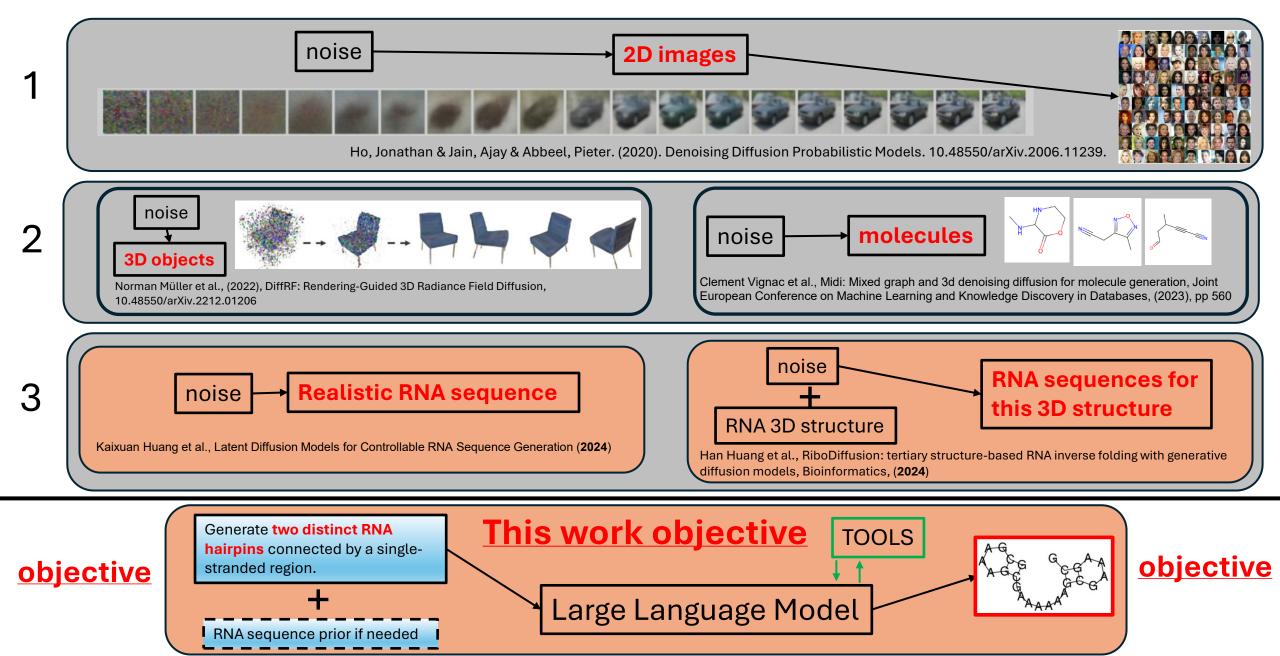
# LLM Agent–Driven ncRNA Design via Intrinsic Features and Structure-Guided Feedback

Work in Progress, Exploratory Analysis
Preliminary Data — Subject to Discussion and Revision

### Inspiration (from Diffusion Models) and the Objective



### **OBJECTIVES**

Develop an LLM agent-driven pipeline that generates and analyzes RNA sequences, under guidance from human-designed prompts. Each pipeline automatically integrates RNA structural features, conditional diffusion-based generative model 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, that proposes alternative RNA sequences with the same 3D structure as the reference one.

#### 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)

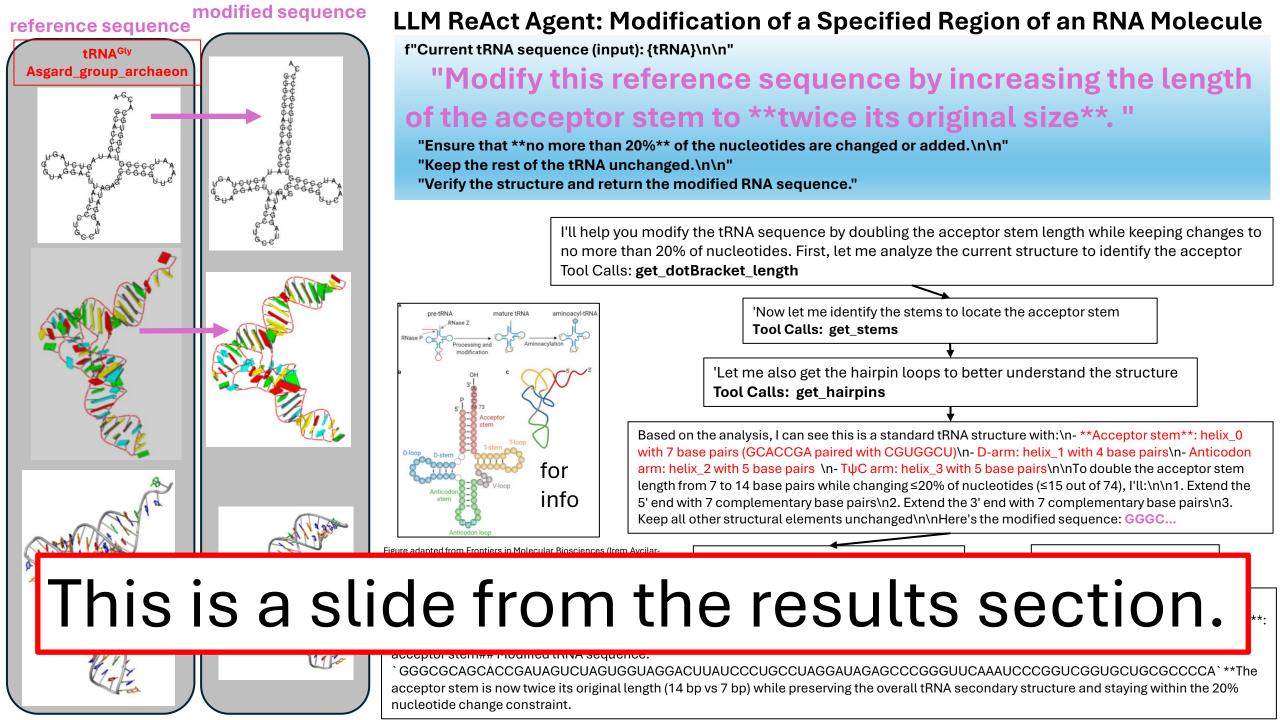
LangGraph (2024): Low-level orchestration framework for building, managing, and deploying long-running, stateful agents, <a href="https://github.com/langchain-ai/langgraph">https://github.com/langchain-ai/langgraph</a>

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)



Small Mutation, Significant Impact: tRNA Point Mutation (A3243G) Linked to MELAS Syndrome DRfold2 3D reconstruction human mt tRNA<sup>Leu(UUR)</sup> AHG wild type cells human mt tRNA<sup>Leu(UUR)</sup> with the A3243G mutation Yohei Kirino et al., Codon-specific translational defect caused by a wobble modification deficiency in mutant tRNA from a human mitochondrial disease, Proc Natl Acad Sci (2004)

### This is another slide from the results section.

### **OUTLINE**

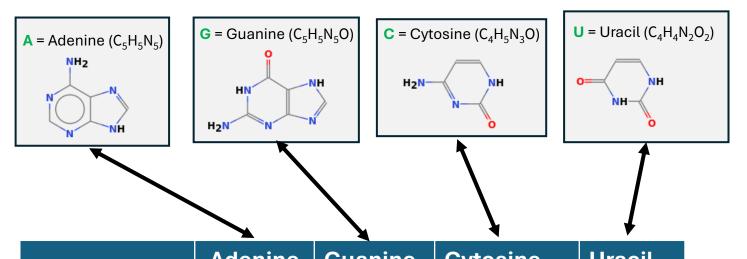
#### Dataset

- What is RNA molecule
- RNAcental database

#### **❖** Tools description

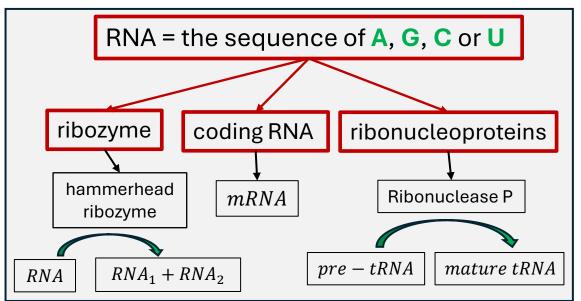
- RNA alignment Tool
- RNA 3D structure parameters Tool
- The structure-guided feedback Tool
- **❖ LLM Tool-Agent Computational Pipeline**
- \* Results for LLM Tool-Agent
  - Speculative simulation of reverse evolution of tRNA<sup>Gly</sup> (6 slides)
- **❖ LLM ReAct-Agent Computational Pipeline**
- **❖** Results for LLM ReAct-Agent
  - Comparison of two RNAs
  - Single point mutation in tRNA associated with MELAS syndrome
  - 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?



GGCGAUCUAGCGCGAUAC	GGUAGCUUAGCGA
C G A U C Bases  C G A G G G G G G G G G G G G G G G G G	RNA Secondary Structure  G G G G G G G G G G G G G G G G G G G

	Adenine	Guanine	Cytosine	Uracii
0	0	1	1	2
N	5	5	3	2
С	5	5	4	4
н	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

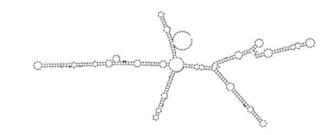


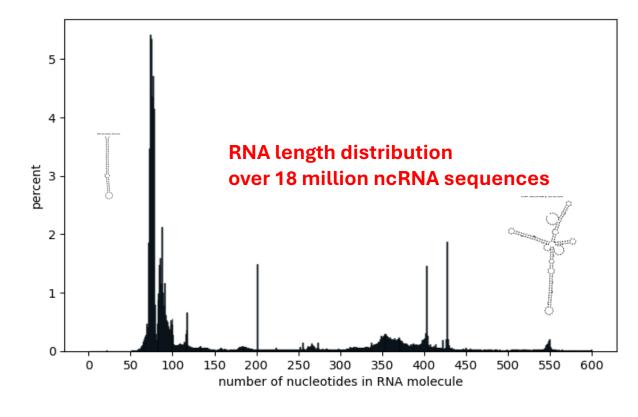
### The dataset: RNAcentral database

#### **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" ] }, ...
                  antisense_RNA
                                        miRNA
      IncRNA
                                ncRNA
   ribozyme
                                       other
            harnmerhead ribozyme
   snoRNA
                                     SRP RNA
             rRNA
                    rna type
     tmRNA
                                      snRNA
                                tRNA
    misc_RNA
                                     scaRNA
                 RNase_MRP_RNA
  pre miRNA
           sRNA
                                  RNase P RNA
                  telomerase RNA
     circRNA
```





#### Objectives

### OUTLINE

- Dataset
  - What is RNA molecule
  - RNAcental database

#### **❖** Tools description

- RNA alignment Tool
- RNA 3D structure parameters Tool
- The structure-guided feedback Tool
- **❖ LLM Tool-Agent Computational Pipeline**
- \* Results for LLM Tool-Agent
  - Speculative simulation of reverse evolution of tRNA<sup>Gly</sup> (6 slides)
- ❖ LLM ReAct-Agent Computational Pipeline
- **❖** Results for LLM ReAct-Agent
  - Comparison of two RNAs
  - Single point mutation in tRNA associated with MELAS syndrome
  - 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?

### **RNA** alignment Tool:

### RNA Prompt analysis using Large Language Model

#### 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.

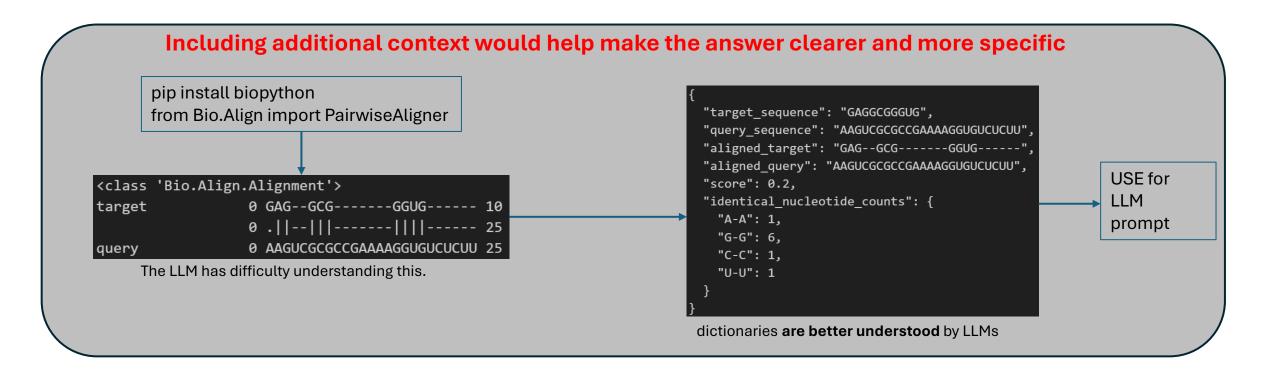
Sequence 1: GGGGGGG

Sequence 2: AAGUCGCGCCGAAAAGGUGUCUCUU

Question: Which sequence is more biologically realistic?

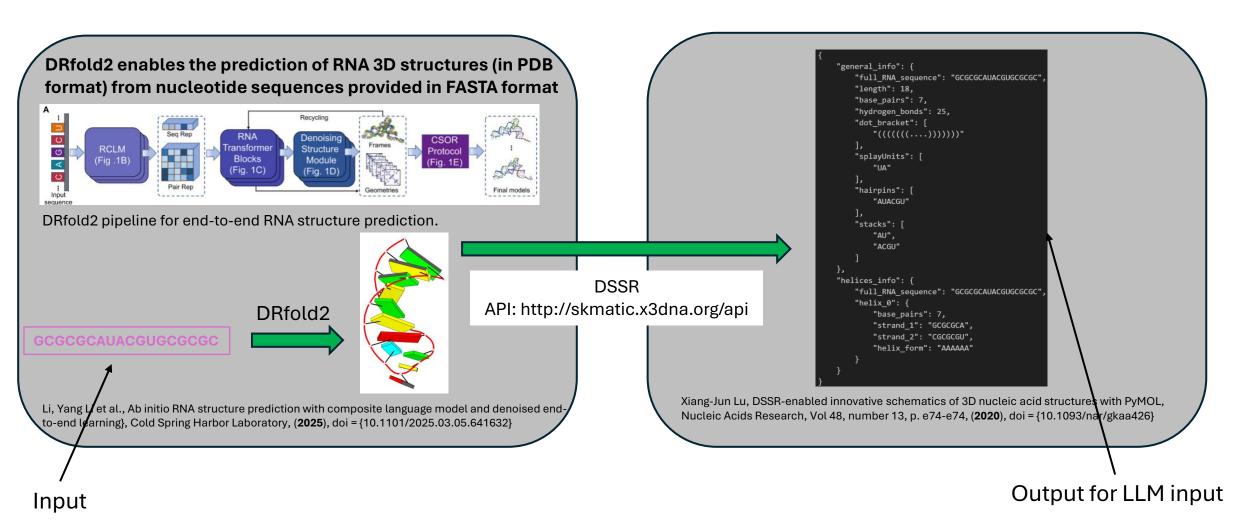
Answer format: One sentence explaining which sequence is more likely and why.

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



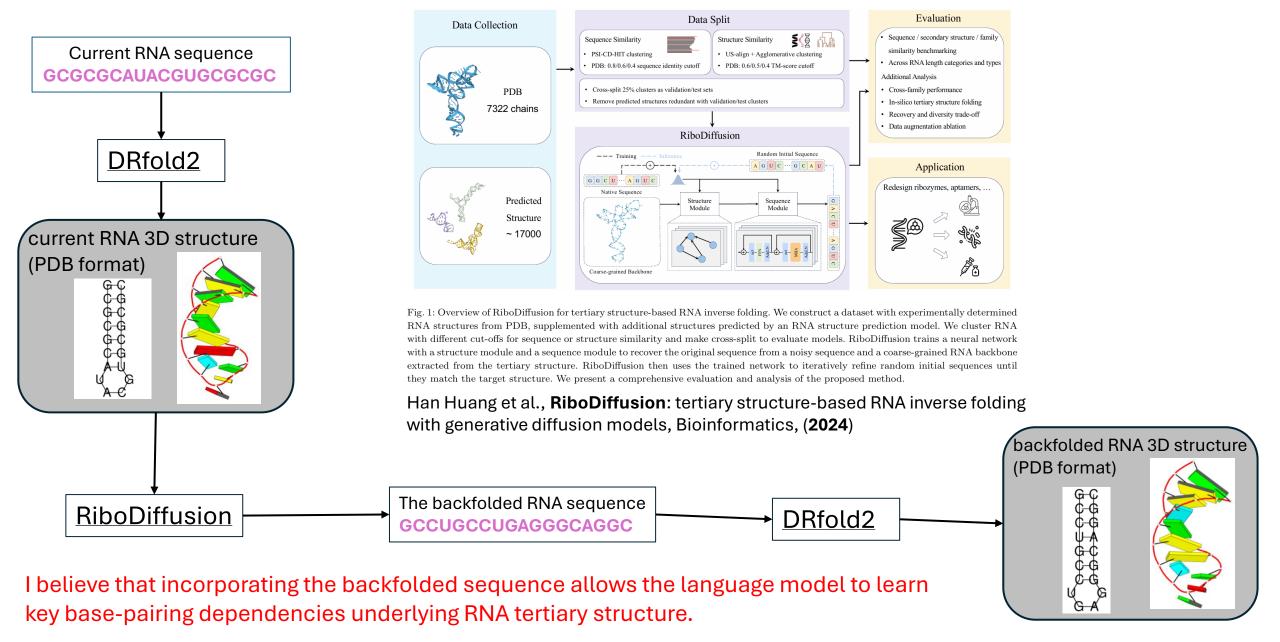
### **RNA 3D structure parameters Tool:**

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



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

### The Structure-Guided Feedback Tool: the inverse folding problem



#### Objectives

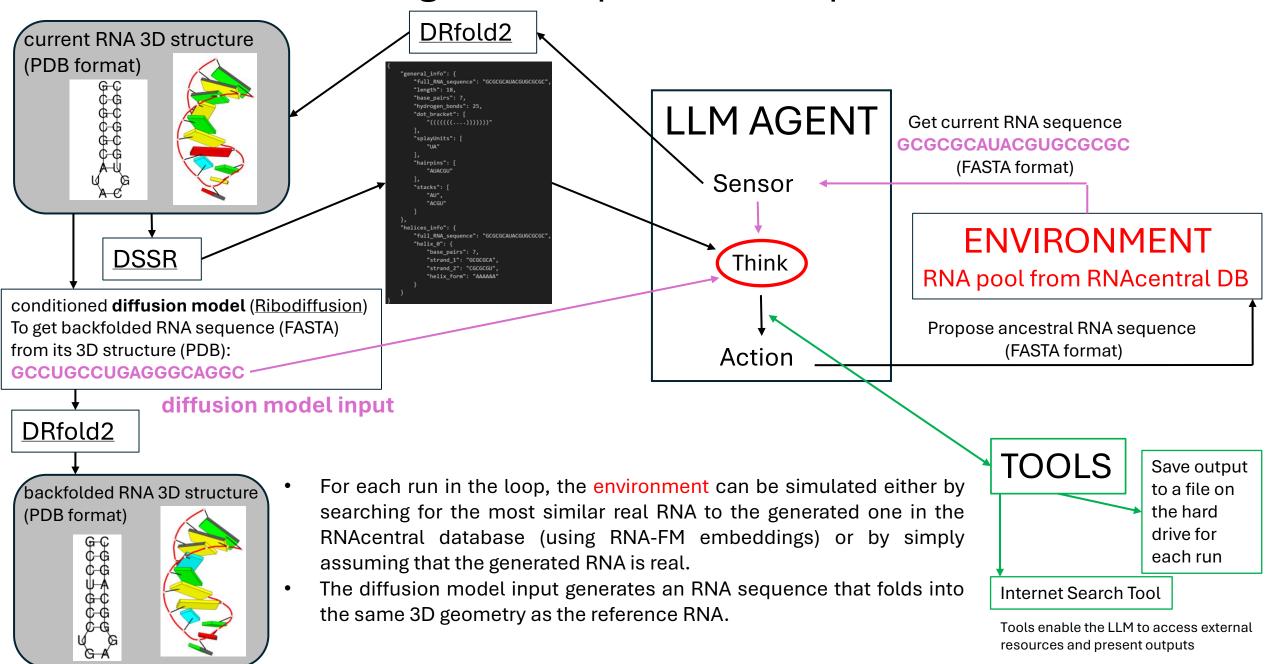
### OUTLINE

- ❖ Dataset
  - What is RNA molecule
  - RNAcental database

#### ❖ Tools description

- RNA alignment Tool
- RNA 3D structure parameters Tool
- The structure-guided feedback Tool
- LLM Tool-Agent Computational Pipeline
- Results for LLM Tool-Agent
  - Speculative simulation of reverse evolution of tRNA<sup>Gly</sup> (6 slides)
- **❖ LLM ReAct-Agent Computational Pipeline**
- **❖** Results for LLM ReAct-Agent
  - Comparison of two RNAs
  - Single point mutation in tRNA associated with MELAS syndrome
  - 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?

### LLM Tool-Agent Computational Pipeline

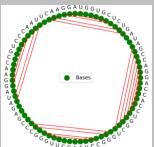


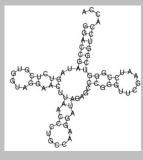
### Prompt Engineering: Designing Effective Inputs for LLMs

```
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}"),
                                                    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"
     "human".
                                                        f"and begin each output in the file with the following section header: evolution_step_{i}")
      """{query}
                                                        raw_response = agent_executor.invoke({"query": query,
                                                                        "current_sequence": current_sequence,
                                                                        "backfolded_sequence": backfolded_sequence,
{current_sequence}
                                                                        "rna structure block": rna structure block})
{rna_structure_block}
{backfolded_sequence}
   ("placeholder", "{agent_scratchpad}"),
).partial(format_instructions=parser.get_format_instructions())
```

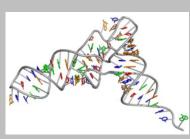
No conditions were imposed on the generated ancestral RNA sequence (e.g., limits on length or number of mutation events per step)!

## STEP 1 GGACCGAUAGUCUCGUGGUAGGAACUUAACCCUGCCAAGG AUAGAGCCCGGGUUCGAAUCCGGGUCGGUCCACCA









#### **DSSR** output

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

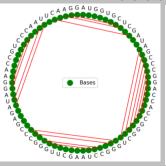
- Length: 75 nucleotides
- Base pairs: 29
- Dot-bracket notation showing complex stem-loop structures
- Multiple hairpins and junction regions
- Two coaxial stacks indicating sophisticated tertiary interactions

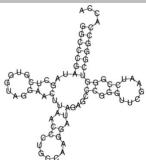
The first two images in each panel were generated using simplified 2D structures for rough visualization purposes only.

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

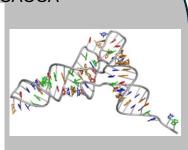
#### **ANCESTRAL\_SEQUENCE:**

#### GGCCCGAUAGCUCGUGGUAGGAACUUAACCCUGCCAAGGAUAGAG CCCGGGUUCGAAUCCGGGUCGGCCACCA









#### ## Rationale for Ancestral Sequence Reconstruction

- \*\*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.
- \*\*Reduced Complexity\*\*: Early RNA molecules likely had simpler tertiary structures with fewer modified nucleotides and less sophisticated coaxial stacking arrangements.
   \*\*Enhanced Thermostability\*\*: Ancestral RNA sequences typically exhibited greater thermodynamic stability to survive in harsh primordial conditions, achieved through strong
- 1. \*\*Functional Core Preservation\*\*: The proposed ancestral sequence maintains the essential functional domains including the acceptor stem, anticodon loop region, and critical
- \*\*Evolutionary Parsimony\*\*: The reconstruction follows the principle of minimal evolutionary changes while preserving essential structural and functional characteristics requir rearly protein synthesis machinery.

ne proposed ancestral sequence represents a plausible evolutionary precursor that could have given rise to the current RNA through gradual structural refinements and optimization r 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.

or the ancestral reconstruction. I considered several evolutionary principles:

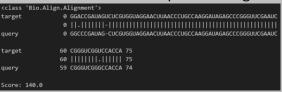
- 1. Ancient RNA sequences likely had simpler secondary structures with fewer complex loops
- 2. Early RNA would have had stronger base-pairing stability in core regions
- . Primitive sequences would have contained more canonical Watson-Crick pairs
- I. The ancestral form would retain essential structural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

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

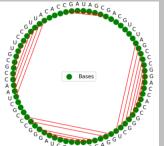
- Simplified hairpin loops with more stable base-pairing
- Reduced junction complexity
- Stronger stem regions with canonical base pair
- Shorter variable loop
- Enhanced stability through increased GC content in critical structural regions

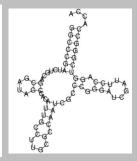
his ancestral form would represent an evolutionary intermediate that possessed the fundamental structural framework while being simpler and more thermodynamically stable that ne modern sequence.

#### Current-Ancestral RNA sequences alignment



# STEP 6 GGCCCGAUCUGCAGCGAUAGCCACAUUGCCUUGCGCCAAUCGCC CGGGAUCGAUUCCAGGUCGGCCACCA









#### **DSSR** output

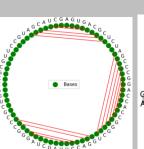
Based on the analysis of the current RNA sequence (71 nucleotides) and its structural properties, I propose an ancestral RNA sequence that maintains the critical structural elements while reflecting probable evolutionary simplifications. The current RNA displays a complex secondary structure with 32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

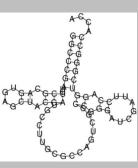
#### Key observations from the current sequence:

- 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

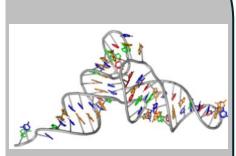
#### ANCESTRAL\_SEQUENCE:

GGCCCGAUCUCGCAGUGAGCUACGAUGCCUUGCGCCAGUCGCCC
GGGAUCGAUUCCAGGUCGGCCACCA









#### 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

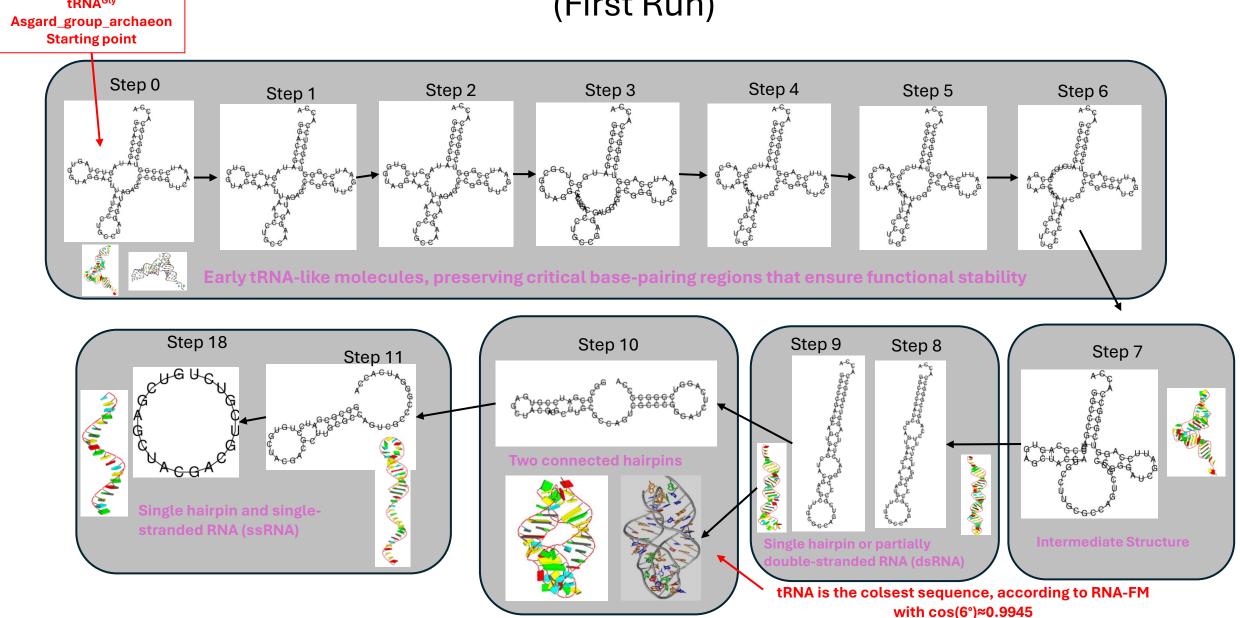
#### The ancestral reconstruction considers:

- 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

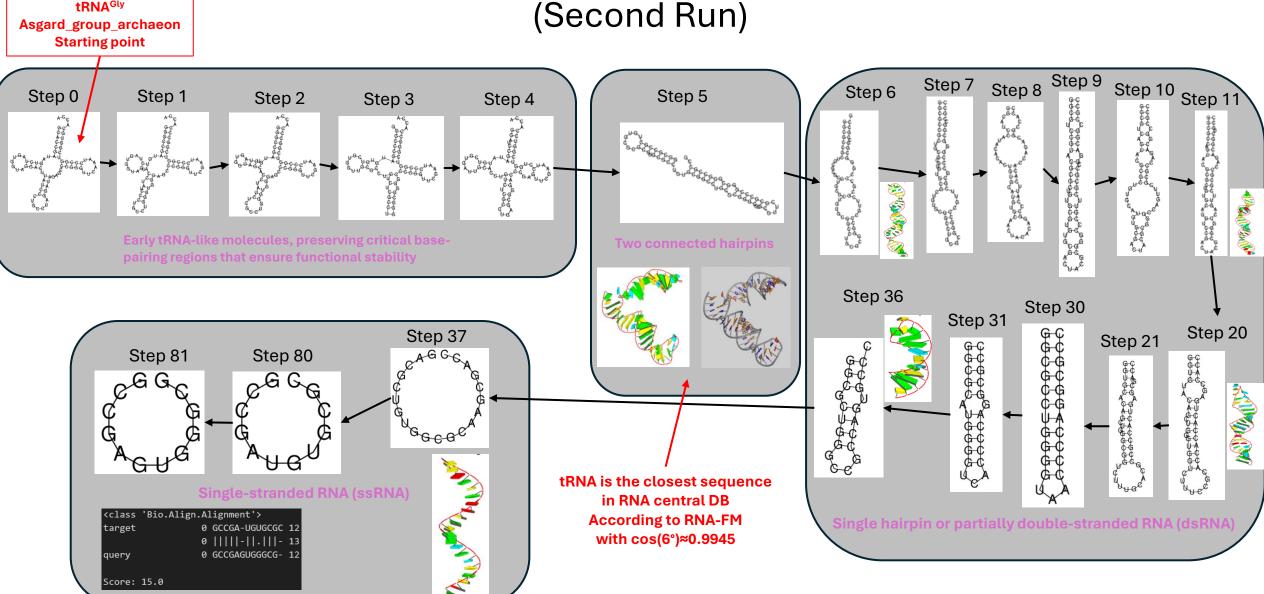
The first two images in each panel were generated using simplified 2D structures for rough visualization purposes only.

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

# Results of Simulated Reverse Evolution of tRNA<sup>Gly</sup> Asgard\_group\_archaeon (First Run)

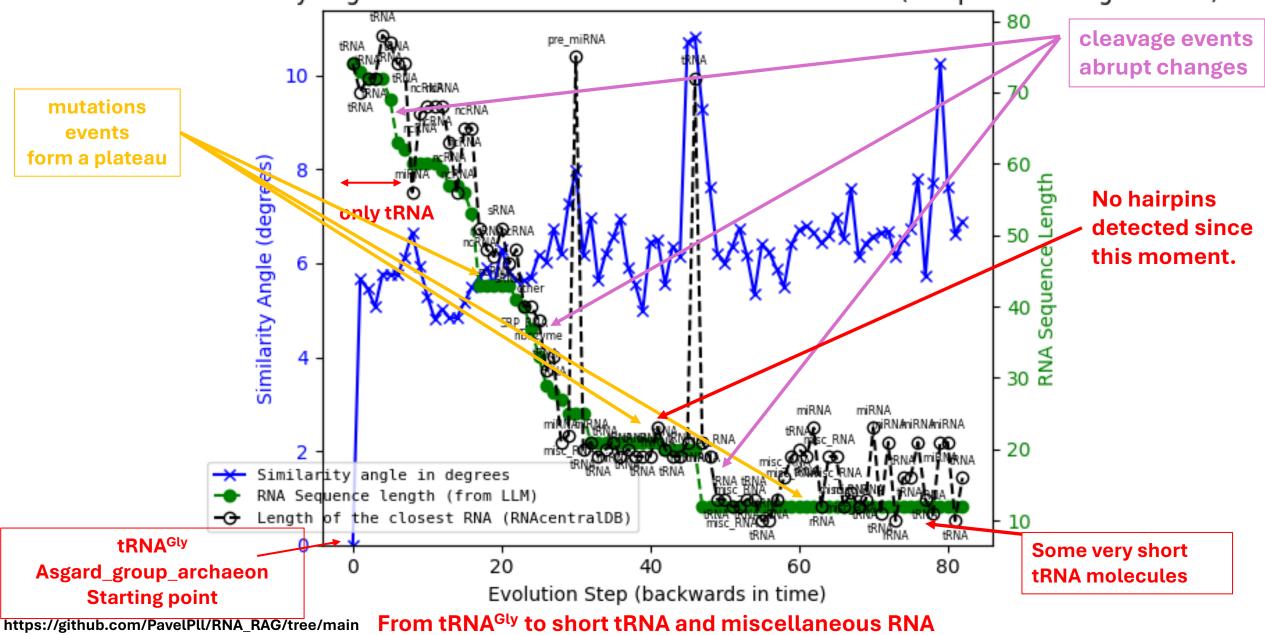


# Results of Simulated Reverse Evolution of tRNA<sup>Gly</sup> Asgard\_group\_archaeon (Second Run)

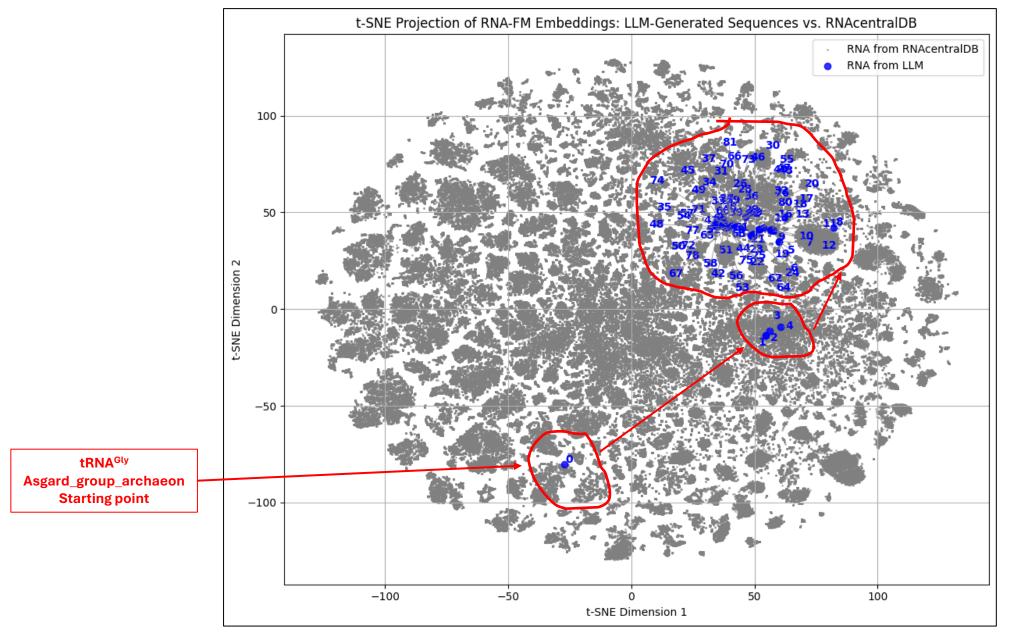


### Using Large Language Models to Simulate the Reverse Evolution of tRNA<sup>Gly</sup>

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



### How Accurately do LLM-Designed RNAs Represent Real RNA?



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

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

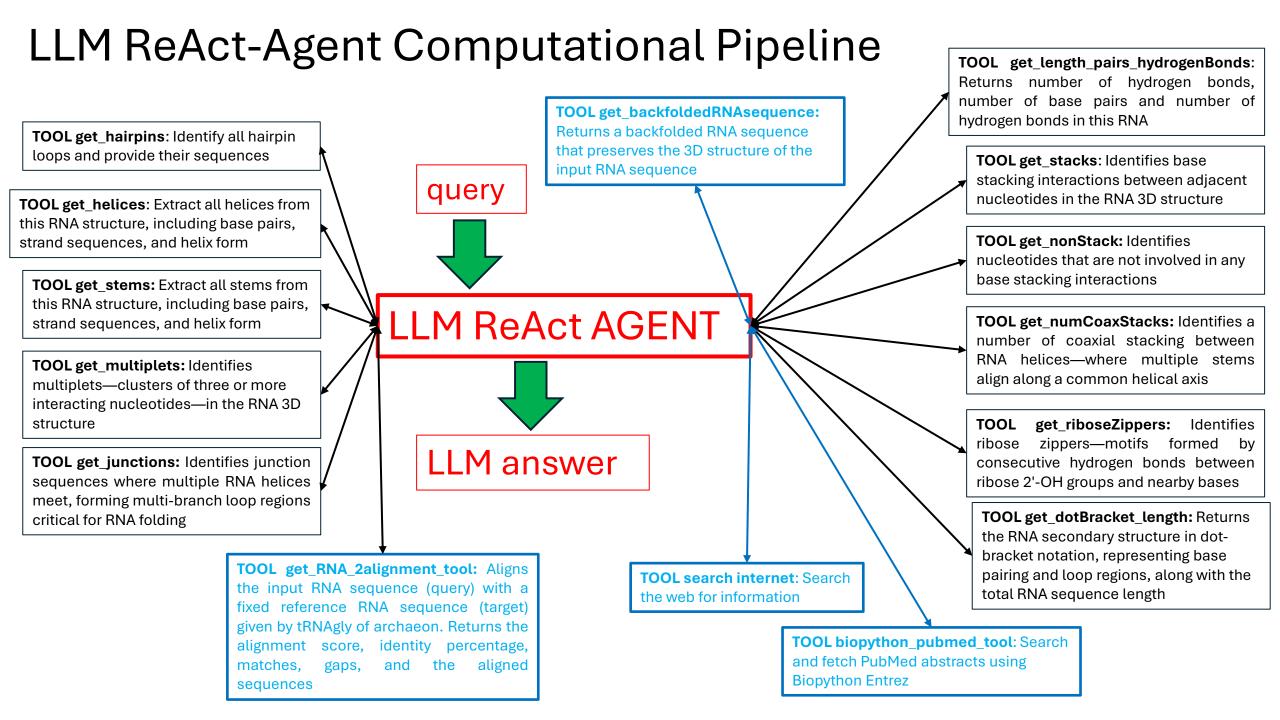
#### Objectives

### **OUTLINE**

- Dataset
  - What is RNA molecule
  - RNAcental database

#### **❖** Tools description

- RNA alignment Tool
- RNA 3D structure parameters Tool
- The structure-guided feedback Tool
- **❖ LLM Tool-Agent Computational Pipeline**
- ❖ Results for LLM Tool-Agent
  - Speculative simulation of reverse evolution of tRNA<sup>Gly</sup> (6 slides)
- LLM ReAct-Agent Computational Pipeline
- \* Results for LLM ReAct-Agent
  - Comparison of two RNAs
  - Single point mutation in tRNA associated with MELAS syndrome
  - 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?



# LLM ReAct Agent: Comparison of Two RNA Sequences

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 Tool Calls: internet search Paper analysis Tool Calls: get\_get\_length\_pairs\_hydrogenBonds for target sequence Tool Calls: get\_get\_length\_pairs\_hydrogenBonds

for query sequence

- 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.

#### # Detailed RNA Molecule Comparison Analysis

#### ## 1. Sequence Alignment Results

- \*\*Query\*\*: AGAGAGAGAGAGAGAGAGAG (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)<sub>10</sub>

#### ### Target Sequence:

- \*\*Length\*\*: 74 nucleotides
- \*\*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

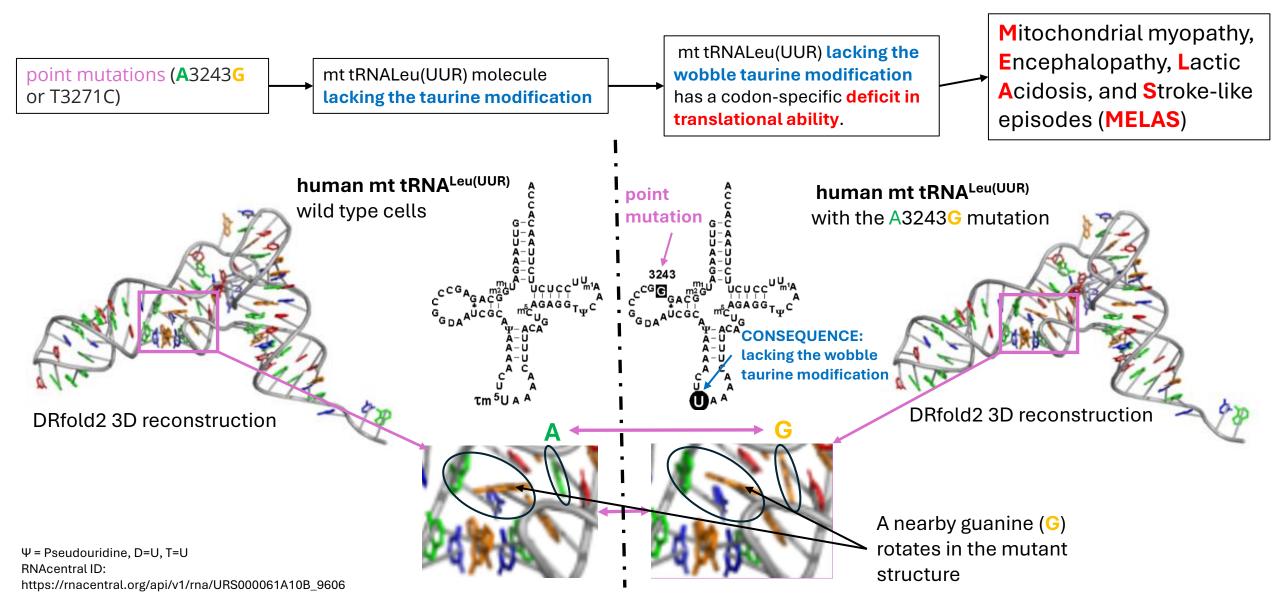
#### ## 4. Similarities and Differences

#### ### 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:

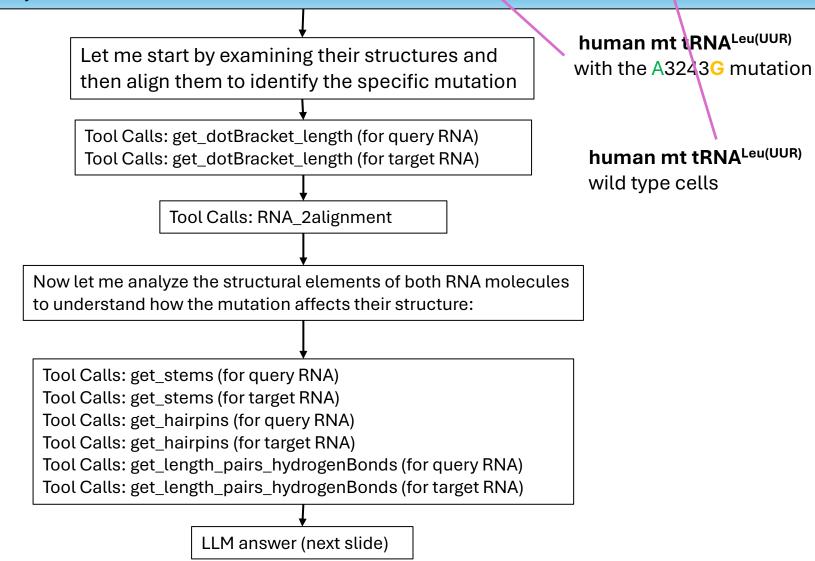
### Single Point Mutation in tRNA Associated with MELAS Syndrome



Yohei Kirino et al., Codon-specific translational defect caused by a wobble modification deficiency in mutant tRNA from a human mitochondrial disease, Proc Natl Acad Sci (2004)

### Single Point Mutation in tRNA Associated with MELAS Syndrome

f"Provide a detailed comparison of the following two RNA molecules: query ({tRNA\_A2G}) and target ({tRNA\_healthy}).\n\n" "Identify any diseases that may be associated with the mutation in the query RNA.\n\n" "Provide the output in clearly labeled sections."



### Single Point Mutation in tRNA Associated with MELAS Syndrome

```
# **Detailed RNA Comparison Analysis**
## **1. Sequence Alignment Results**
### **Basic Statistics:**
- **Sequence Length:** 78 nucleotides (both sequences)
- **Alignment Score:** 153.0
- **Identity: ** 98.72% (77/78 matches)
- **Mismatches:** 1
- **Gaps:** 0 in both sequences
### **Mutation Identification:**
- **Position:** 14 (counting from 5' end)
- **Mutation Type: ** G→A transition
- **Query sequence:**
GUUAAGAUGGCAG**G**GCCCGGUAAUCGCAUAAAACUUAAAACUUUACAGUCAGAGGUUCAAUUCCUC
UUCUUAACACCA
- **Target sequence:**
GUUAAGAUGGCAG**A**GCCCGGUAAUCGCAUAAAACUUAAAACUUUACAGUCAGAGGUUCAAUUCCUC
UUCUUAACACCA
## **2. Secondary Structure Comparison**
### **Dot-Bracket Notation:**
Both RNA molecules have identical secondary structure patterns:
### **Structural Elements:**
**Query RNA (with G at position 14):**
- **Stems: ** 4 helical regions with identical base pairing patterns
- **Hairpin Loops:**
- Loop 1: CAGGGCCCGGUAAUCG (16 nucleotides)
- Loop 2: AACUUAAAACU (11 nucleotides)
- Loop 3: GUUCAAUUC (9 nucleotides)
- **Base Pairs:** 34
- **Hydrogen Bonds:** 106
**Target RNA (with A at position 14):**
- **Stems: ** 4 helical regions with identical base pairing patterns
- **Hairpin Loops:**
- Loop 1: CAGAGCCCGGUAAUCG (16 nucleotides)
- Loop 2: AACUUAAAACU (11 nucleotides)
- Loop 3: GUUCAAUUC (9 nucleotides)
- **Base Pairs:** 37
- **Hvdrogen Bonds:** 108
```

#### ## \*\*3. Structural Impact Analysis\*\*

#### ### \*\*Key Differences:\*\*

- 1. \*\*Base Pairs: \*\* Target RNA has 3 additional base pairs (37 vs 34)
- 2. \*\*Hydrogen Bonds:\*\* Target RNA has 2 additional hydrogen bonds (108 vs 106)
- 3. \*\*Loop Sequence:\*\* The first hairpin loop differs by one nucleotide (GGG vs GAG)

#### ### \*\*Structural Stability:\*\*

The G→A mutation appears to \*\*increase structural stability\*\* in the target RNA by:

- Creating additional base pairing opportunities
- Increasing the total number of hydrogen bonds
- Potentially stabilizing the first hairpin loop region

#### ## \*\*4. RNA Identity and Disease Associations\*\*

Based on the sequence characteristics and structure, this RNA appears to be a \*\*transfer RNA (tRNA)\*\* molecule, given its:

- Length (~78 nucleotides)
- Secondary structure with multiple stems and loops
- Characteristic tRNA-like folding pattern

#### ### \*\*Potential Disease Associations:\*\*

\*\*G>A transitions in tRNA molecules\*\* are commonly associated with:

#### 1. \*\*Mitochondrial Diseases: \*\*

#### - MELAS (Mitochondrial Encephalomyopathy, Lactic Acidosis, and Stroke-like episodes)

- MERRF (Myoclonus Epilepsy with Ragged Red Fibers)
- Various mitochondrial myopathies
- 2. \*\*Metabolic Disorders:\*\*
- Respiratory chain deficiencies
- Energy metabolism dysfunction
- 3. \*\*Neurological Conditions:\*\*
- Progressive external ophthalmoplegia
- Leigh syndrome variants
- \*\*Note:\*\* The specific disease association would depend on:
- Which specific tRNA this represents
- The cellular location (mitochondrial vs cytoplasmic)
- The functional impact on amino acid charging or translation

#### ## \*\*5. Functional Implications\*\*

The mutation from G to A at position 14:

- \*\*Increases structural stability\*\* (more base pairs and H-bonds)

### 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\*\*.

Tool Calls: get\_hairpins
get\_hairpins CAAAG, UUAA

Tool Calls: get\_dotBracket\_length
get\_dotBracket\_length2D
((((((...)))))).......(((((...)))))
RNA length: 35 nucleotides

-----Generate two distinct RNA hairpins connected by a \*\*singlestranded RNA linker\*\*.-----Ai Message [om------[ftext:"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```\nGGGCCCAAAGGGCCCUUUUUUUUGGGUUUAAACCC\n```\n\nThis consists of:\n- \*\*Hairpin 1:\*\* `GGGCCCAAAGGCCC` (GGG-CCC stem with AAA loop)\n- \*\*Linker:\*\* `UUUUUUUUU` (8 uracils as singlestranded connector)\n- \*\*Hairpin 2:\*\* `GGGUUUAAACCC` (GGG-CCC stem with UUU-AAA loop)\n\nNow let me analyze this structure to verify the hairpins: "type: 'text', ('id': 'toolu 010645amkxCPhDjA4qxe0bVH', 'input': [\_arg1': 'GGGCCAAAGGGCCCUUUUUUUUGGGUUUAAACCC', 'name': 'get hairpins', 'type': tool use'] Tool Calls: get hairpins (toolu 010645amkxCPhDjA4qxeQbVH) Call ID: toolu 010645amkxCPhDjA4qxeQbVH Args: \_arg1 the dot-bracket notation to visualize the secondary structure: ', 'type: 'text'}, (id: 'toolu\_0149fG8yTvHaMattqfVq44nJ, 'input': {\_arg1': 'GGGCCCAAAGGGCCCUUUUUUUUGGGUUUAACCC'), 'name': 'get\_dotBracket\_length', 'type': 'tool\_use'}Tool Calls: get\_dotBracket\_length (toolu\_0149fG8yTvHaMatfqfVq44nJ)Call ID: toolu\_0149fG8yTvHaMatfqVq44nJ structure in dot-bracket notation: ((((((...)))))).......(((((...))))) RNA length: 35 nucleotides...... Message [m------Perfect! The analysis confirms our design. Here's the complete structure:## Generated RNA Structure with Two Hairpins and Linker\*\*RNA Sequence:\*\* `GGGCCCAAAGGGCCCUUUUUUUUGGGUUUAAACCC`\*\*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

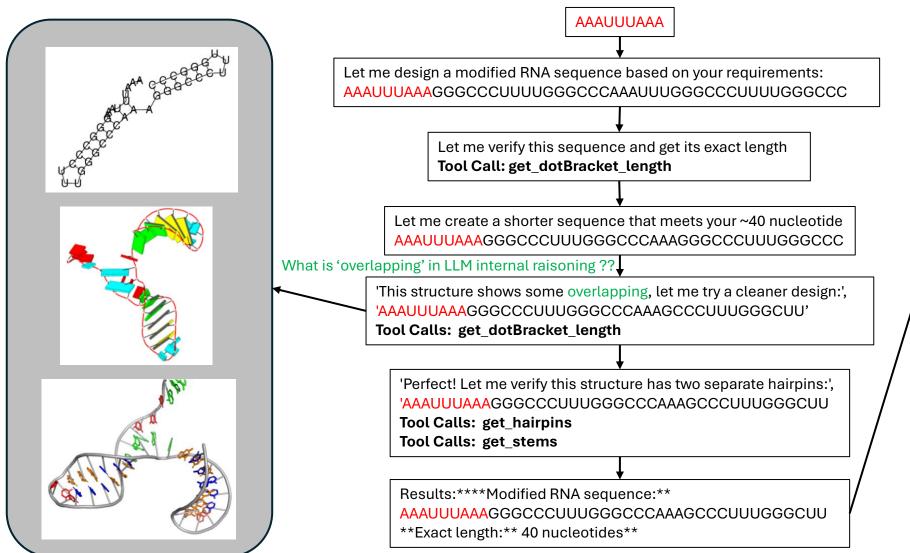
LLM could have called the **get\_stems** tool to verify the stems, but it didn't. Still, not bad for such a simple prompt.

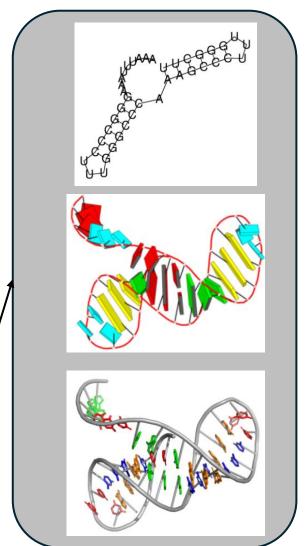
### LLM ReAct agent: Generate Two Connected RNA Hairpins with a Prior RNA Template

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.





# modified sequence reference sequence **tRNA**Gly Asgard group archaeon

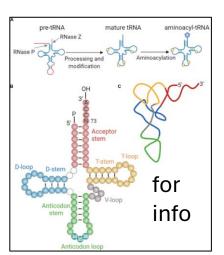
### LLM ReAct Agent: Targeted modification of a specific region of an RNA 1

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):

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-  $T\psi C$  arm: helix\_3 with 5 base pairs\n\nTo double the acceptor stem length from 7 to 14 base pairs while changing  $\leq 20\%$  of nucleotides ( $\leq 15$  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, T\u00fcC arm) remain unchanged- \*\*Base pair complementarity\*\*: Maintained throughout the extended acceptor stem## Modified tRNA sequence:\*\*

modified sequence LLM ReAct Agent: Targeted modification of a specific region of an RNA 2 reference sequence f"Current tRNA sequence (input): {tRNA}\n\n" It does not **tRNA**Gly "Modify this reference sequence by increasing the length of the **D-stem** to \*\*twice its original size\*\*." work Asgard group archaeon "Use backfolded sequence which forms the same 3D structure. \n\n" without "Keep the rest of the tRNA sequence and its cloverleaf structure unchanged.\n\n" backfolded "Verify the structure and return the modified RNA sequence." 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 T\u00fcC 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 get\_dotBracket\_length Tool Calls: get\_stems '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 Tool Calls: get\_stems 'Getting closer! Let me make one more adjustment to achieve exactly 8 base pairs in the D-Tool Calls: get\_stems 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 Name: get\_dotBracket\_length I have successfully modified the tRNA sequence by doubling the D-stem length from 4 to 8 base pairs.

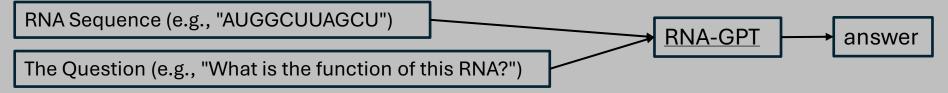
### CONCLUSIONS

- ❖ 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).

### 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. Comprehensive literature research
- 4. Add more Tools

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



The RNA-FM sequence encoder, which accounts for 3D structure, was used to embed RNA sequences for alignment with natural language, at this time, RNA-GPT is not available for use

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

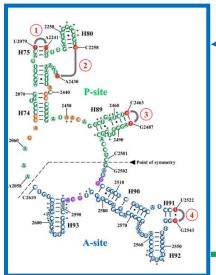
#### 6. Fundamental Studies?

- The oldest part of ribosome: pseudosymmetrical region (~90+90nt)
- The Spiegelman monster: the shortest "self-replicating" RNA molecule (~220nt)

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

Complex cells, including all plants and animals

EUKARYOTES

ARCHAEA

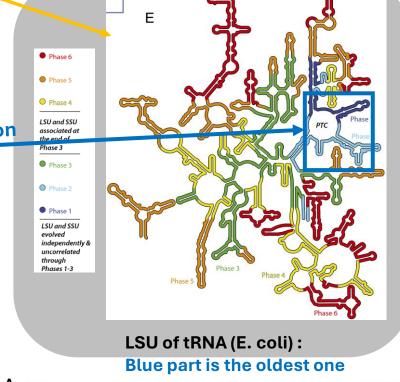
Evolution time

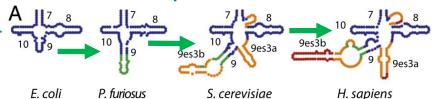
The pseudosymmetrical region of rRNA (**SymR**; *Agmon et al., 2005*), derived from the LSU secondary structure of Thermus thermophilus (*Petrov et al., 2013*).

(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**.

### Peptidyl Transferase Center (PTC) Sequences

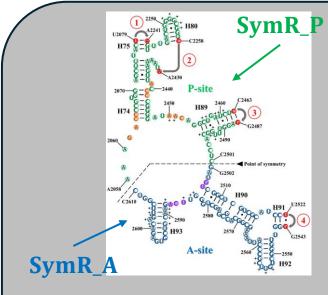
#### the idea

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



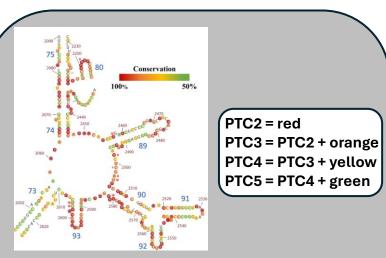
#### $SymR_PA = SymR_P + SymR_A$

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)

**SymR\_P** is older than **SymR\_PA**PTC2 is older than PTC3, PTC4, PTC5

#### PTC



#### **Nucleotide CONSERVATION level:**

Red circles: 100% conservation (78 nt).

Drange 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)