TA to identify similar fragments in RNA 3D structures. *BMC Bioinformatics* 18(1):1-13 (doi:10.1186/s12859-017-1867-6).

- Konsultacje prowadzonych prac
- Walidacja zaproponowanego algorytmu i przygotowanie przykładowej wizualizacji drzewa wykonania algorytmu
- Wsparcie tworzenia manuskryptu

[3] Wiedemann J, Kaczor J, Milostan M, Zok T, Blazewicz J, Szachniuk M, Antczak M (2022) RNAloops: a database of RNA multiloops. *Bioinformatics*, submitted for publication.

- Udostępnienie środowiska serwerowego
- Konsultacje na wczesnym etapie prac
- Weryfikacja manuskryptu

Dr inż. Maciej Milostan



Poznan, 12.04.2022

Declaration of a co-author

I declare the following contributions to publications co-authored with Jakub Wiedemann:

- [1] J. Wiedemann, T. Zok, M. Milostan, M. Szachniuk (2017) LCS-TA to identify similar fragments in RNA 3D structures. *BMC Bioinformatics* 18(1):1–13.
 - project initiation
 - coordination of the teamwork
 - participation in manuscript writing
- [2] M. Magnus, M. Antczak, T. Zok, J. Wiedemann, P. Lukasiak, Y. Cao, J. Bujnicki, E. Westhof, M. Szachniuk, Z. Miao Z (2020) RNA-Puzzles toolkit: a computational resource of RNA 3D structure benchmark datasets, structure manipulation, and evaluation tools. *Nucleic Acids Research* 48(2):576–588.
 - coordination of the teamwork
 - participation in manuscript writing
- [3] J. Wiedemann, J. Kaczor, M. Milostan, T. Zok, J. Blazewicz, M. Szachniuk, M. Antczak (2022) RNAloops: a database of RNA multiloops. *submitted for publication*.
 - participation in system design
 - web application testing
 - participation in manuscript writing

Marta Szachniuk

Prof. dr hab. inż. Marta Szachniuk



Strasbourg, le 13 Avril 2022

To Whom It May Concern

In connection with the process of the doctoral degree conferment to Mr Jakub Wiedemann:

M. Magnus, M. Antczak M, T. Zok, J. Wiedemann, P. Lukasiak, Y. Cao, J.M. Bujnicki, E. Westhof, M. Szachniuk, and Z. Miao (2020) *Nucleic Acids Research* 48(2):576-588

I supervised the entire project and drafted the manuscript. If you require any more information, please do not hesitate to contact me.

Yours faithfully,

Sous la tutelle du



Associé à





Poznan, 12.04.2022

Declaration of a co-author

I declare the following contributions to publications co-authored with Jakub Wiedemann:

1. LCS-TA to Identify Similar Fragments in RNA 3D Structures. J. Wiedemann, T. Zok, M. Milostan and M. Szachniuk. *BMC Bioinformatics*. 2017. 18(1):456. doi:10.1186/s12859-017-1867-6
 - Preparation of library to analyze RNA 3D structures in torsion angle space
 - Participation in discussions regarding the algorithm design and manuscript preparation
2. RNA-Puzzles Toolkit: A Computational Resource of RNA 3D Structure Benchmark Datasets, Structure Manipulation and Evaluation Tools. M. Magnus, M. Antczak, T. Zok, J. Wiedemann, P. Lukasiak, Y. Cao, J.M. Bujnicki, E. Westhof, M. Szachniuk and Z. Miao. *Nucleic Acids Research*. 2020. 48(2):576–588. doi:10.1093/nar/gkz1108
 - Participation in the initial discussion and planning of the article and its scope
 - Integration of MCQ4Structures application in the RNA-Puzzles toolkit
 - Creation of Docker image with the RNA-Puzzles toolkit
 - Description of the above in the manuscript and critical review of the whole
3. RNAloops: a database of RNA multiloops. J. Wiedemann, J. Kaczor, M. Milostan, T. Zok, J. Blazewicz, M. Szachniuk, M. Antczak. *Submitted for publication*. 2022
 - Principal Investigator of a *Młoda Kadra* grant with Jakub Wiedemann as its executor
 - Testing and review of the RNAloops webserver

dr inż. Tomasz Żok

Extended abstract in Polish

Niniejsza praca doktorska poświęcona jest badaniom nad strukturą cząsteczek RNA i tworzeniu metod, głównie kombinatorycznych, pozwalających ją analizować. W badaniach skoncentrowano się na analizie charakterystyk kątowych struktur przestrzennych (3D) RNA w kontekście porównywania struktur realizowanego w ramach ewaluacji modeli uzyskiwanych z wykorzystaniem metod obliczeniowych, a także identyfikacji specyficznych motywów strukturalnych zwanych pętlami wieloramiennymi. Prezentowane w pracy badania rozpoczęły się od analizy związku pomiędzy sekwencją cząsteczki RNA, a jej strukturą trzeciorzędową. Uzyskane wyniki wykazały, że stosunkowo wysokie podobieństwo sekwencyjne nie zawsze gwarantuje zachowanie kształtu przestrzennego cząsteczek. Ponadto badania przeprowadzone podczas opracowywania narzędzia *StructAnalyzer* pozwoliły zaobserwować, że struktury, które w ujęciu globalnym różnią się znacząco wykazują często wyraźne podobieństwo w ujęciu lokalnym. Potwierdziło to istotność problemu wyszukiwania lokalnych, konserwatywnych motywów pomiędzy pozornie różniącymi się od siebie strukturami. Konserwatywne motywy wskazują zwykle, że struktury, które je zawierają prawdopodobnie pełnią zbliżoną funkcję biologiczną.

Scharakteryzowane krótko wyniki zainspirowały opracowanie nowej miary, która ocenia podobieństwo struktur 3D RNA z lokalnej perspektywy. Mnogość struktur oraz złożoność procesu porównywania wielu struktur

skierowały nas w kierunku wykorzystania reprezentacji kątowej do opisywania struktur przestrzennych, gdyż ta reprezentacja pozwala na pominięcie procesu uliniawiania struktur przestrzennych RNA. Opracowany algorytm *LCS-TA* pozwala na identyfikację najdłuższych ciągłych segmentów charakteryzujących się określonym współczynnikiem podobieństwa (bazującym na mierze *MCQ*) nie przekraczającym oczekiwanego przez użytkownika odcięcia pomiędzy porównywaną parą struktur 3D RNA. W rezultacie zwracana jest lokalizacja dopasowanych do siebie segmentów i ich długość (będącą dość intuicyjną miarą podobieństwa analizowanych struktur). Zaproponowany algorytm został wykorzystany do oceny modeli 3D RNA zgłoszonych w rundzie IV konkursu RNA-Puzzles oraz udostępniony w ramach zestawu narzędzi wykorzystywanych przez społeczność RNA-Puzzles. Analizy wyników uzyskiwanych w ramach konkursu RNA-Puzzles wskazują jednoznacznie na istnienie motywów, które są szczególnie trudne do przewidzenia obliczeniowo. Jednym z nich są pętle wieloramienne (ang. *N-way junctions*). Motywy te, które często obejmują niekanoniczne pary zasad, znacząco wpływają na proces fałdowania cząsteczek RNA i ich ostateczny kształt. Ponadto analiza dostępnych źródeł wskazywała na brak na bieżąco uaktualnianych repozytoriów udostępniających znane konformacje obserwowane w eksperymentalnie określonych strukturach 3D RNA, co sprawia, że ich wykorzystanie chociażby podczas modelowania homologicznego często nastrocza trudności. W celu wypełnienia zidentyfikowanej luki opracowano bazę danych *RNAloops*, aby w sposób pełni zautomatyzowany gromadzić i udostępniać w ramach jednego repozytorium informacje o multipętlach zidentyfikowanych w eksperymentalnie określonych strukturach przestrzennych cząsteczek RNA. Baza wyposażona w przyjazny użytkownikowi interfejs udostępnia następujące informacje o multipętlach: sekwencja, struktury 2D i 3D, oraz relacje zachodzące pomiędzy kolejnymi parami sąsiadujących rozgałęzień z wykorzystaniem

trzech kątów Eulera oraz jednego kąta płaskiego. Tego typu przestrzenne zależności mogą zostać wykorzystane podczas eksperckiego modelowania 3D RNA, które zakłada wyszukiwanie obiecujących multipętli je spełniających. Korzyścią z wykorzystania kątowej reprezentacji w tym zastosowaniu jest większa uniwersalność procedury wyszukiwania obiecujących motywów. Platforma ponadto umożliwia użytkownikom elastyczne wyszukiwanie interesujących rekordów na podstawie szeregu kryteriów, m.in., sekwencji, struktury drugorzędowej, czy liczby rozgałęzień poszukiwanej pętli. Funkcjonalności te wspierają, m.in., ekstrakcję motywów o określonych cechach, ich analizę porównawczą czy modelowanie struktur przestrzennych charakteryzujących się określonymi własnościami np. podczas projektowania rozwiązań terapeutycznych.

Appendices

APPENDIX A

Participation in research projects

- Theme of the project: Genomic Map of Poland
Grant number: POIR.04.02.00-30-A004/16
Participation period: 01.04.2017 – 31.12.2017
Principal investigator: prof. Jacek Błażewicz

- Theme of the project: RNAPolis - methods and algorithms to model and analyze the RNA structure
Grant number: 2016/23/B/ST6/03931
Participation period: 01.09.2017 – 19.11.2020
Principal investigator: prof. Marta Szachniuk

- Theme of the project: RNA multi-loop analysis tool and database
Grant number: 0311/SBAD/0705 (Młoda Kadra)
Participation period: 01.06.2020 – 01.06.2021
Principal investigator: dr Tomasz Żok

- Theme of the project: Feature exploration and modelling of quadruplex structures
Grant number: 2019/35/B/ST6/03074
Participation period: since 14.07.2020
Principal investigator: prof. Marta Szachniuk

APPENDIX B

Conference presentations

During my Ph.D. study, I gave 27 presentations (talks and posters) at national and international scientific conferences and seminars:

1. *Parallel Approach to Multiple Sequence/Structure Comparative Analysis*, Seminar of the Laboratory of Algorithm Design and Data Structures, Institute of Computing Science, Poznan University of Technology, June 2016, Poznan, Poland.
2. *StructAnalyzer - a tool for sequence vs structure similarity analysis*, BIT'16: Bioinformatics in Torun, June 2016, Torun, Poland.
3. *StructAnalyzer - a tool for sequence vs structure similarity analysis*, EURO2016 - 28th European Conference on Operational Research, July 2016, Poznan, Poland.
4. *StructAnalyzer - a tool for sequence vs structure similarity analysis*, 9th Symposium of the Polish Bioinformatics Society, September 2016, Bialystok, Poland.
5. *Structular Analysis of RNA with sequential context*, ICOLE16, Lessach, September 2016, Austria.
6. *LCS-TA to identify similarities in RNA 3D structures*, IBCH PAS seminar, April 2017, Poznan, Poland.

7. *LCS-TA to identify similarity in molecular structures*, Joint EURO ORSC ECCO Conference 2017 on Combinatorial Optimization in Koper, May 2016, Koper, Slovenia.
8. *LCS-TA to identify similarity in molecular structures*, BIT'17: Bioinformatics in Torun, June 2017, Torun, Poland.
9. *LCS-TA to identify similarity in molecular structures*, X Symposium of the Polish Bioinformatics Society, September 2017, Uniejow, Poland.
10. *A new approach in RNA similar fragments identification*, ICOLE17, September 2017, Lessach, Austria.
11. *LCS-TA to identify similar fragments in RNA 3D structures*, Seminar of the Laboratory of Algorithm Design and Data Structures, Institute of Computing Science, Poznan University of Technology, September 2017, Poznan, Poland.
12. *LCS-TA to identify similar fragments in RNA 3D structures*, RECOMB 2018, April 2018, Paris, France.
13. *Euler angles in representation and analysis of biomolecules*, RNA & Computing, May 2018, Biedrusko, Poland.
14. *N-way junction modeling and analysis*, ECCO XXXI, June 2018, Fribourg, Switzerland.
15. *Euler angles in n-way junction modeling and analysis*, BIT'18: Bioinformatics in Torun, June 2018, Torun, Poland
16. *N-way junction modeling and analysis*, XXIX EURO: 29th European Conference on Operational Research, July 2018, Valencia, Spain.

17. *Bioinformatic study of RNA multi-branched loops*, XI Convention of the Polish Bioinformatics Society, September 2018, Wroclaw, Poland.
18. *LCS-TA locates RNA 3D fragments with similar folds*, RNA: WIN, September 2018, Obrzycko, Poland.
19. *Torsion angle-driven evaluation and ranking*, 2nd RNA Puzzles meeting, December 2018, Warszawa, Poland.
20. *Assessing similarity in the set of RNA 3D structures*, BIT'19: Bioinformatics in Torun, June 2019, Torun, Poland
21. *Bioinformatic analysis and representation of multibranched loops*, The FEBS Congress, July 2019, Krakow, Poland
22. *New approach to find structural patterns in RNAs*, XII Symposium of the Polish Bioinformatics Society, September 2019, Krakow, Poland.
23. *Bioinformatic analysis of multibranched loops*, ICOLE19, September 2019, Lessach, Austria.
24. *Bioinformatic analysis of multibranched loops*, 3rd Workshop "BioInformatics meets Machine Learning", December 2019, Mittweida, Germany.
25. *New approach to find structural patterns in RNAs*, IBCH PAS seminar, February 2020, Poznan, Poland
26. *RNA junctions from a 3D structure perspective*, Autumn Workshop of the Polish Bioinformatics Society, November 2020, online event.
27. *RNA junctions from a 3D structure perspective*, Seminar of the Laboratory of Algorithm Design and Data Structures, Institute of Computing Science, Poznan University of Technology, December 2020, Poznan, Poland.

APPENDIX C

Awards and distinctions

- ▶ Pro-quality scholarship for the best Ph.D. students granted by the Rector of Poznan University of Technology (2016/2017, 2017/2018, 2018/2019, 2019/2020, 2020/2021).
- ▶ Scholarship for the best Ph.D. students in the Faculty of Computing granted by the Rector of Poznan University of Technology (2016/2017, 2017/2018, 2018/2019, 2019/2020, 2020/2021).
- ▶ PTBioch FEBS conference scholarship, Kraków, Poland (2019).
- ▶ Laureate of the 240+ incentive programme at the Institute of Computer Science, Poznan University of Technology (2021).

Awards for works I co-authored or co-supervised:

- ▶ Best poster award at the 2021 meeting of the RNA Society. Awarded poster *RNAloops, a Database of RNA Multiloops*. Presented by T. Zok.
- ▶ Best M.Sc. Thesis Award given by the Polish Bioinformatics Society in 2021. Awarded thesis: *RNAloops: a database of RNA multi-branched loops* (RNAloops: baza struktur wieloramiennych pętli RNA). Authored by J. Kaczor, supervised by M. Antczak.

APPENDIX D ---

Organizational activities

- ▶ Member of the Self-Governance of Ph.D. Students at the Poznan University of Technology (2016-2020).
- ▶ Member of the Organizing Committee of RNA & Computing conference held in Biedrusko (2018).
- ▶ Member of the local Organizing Committee of the X Symposium of the Polish Bioinformatics Society held in Uniejow (2017).