

LLM Agent-Based Modeling of RNA Ancestral State Reconstruction

Work in Progress, preliminary results

OBJECTIVES

Construct an LLM agent to describe RNA evolution over time using the following contextual sources:

- **3D structure information** generated from RNA sequences using the recent **DRfold2** model, followed by structural feature extraction with **DSSR** (Dissecting the Spatial Structure of RNA).
- **Structural refinement** through RNA generation guided by the **conditional diffusion model RiboDiffusion**, enabling mutual correction between RNA sequences and predicted 3D structures via bidirectional inference.

This work attempts to formally model approximate RNA evolution over time by leveraging a large language model (LLM) integrated with structured contextual data and agent tools.

Literature used:

LLM: Anthropic. Claude Sonnet 4 (20250514). <https://www.anthropic.com>. (Mai **2025**)

DRfold2: Li, Yang Li et al., Ab initio RNA structure prediction with composite language model and denoised end-to-end learning, Cold Spring Harbor Laboratory, (**2025**)

Ribodiffusion: Han Huang et al., RiboDiffusion: tertiary structure-based RNA inverse folding with generative diffusion models, Bioinformatics, (**2024**)

RNA-FM: Jiayang Chen et al., Interpretable RNA Foundation Model from Unannotated Data for Highly Accurate RNA Structure and Function Predictions, bioRxiv, (**2022**)

DSSR: Xiang-Jun Lu, DSSR-enabled innovative schematics of 3D nucleic acid structures with PyMOL, Nucleic Acids Research, (**2020**)

OUTLINE

❖ Objectives

❖ Dataset

- What is RNA molecule
- RNACentral database

❖ LLM Agent-Based Computational Pipeline

- Simple LLM Prompt and Context example for RNA Sequences
- Context Enrichment
- Pipeline Scheme
- Prompt Design

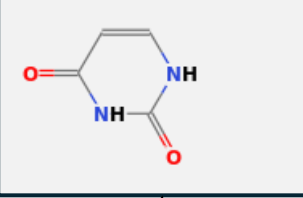
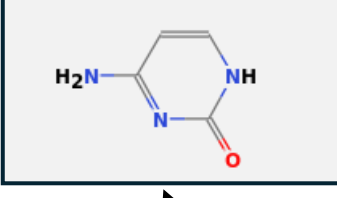
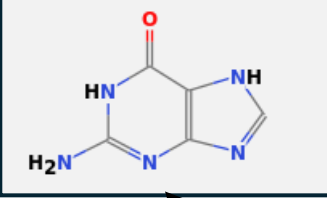
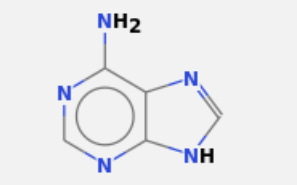
❖ Results

❖ Conclusions

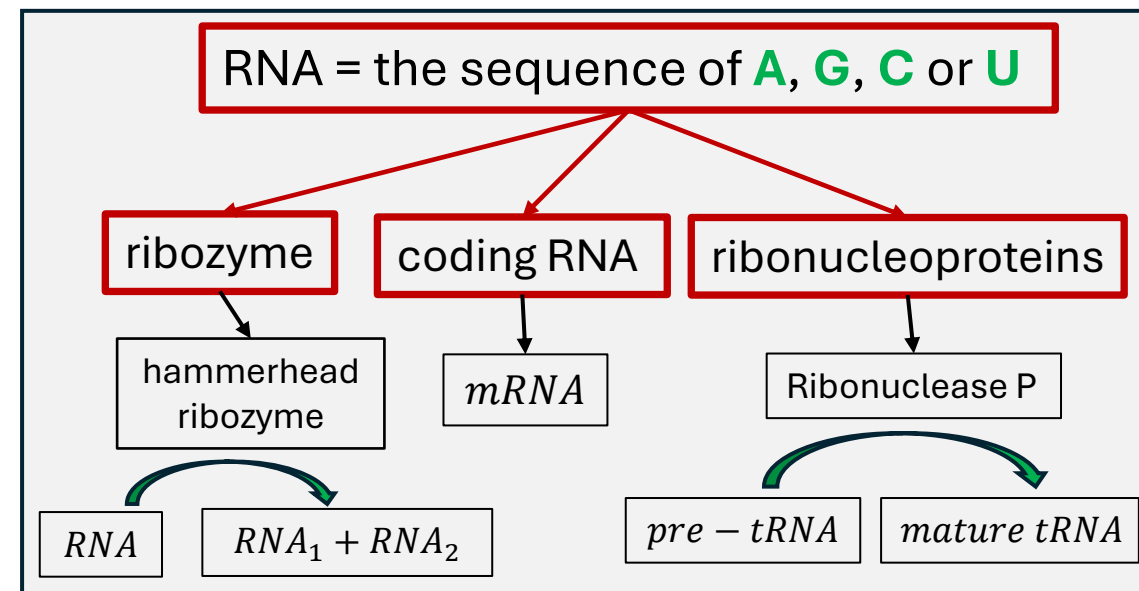
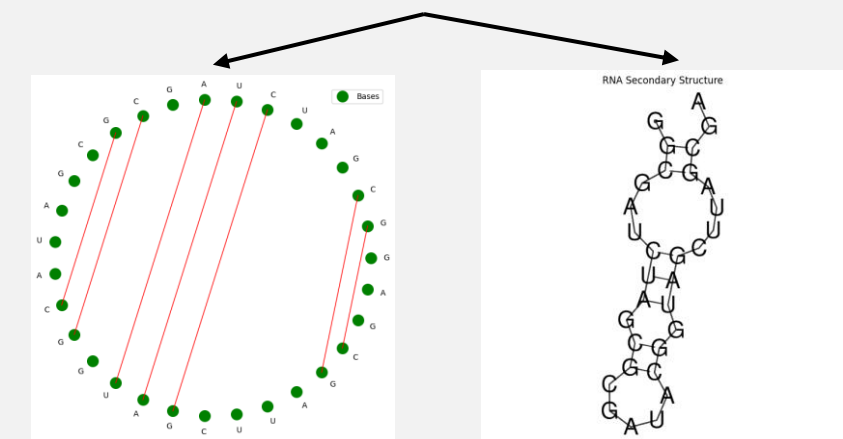
❖ 1. What's next

❖ 2. What is possible destination: Hairpins or ssRNA ?

What is RNA molecule ?



	Adenine	Guanine	Cytosine	Uracil
O	0	1	1	2
N	5	5	3	2
C	5	5	4	4
H	5	5	5	4
$\Delta_f H_{solid}^0, kJ/mol$	96.9	-183.9	-221	-424.4
$\Delta_c H_{solid}^0, kJ/mol$	-2779.0	-2498.2	-2067	-1721.3
$M_w, g/mol$	135	151	111	112
Hydrogen bonds	2	3	3	2

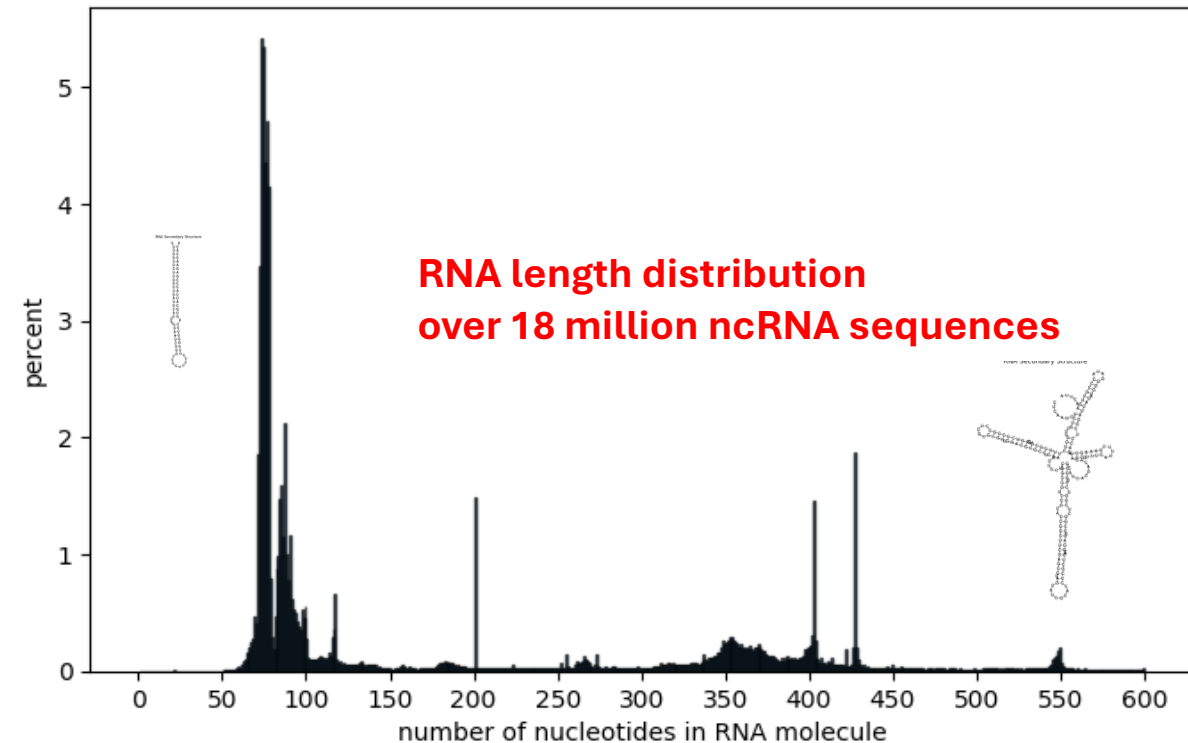
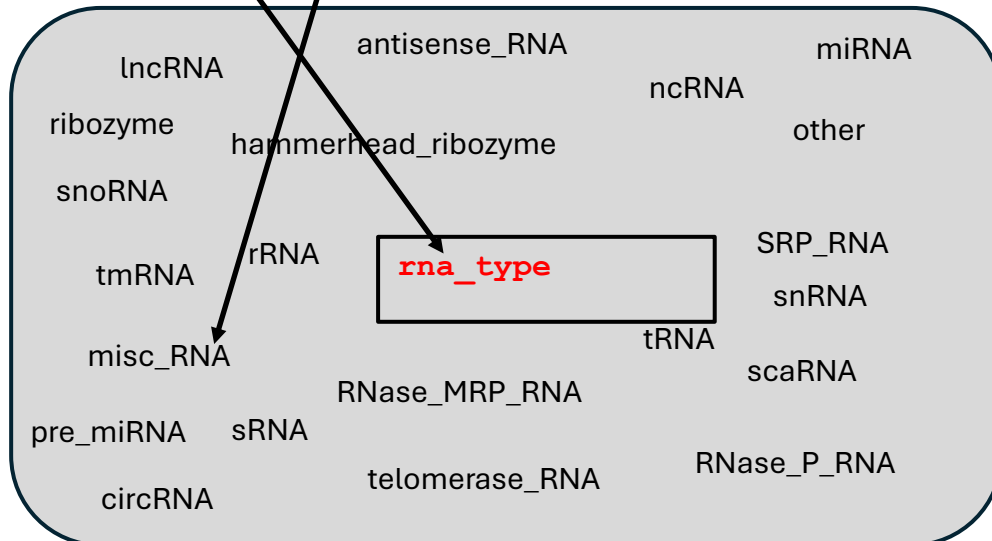
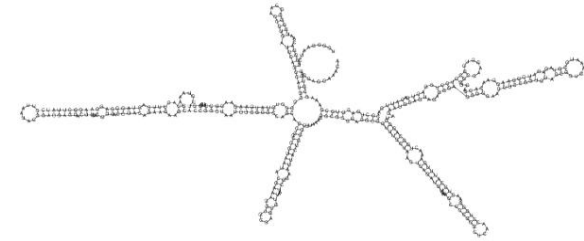


The dataset: RNACentral database

RNACentral Browsable API

https://rnacentral.org/api/v1/rna/?page=3&page_size=100 gives:

```
{ "url": "http://rnacentral.org/api/v1/rna/URS0002915621",
  "rnacentral_id": "URS0002915621",
  "md5": "fee3fe68dbd91ee898bffd9d4b89b2e9",
  "sequence": "AUGGAUGGUUGAUCAGAGAACGUACAUUUUAUAAAUGGUGUAUGUCAAUUGAUCCACAGUCCCU",
  "length": 64,
  "xrefs": "http://rnacentral.org/api/v1/rna/URS0002915621/xrefs",
  "publications": "http://rnacentral.org/api/v1/rna/URS0002915621/publications",
  "is_active": true,
  "description": "pre_miRNA from 0 species",
  "rna_type": "pre_miRNA",
  "count_distinct_organisms": 4,
  "distinct_databases": [ "Rfam" ] }, ...
```



RNA Prompt analysis using Large Language Model (RNA sequence comparison)

The answer is fully correct.

You are an expert in molecular biology. Your task is to analyze two RNA sequences and determine which is more likely to be found in a real biological organism. Justify your answer in one sentence.

Sequence 1: GGGGGGG

Sequence 2: AAGUCGCGCCGAAAAGGUGUCUCUU

Question: Which sequence is more biologically realistic?

Answer format: One sentence explaining which sequence is more likely and why.

1. Sequence 2 is more biologically realistic because it has a more diverse nucleotide composition and is not a repetition of a single nucleotide.

Including additional context would help make the answer clearer and more specific

```
pip install biopython
from Bio.Align import PairwiseAligner
```

```
<class 'Bio.Align.Alignment'>
target      0 GAG--GCG-----GGUG----- 10
            0 .||--|||-----|||----- 25
query       0 AAGUCGCGCCGAAAAGGUGUCUCUU 25
```

The LLM has difficulty understanding this.

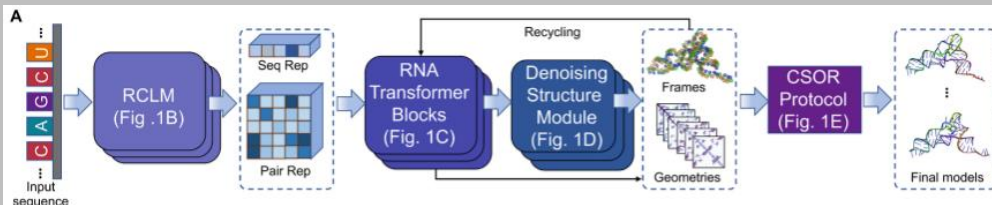
```
{
  "target_sequence": "GAGGCGGGUG",
  "query_sequence": "AAGUCGCGCCGAAAAGGUGUCUCUU",
  "aligned_target": "GAG--GCG-----GGUG-----",
  "aligned_query": "AAGUCGCGCCGAAAAGGUGUCUCUU",
  "score": 0.2,
  "identical_nucleotide_counts": {
    "A-A": 1,
    "G-G": 6,
    "C-C": 1,
    "U-U": 1
  }
}
```

dictionaries **are better understood** by LLMs

USE for
LLM
prompt

Go further with LLM context enrichment: combine DRfold2 and DSSR to extract 3D structural parameters of RNA

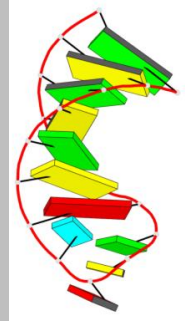
DRfold2 enables the prediction of RNA 3D structures (in PDB format) from nucleotide sequences provided in FASTA format



DRfold2 pipeline for end-to-end RNA structure prediction.

GCGCGCAUACGUGCGCGC

DRfold2



Li, Yang Li et al., Ab initio RNA structure prediction with composite language model and denoised end-to-end learning, Cold Spring Harbor Laboratory, (2025), doi = {10.1101/2025.03.05.641632}

DSSR

API: <http://skmatic.x3dna.org/api>

```
{
  "general_info": {
    "full_RNA_sequence": "GCGCGCAUACGUGCGCGC",
    "length": 18,
    "base_pairs": 7,
    "hydrogen_bonds": 25,
    "dot_bracket": [
      "(((((((.....)))))))))"
    ],
    "splayUnits": [
      "UA"
    ],
    "hairpins": [
      "AUACGU"
    ],
    "stacks": [
      "AU",
      "ACGU"
    ]
  },
  "helices_info": {
    "full_RNA_sequence": "GCGCGCAUACGUGCGCGC",
    "helix_0": {
      "base_pairs": 7,
      "strand_1": "GCGCGCA",
      "strand_2": "CGCGCGU",
      "helix_form": "AAAAAA"
    }
  }
}
```

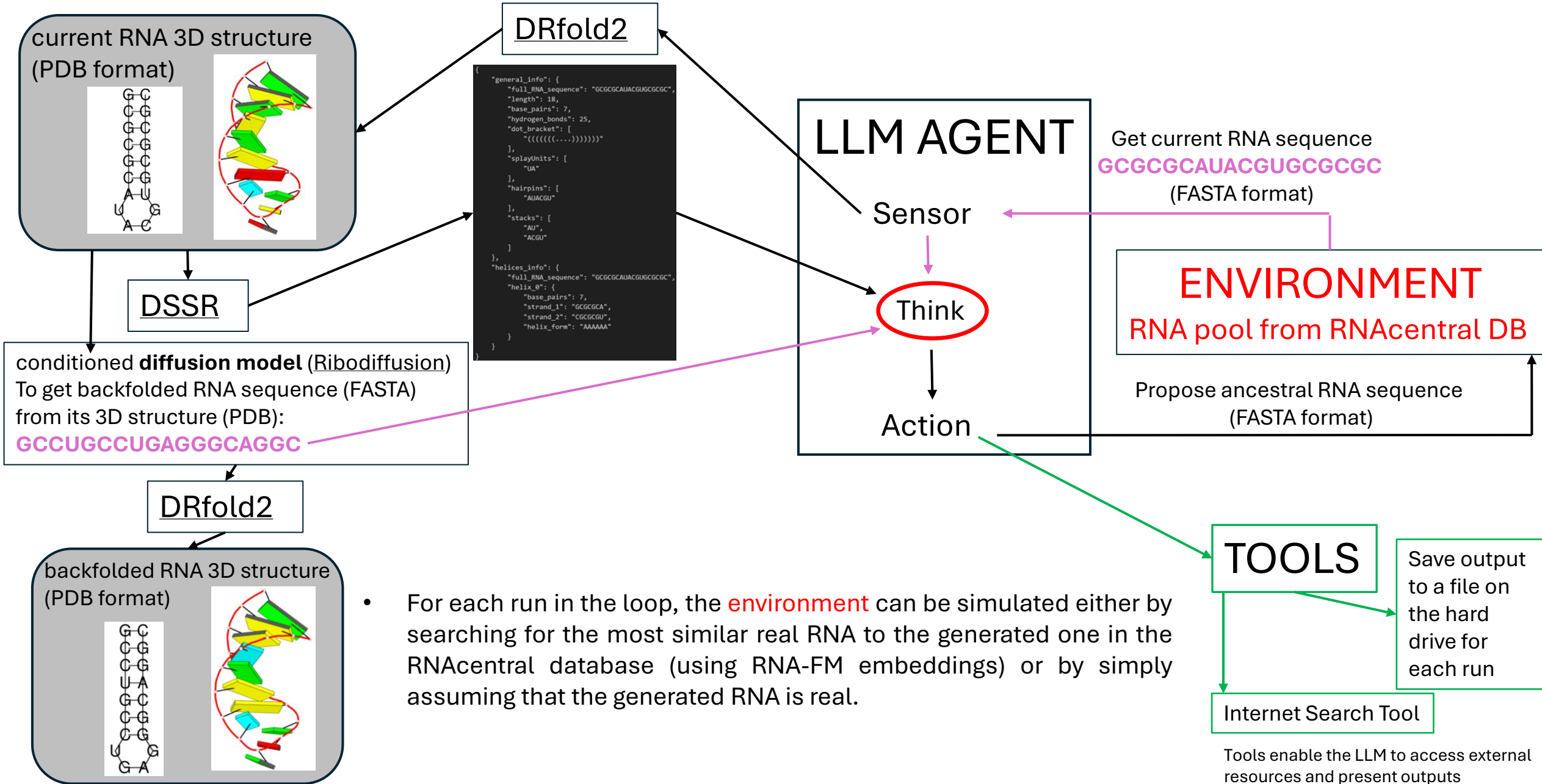
Xiang-Jun Lu, DSSR-enabled innovative schematics of 3D nucleic acid structures with PyMOL, Nucleic Acids Research, Vol 48, number 13, p. e74-e74, (2020), doi = {10.1093/nar/gkaa426}

Input

Output for LLM

From RNA sequence to a parametric description of its 3D structure

Computational Pipeline: LLM Agent-Based Simulation of Ancestral RNA Sequences



Prompt Engineering: Designing Effective Inputs for LLMs

```
prompt = ChatPromptTemplate.from_messages(  
    [  
        (  
            "system",  
            """  
            You are a research assistant that will help generate a research paper.  
            Answer the user query and use necessary tools.  
            Wrap the output in this format and provide no other text:  
            {format_instructions}  
            Do not include any other commentary or explanation — output only the JSON.  
            """,  
        ),  
        ("placeholder", "{chat_history}"),  
        (  
            "human",  
            """{query}""",  
        ),  
    ],  
    partial(format_instructions=parser.get_format_instructions())  
)
```

Current RNA sequence:
{current_sequence}

3D Structural Parameters of the current RNA:
{rna_structure_block}

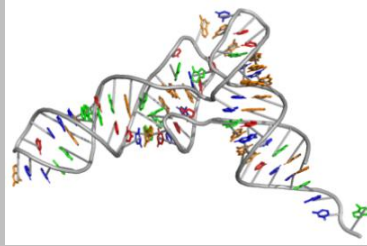
Backfolded RNA sequence from 3D structure of the current RNA:
{backfolded_sequence}

```
query = (  
    f"Propose a plausible ancestral RNA sequence based on the current RNA sequence ant its properties."  
    f"Save both a summary and the proposed ancestral_sequence to a file"  
    f"and begin each output in the file with the following section header: evolution_step_{i}"  
    raw_response = agent_executor.invoke({"query": query,  
        "current_sequence": current_sequence,  
        "backfolded_sequence": backfolded_sequence,  
        "rna_structure_block": rna_structure_block})  
)
```

No conditions were imposed on the generated ancestral RNA sequence (e.g., limits on length or number of mutation events per step)!

STEP 1

GGACCGAUAGUCUCGUGGUAGGAACUUAACCCUGCCAAGG
AUAGAGCCCGGGUUCGAAUCCGGGUCGGUCCACCA



The provided 75-nucleotide RNA sequence exhibits characteristic tRNA-like structural features with a cloverleaf secondary structure. Key structural parameters include:

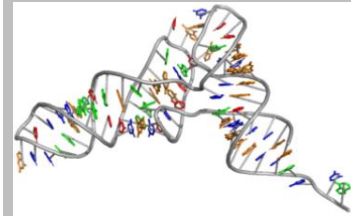
- Length: 75 nucleotides
- Base pairs: 29
- Dot-bracket notation showing complex stem-loop structures
- Multiple hairpins and junction regions
- Two coaxial stacks indicating sophisticated tertiary interactions

The first two images in each panel were generated using ViennaRNA package for rough visualization purposes only.

The last two images depict the actual 3D structures, predicted by DRfold2 and further processed by DSSR to extract structural parameters for LLM input.

ANCESTRAL_SEQUENCE:

GGCCCGAUAGCUCGUGGUAGGAACUUAACCCUGCCAAGGAUAGAG
CCCGGGUUCGAAUCCGGGUCGGGCCACCA



Rationale for Ancestral Sequence Reconstruction

1. ****Structural Conservation****: The ancestral sequence maintains the essential cloverleaf structure characteristic of early tRNA-like molecules, preserving critical base-pairing regions that ensure functional stability.
2. ****Reduced Complexity****: Early RNA molecules likely had simpler tertiary structures with fewer modified nucleotides and less sophisticated coaxial stacking arrangements.
3. ****Enhanced Thermostability****: Ancestral RNA sequences typically exhibited greater thermodynamic stability to survive in harsh primordial conditions, achieved through stronger base-pairing interactions.
4. ****Functional Core Preservation****: The proposed ancestral sequence maintains the essential functional domains including the acceptor stem, anticodon loop region, and critical junction sequences necessary for aminoacylation and ribosome binding.
5. ****Evolutionary Parsimony****: The reconstruction follows the principle of minimal evolutionary changes while preserving essential structural and functional characteristics required for early protein synthesis machinery.

The proposed ancestral sequence represents a plausible evolutionary precursor that could have given rise to the current RNA through gradual structural refinements and optimization for specific cellular environments.

Based on analysis of the current RNA sequence (75 nucleotides) with its complex secondary structure containing 29 base pairs, multiple hairpins, and helical regions, I propose an ancestral sequence that would have been simpler yet structurally viable. The current sequence shows sophisticated features including two major helices, multiple hairpin loops, and complex junction regions.

For the ancestral reconstruction, I considered several evolutionary principles:

1. Ancient RNA sequences likely had simpler secondary structures with fewer complex loops
2. Early RNA would have had stronger base-pairing stability in core regions
3. Primitive sequences would have contained more canonical Watson-Crick pairs
4. The ancestral form would retain essential structural motifs but with reduced complexity

The proposed ancestral sequence maintains the basic cloverleaf-like structure but with:

- Simplified hairpin loops with more stable base-pairing
- Reduced junction complexity
- Stronger stem regions with canonical base pairs
- Shorter variable loops
- Enhanced stability through increased GC content in critical structural regions

This ancestral form would represent an evolutionary intermediate that possessed the fundamental structural framework while being simpler and more thermodynamically stable than the modern sequence.

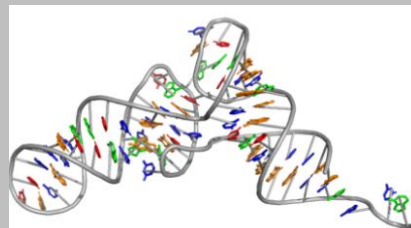
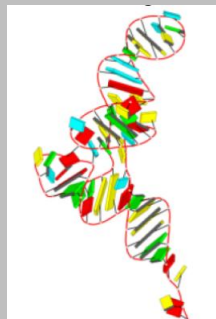
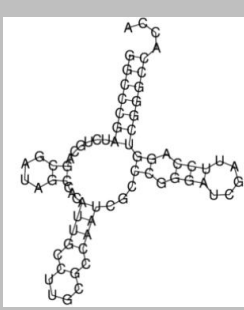
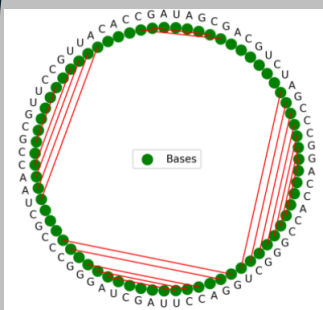
```
<class 'Bio.Align.Alignment'>  
target      0   GGACCGAAUAGCUCUGGUJGGAAGA CUUAACC CUGCCAAGGAUAGAGSC CGGGUUGC AAL  
               |||...|||.....  
query       0   GGCCTCAUAG -CUCUGGUJAGGA AC UUAACC CUGCCAAGGAUAGAGSC CGGGUUGC AAL  
               |..-.-....  
  
target     60  CGGGUCGGUCCACCA    75  
              |||||||.----- 75  
query      59  CGGGUCGGGCCACCA    74
```

Score: 140.0

To improve the reliability of these evolutionary trajectories, it is necessary to introduce constraints on the number of mutation events per step.

STEP 6

GGCCCGAUCUGCAGCGAUAGCCACAUUGCCUUGCGCCAAUCGCC
CGGGAUCGAUUCCAGGUCGGGCCACCA



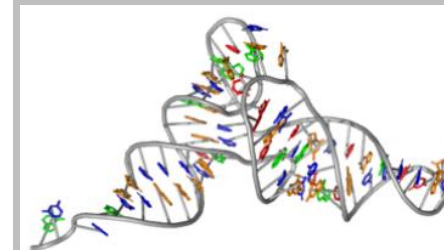
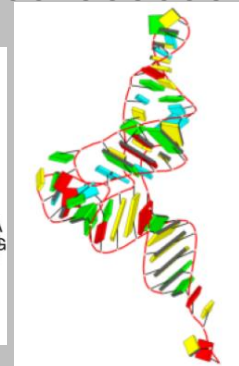
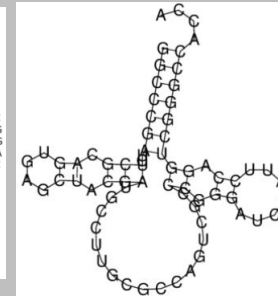
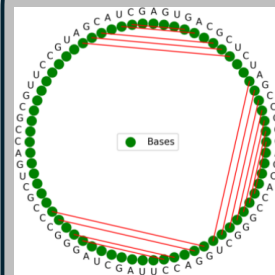
Based on the analysis of the current RNA sequence (71 nucleotides) and its structural properties, I propose an ancestral RNA sequence that maintains the critical structural elements while reflecting probable evolutionary simplifications. The current RNA displays a complex secondary structure with 32 base pairs, 3 major helices, and multiple structural motifs including hairpins, bulges, and junctions.

Key observations from the current sequence:

- Length: 71 nucleotides
- Three major helices (13, 8, and 5 base pairs respectively)
- Complex dot-bracket notation indicating nested stem-loop structures
- Multiple hairpin loops (CGAUAG, CUUGCG, AUCGAUU)
- Junction regions and single-stranded segments
- Strong GC content in stem regions providing stability

ANCESTRAL_SEQUENCE:

GGCCCGAUCUCGCAGUGAGCUACGAUGCCUUGCGCCAGUCGCCC
GGGAUCGAUUCCAGGUCGGGCCACCA



The proposed ancestral sequence likely represents a simpler evolutionary precursor that:

1. Maintains the core three-helix architecture
2. Has simplified loop regions with fewer complex motifs
3. Retains essential base-pairing patterns for structural stability
4. Shows reduced sequence complexity while preserving functional domains
5. Exhibits characteristics consistent with early RNA world molecules

The ancestral reconstruction considers:

- Conservation of critical structural elements
- Simplification of complex motifs that likely evolved later
- Maintenance of thermodynamic stability
- Preservation of potential catalytic or binding sites
- Evolutionary parsimony in sequence changes

```
<class 'Bio.Align.Alignment'>
target      0 GGCCCGAUCU-GCAGCGAUAGCCACAUUGCCUUGCGCCAAUCGCCCGGGAUCGAUUCCAG
              |||||
query       0 GGCCCGAUCUCGAGUG--AGCUACGAUGCCUUGCGCCAGUCGCCCGGGAUCGAUUCCAG

target      59 GUCGGGCCACCA 71
              |||||
query       58 GUCGGGCCACCA 70

Score: 118.5
```

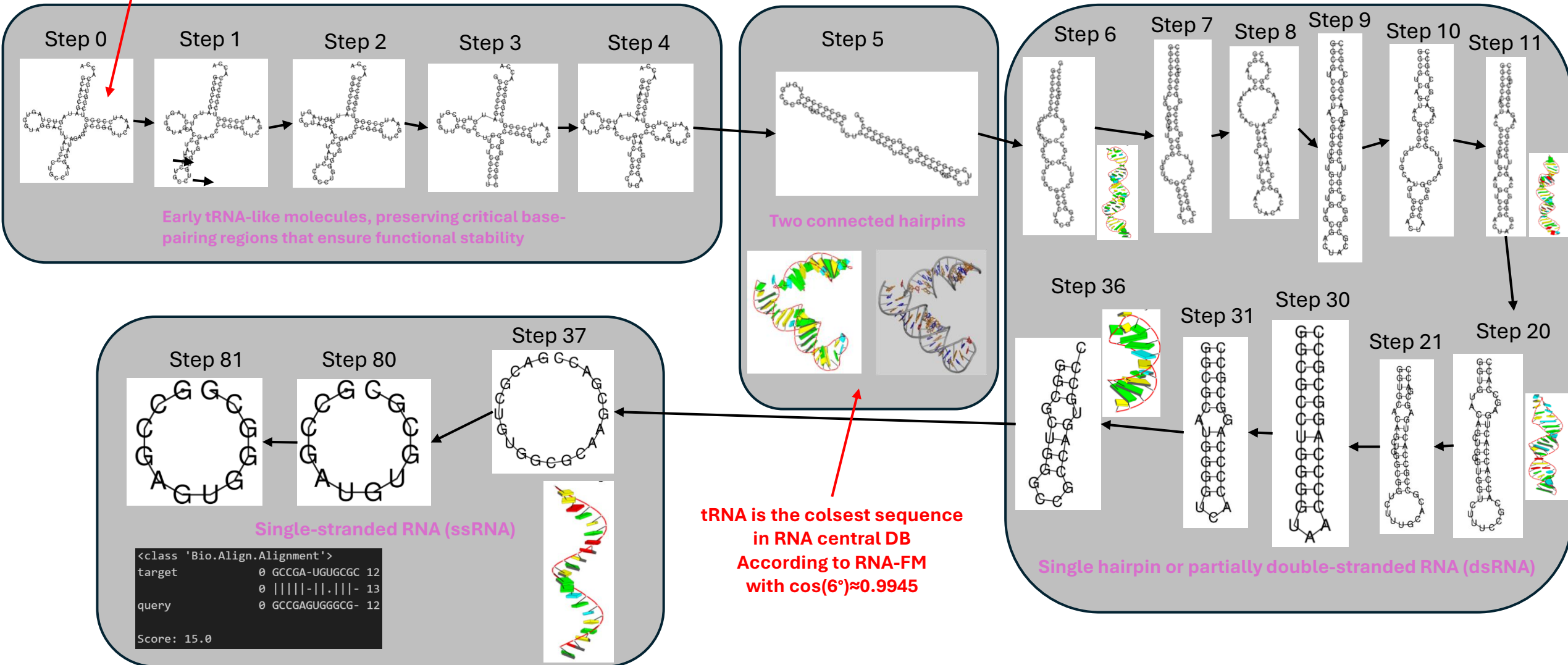
The first two images in each panel were generated using ViennaRNA package for rough visualization purposes only.

The last two images depict the actual 3D structures, predicted by DRfold2 and further processed by DSSR to extract structural parameters for LLM input

To improve the reliability of these evolutionary trajectories, it is necessary to introduce constraints on the number of mutation events per step.

Results of Simulated Reverse Evolution of tRNA^{Gly} Asgard_group_archaeon (Second Run Using Internet Search Tool)

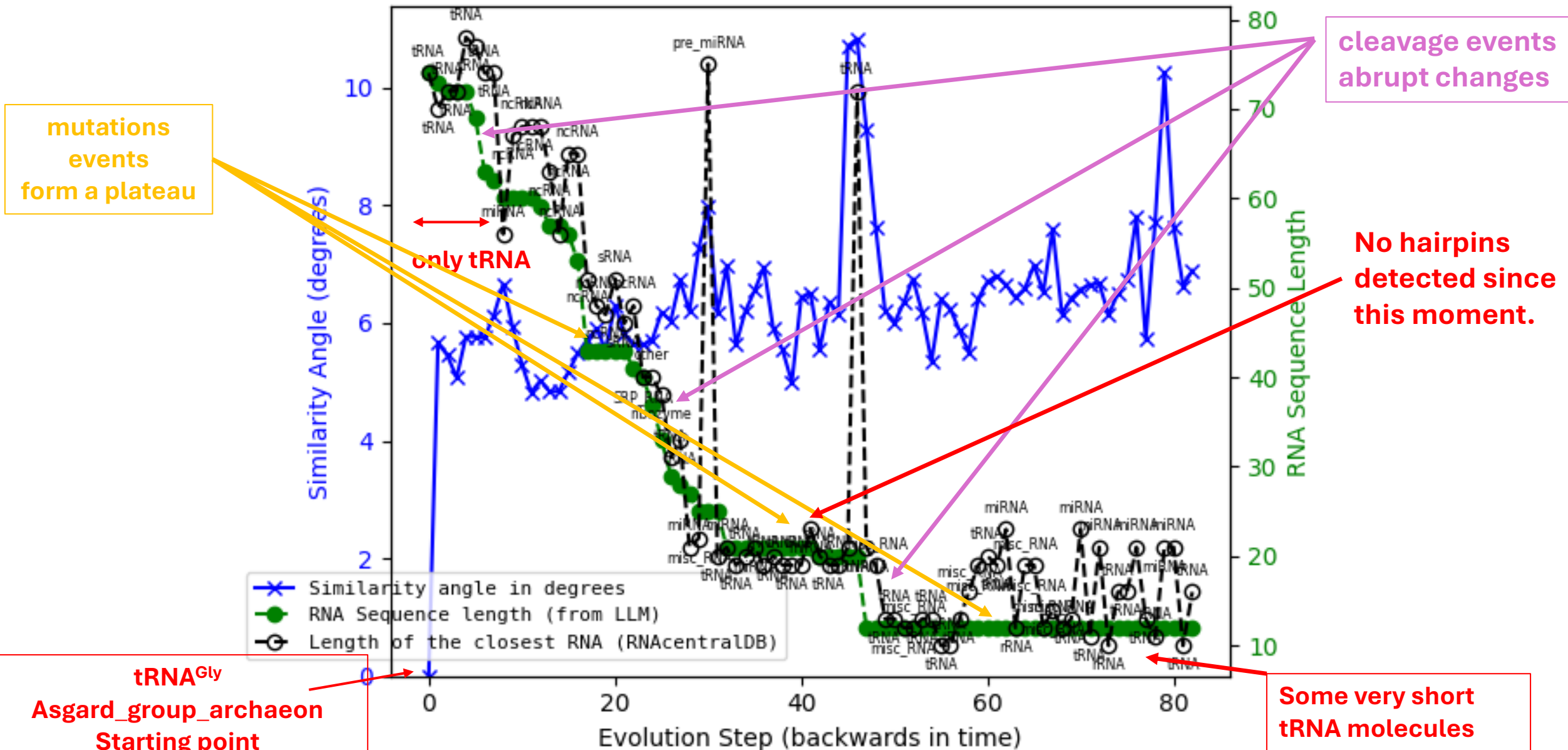
tRNA^{Gly}
Asgard_group_archaeon
Starting point



To improve the reliability of these evolutionary trajectories, it is necessary to introduce constraints on the number of mutation events per step.

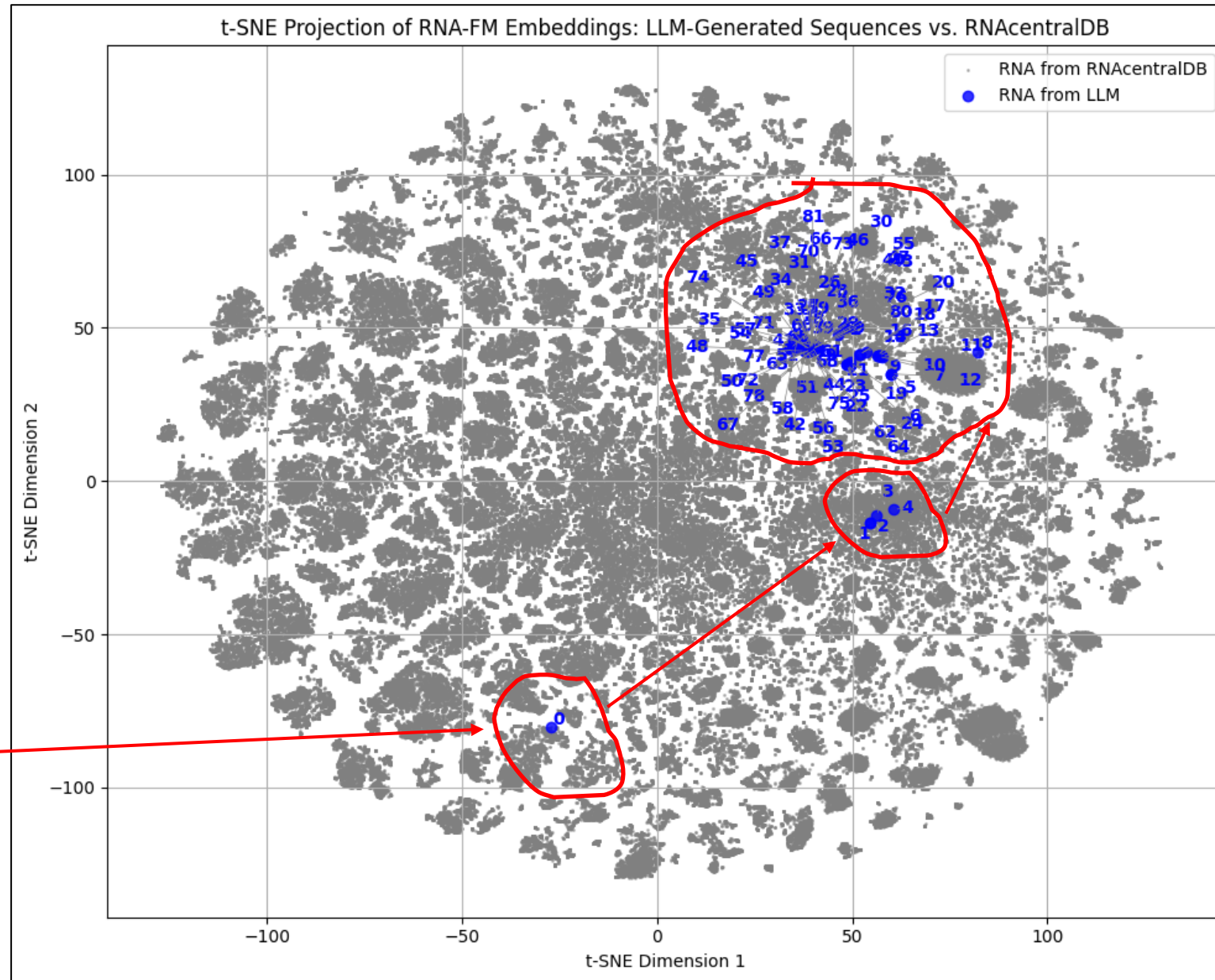
Using Large Language Models to Simulate the Reverse Evolution of tRNA^{Gly}

Smallest similarity angle between RNAs from LLM and RNACentralDB (comparison using RNA-FM)



From tRNA^{Gly} to short tRNA and miscellaneous RNA

How Accurately do LLM-Designed RNAs Represent Real RNA?



0, 1, ... are
evoution step
(backwards in
time)

tRNA^{Gly}
Asgard_group_archaeon
Starting point

400,000 tRNA sequences from the RNAcentral database with lengths less than 80 nucleotides

CONCLUSIONS

- ❖ **LLM agent can be used to generate new RNA sequences, potentially providing an advantage when guiding RNA generation only with conditional diffusion models**
- ❖ **A preliminary attempt to model RNA reverse evolution shows a decrease in RNA length and identifies three distinct evolutionary patterns:**
 - The ancestral sequence maintains the essential cloverleaf structure characteristic of early tRNA-like molecules, preserving critical base-pairing regions.
 - The tRNA-like molecule transforms into a single hairpin or partially double-stranded RNA (dsRNA) through intermediate structures (e.g., two connected hairpins).
 - Finally, the hairpin or dsRNA transforms into single-stranded RNA (ssRNA).
- ❖ **To improve the reliability of these evolutionary trajectories, it is necessary to introduce **constraints** on the number of mutation events per step.**

1. What's next ?

1. Introduce **limits** on the length or **the number of mutation events per step**
2. Search for mistakes in the answers and correct the prompt accordingly
3. Try different agent architectures for selective and iterative pipelines

4. More context from answering questions about RNA sequences

They effectively used the RNA-FM sequence encoder to embed RNA sequences for alignment with natural language

Only RNA Sequence (e.g., "AUGGCUUAGCU...")

The Question (e.g., "What is the function of this RNA?")

RNA-GPT

answer

at this time, **RNA-GPT** is not freely available for use

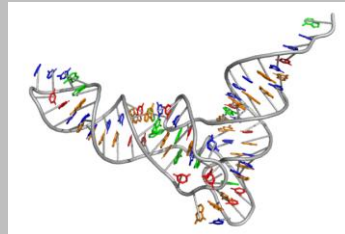
Yijia Xiao et al. , RNA-GPT: Multimodal Generative System for RNA Sequence Understanding, **2025**

5. More context from answering questions about RNA images?

HealthGPT is a medical large vision-language model for unifying comprehension and generation via heterogeneous knowledge adaptation

HealthGPT **understands medical images** (e.g. answer questions about them)

Q: What type of RNA molecule is shown in the image?



HealthGPT

This is tRNA not rRNA...

HealthGPT-M3 is available

HealthGPT-L14 weights are still not released yet

ANSWER of HealthGPT-M3 : The image shows a **ribosomal RNA (rRNA)** molecule. Ribosomal RNA is a type of RNA that is a key component of the ribosome, the cellular machinery responsible for protein synthesis. The image depicts the secondary structure of the rRNA molecule, which is formed by the folding and pairing of the RNA strands.

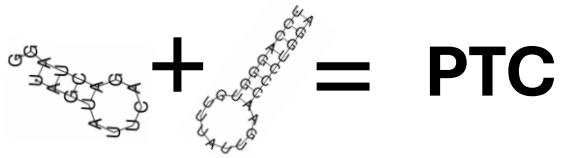
Tianwei Lin et al., HealthGPT: A Medical Large Vision-Language Model for Unifying Comprehension and Generation via Heterogeneous Knowledge Adaptation, **2025**

2. What is possible destination: Hairpins ?

Peptidyl Transferase Center (PTC) Sequences

the idea

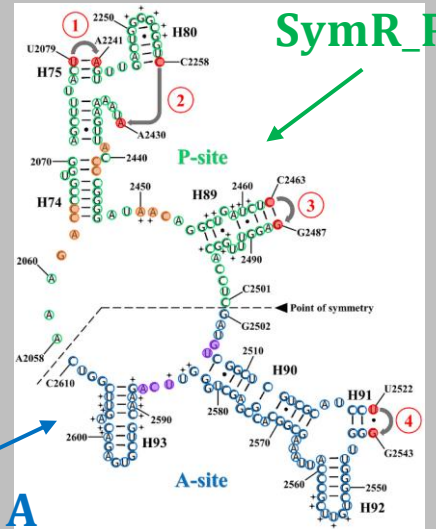
the dimerization of two similar RNA structures



“The peptidyl transferase center (PTC) evolved from a primitive system in the RNA world comprising tRNA-like molecules formed by **duplication of minihelix-like small RNA**”

Tamura, J. Biosci, 2011

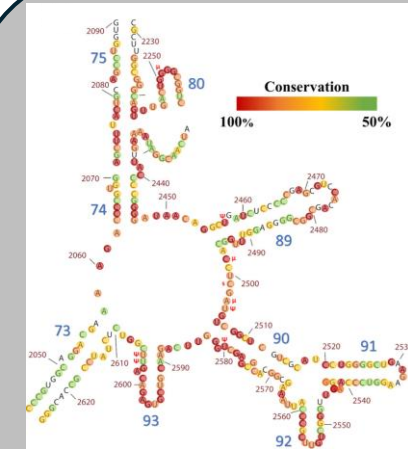
pseudosymmetrical region



$$\text{SymR}_{\text{PA}} = \text{SymR}_{\text{P}} + \text{SymR}_{\text{A}}$$

Secondary structure of the pseudosymmetrical region (**SymR**; Agmon et al., 2005), derived from the LSU secondary structure of *Thermus thermophilus* (Petrov et al., 2013). (Madhan R. Tirumalai et al., 2021)

PTC



PTC2 = red
PTC3 = PTC2 + orange
PTC4 = PTC3 + yellow
PTC5 = PTC4 + green

Nucleotide CONSERVATION level:

Red circles: 100% conservation (78 nt).
Orange circles: 90 to 99.9% conservation (68 nt)
Yellow circles: 70 to 89.9% (52nt)
Green circles: 50 to 69.9% conservation (49nt)
Black letters: less than 50% conservation (35nt)

(Bernier et al.; Faraday Discuss, 2014)
(Madhan R. Tirumalai et al., 2021)

SymR_P is older than SymR_PA
PTC2 is older than PTC3, PTC4, PTC5

rRNA evolution: Age Variability Across Different Regions



Last Universal Common Ancestor (LUCA)

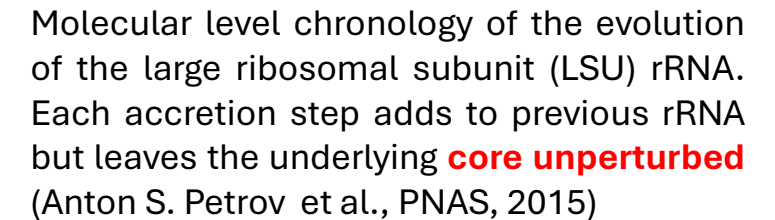
Diagram E illustrates the LSU of tRNA (E. coli) with various phases and a highlighted PTC region. The diagram shows a complex network of colored lines representing different phases of the LSU. A blue box highlights a specific region labeled 'PTC' (Phase Transition Core) and 'Phase 1'. A blue arrow points from the text 'LSU and SSU associated at the end of Phase 3' to this region. The legend indicates the following phases:

- Phase 6 (Red)
- Phase 5 (Orange)
- Phase 4 (Yellow)
- Phase 3 (Green)
- Phase 2 (Light Blue)
- Phase 1 (Dark Blue)

Text labels within the diagram include 'Phase 3', 'Phase 4', 'Phase 5', and 'Phase 6' near specific clusters of lines. The text 'LSU and SSU associated at the end of Phase 3' is located near the highlighted PTC region. The text 'LSU and SSU evolved independently & uncorrelated through Phases 1-3' is located below the legend.

LSU of tRNA (E. coli):

Blue part is the oldest one



This symmetry (**SymR**) suggests that the ancient ribosome may have been **a dimer of identical or nearly identical RNA molecules**, later evolving into the asymmetrical modern ribosome with **PTC**.