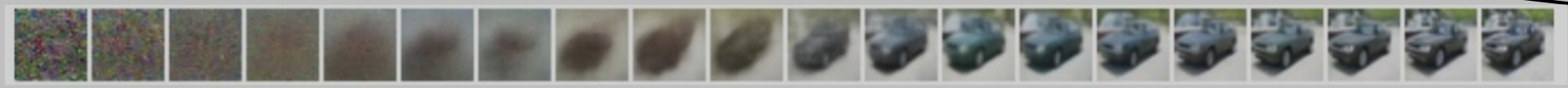
LLM Agent–Driven ncRNA Design via Structural Feedback and ncRNA Diffusion Model

Work in Progress, preliminary results

Inspiration from Diffusion Models for RNA Sequence Generation

1

noise → **2D images**



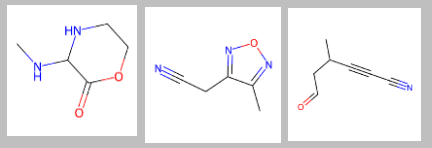
Ho, Jonathan & Jain, Ajay & Abbeel, Pieter. (2020). Denoising Diffusion Probabilistic Models. 10.48550/arXiv.2006.11239.

2

noise → **3D objects**

Norman Müller et al., (2022), DiffRF: Rendering-Guided 3D Radiance Field Diffusion, 10.48550/arXiv.2212.01206

noise → **molecules**



Clement Vignac et al., Midi: Mixed graph and 3d denoising diffusion for molecule generation, Joint European Conference on Machine Learning and Knowledge Discovery in Databases, (2023), pp 560

3

noise → **Realistic RNA sequence**

Kaixuan Huang et al., Latent Diffusion Models for Controllable RNA Sequence Generation (**2024**)

noise + RNA 3D structure → **RNA sequences for this 3D structure**

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

objective

Generate **two distinct RNA hairpins** connected by a single-stranded region.

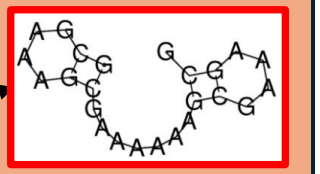
\+

RNA sequence prior if needed

This work objective

TOOLS

Large Language Model



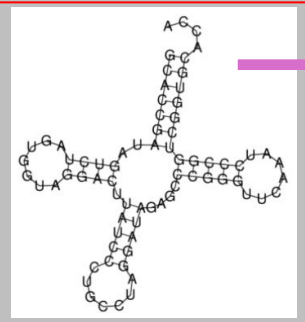
objective

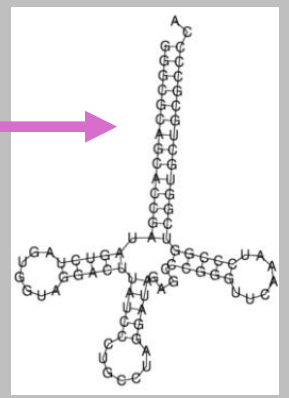
LLM ReAct Agent: Modification of a Specified Region of an RNA Molecule

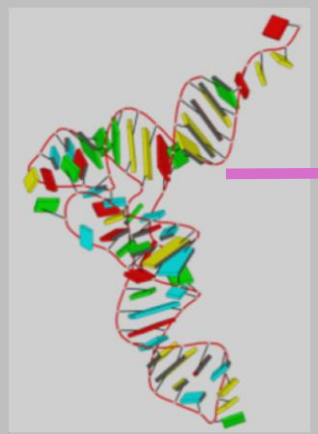
reference sequence

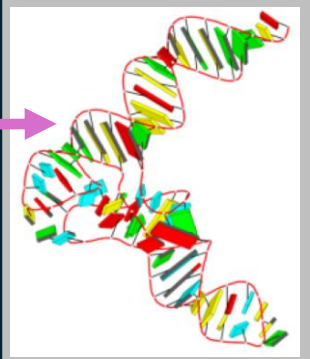
modified sequence

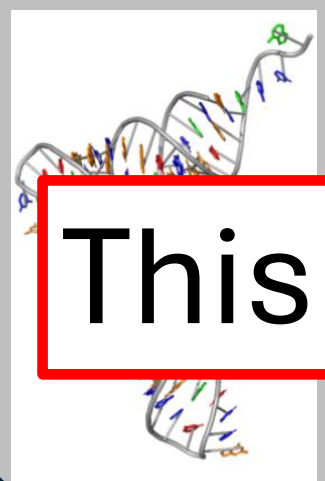
tRNA^{Gly}
Asgard_group_archaeon







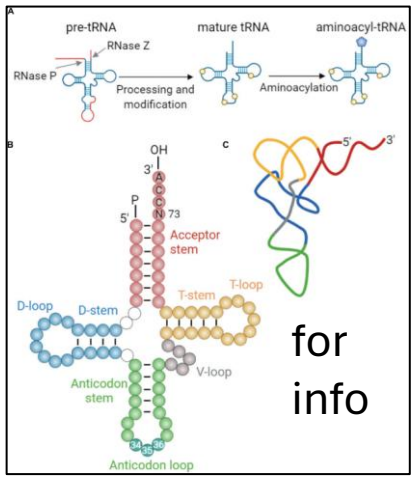




f"Current tRNA sequence (input): {tRNA}\n\n"
"Modify this reference sequence by increasing the length of the acceptor stem to **twice its original size. "**
"Ensure that **no more than 20% of the nucleotides are changed or added.\n\n"**
"Keep the rest of the tRNA unchanged.\n\n"
"Verify the structure and return the modified RNA sequence."

Current tRNA sequence (input):
GCACCGAUAGUCUAGUGGUAGGACUUAUCCCUGCCUAGGAUAGAGCCCGGGUUCAAAUCCCGGUCGGUGCACCA

I'll help you modify the tRNA sequence by doubling the acceptor stem length while keeping changes to no more than 20% of nucleotides. First, let me analyze the current structure to identify the acceptor
Tool Calls: **get_dotBracket_length**



'Now let me identify the stems to locate the acceptor stem
Tool Calls: get_stems

'Let me also get the hairpin loops to better understand the structure
Tool Calls: get_hairpins

Based on the analysis, I can see this is a standard tRNA structure with:\n- **Acceptor stem**: helix_0 with 7 base pairs (GCACCGA paired with CGUGGCU)\n- D-arm: helix_1 with 4 base pairs\n- Anticodon arm: helix_2 with 5 base pairs \n- TψC arm: helix_3 with 5 base pairs\n\nTo double the acceptor stem length from 7 to 14 base pairs while changing ≤20% of nucleotides (≤15 out of 74), I'll:\n\n1. Extend the 5' end with 7 complementary base pairs\n2. Extend the 3' end with 7 complementary base pairs\n3. Keep all other structural elements unchanged\n\nHere's the modified sequence: GGGC...

Figure adapted from Frontiers in Molecular Biosciences (Irem Avcilar-

This is a slide from the results section.

**:

acceptor stem## Modified tRNA sequence:
`GGGCGCAGCACCGAUAGUCUAGUGGUAGGACUUAUCCCUGCCUAGGAUAGAGCCCGGGUUCAAAUCCCGGUCGGUGCUGCGCCCCA` **The acceptor stem is now twice its original length (14 bp vs 7 bp) while preserving the overall tRNA secondary structure and staying within the 20% nucleotide change constraint.

OBJECTIVES

Develop two LLM agent–driven pipelines for RNA sequence generation, guided by human-designed prompts, that automatically integrate RNA structural feedback, diffusion-based generative modeling, and internet search.

- **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** is guided by the **conditional diffusion model RiboDiffusion**, enabling mutual correction between RNA sequences and predicted 3D structures via bidirectional inference.

Literature used:

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

***LLM ReAct agent**: Shunyu Yao et al., ReAct: Synergizing Reasoning and Acting in Language Models, (**2023**)*

***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
  - RNAcental database

- **Tools description**
  - RNA alignment Tool
  - RNA 3D structure parameters Tool

- **LLM Tool-Agent Computational Pipeline**
- **Results for LLM Tool-Agent**
  - Speculative simulation of reverse evolution of $tRNA^{Gly}$ (6 slides)

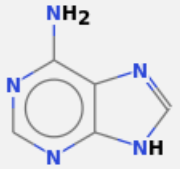
- **LLM ReAct-Agent Computational Pipeline**
- **Results for LLM ReAct-Agent**
  - Comparison of two RNAs
  - Generate two connected RNA hairpins without RNA template
  - Generate two connected RNA hairpins from a given RNA template
  - Targeted modification of a specific region of an RNA molecule

- **Conclusions**

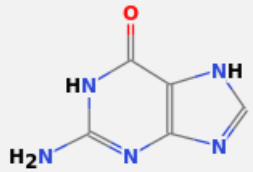
- **What's next ?**

What is RNA molecule ?

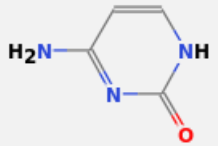
A = Adenine ($C_5H_5N_5$)



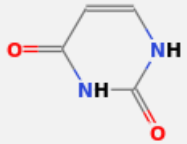
G = Guanine ($C_5H_5N_5O$)



C = Cytosine ($C_4H_5N_3O$)

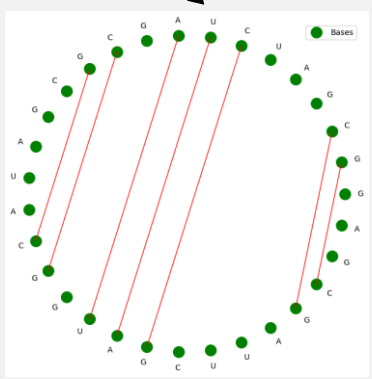


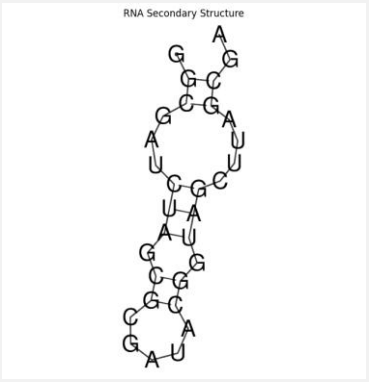
U = Uracil ($C_4H_4N_2O_2$)



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

GGCGAUCUAGCGCGAUACGGUAGCUUAGCGA





RNA = the sequence of **A**, **G**, **C** or **U**

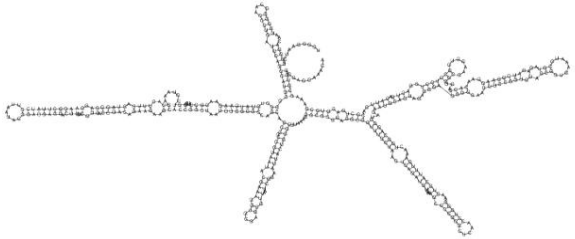
- ribozyme
  - hammerhead ribozyme
    - RNA
    - $RNA_1 + RNA_2$
- coding RNA
  - $mRNA$
- ribonucleoproteins
  - Ribonuclease P
    - $pre-tRNA$
    - $mature\ tRNA$

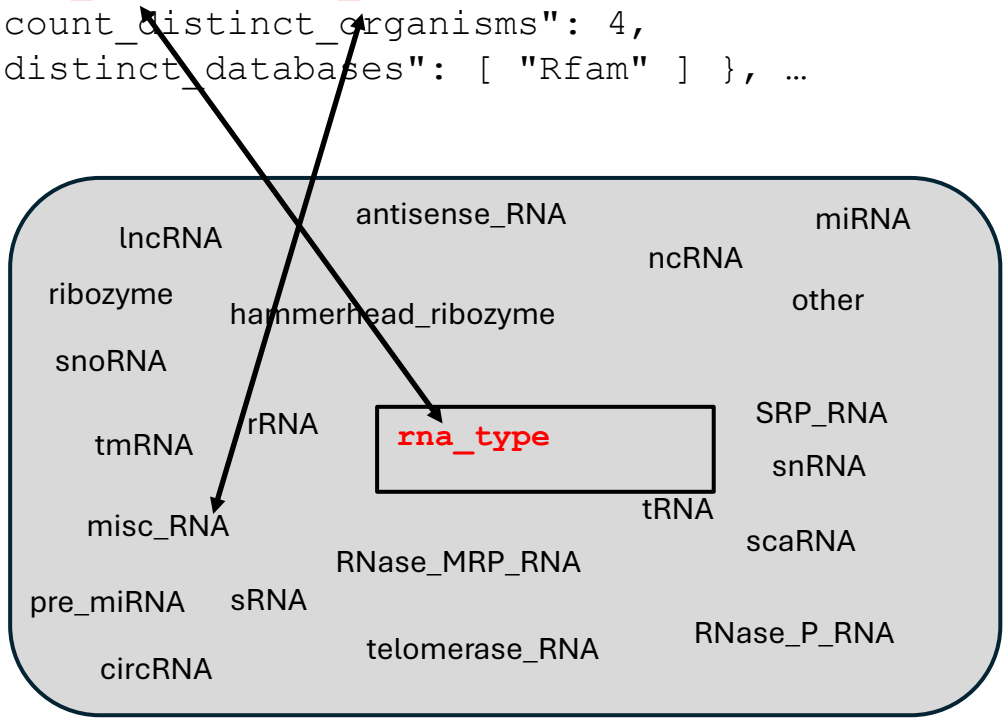
The dataset: RNAcentral database

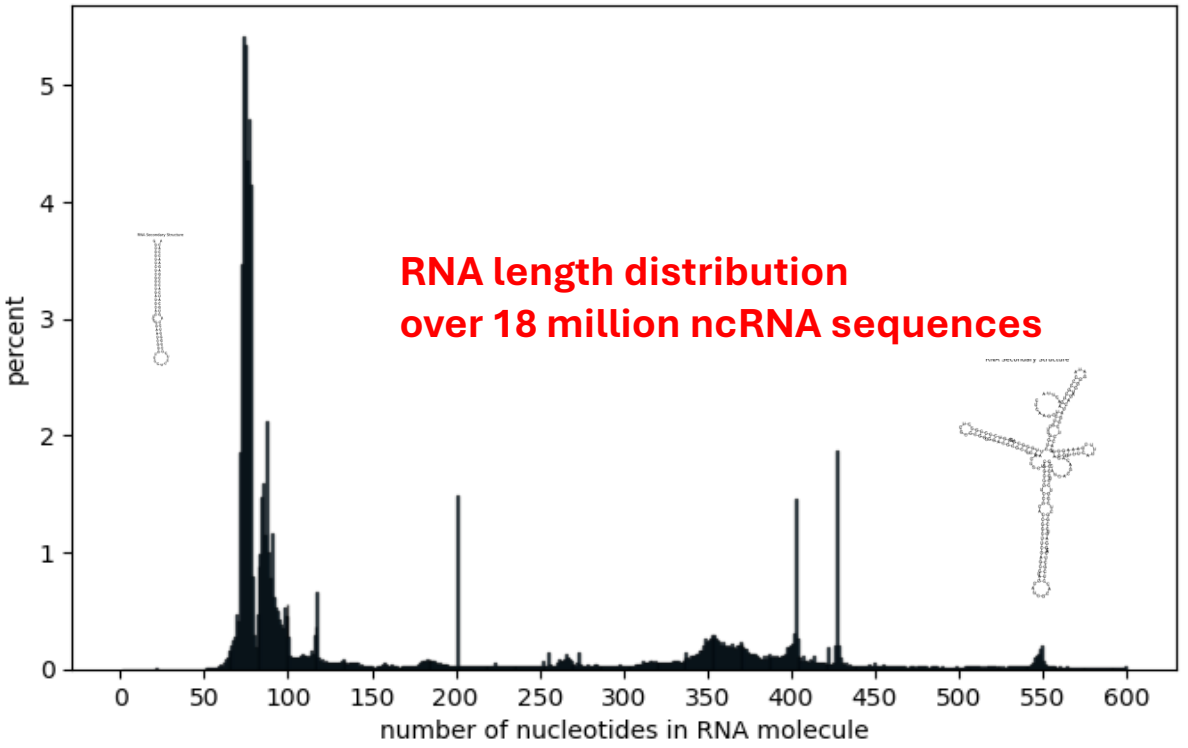
RNAcentral Browsable API

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 alignment Tool** for 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
  }
}
```

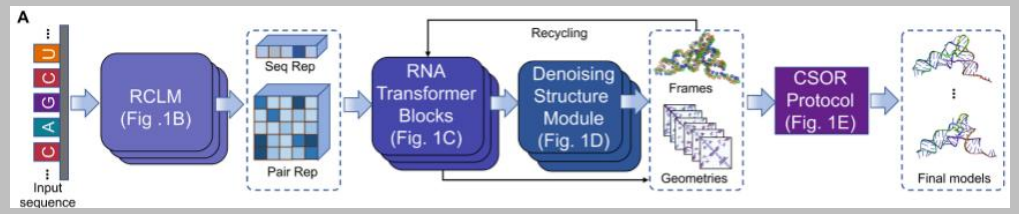
dictionaries **are better understood** by LLMs

USE for LLM prompt

RNA 3D structure parameters Tool:

combine DRfold2 and DSSR to extract 3D structural parameters of RNA

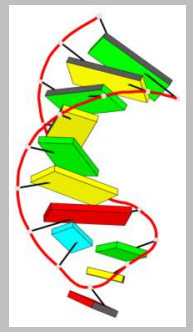
DRfold2 enables the prediction of RNA 3D structures (in PDB format) from nucleotide sequences provided in FASTA format



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

GCGCGCAUACGUGCGCGC

DRfold2



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"
        }
    }
}
```


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}

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}

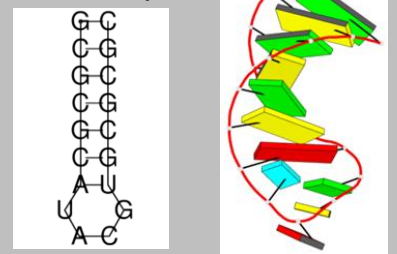
Input

Output for LLM input

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

LLM Tool-Agent Computational Pipeline

current RNA 3D structure (PDB format)



DRfold2

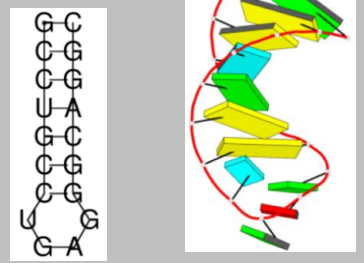
DSSR

conditioned **diffusion model** (Ribodiffusion)
To get backfolded RNA sequence (FASTA)
from its 3D structure (PDB):
GCCUGCCUGAGGGCAGGC

diffusion model input

DRfold2

backfolded RNA 3D structure (PDB format)



LLM AGENT

Sensor

Think

Action

Get current RNA sequence
GCGCGCAUACGUGCGCGC
(FASTA format)

ENVIRONMENT
RNA pool from RNAcentral DB

Propose ancestral RNA sequence
(FASTA format)

TOOLS

Save output to a file on the hard drive for each run

Internet Search Tool

Tools enable the LLM to access external resources and present outputs

- For each run in the loop, the environment can be simulated either by searching for the most similar real RNA to the generated one in the RNAcentral database (using RNA-FM embeddings) or by simply assuming that the generated RNA is real.
- The diffusion model input generates an RNA sequence that folds into the same 3D geometry as the reference RNA.

Prompt Engineering: Designing Effective Inputs for LLMs

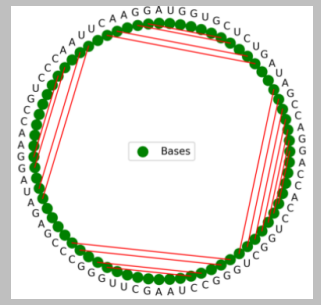
```
prompt = ChatPromptTemplate.from_messages(
  [
    (
      "system",
      """
        You are a research assistant that will help generate a research paper.
        Answer the user query and use necessary tools.
        Wrap the output in this format and provide no other text:
        {format_instructions}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

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}
""",
    ),
    ("placeholder", "{agent_scratchpad}"),
  ]
).partial(format_instructions=parser.get_format_instructions())
```

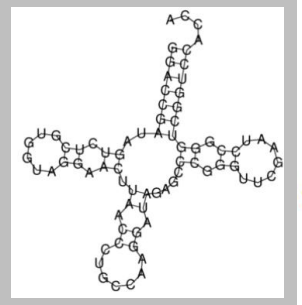
```
query = (
        f"Propose a plausible ancestral RNA sequence based on the current RNA sequence ant its properties."
        f"Save both a summary and the proposed ancestral_sequence to a file"
        f"and begin each output in the file with the following section header: evolution_step_{i}")
 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





DSSR output

```
length: 75
base_pairs: 29
hydrogen_bonds: 112
dot_bracket: (((((((..((.((......)).)).((.((.......)).))....((.((.......)).)).........)))))))....
multiplets: UUG, CGG
splayUnits: AU, UG, GUA, CC, AA, CA
ssSegments: ACCA
junctions: AUAGCUUAGAGCCGU
hairpins: CGUGGUAG, CCUGCCAAG, GUUCGAAUC
nonStack: UGUUGA
stacks: AC, GG, CU, GC, CG, AC, GU, UCU, UGG, GAA, AAC, CCU, GAU, GGG, GUU, GACC, AAGA, ACAU, CAAG
num_coaxStacks: 2
helix_0: base_pairs=12, strand_1=GGACCGACCGGG, strand_2=CCUGGCUGGGCC, helix_form=AAAAAAxAA.A
helix_1: base_pairs=13, strand_1=CAGGAUAGUCUCG, strand_2=UCCCAAUCAAGGA, helix_form=x.A..AxAA..x
```

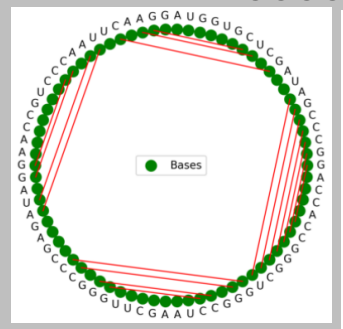
The provided 75-nucleotide RNA sequence exhibits characteristic tRNA-like structural features with a cloverleaf secondary structure. Key structural parameters include:

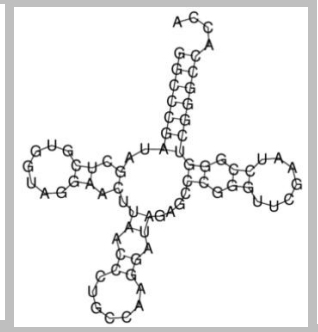
- Length: 75 nucleotides
- Base pairs: 29
- Dot-bracket notation showing complex stem-loop structures
- Multiple hairpins and junction regions
- Two coaxial stacks indicating sophisticated tertiary interactions

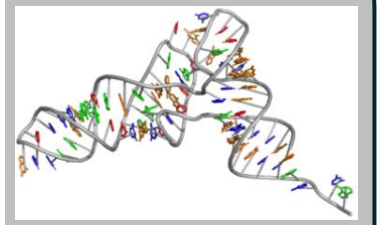
The first two images in each panel were generated using simplified 2D structures for rough visualization purposes only.
The last two images depict the actual 3D structures, predicted by DRfold2 and further processed by DSSR to extract structural parameters for LLM input.

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.

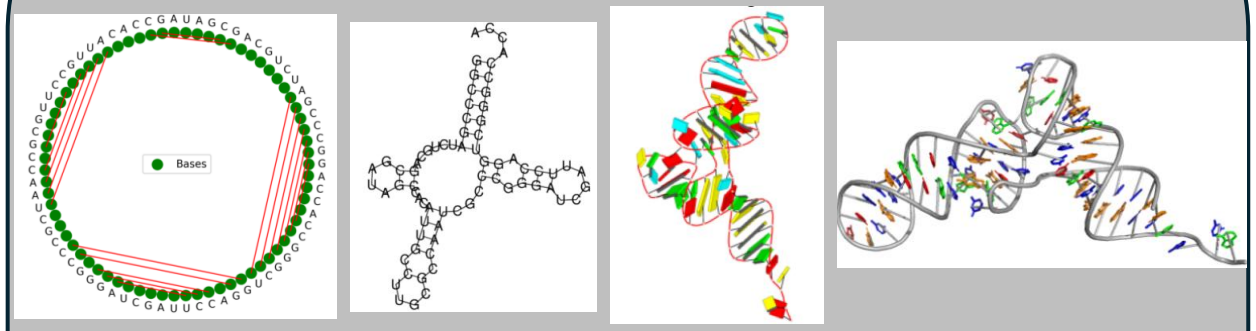
Current-Ancestral RNA sequences alignment

```
<class 'Bio.Align.Alignment'>
target            0 GGACCGAUAGUCUCGUGGUAGGAACUUAACCCUGCCAAGGAUAGAGCCCGGGUUCGAAUC
                  0 ||.|||||||-|||||||||||||||||||||||||||||||||||||||||||||||||
query             0 GGCCCGAUAG-CUCGUGGUAGGAACUUAACCCUGCCAAGGAUAGAGCCCGGGUUCGAAUC

target           60 CGGGUCGGUCCACCA 75
                 60 |||||||||.|||||| 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



DSSR output

```
length: 71
base_pairs: 32
hydrogen_bonds: 111
dot_bracket: (((((((.....((....)).((((.(....).)))).((.(((.....))).)))))))))....
multiplets: CGC, UGCG
splayUnits: UC, GAUA, UU, CC, UU, CC
ssSegments: ACCA
junctions: AUCUGCAGCCACAUCGCCGU
hairpins: CGAUAG, CUUGCG, AUCGAUU
nonStack: AUAU
stacks: CU, UC, CG, AU, CA, GU, CA, GAC, AUC, GGC, CAG, CGU, GGG, CCAC, GCCU, UGCGCC
num_coaxStacks: 1
helix_0: base_pairs=13, strand_1=GGCCCGACCGGGA, strand_2=CCGGGCUGGACCU, helix_form=AAAAAAxA..AA
helix_1: base_pairs=8, strand_1=UGAUUGCU, strand_2=GCUAACGC, helix_form=.xAAAx.
helix_2: base_pairs=5, strand_1=CAGCG, strand_2=ACCGC, helix_form=..Ax
```

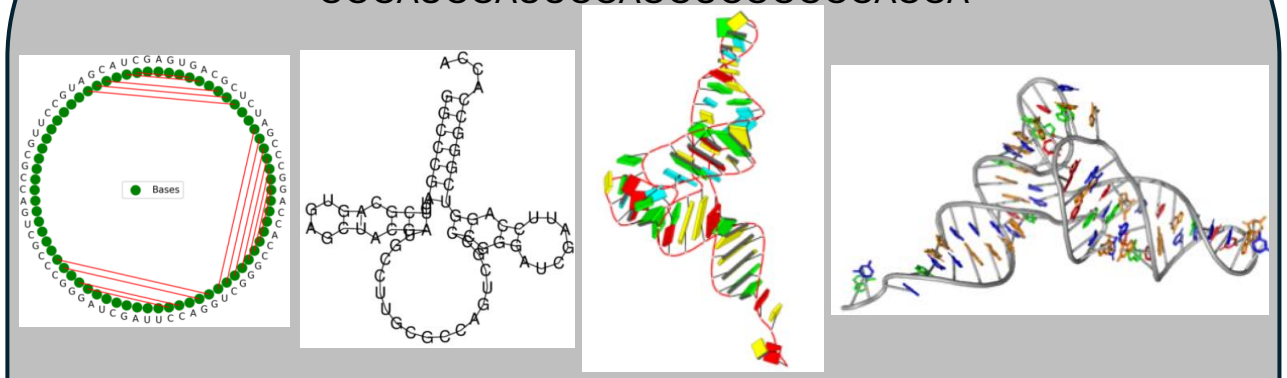
Based on the analysis of the current RNA sequence (71 nucleotides) and its structural properties, I propose an ancestral RNA sequence that maintains the critical structural elements while reflecting probable evolutionary simplifications. The current RNA displays a complex secondary structure with 32 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

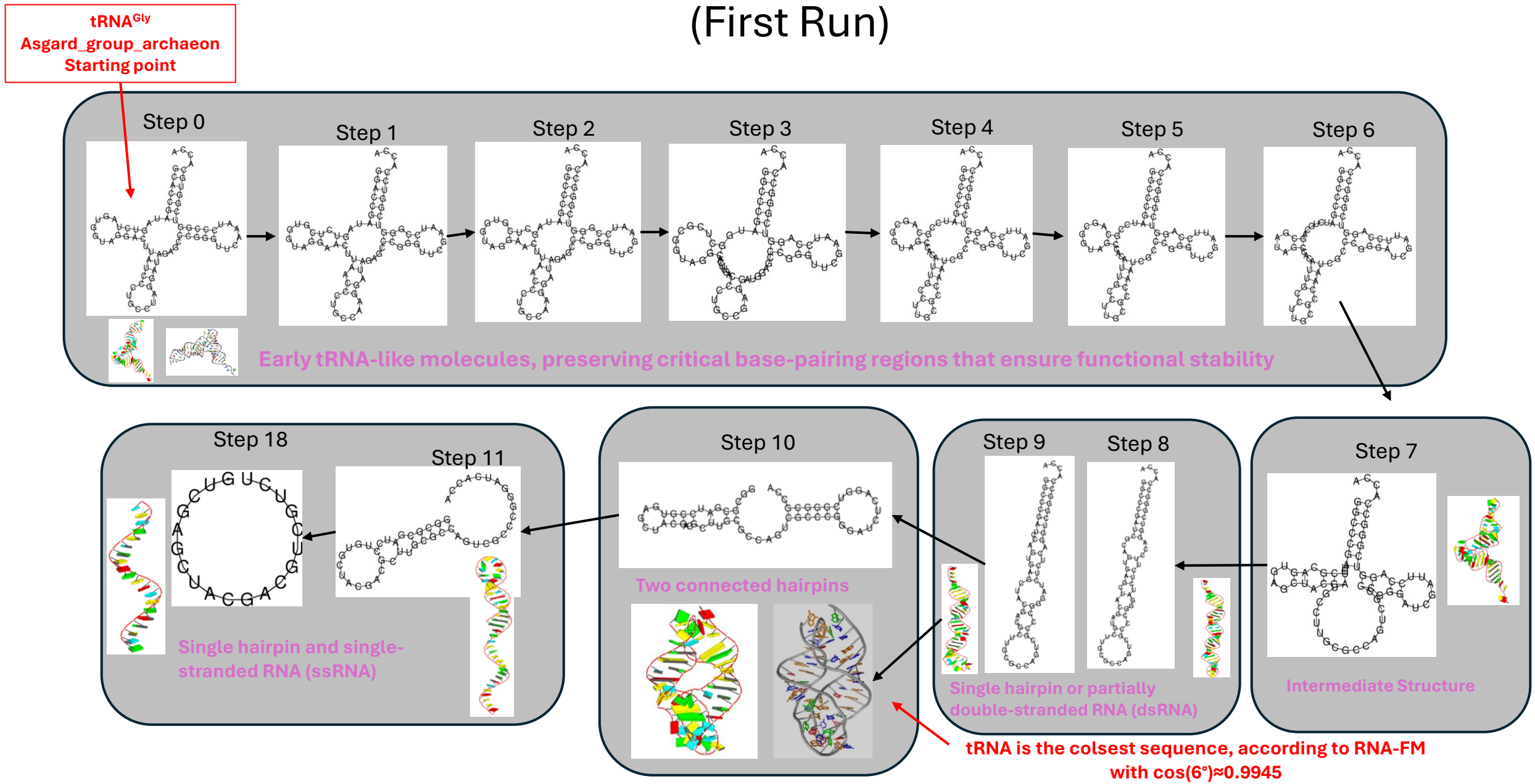
```
<class 'Bio.Align.Alignment'>
target            0 GGCCCGAUCU-GCAGCGAUAGCCACAUUGCCUUGCGCCAAUCGCCCGGGAUCGAUUCCAG
                  0 ||||||||||-||||.|--|||.||..||||||||||.||||||||||||||||||||
query             0 GGCCCGAUCUCGCAGUG--AGCUACGAUGCCUUGCGCCAGUCGCCCGGGAUCGAUUCCAG

target           59 GUCGGGCCACCA 71
                 60 |||||||||||| 72
query            58 GUCGGGCCACCA 70

Score: 118.5
```

The first two images in each panel were generated using simplified 2D structures 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 (First Run)

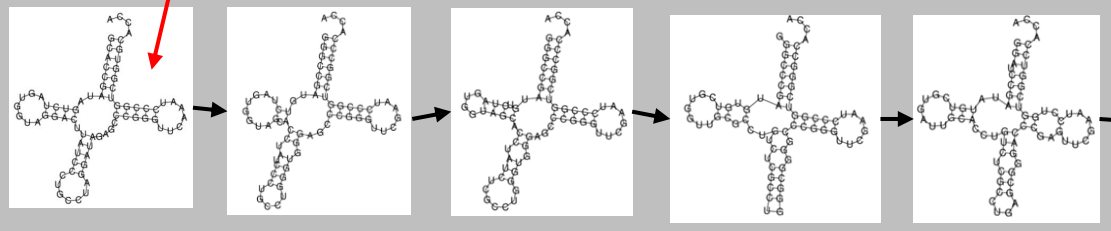


To improve the reliability of these evolutionary trajectories, it is necessary to introduce constraints on the number of mutation events per step.

Results of Simulated Reverse Evolution of $tRNA^{Gly}$ Asgard_group_archaeon (Second Run)

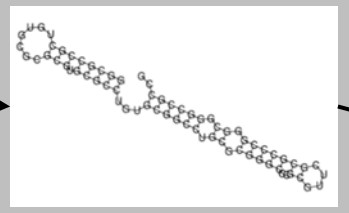
$tRNA^{Gly}$ Asgard_group_archaeon Starting point

Step 0 Step 1 Step 2 Step 3 Step 4

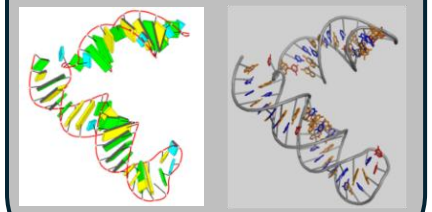


Early tRNA-like molecules, preserving critical base-pairing regions that ensure functional stability

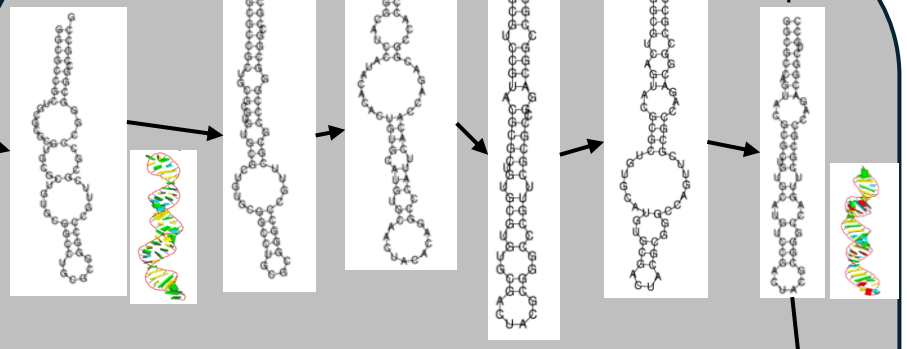
Step 5



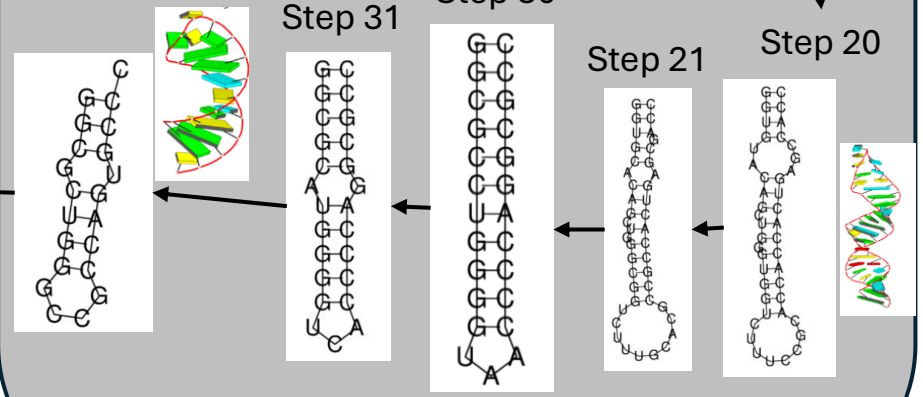
Two connected hairpins



Step 6 Step 7 Step 8 Step 9 Step 10 Step 11



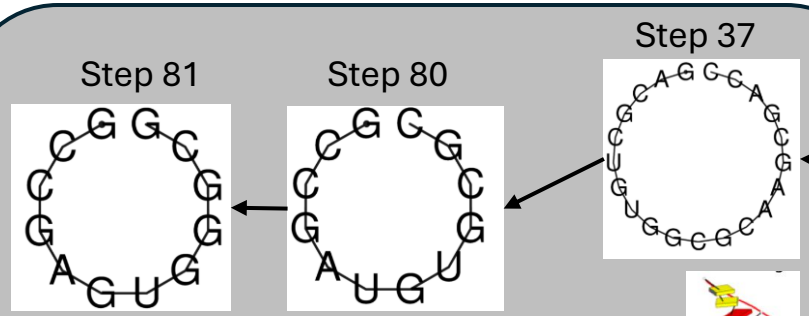
Step 20 Step 21 Step 30 Step 31 Step 36



Single hairpin or partially double-stranded RNA (dsRNA)

tRNA is the closest sequence in RNA central DB According to RNA-FM with cos(6°)≈0.9945

Step 37 Step 80 Step 81

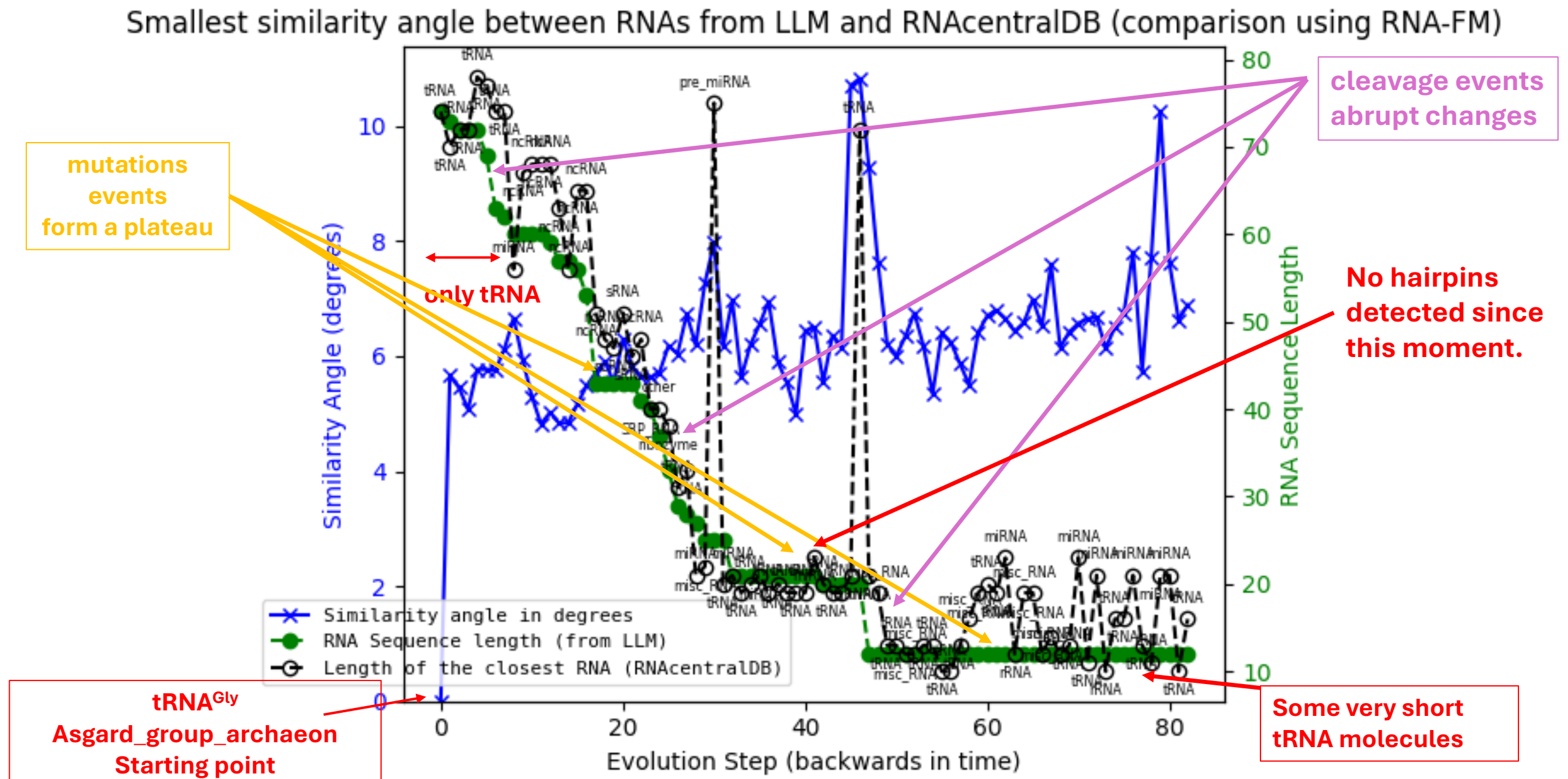


Single-stranded RNA (ssRNA)

```
<class 'Bio.Align.Alignment'>
target            0 GCCGA-UGUGCGC 12
                  0 |||||-||.|||- 13
query             0 GCCGAGUGGGCG- 12

Score: 15.0
```

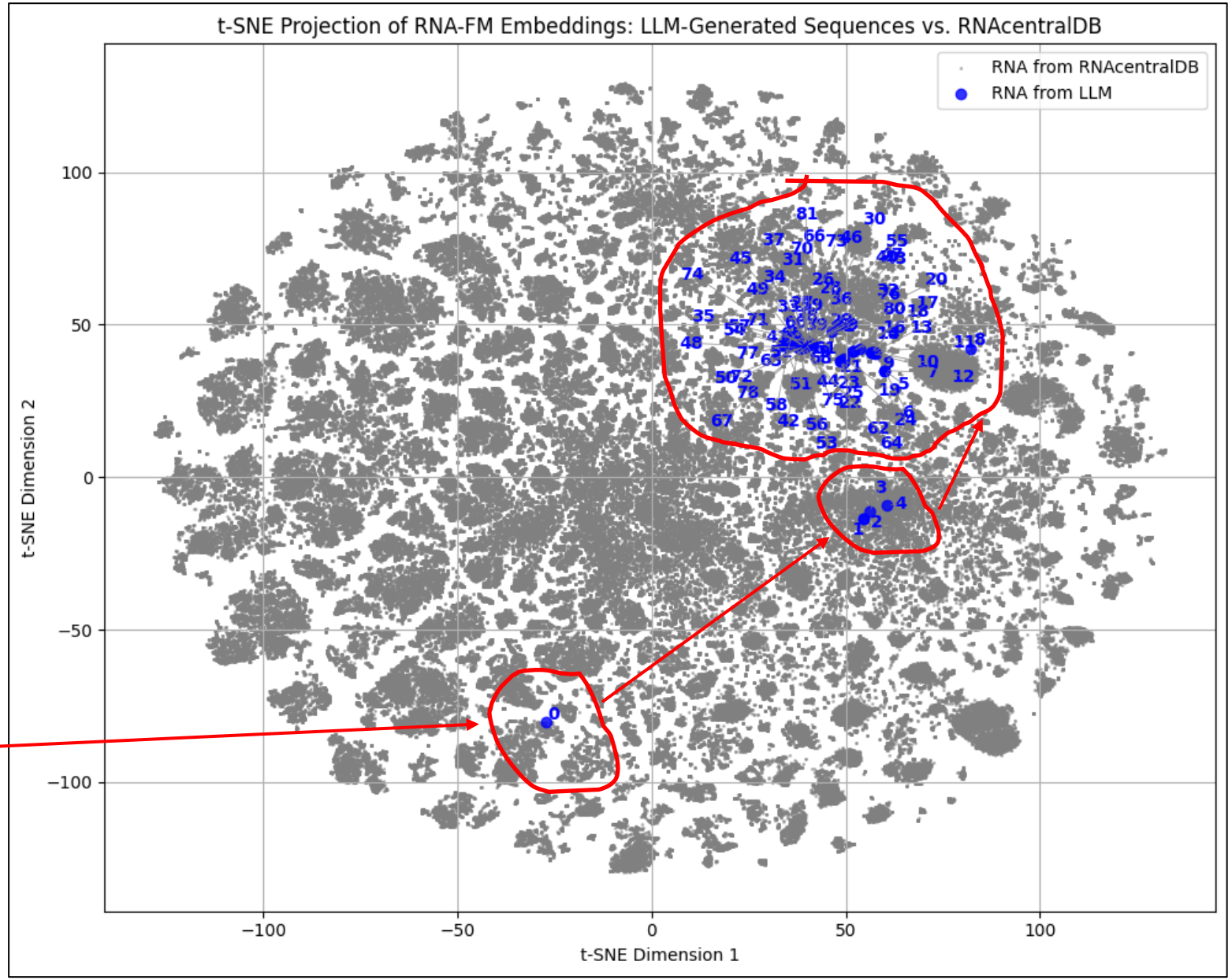
To improve the reliability of these evolutionary trajectories, it is necessary to introduce constraints on the number of mutation events per step.

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



https://github.com/PavelPll/RNA_RAG/tree/main

How Accurately do LLM-Designed RNAs Represent Real RNA?

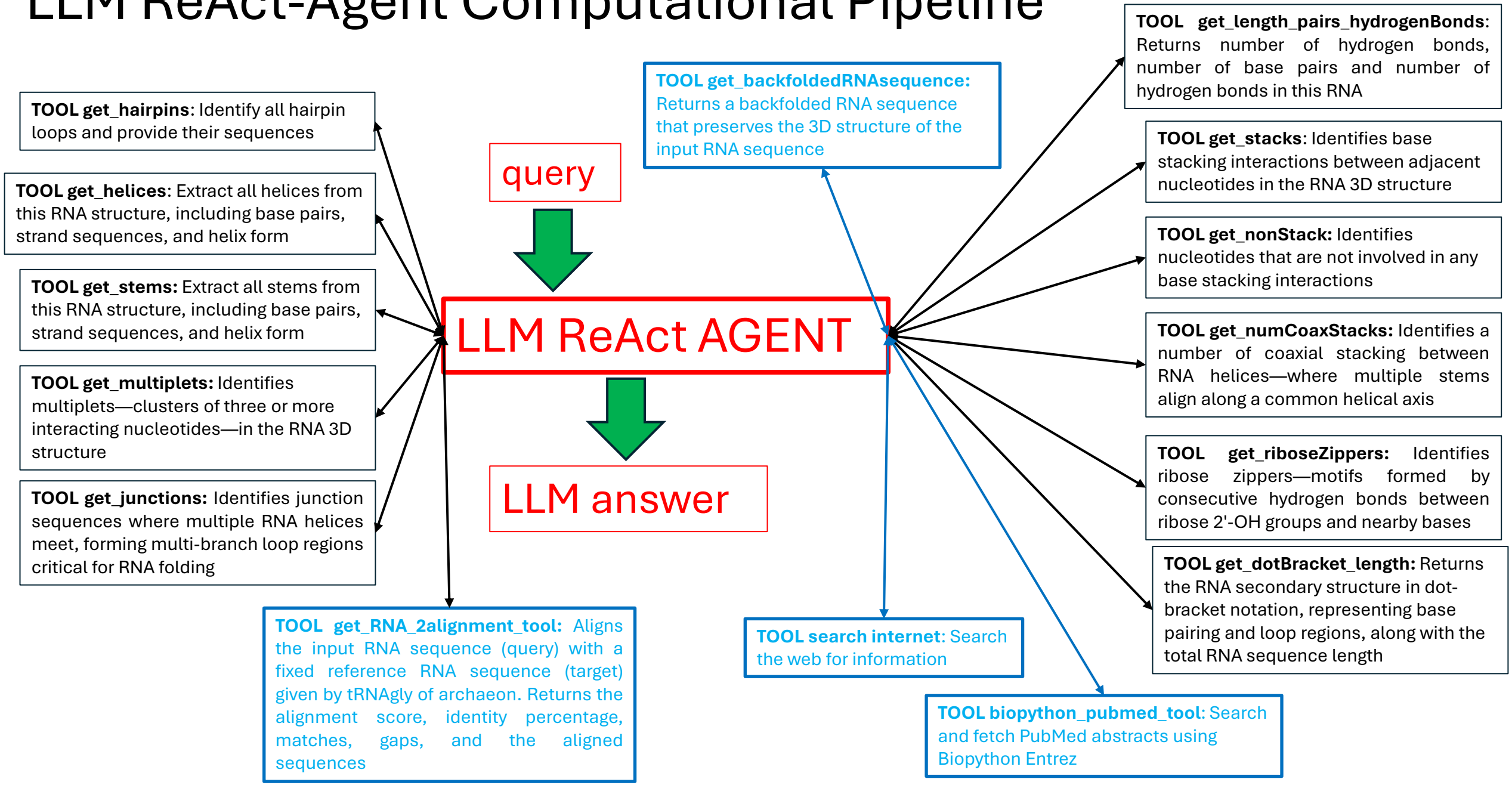


tRNA^{Gly}
Asgard_group_archaeon
Starting point

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

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

LLM ReAct-Agent Computational Pipeline



LLM ReAct Agent: Comparison of Two RNA Sequences

Provide a detailed comparison of the following two RNA molecules: query AGAGAGAGAGAGAGAGAG and target GCACCGAUAGUCUAGUGGUAGGACUUAUCCCUGCCUAGGAUAGAGCCCGGGUUCAAAUCCCGGUCGGUGCACCA

1. Run a simple sequence alignment between them.
2. Analyze and compare their secondary and tertiary structures.
3. Conduct a comprehensive search of scientific literature to identify known information about these sequences.
4. Highlight their similarities and differences.

Provide the output in clearly labeled sections.

Tool Calls: RNA_2alignment

Tool Calls: get_dotBracket_length for query sequence

Tool Calls: get_dotBracket_length for target sequence

Tool Calls: get_stems for target sequence

Tool Calls: get_hairpins for target sequence

Tool Calls: get_junctions for target sequence

Get abstracts → Tool Calls: biopython_pubmed

Search for papers → Tool Calls: internet search

Paper analysis → Tool Calls: internet search

Tool Calls: get_get_length_pairs_hydrogenBonds for target sequence

Tool Calls: get_get_length_pairs_hydrogenBonds for query sequence

Detailed RNA Molecule Comparison Analysis

1. Sequence Alignment Results
Query: AGAGAGAGAGAGAGAGAG (20 nucleotides)
Target:
GCACCGAUAGUCUAGUGGUAGGACUUAUCCCUGCCUAGGAUAGAGCCCGGGUUC
AAAUCCCGGUCGGUGCACCA (74 nucleotides)
Alignment Statistics:
- **Alignment Score**: 19.0
- **Identity**: 25.33%
- **Matches**: 19 (10 A-A matches, 9 G-G matches)
- **Mismatches**: 0
- **Gaps**: 56 total (1 in target, 55 in query)
- **Alignment Length**: 75 nucleotides
The alignment shows limited sequence similarity, with the query's alternating AG pattern only partially matching scattered positions in the target sequence.

2. Secondary and Tertiary Structure Analysis
Query Sequence (AGAGAGAGAGAGAGAGAG):
- **Length**: 20 nucleotides
- **Secondary Structure**: `.....` (completely unstructured)
- **Base Pairs**: 0
- **Hydrogen Bonds**: 0
- **Structural Elements**: None detected
- **Characteristics**: Simple alternating purine-purine repeat $(AG)_{10}$
Target Sequence:
- **Length**: 74 nucleotides
- **Secondary Structure**: `(((((((..((((.....)))).(((((.....)))))....(((((.....)))))))))))....`
- **Base Pairs**: 33
- **Hydrogen Bonds**: 123
- **Structural Complexity**: Highly structured with multiple secondary elements
Structural Elements of Target:
Stems/Helices (4 major stems):
1. **Helix 0**: 7 base pairs (GCACCGA/CGUGGCU)
2. **Helix 1**: 4 base pairs (GUCU/CAGG)
3. **Helix 2**: 5 base pairs (UAUCC/AUAGG)
4. **Helix 3**: 5 base pairs (CCGGG/GGCCC)
Hairpin Loops (3 loops):
- UAGUGGUAG
- CCUGCCUAG
- GUUCAAAUC
Junction Region: AUAGCUUAGAGCCGU

3. Literature Search Results

Target Sequence:
The target sequence shows structural characteristics typical of **transfer RNA (tRNA)**:
- 74 nucleotides (close to canonical tRNA length of ~75-90 nt)
- Complex secondary structure with multiple stems and loops
- Cloverleaf-like folding pattern consistent with tRNA architecture
- Literature confirms this matches tRNA structural motifs, particularly the cloverleaf model

Query Sequence:
The alternating AG repeat sequence appears in recent literature as:
- **Synthetic oligonucleotide** used in RNA-protein interaction studies
- **Model sequence** for studying purine-purine stacking interactions
- **Control sequence** in fluorescence and binding assays
- No natural biological function identified for this specific repeat

4. Similarities and Differences

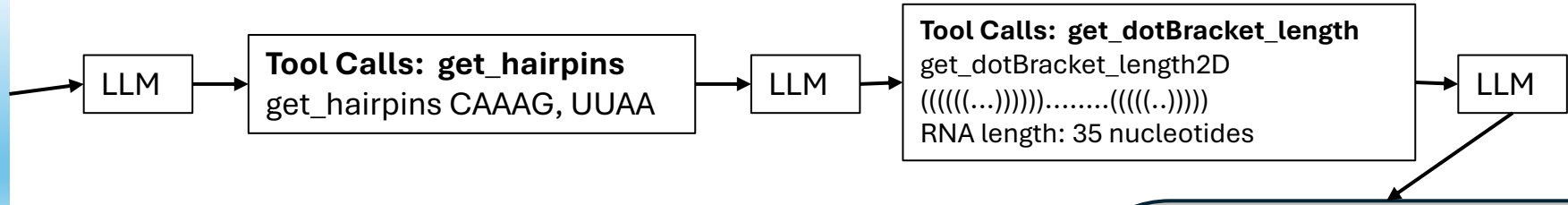
Similarities:
1. **Composition**: Both are RNA sequences (contain U instead of T)
2. **Purine Content**: Both are purine-rich
3. **Some Alignment**: 25.33% sequence identity with 19 matching positions
4. **Chemical Nature**: Both can participate in Watson-Crick base pairing (when complementary partners are present)

Major Differences:

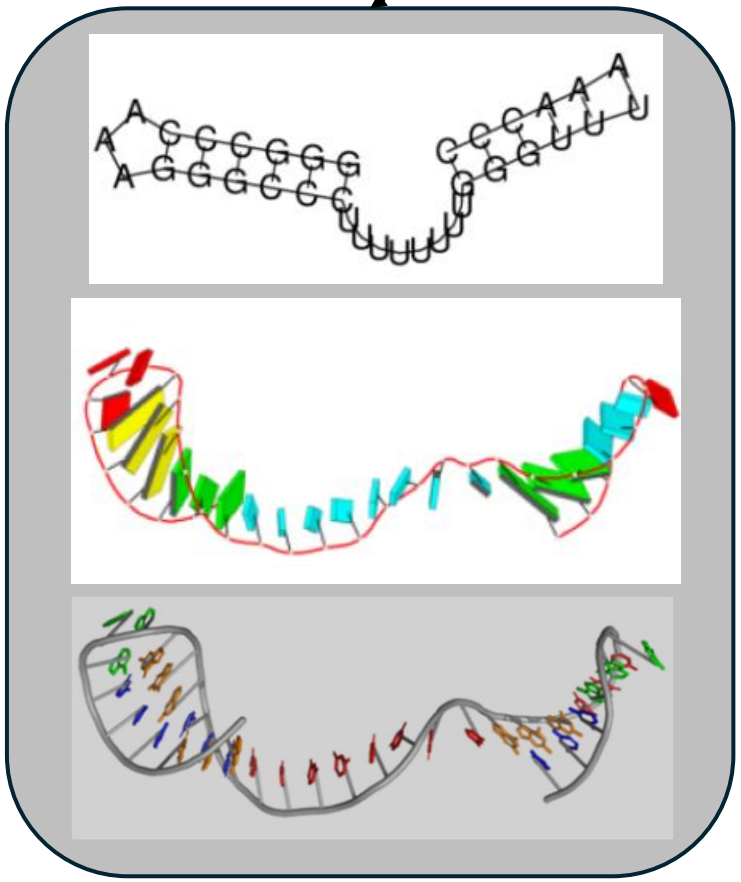
Aspect	Query (AG repeat)	Target (tRNA-like)
Length	20 nt	74 nt
Structure	Completely unstructured	Highly structured (33 base pairs)
Complexity	Simple alternating repeat	Complex with 4 helices, 3 hairpins
Function	Synthetic/artificial	Functional tRNA-like molecule
Stability	Low (no base pairs)	High (123 hydrogen bonds)
Evolution	Likely artificial	Biologically evolved structure
Diversity		

LLM ReAct agent : Generate Two Connected RNA Hairpins without a Prior RNA Template

Human Message
Generate two distinct RNA hairpins connected by a **single-stranded RNA linker**.



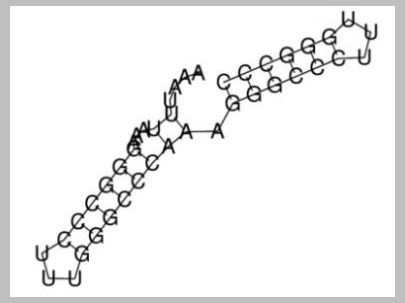
==========[1m **Human Message** [0m==========Generate two distinct RNA hairpins connected by a **single-stranded RNA linker**.==========[1m **Ai Message** [0m==========[{'text': "I'll help you generate two distinct RNA hairpins connected by a single-stranded linker. Let me create this structure and then analyze it to confirm the hairpin formations.\n\nFirst, let me design the RNA sequence with two hairpins and a linker:\n\n**Designed RNA sequence:**\n```\nGGGCCCAAAGGGCCCUUUUUUUUUGGGUUUAAACCC\n```\n\nThis consists of:\n- **Hairpin 1:** `GGGCCCAAAGGGCCC` (GGG-CCC stem with AAA loop)\n- **Linker:** `UUUUUUUU` (8 uracils as single-stranded connector)\n- **Hairpin 2:** `GGGUUUAAACCC` (GGG-CCC stem with UUU-AAA loop)\n\n**Now let me analyze this structure to verify the hairpins**:', 'type': 'text'}, {'id': 'toolu_01Q645amkxCPhDjA4qxeQbVH', 'input': {'__arg1': 'GGGCCCAAAGGGCCCUUUUUUUUUGGGUUUAAACCC'}, 'name': 'get_hairpins', 'type': 'tool_use'}] **Tool Calls: get_hairpins** (toolu_01Q645amkxCPhDjA4qxeQbVH) Call ID: toolu_01Q645amkxCPhDjA4qxeQbVH Args: __arg1: GGGCCCAAAGGGCCCUUUUUUUUUGGGUUUAAACCC c:\Users\User\work\rna_agent\RiboDiffusion\utils.py:16: loaded_state = torch.load(ckpt_dir, map_location=device) ==========[1m Tool Message [0m==========Name: **get_hairpinsCAAAG, UUAA** ==========[1m Ai Message [0m==========[{'text': **'Let me also get the dot-bracket notation to visualize the secondary structure:**', 'type': 'text'}, {'id': 'toolu_0149fG8yTvHaMatfqfVq44nJ', 'input': {'__arg1': 'GGGCCCAAAGGGCCCUUUUUUUUUGGGUUUAAACCC'}, 'name': 'get_dotBracket_length', 'type': 'tool_use'}] **Tool Calls: get_dotBracket_length** (toolu_0149fG8yTvHaMatfqfVq44nJ) Call ID: toolu_0149fG8yTvHaMatfqfVq44nJ Args: __arg1: GGGCCCAAAGGGCCCUUUUUUUUUGGGUUUAAACCC ==========[1m Tool Message [0m==========Name: **get_dotBracket_length2D RNA structure in dot-bracket notation: ((((((...))))))........(((((..))))) RNA length: 35 nucleotides** ==========[1m Ai Message [0m==========Perfect! The analysis confirms our design. Here's the complete structure:## Generated RNA Structure with Two Hairpins and Linker**RNA Sequence:** `GGGCCCAAAGGGCCCUUUUUUUUUGGGUUUAAACCC` **Secondary Structure:** `((((((...))))))........(((((..)))))` ### Structure Breakdown:1. **First Hairpin:** - Stem: GGG-CCC/GGG-CCC (6 base pairs) - Loop: CAAAG (5 nucleotides) - Positions: 1-152. **Single-Stranded Linker:** - Sequence: UUUUUUUU (8 uracils) - Positions: 16-23 - Represented by dots (.) in dot-bracket notation3. **Second Hairpin:** - Stem: GGG-CCC (3 base pairs) - Loop: UUAA (4 nucleotides) - Positions: 24-35**Total Length:** 35 nucleotidesThe structure successfully contains two distinct hairpin loops connected by an 8-nucleotide single-stranded RNA linker, exactly as requested. The linker region provides flexibility between the two structured domains and could serve as a hinge region in functional applications.

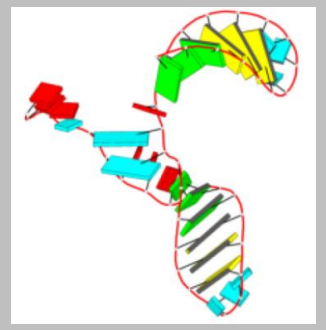


LLM could have called the **get_stems** tool to verify the stems, but it didn't. Still, not bad for such a simple prompt.

LLM ReAct agent : Generate Two Connected RNA Hairpins with a Prior RNA Template

The reference sequence (input): AAAUUUAAA
Modify this reference sequence to form a structure without **pseudoknot base pairs** consisting of **two separate hairpins** connected by a **single-stranded RNA linker**.
The modified RNA sequence must be about 40 nucleotides long.Return the modified RNA sequence and its exact length.





AAAUUUAAA

Let me design a modified RNA sequence based on your requirements:
AAAUUUAAAGGGCCCUUUGGGCCCAAAUUUGGGCCCUUUGGGCCC

Let me verify this sequence and get its exact length
Tool Call: get_dotBracket_length

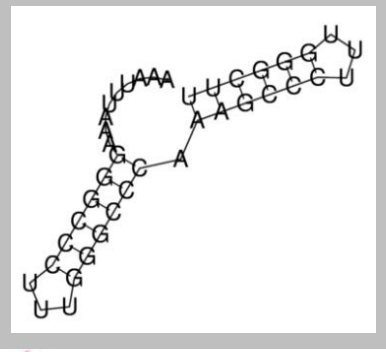
Let me create a shorter sequence that meets your ~40 nucleotide
AAAUUUAAAGGGCCCUUUGGGCCCAAAGGGCCCUUUGGGCCC

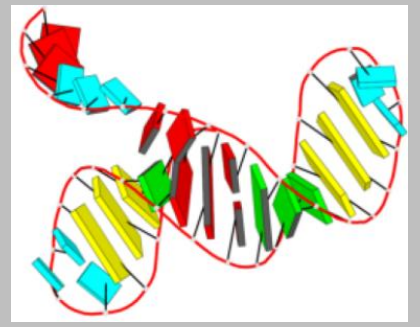
What is 'overlapping' in LLM internal raisoning ??

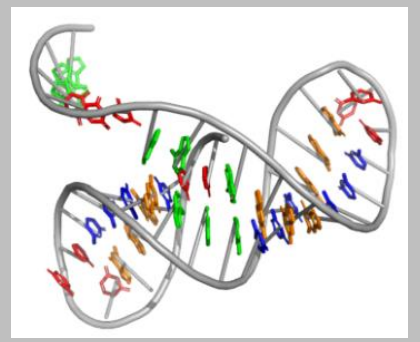
'This structure shows some overlapping, let me try a cleaner design:',
'AAAUUUAAAGGGCCCUUUGGGCCCAAAGCCCUUUGGGCUU'
Tool Calls: get_dotBracket_length

'Perfect! Let me verify this structure has two separate hairpins:',
'AAAUUUAAAGGGCCCUUUGGGCCCAAAGCCCUUUGGGCUU
Tool Calls: get_hairpins
Tool Calls: get_stems

Results:****Modified RNA sequence:**
AAAUUUAAAGGGCCCUUUGGGCCCAAAGCCCUUUGGGCUU
Exact length: 40 nucleotides**





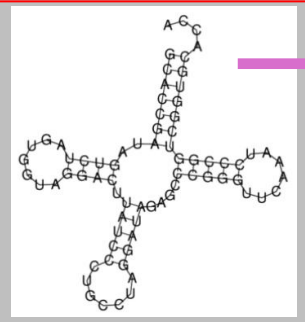


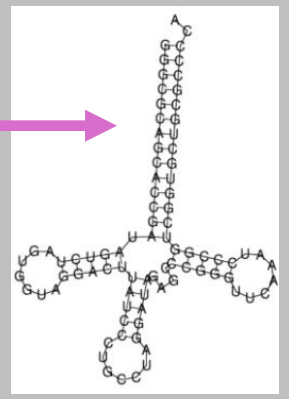
LLM ReAct Agent: Targeted modification of a specific region of an RNA 1

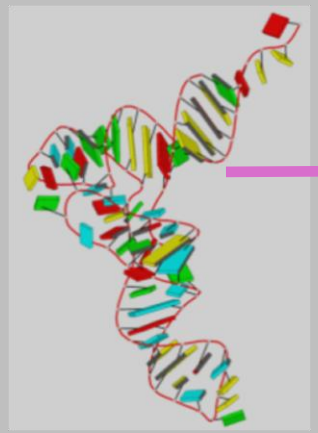
reference sequence

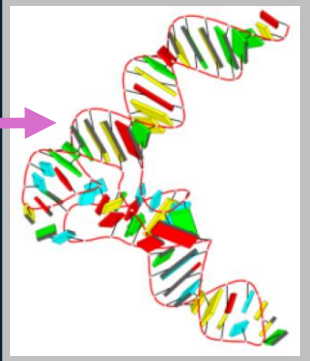
modified sequence

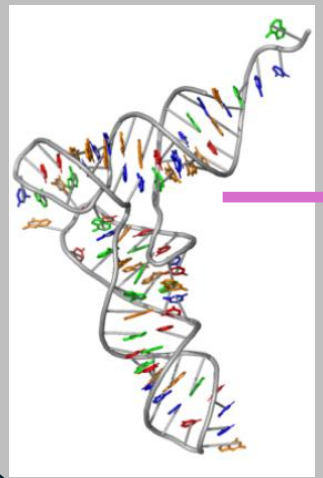
tRNA^{Gly} Asgard_group_archaeon

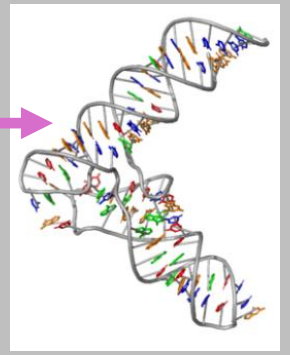












f"Current tRNA sequence (input): {tRNA}\n\n"
"Modify this reference sequence by increasing the length of the acceptor stem to **twice its original size. "**
"Ensure that **no more than 20% of the nucleotides are changed or added.\n\n"**
"Keep the rest of the tRNA unchanged.\n\n"
"Verify the structure and return the modified RNA sequence."

Current tRNA sequence (input):
GCACCGAUAGUCUAGUGGUAGGACUUAUCCCUGCCUAGGAUAGAGCCCGGGUUCAAAUCCCGGUCGGUGCACCA

I'll help you modify the tRNA sequence by doubling the acceptor stem length while keeping changes to no more than 20% of nucleotides. First, let me analyze the current structure to identify the acceptor
Tool Calls: **get_dotBracket_length**

'Now let me identify the stems to locate the acceptor stem
Tool Calls: get_stems

'Let me also get the hairpin loops to better understand the structure
Tool Calls: get_hairpins

Based on the analysis, I can see this is a standard tRNA structure with:\n- **Acceptor stem**: helix_0 with 7 base pairs (GCACCGA paired with CGUGGCU)\n- D-arm: helix_1 with 4 base pairs\n- Anticodon arm: helix_2 with 5 base pairs \n- TψC arm: helix_3 with 5 base pairs\n\nTo double the acceptor stem length from 7 to 14 base pairs while changing ≤20% of nucleotides (≤15 out of 74), I'll:\n\n1. Extend the 5' end with 7 complementary base pairs\n2. Extend the 3' end with 7 complementary base pairs\n3. Keep all other structural elements unchanged\n\nHere's the modified sequence: GGGC...

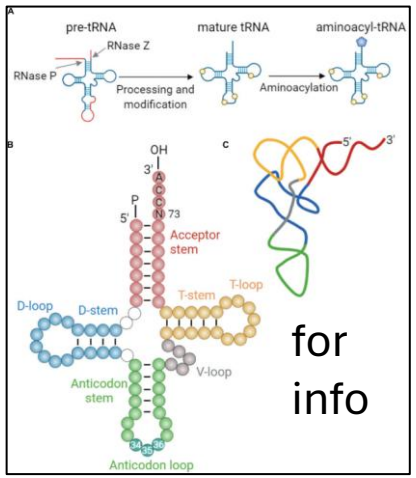


Figure adapted from Frontiers in Molecular Biosciences (Irem Avcilar-Kucukgoze et al., 2020), CC BY 4.0. DOI: 10.3389/fmolb.2020.610617

Tool Calls: get_dotBracket_length → **Tool Calls: get_stems**

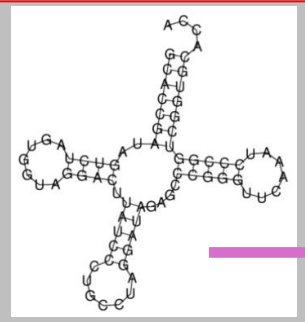
Perfect! The modification is successful:## Summary:- **Original length**: 74 nucleotides- **Modified length**: 86 nucleotides (12 nucleotides added = 16.2% increase, well under 20% limit)- **Acceptor stem**: **Successfully doubled from 7 to 14 base pairs**- **Structure integrity**: All other arms (D-arm, anticodon arm, TψC arm) remain unchanged- **Base pair complementarity**: Maintained throughout the extended acceptor stem## Modified tRNA sequence:**
`GGGCGCAGCACCGAUAGUCUAGUGGUAGGACUUAUCCCUGCCUAGGAUAGAGCCCGGGUUCAAAUCCCGGUCGGUGCUGCGCCCCA` **The acceptor stem is now twice its original length (14 bp vs 7 bp) while preserving the overall tRNA secondary structure and staying within the 20% nucleotide change constraint.

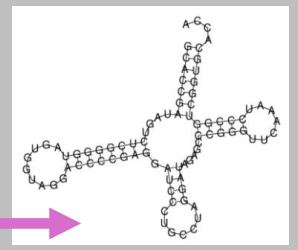
LLM ReAct Agent: Targeted modification of a specific region of an RNA 2

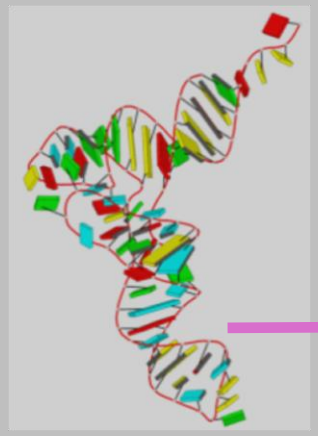
reference sequence

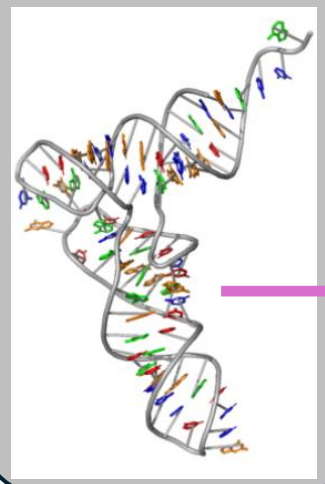
modified sequence

tRNA^{Gly} Asgard_group_archaeon









f"Current tRNA sequence (input): {tRNA}\n\n"
"Modify this reference sequence by increasing the length of the **D-stem** to **twice its original size**. "
"Use backfolded sequence which forms the same 3D structure. \n\n"
"Keep the rest of the tRNA sequence and its cloverleaf structure unchanged.\n\n"
"Verify the structure and return the modified RNA sequence."

Let me start by analyzing the current structure to identify the D-stem and then create the modified sequence.
Tool Calls: get_stems
get_dotBracket_length
get_backfoldedRNAsequence

Looking at the structure, the D-stem is located at positions approximately 10-13 and 22-25 in the original sequence. To double the D-stem length while preserving the overall tRNA structure, I'll extend both strands of the D-stem by 4 additional base pairs.\n\nHere's my strategy:\n1. Keep the acceptor stem, anticodon stem, and TψC stem unchanged\n2. Extend the D-stem from 4 to 8 base pairs by adding complementary sequences\n3. Use the backfolded approach to maintain 3D structure\n\nLet me create the modified sequence by extending the D-stem:\n\nOriginal D-stem: GUCU...CAGG (4 bp)\nExtended D-stem: GUCUGGCC...CCAGGACG (8 bp)\n\nThe modified tRNA sequence would be:\n
GCACCGAUAGUCUGGCCAGUGGUAGGACGCCAGGAUCCCUGCCUAGGAUAGAGCCCGGGUUCAAAUCCCGGUCGGUGCACCA

Tool Calls: get_stems

get_dotBracket_length

'Good! The D-stem is now 5 base pairs, but I need it to be 8 (double the original 4). Let me create a better modification:\n\nThe modified sequence with D-stem extended to 8 base pairs:\n
GCACCGAUAGUCUCGGGCAGUGGUAGGACGCCCGAGGAUCCCUGCCUAGGAUAGAGCCCGGGUUCAAAUCCCGGUCGGUGCACCA'

Tool Calls: get_stems

'Getting closer! Let me make one more adjustment to achieve exactly 8 base pairs in the D-stem:\n\nGCACCGAUAGUCUCGGGGAGUGGUAGGACCCCCGAGGAUCCCUGCCUAGGAUAGAGCCCGGGUUCAAAUCCCGGUCGGUGCACCA'

Tool Calls: get_stems

Let me try a more precise approach to get exactly 8 base pairs:\n\nGCACCGAUAGUCUCGGGGUAGUGGUAGGACCCCGAGGAUCCCUGCCUAGGAUAGAGCCCGGGUUCAAAUCCCGGUCGGUGCACCA'

Tool Calls: get_stems

'Perfect! Now I have the D-stem with exactly 8 base pairs (double the original 4). Let me get the backfolded sequence and verify the complete structure:
Tool Calls: get_backfoldedRNAsequence
GCCCCAGUAGAGGCCGGGCUUAGGUGUCACCCGACCGGUGGCUCAGACCCAUGACCCACGGGUUAAAGUCCCGUCUGGGGCACCA

GCACCGAUAGUCUCGGGGUAGUGGUAGGACCCCGAGGAUCCCUGCCUAGGAUAGAGCCCGGGUUCAAAUCCCGGUCGGUGCACCA
================================= Tool Message =================================
Name: get_dotBracket_length

I have successfully modified the tRNA sequence by doubling the D-stem length from 4 to 8 base pairs.

CONCLUSIONS

- **An LLM agent can be used not only to generate new RNA sequences but also to modify existing ones, offering an advantage over conditional diffusion models alone when guiding RNA design.**

- **Without external tools, an LLM is not even able to reliably count the number of nucleotides in an RNA sequence. A backfolded sequence turns out to be an effective, LLM-comprehensible alternative to 3D structural parameters for preserving the original structure of an input (template) sequence during partial modification.**

- **A preliminary, speculative attempt to model tRNA reverse evolution captures certain trends. RNA length decreases over time, and three distinct evolutionary patterns have been identified:**
  - *The ancestral sequence maintains the essential cloverleaf structure characteristic of early tRNA-like molecules, preserving critical base-pairing regions.*
  - *A cloverleaf structure 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. Additional ideas are presented on the next three slides.**

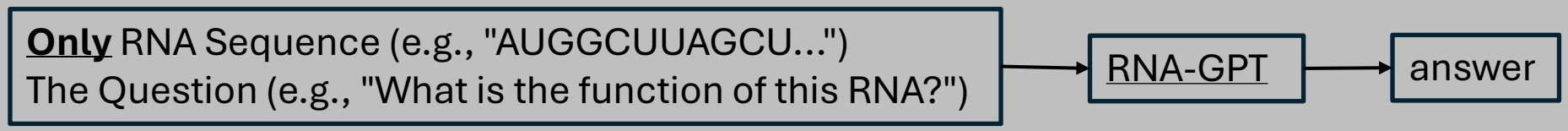
1. What's next ?

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. **Use RNA-GPT to get more context from answering questions about given RNA sequences**

RNA-FM sequence encoder was used to embed RNA sequences for alignment with natural language



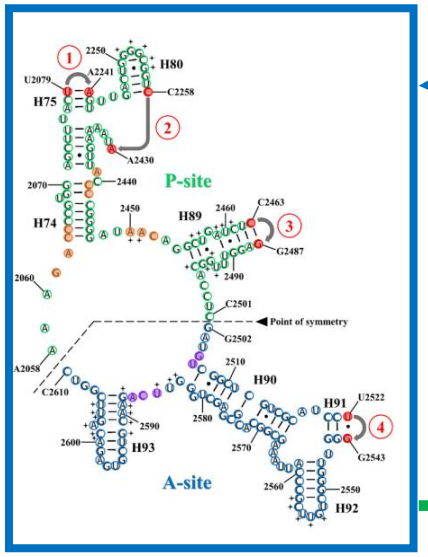
at this time, **RNA-GPT** is not freely available for use

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

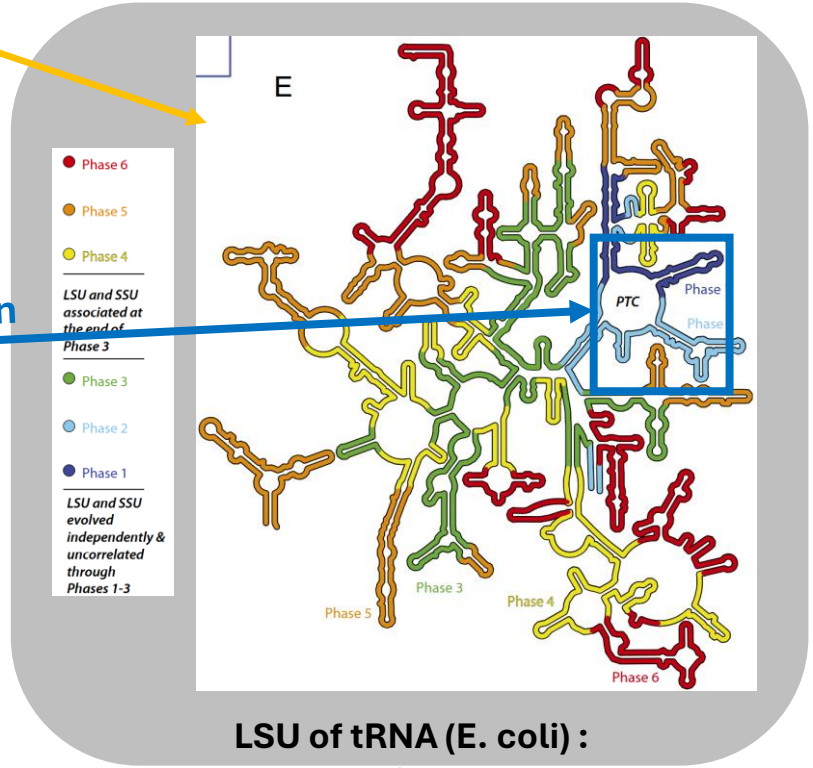
large ribosomal subunit (LSU) rRNA

2. What is the oldest know ncRNA ?

rRNA evolution: Age Variability Across Different Regions

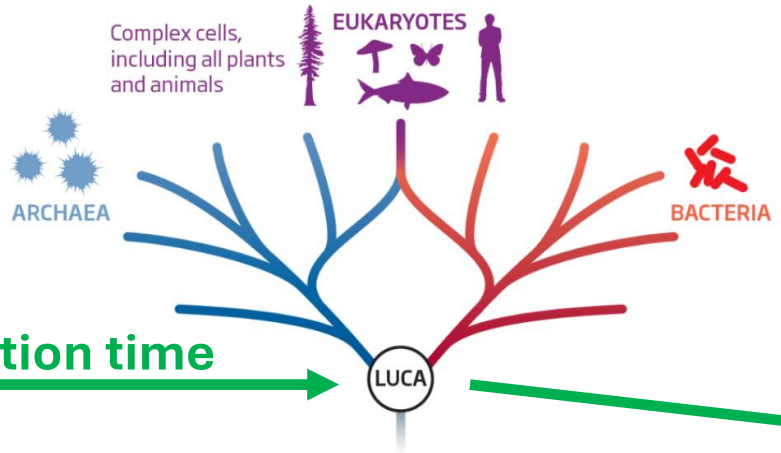


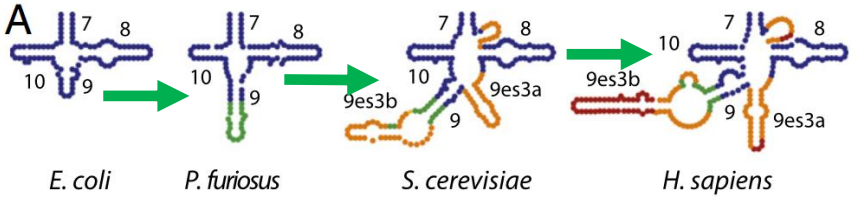
the oldest functional part, conserved throughout evolution



LSU of tRNA (E. coli) :
Blue part is the oldest one

evolution time





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

Last Universal Common Ancestor (LUCA)

Image from:
https://www.pulseheadlines.com/earths-universal-common-ancestor-volcanic-origins/43890/

Molecular level chronology of the evolution of the large ribosomal subunit (LSU) rRNA. Each accretion step adds to previous rRNA but leaves the underlying **core unperturbed**
(Anton S. Petrov et al., PNAS, 2015)

Peptidyl Transferase Center (PTC) is the oldest part of ribosomes and contains no proteins

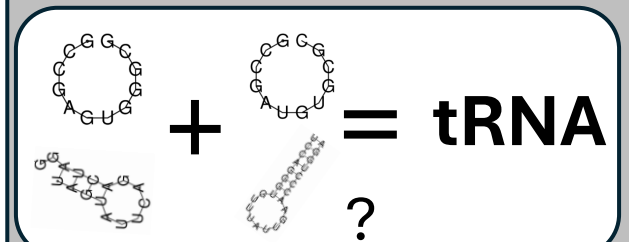
This symmetry (**SymR**) suggests that the ancient ribosome may have been **a dimer of identical or nearly identical RNA molecules**, later evolving into the asymmetrical modern ribosome with **PTC**.

3. What is possible destination in reverse evolution: Hairpins ?

Peptidyl Transferase Center (PTC) Sequences

the idea

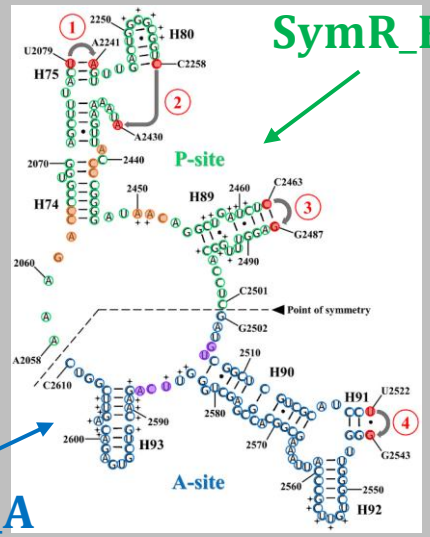
the dimerization of two similar RNA structures



"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

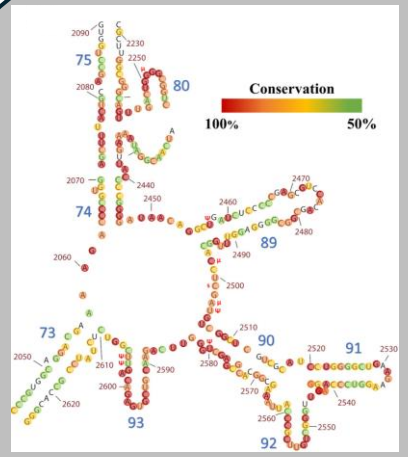
pseudosymmetrical region



SymR_PA = SymR_P + SymR_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