High-quality customizable algorithms for RNA 3D structure alignment

Maciej Antczak[1,2],
Michal Zurkowski[1], Marta Szachniuk[1,2]

1) Institute of Computing Science, PUT
2) Institute of Bioorganic Chemistry, PAS

Benasque 2022, 07.08-20.08.2022

Outline

- Introduction
- Algorithms description
- Experimental results
- Conclusions

Why does reliable 3D structure alignment matter?

- The **alignment** of **evolutionary-related structures** reveals
  - **a correspondence** between **conserved residues** and **motifs**,
  - that **may be indicative** of **common biological functions**.

- **3D structure alignment** is **valuable** in **various applications**, e.g.:
  - homology modeling,
  - structural classification,
  - function prediction, etc.

When the problem is hard to solve?

- While 3D structures:
  - **differ** in **the chain(s) length** and/or **the sequence**,
  - **differ** in **structural complexity** and/or **topology**,
  - **exhibit conformational changes**.

- When one is interested **not in some feasible alignment** but **the longest alignment** of **the expected accuracy** (i.e., the score computed for the particular residue alignment cannot exceed some predefined cut-off value).

Root-Mean-Square Deviation (RMSD) [1]

- It represents **a distance** between two compared **atom sets of the same cardinality** after superposition, where $\boldsymbol{d(ai, bi)}$ is **the Euclidean distance** between the particular atom pair:

$$RMSD(A,B) = \sqrt{\frac{1}{N}\sum_{i=1}^{N} d(a_i, b_i)^2}$$

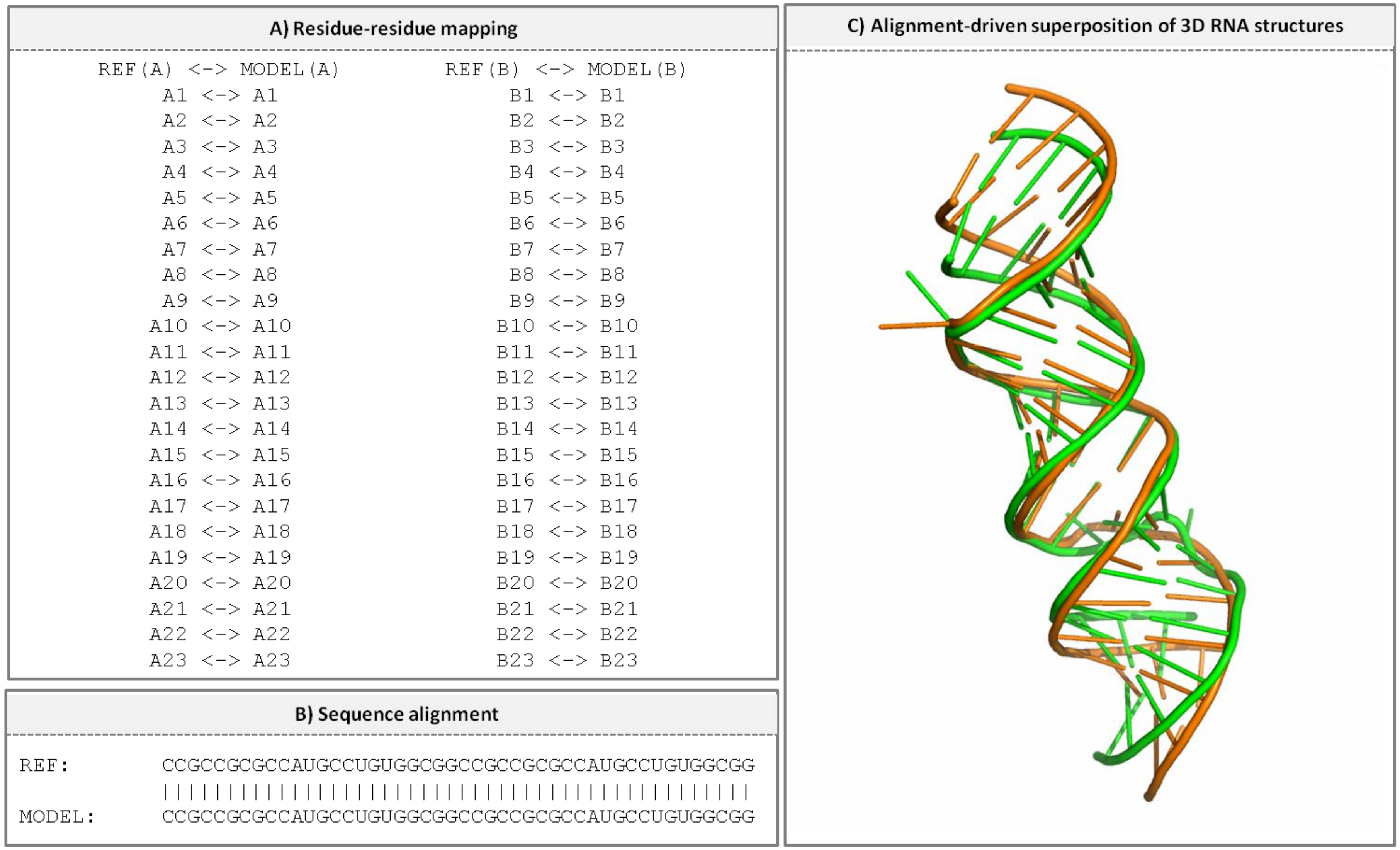
- It is the **standard measure**.

- It is **sequence length-dependent** score.
- It is **very sensitive**, e.g., on slight differences of torsion angles.

[1] W. Kabsch. A discussion of the solution for the best rotation to relate two sets of vectors. Acta Crystallo- graphica Section A, 34(5):827-828, Sep 1978.

The *solution* is…

- The **longest alignment** whose **RMSD score does not exceed** the predefined **cut-off** set by the user (e.g., 3.5 Å).
  - It could **consist of** a set of **discontinuous fragments**.



A) Residue-residue mapping

```
REF(A) <-> MODEL(A)        REF(B) <-> MODEL(B)
    A1 <-> A1                  B1 <-> B1
    A2 <-> A2                  B2 <-> B2
    A3 <-> A3                  B3 <-> B3
    A4 <-> A4                  B4 <-> B4
    A5 <-> A5                  B5 <-> B5
    A6 <-> A6                  B6 <-> B6
    A7 <-> A7                  B7 <-> B7
    A8 <-> A8                  B8 <-> B8
    A9 <-> A9                  B9 <-> B9
   A10 <-> A10                B10 <-> B10
   A11 <-> A11                B11 <-> B11
   A12 <-> A12                B12 <-> B12
   A13 <-> A13                B13 <-> B13
   A14 <-> A14                B14 <-> B14
   A15 <-> A15                B15 <-> B15
   A16 <-> A16                B16 <-> B16
   A17 <-> A17                B17 <-> B17
   A18 <-> A18                B18 <-> B18
   A19 <-> A19                B19 <-> B19
   A20 <-> A20                B20 <-> B20
   A21 <-> A21                B21 <-> B21
   A22 <-> A22                B22 <-> B22
   A23 <-> A23                B23 <-> B23
```

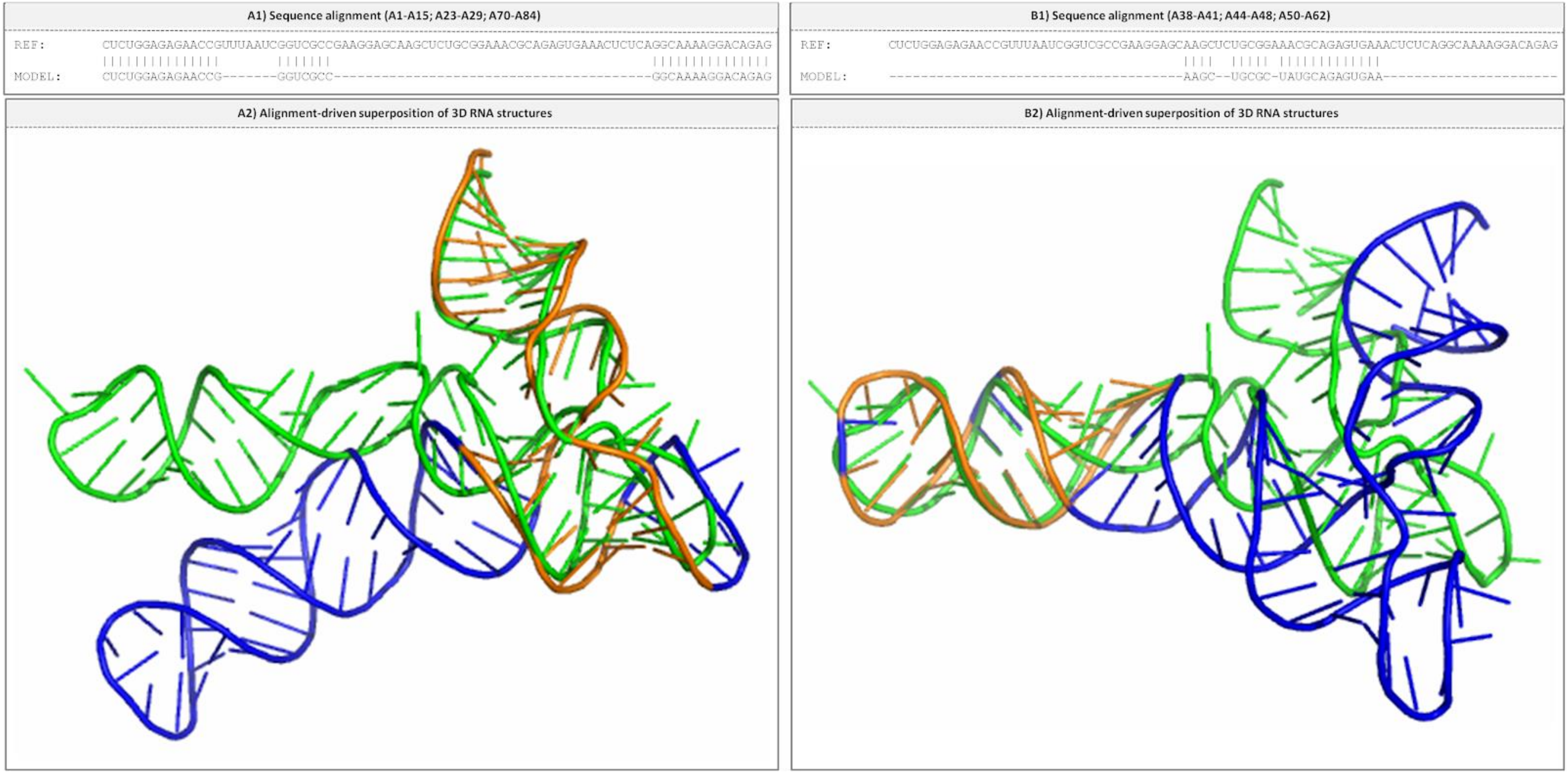
B) Sequence alignment

```
REF:      CCGCCGCGCCAUGCCUGUGGCGGCCGCCGCGCCAUGCCUGUGGCGG
          ||||||||||||||||||||||||||||||||||||||||||||||
MODEL:    CCGCCGCGCCAUGCCUGUGGCGGCCGCCGCGCCAUGCCUGUGGCGG
```

C) Alignment-driven superposition of 3D RNA structures

#3 Das model superimposed into the reference structure (3MEI)

The example solutions (PZ03 – sequence-dependent mode)



A1) Sequence alignment (A1-A15; A23-A29; A70-A84)

```
REF:   CUCUGGAGAGAACCGUUUAAUCGGUCGCCGAAGGAGCAAGCUCUGCGGAAACGCAGAGUGAAACUCUCAGGCAAAAGGACAGAG
       |||||||||||||||       |||||||                                        |||||||||||||||
MODEL: CUCUGGAGAGAACCG-------GGUCGCC----------------------------------------GGCAAAAGGACAGAG
```

A2) Alignment-driven superposition of 3D RNA structures

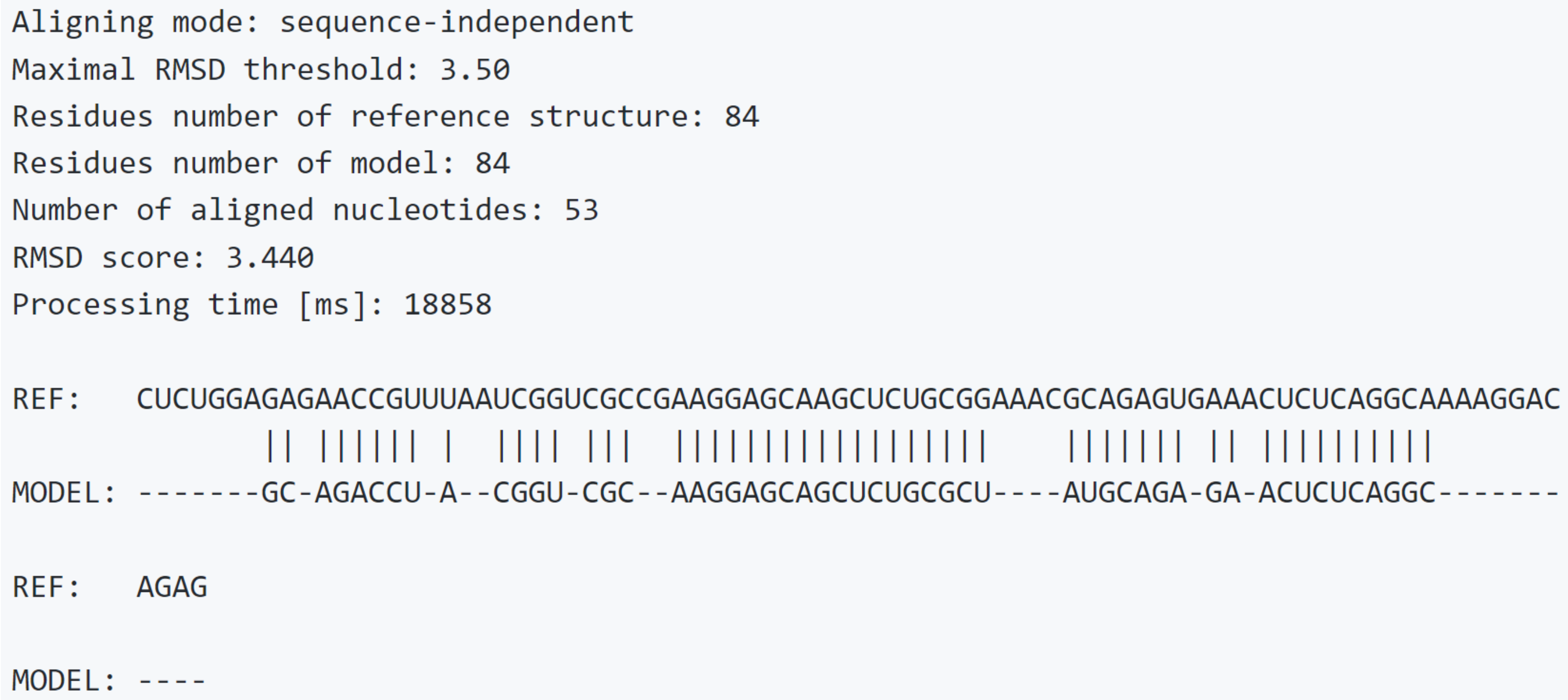
B1) Sequence alignment (A38-A41; A44-A48; A50-A62)

```
REF:   CUCUGGAGAGAACCGUUUAAUCGGUCGCCGAAGGAGCAAGCUCUGCGGAAACGCAGAGUGAAACUCUCAGGCAAAAGGACAGAG
                                            ||||  ||||| |||||||||||||
MODEL: -------------------------------------AAGC--UGCGC-UAUGCAGAGUGAA----------------------
```

B2) Alignment-driven superposition of 3D RNA structures

#1 Chen model superimposed into the reference structure (3OWZ), 3.0 Å

The example solution (PZ03 – sequence-independent mode)

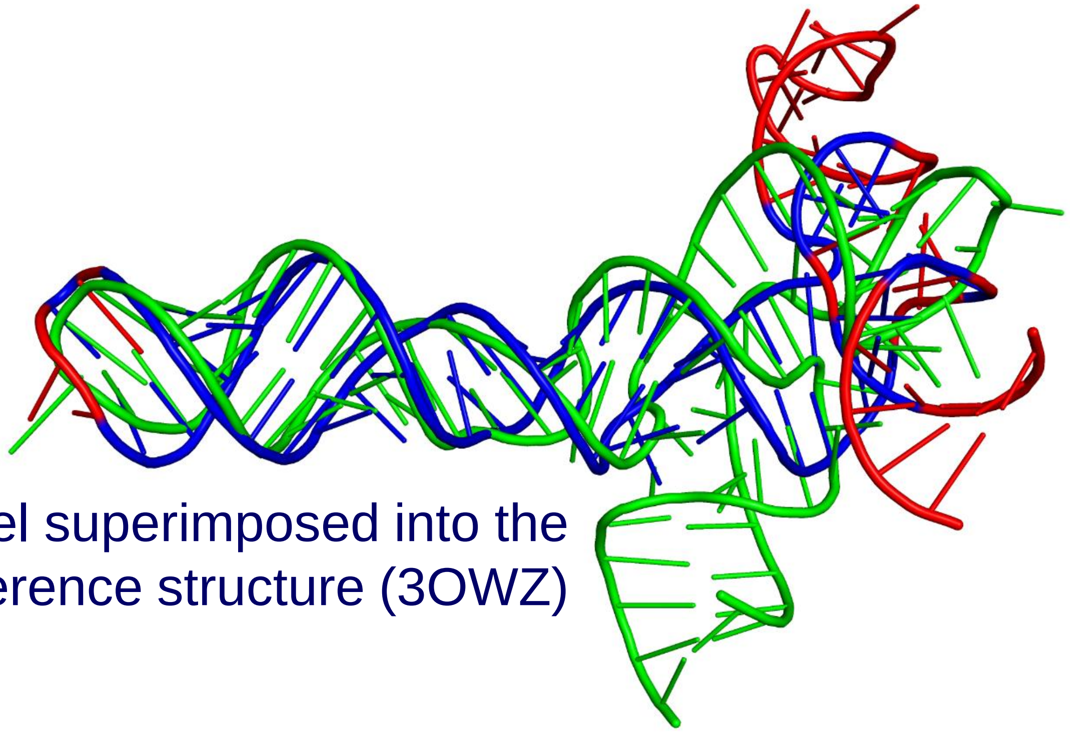


```
Aligning mode: sequence-independent
Maximal RMSD threshold: 3.50
Residues number of reference structure: 84
Residues number of model: 84
Number of aligned nucleotides: 53
RMSD score: 3.440
Processing time [ms]: 18858

REF:   CUCUGGAGAGAACCGUUUAAUCGGUCGCCGAAGGAGCAAGCUCUGCGGAAACGCAGAGUGAAACUCUCAGGCAAAAGGAC
              || |||||| |  |||| |||  ||||||||||||||||||    ||||||| || ||||||||||
MODEL: -------GC-AGACCU-A--CGGU-CGC--AAGGAGCAGCUCUGCGCU----AUGCAGA-GA-ACUCUCAGGC-------

REF:   AGAG

MODEL: ----
```



#1 Chen model superimposed into the reference structure (3OWZ)

State-of-the-Art

- There are many solutions that **usually quite well aligning 3D structures**, such as:
  - *RMAlign* [1],
  - *R3DAlign* [2],
  - *SupeRNAlign* [3],

- However, **existing tools do not allow** the user **to filter non-acceptable solutions** (by setting the cut-off value).

[1] Zheng J, Xie J, Hong X, Liu S. RMalign: an RNA structural alignment tool based on a novel scoring function RMscore. *BMC Genomics*. 2019 Apr 8;20(1):276.

[2] Ryan R. Rahrig, Neocles B. Leontis, and Craig L. Zirbel. R3D Align: global pairwise alignment of RNA 3D structures using local superpositions. *Bioinformatics*, 26(21):2689-2697.

[3] Piątkowski, P., Jabłońska, J., Żyła, A., Niedziałek, D., Matelska, D., Jankowska, E., ... & Bujnicki, J. M. (2017). SupeRNAlign: a new tool for flexible superposition of homologous RNA structures and inference of accurate structure-based sequence alignments. *Nucleic acids research*, 45(16), e150-e150.

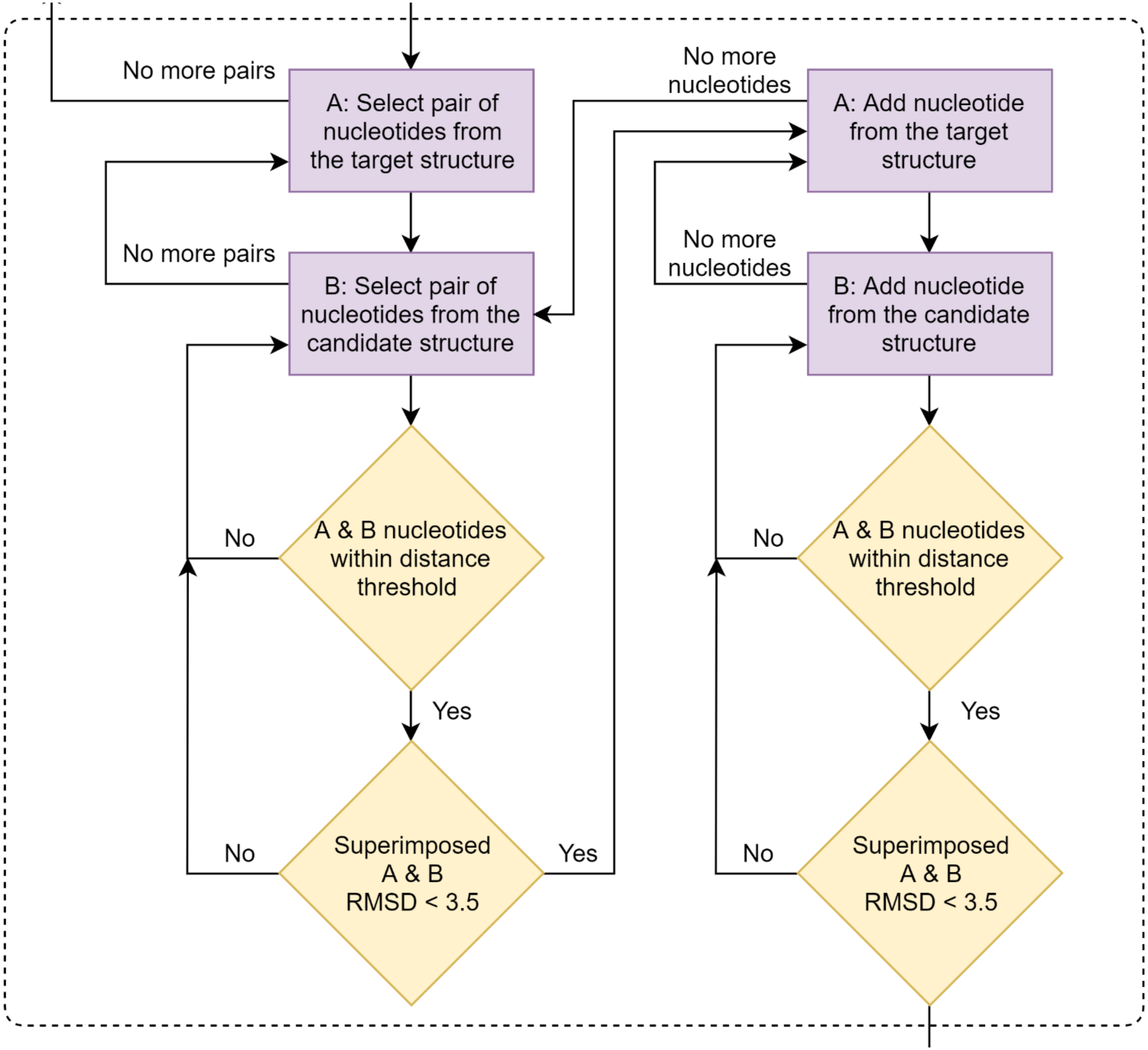
Coarse-grained representation of nucleotide

- **Every nucleotide** in the RNA 3D structure **is described** by the following **representative coordinates**:
  - the **sugar-phosphate backbone** (e.g., P or C5’ atom coordinates),
  - the **ribose atoms** geometric center,
  - the **nucleobase atoms** geometric center.

Geometric search (GEOS): promising kernels identification

- **Identify residue pairs** treated as **preliminary kernels**.

- **Extend preliminary kernels** by **adding another residue pair** close to each other in 3D space to construct **promising kernels**.

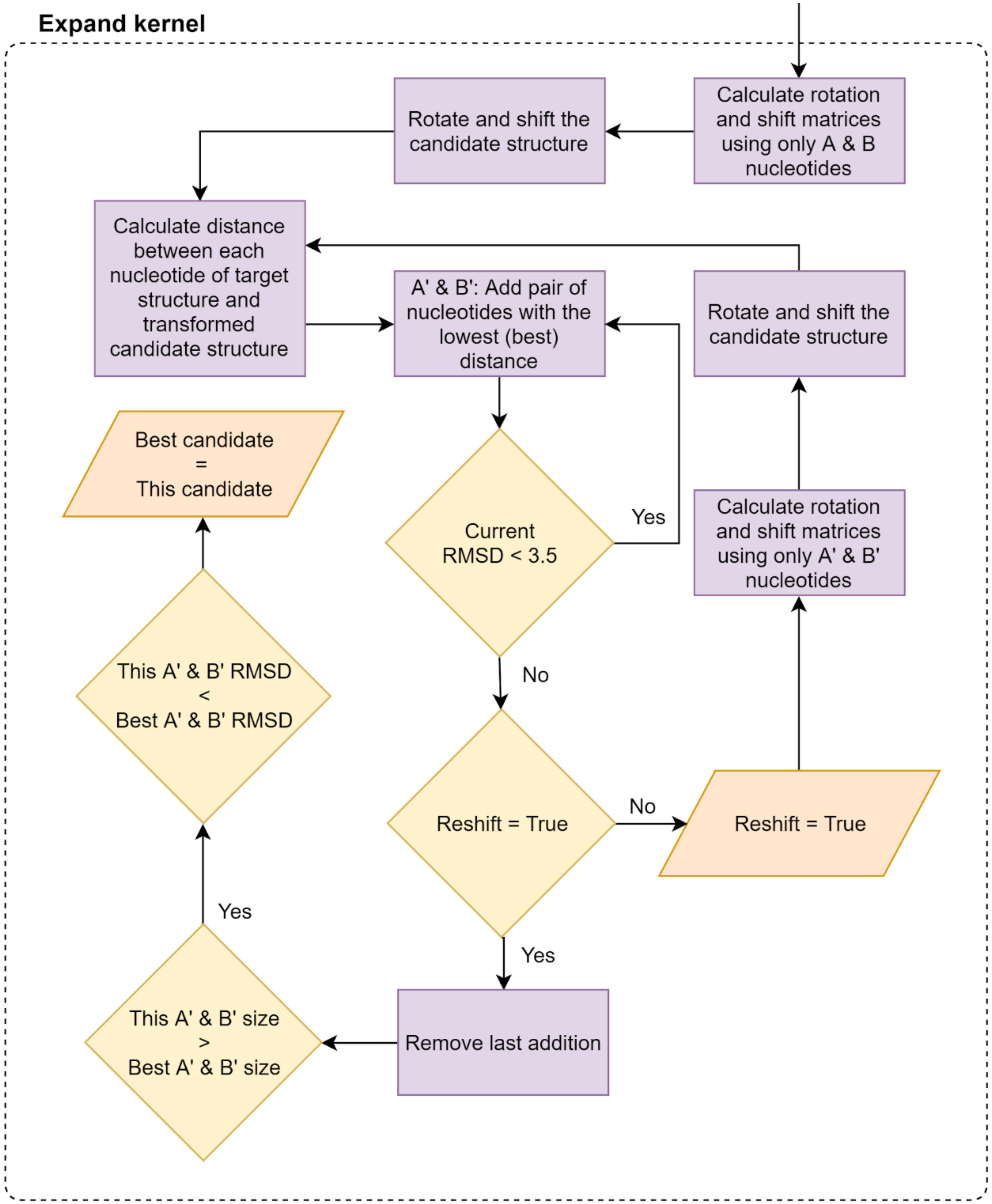


No more pairs
A: Select pair of nucleotides from the target structure
No more pairs
B: Select pair of nucleotides from the candidate structure
No
A & B nucleotides within distance threshold
Yes
No
Superimposed A & B RMSD < 3.5
Yes
No more nucleotides
A: Add nucleotide from the target structure
No more nucleotides
B: Add nucleotide from the candidate structure
No
A & B nucleotides within distance threshold
Yes
No
Superimposed A & B RMSD < 3.5

Geometric search (GEOS): promising kernels expansion

- **Expand promising kernels** by **adding iteratively** the **next residue pairs** close to each other in 3D space.

- **Calculate a superposition** of **the model** into **the reference structure**:
  - **at the beginning** of **the kernel expansion**,
  - when **the current kernel cannot be extended**.



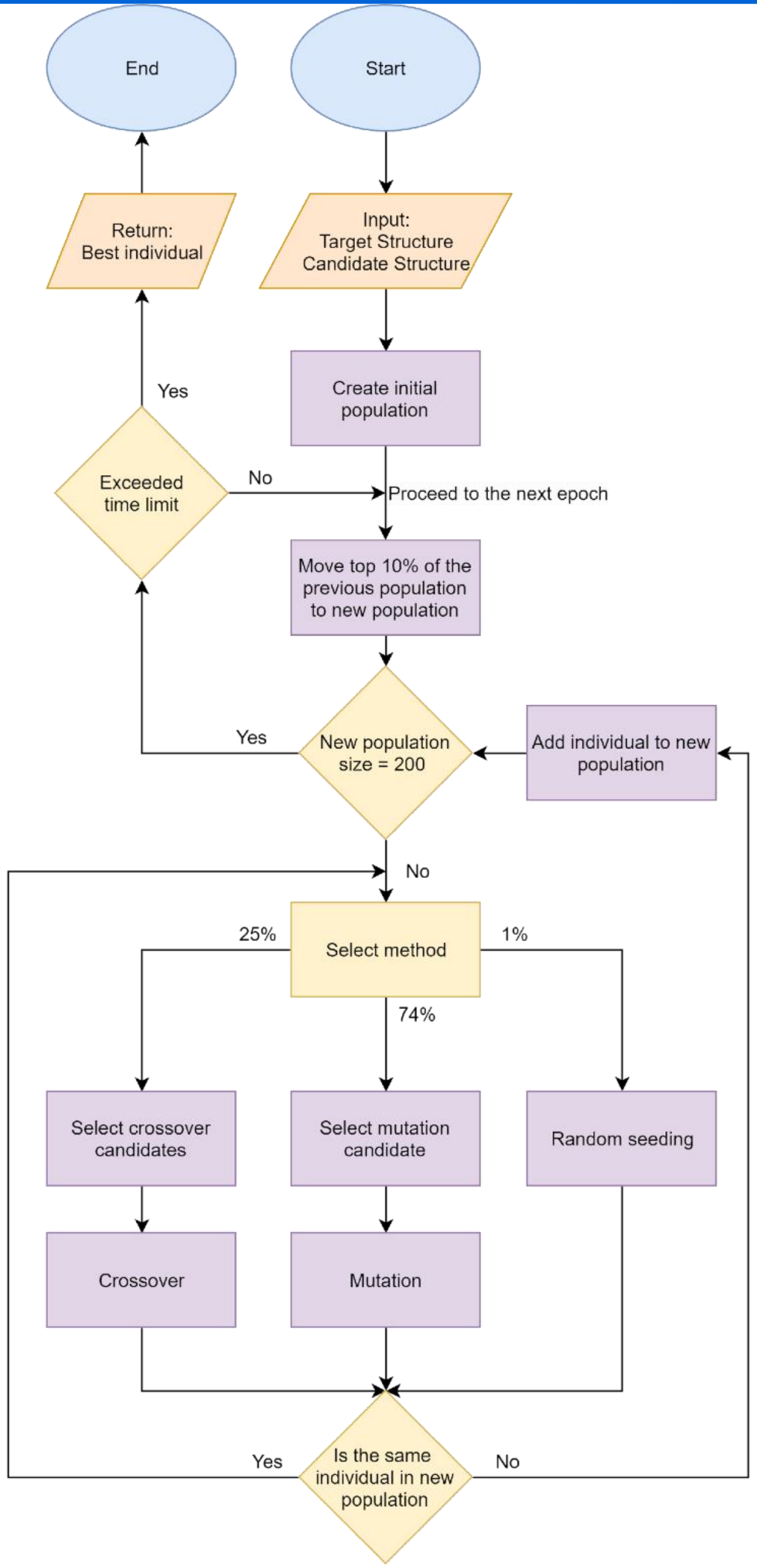
Expand kernel
Calculate rotation and shift matrices using only A & B nucleotides
Rotate and shift the candidate structure
Calculate distance between each nucleotide of target structure and transformed candidate structure
A' & B': Add pair of nucleotides with the lowest (best) distance
Rotate and shift the candidate structure
Best candidate = This candidate
Current RMSD < 3.5
Yes
Calculate rotation and shift matrices using only A' & B' nucleotides
No
This A' & B' RMSD < Best A' & B' RMSD
Reshift = True
No
Reshift = True
Yes
Yes
This A' & B' size > Best A' & B' size
Remove last addition

Geometric search (GEOS): pros and cons

- **Advantages**:
  - Deterministic.
  - Scalability.
- **Disadvantages**:
  - Dedicated heuristic.
  - In the case of some instances, could be computationally expensive.

Genetic search (GENS): initial population

- **Every individual** is represented as **a mapping** of **residues** (i.e., a list of aligned residue pairs) between **the model** and **the reference structure**.
- **Initial population size** and **a minimal number** of **residue pairs** in **individuals** are **configurable parameters**.
- **The top 10%** of **best individuals are preserved** between two **consecutive populations**.

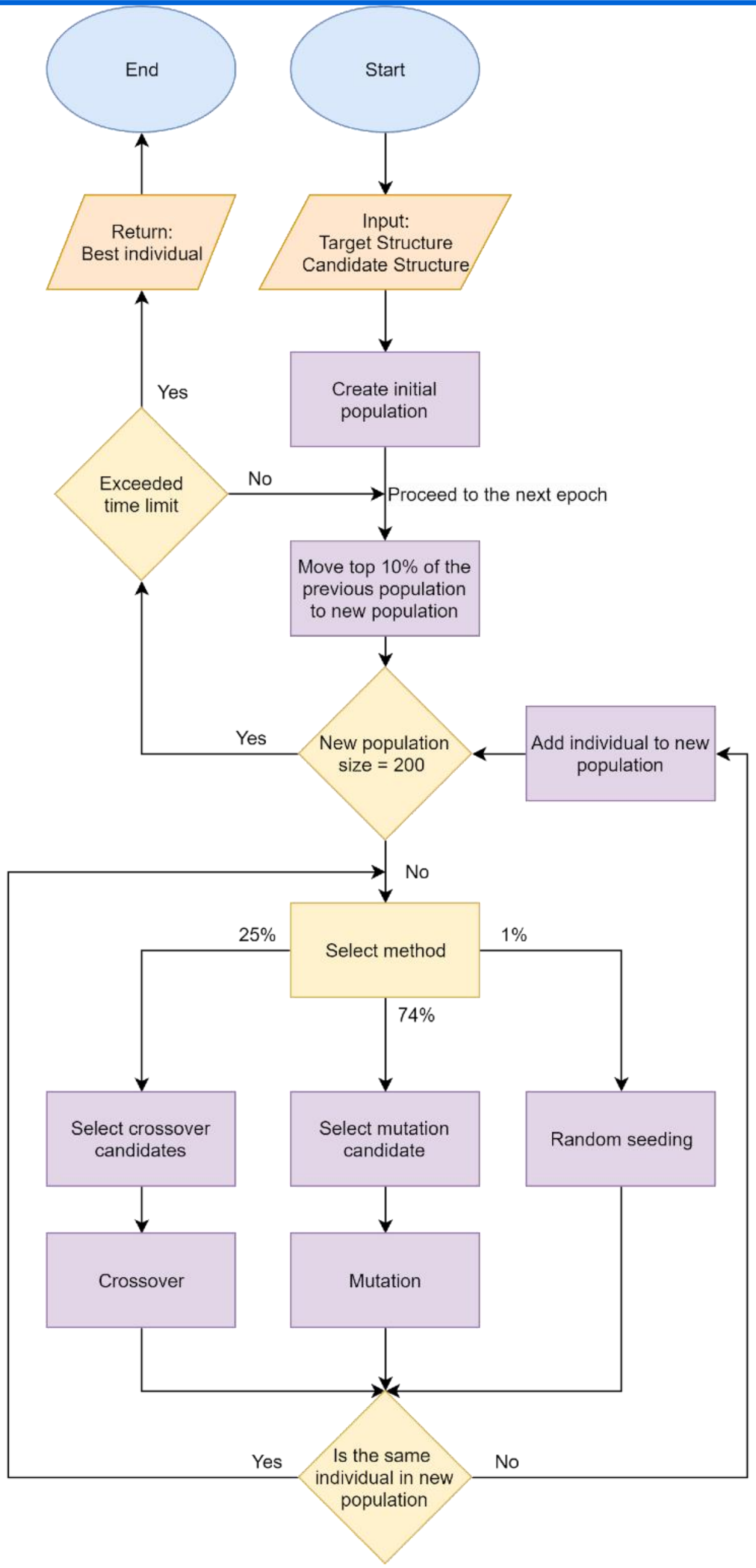

End
Start
Return:
Best individual
Input:
Target Structure
Candidate Structure
Yes
Create initial population
Exceeded time limit
No
Proceed to the next epoch
Move top 10% of the previous population to new population
Yes
New population size = 200
Add individual to new population
No
25%
Select method
1%
74%
Select crossover candidates
Select mutation candidate
Random seeding
Crossover
Mutation
Yes
Is the same individual in new population
No

Genetic search (GENS): operators

- **Mutation** operators of **the individual** (25%):
  - **Add/remove** randomly selected **unused residue pair from** both **the model** and **the reference** to the individual.
  - **Assign** randomly **unused residue** of **the model** to the randomly selected **residue** of **the reference** in the individual.

- **Crossover operators** applied **for** randomly selected **individuals pair** (74%):
  - **Inject** a randomly selected **subset** of **residue pairs** of **one individual** into **another**.
  - **Swap** randomly selected **subset** of **residue pairs** between **a pair** of **individuals**.

- **Addition** of randomly seeded **individual** (1%).

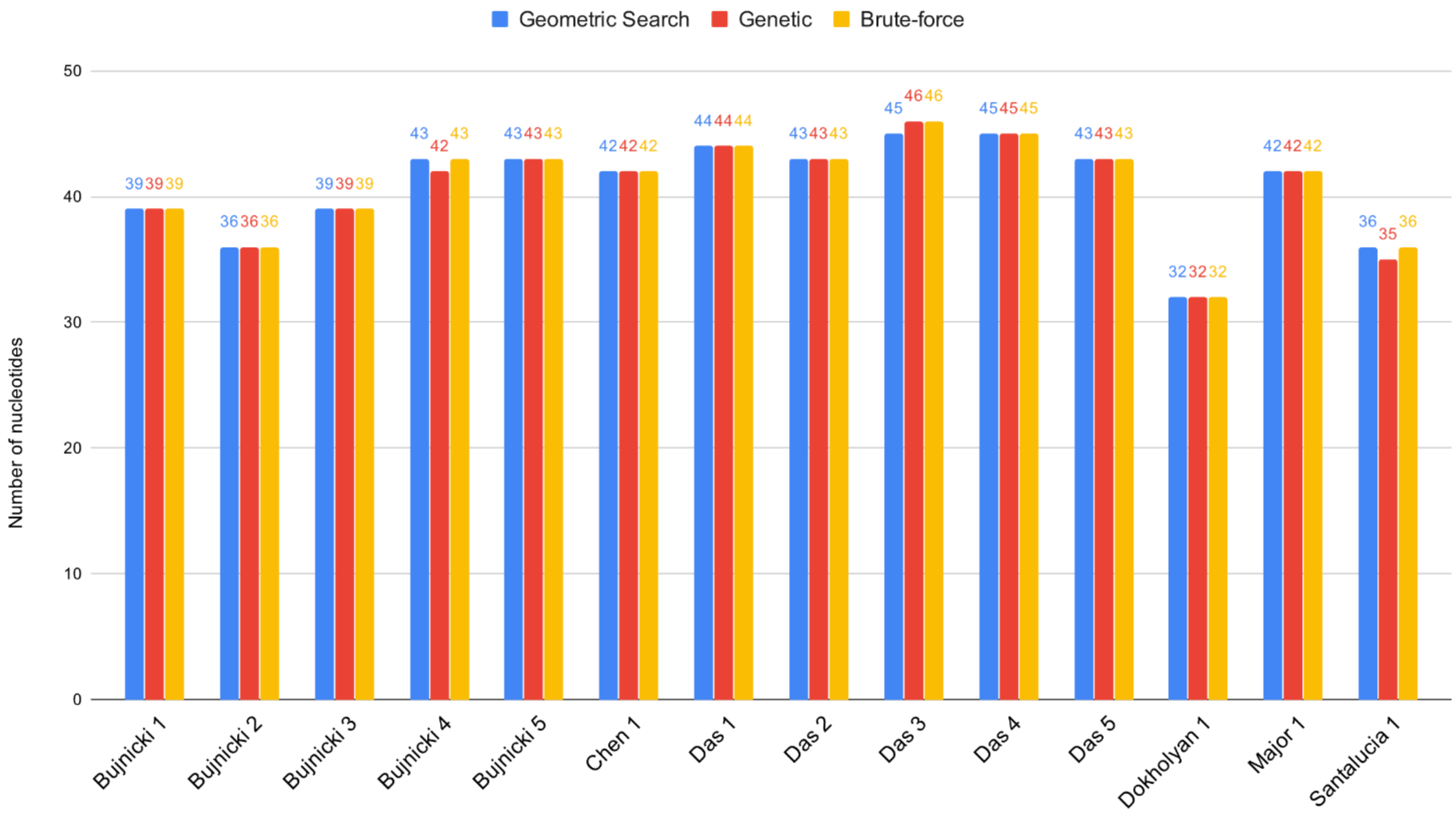

End
Start
Return: Best individual
Input: Target Structure Candidate Structure
Yes
Create initial population
Exceeded time limit
No
Proceed to the next epoch
Move top 10% of the previous population to new population
Yes
New population size = 200
Add individual to new population
No
25%
Select method
1%
74%
Select crossover candidates
Select mutation candidate
Random seeding
Crossover
Mutation
Yes
Is the same individual in new population
No

Genetic search (GENS): pros and cons

- **Advantages**:
  - Is able to find optimal solution.
  - Return many alternative acceptable solutions.
- **Disadvantages**:
  - Non-deterministic.
  - Parameters tuning required.

Computational experiments summary

- A representative set of **22 challenges** published in **the RNA-Puzzles** [1].
- Challenge #1 (46 nts).

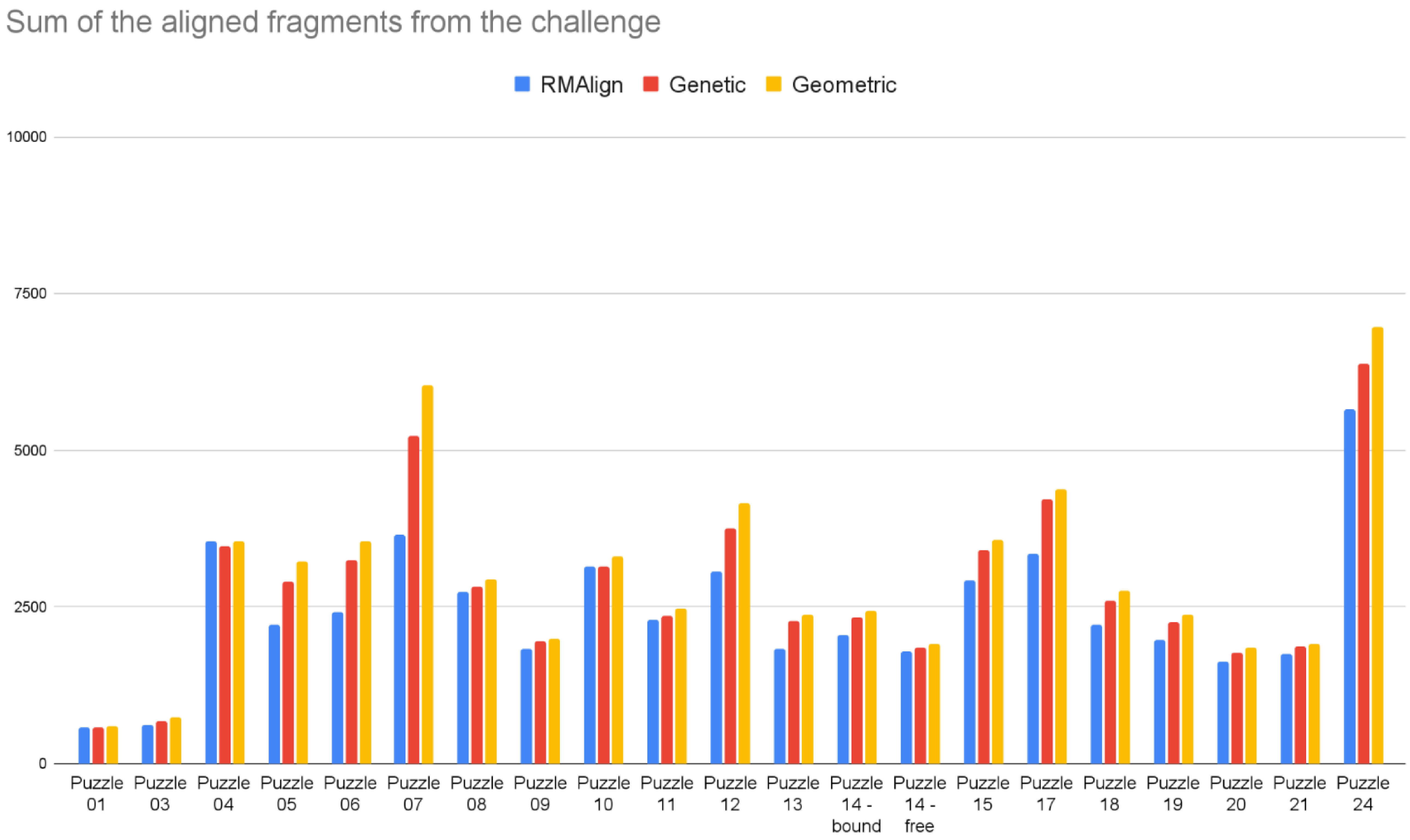


[1] Cruz, J. A., Blanchet, M. F., Boniecki, M., Bujnicki, J. M., Chen, S. J., Cao, S., ... & Westhof, E. (2012). RNA-Puzzles: a CASP-like evaluation of RNA three-dimensional structure prediction. Rna, 18(4), 610-625.

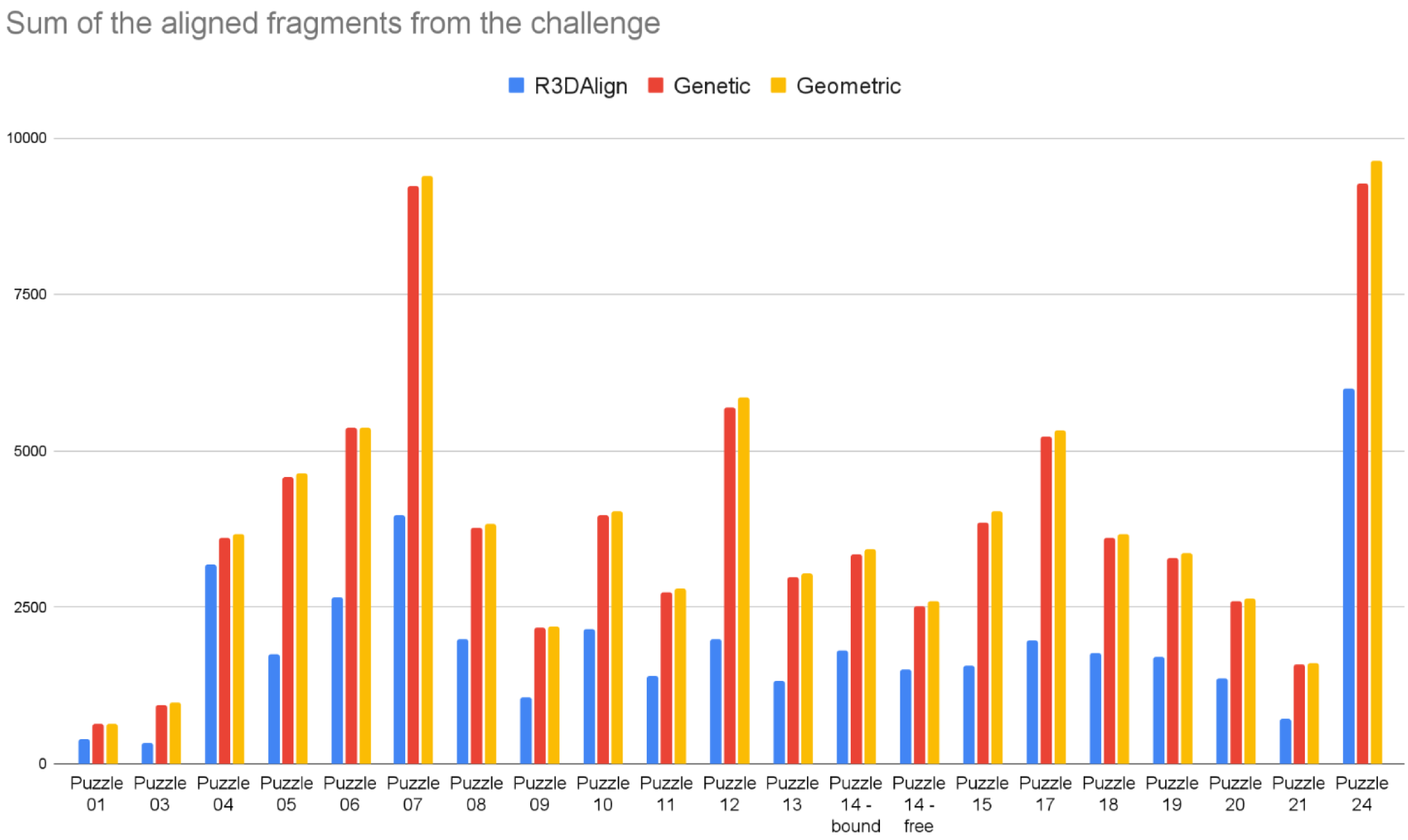
Computational experiments summary (2)

- For every RNA-Puzzles challenge, we computed **RNA 3D structure alignment** between **every 3D submission** and **the corresponding reference structure** using **every considered approach**.
- We **executed the state-of-the-art algorithms** to **get the alignment** and then **computed the RMSD score** for **aligned residues** in the solution. Finally, **the RMSD score was used** as **a cut-off value** applied for the proposed algorithms for this particular model-residue pair.
- **The proposed algorithms were ranked** based on **the total number** of **aligned residues** for **all considered model-reference pairs** in **the particular challenge within the context** of every considered **state-of-the-art algorithm independently**.

Computational experiments summary (3)



Computational experiments summary (4)



Conclusions

- The **algorithms**, i.e., *geometric search heuristic* (GEOS) and *genetic search algorithm* (GENS) **solving the RNA 3D structure alignment** have been proposed.
    - They are freely-available at GitHub (`https://github.com/RNApolis/rnahugs`).
- **Results** of computational experiments **confirming the accuracy** of **the proposed algorithms** have been presented.
- The proposed approaches usually **outperform the state-of-the-art algorithms in terms** of **quality**.
- **Processing efficiency** is **a limitation** of **the GENS**.
- We believe **the accurate RNA 3D structure alignment simplifies**, e.g., the **homologous modeling** of RNA tertiary structures.

Acknowledgments

Michal Zurkowski

Marta Szachniuk

Thank you for your attention!

WWW: `http://www.cs.put.poznan.pl/mantczak`

Contact: `Maciej.Antczak@cs.put.poznan.pl`

This work has been partially supported by the National Science Center, Poland (2019/35/B/ST6/03074) and statutory funds of Poznan University of Technology, Poland.