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

Work in Progress, Exploratory Analysis

Preliminary Data — Subject to Discussion and Revision

Focuses on RNA ribonucleotide sequences; no DNA or proteins explicitly involved

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

1

noise → **2D images**

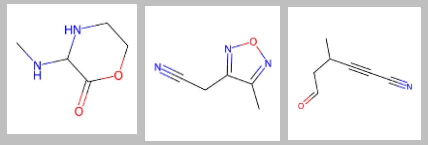
Ho, Jonathan & Jain, Ajay & Abbeel, Pieter. (2020). Denoising Diffusion Probabilistic Models. 10.48550/arXiv.2006.11239.

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

This work objective

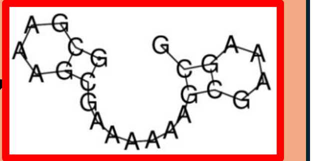
Generate **two distinct RNA hairpins** connected by a single-stranded region.

+

RNA sequence prior if needed

TOOLS

Large Language Model



objective

OBJECTIVES

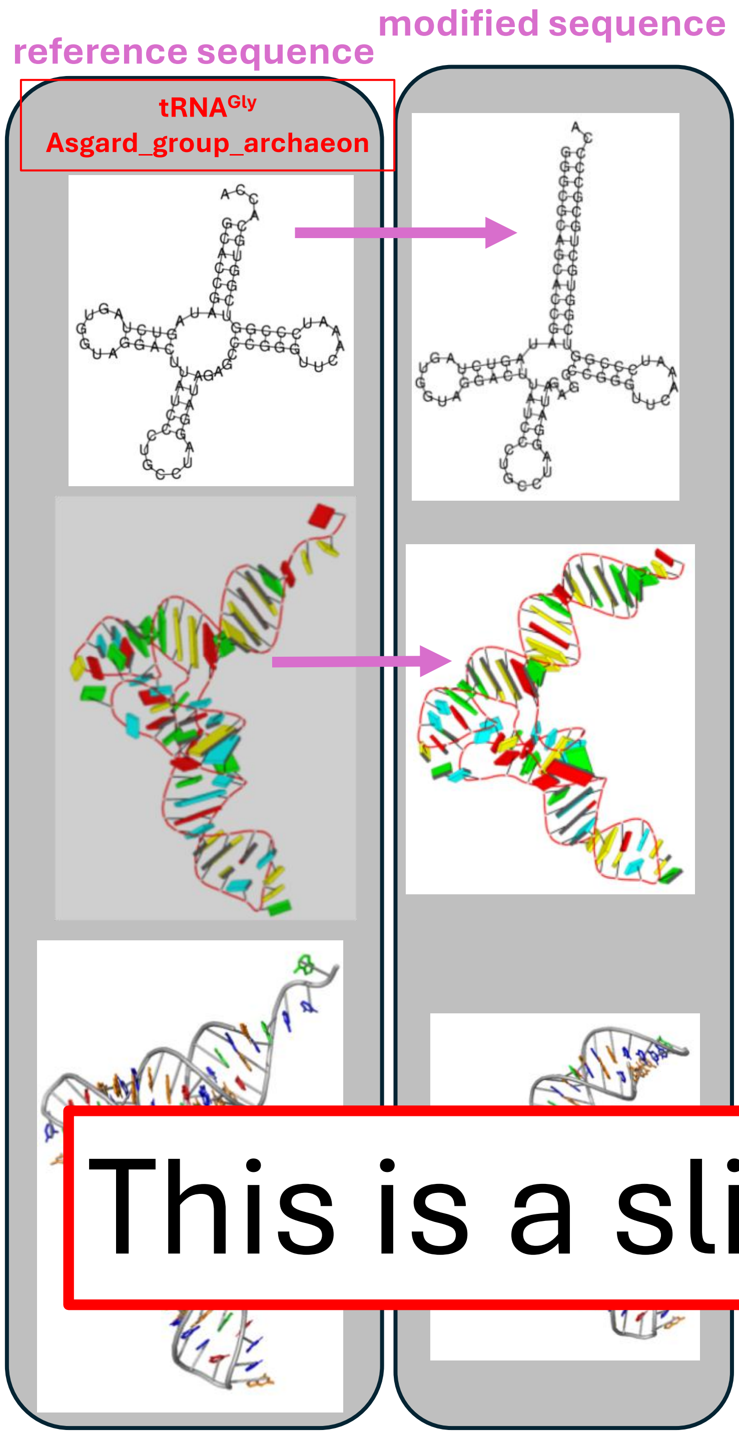
Develop an LLM agent–driven pipeline that generates and analyzes RNA sequences, under guidance from human-designed prompts. Each pipeline automatically integrates RNA structural features, conditional diffusion-based generative model and internet search.

- **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**, which proposes alternative RNA sequences with the same 3D structure as the reference.

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, https://github.com/langchain-ai/langgraph*
***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**)*

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



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. "**

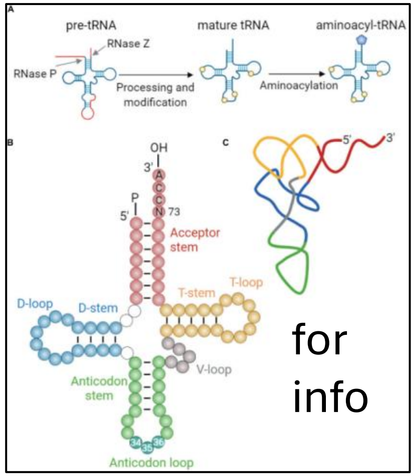
"Ensure that **no more than 20% of the nucleotides are changed or added.\n\n"**
"Keep the rest of the tRNA unchanged.\n\n"
"Verify the structure and return the modified RNA sequence."

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...



for info

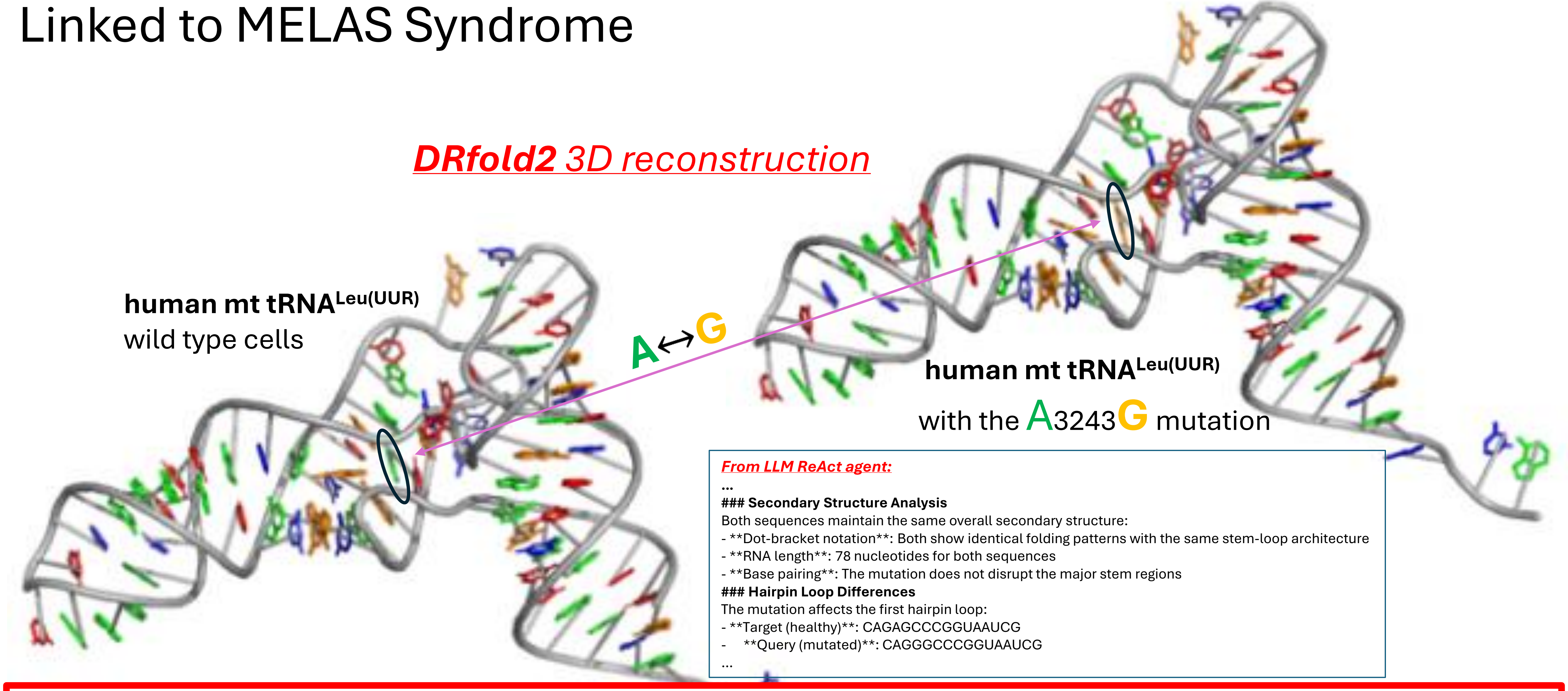
Figure adapted from Frontiers in Molecular Biosciences (Irem Avcilar-

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.

Small Mutation, Significant Impact: tRNA Point Mutation (A3243G) Linked to MELAS Syndrome



DRfold2 *3D reconstruction*

human mt tRNA$^{\text{Leu(UUR)}}$
wild type cells

A↔G

human mt tRNA$^{\text{Leu(UUR)}}$
with the A3243G mutation

From LLM ReAct agent:

```
...
### Secondary Structure Analysis
Both sequences maintain the same overall secondary structure:
- **Dot-bracket notation**: Both show identical folding patterns with the same stem-loop architecture
- **RNA length**: 78 nucleotides for both sequences
- **Base pairing**: The mutation does not disrupt the major stem regions
### Hairpin Loop Differences
The mutation affects the first hairpin loop:
- **Target (healthy)**: CAGAGCCCGGUAAUCG
-    **Query (mutated)**: CAGGGCCCGGUAAUCG
...
```

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

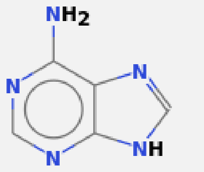
- **LLM ReAct-Agent Computational Pipeline**
- **Results for LLM ReAct-Agent**
  - Comparison of two RNAs
  - Generate two connected RNA hairpins without RNA template
  - Generate two connected RNA hairpins from a given RNA template
  - Targeted modification of a specific region of an RNA molecule

- **Example Use Case**
  - Single point mutation in tRNA associated with MELAS syndrome

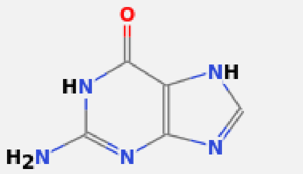
- **Conclusions**
- **What's next ?**

What is RNA molecule ?

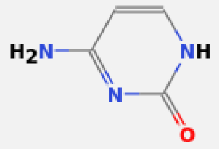
A = Adenine ($C_5H_5N_5$)



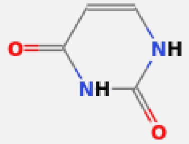
G = Guanine ($C_5H_5N_5O$)



C = Cytosine ($C_4H_5N_3O$)

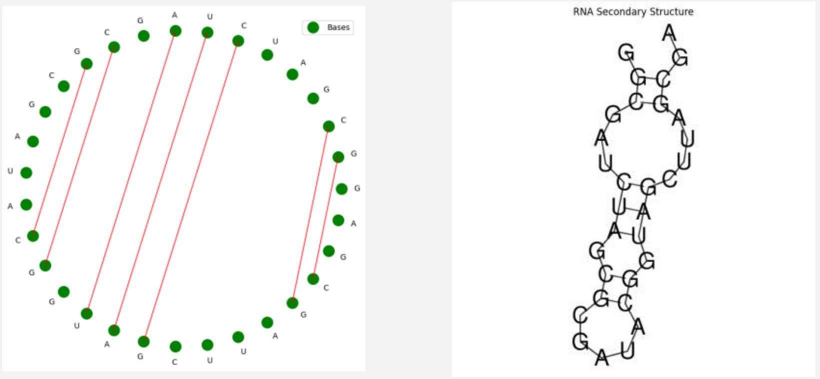


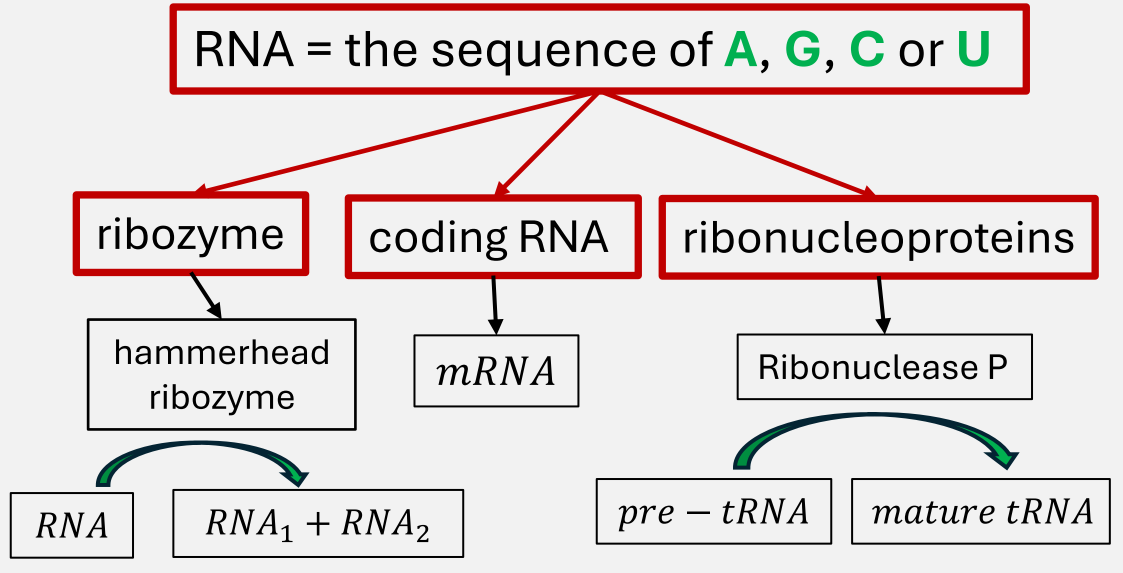
U = Uracil ($C_4H_4N_2O_2$)



	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

$pre-tRNA$

$mature\ tRNA$

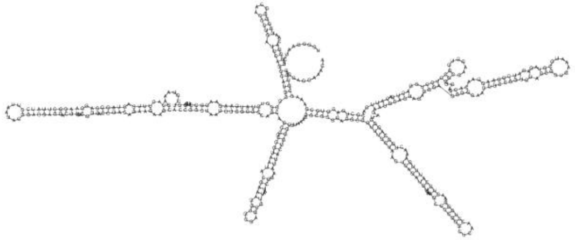
The dataset: RNAcentral database

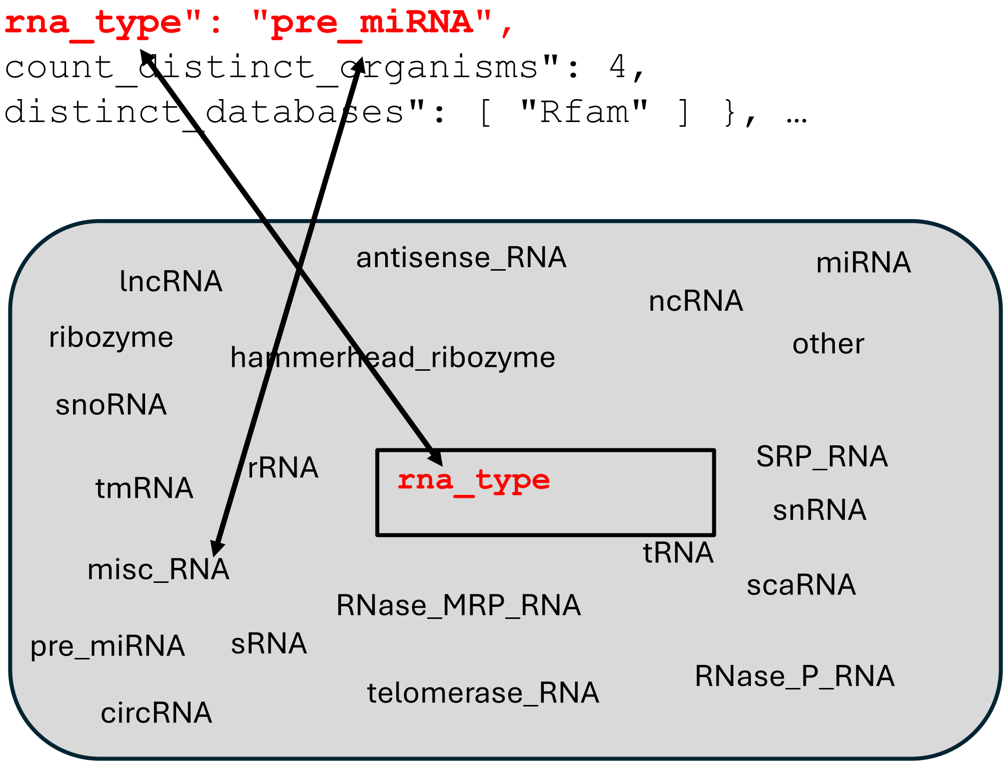
RNAcentral Browsable API

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

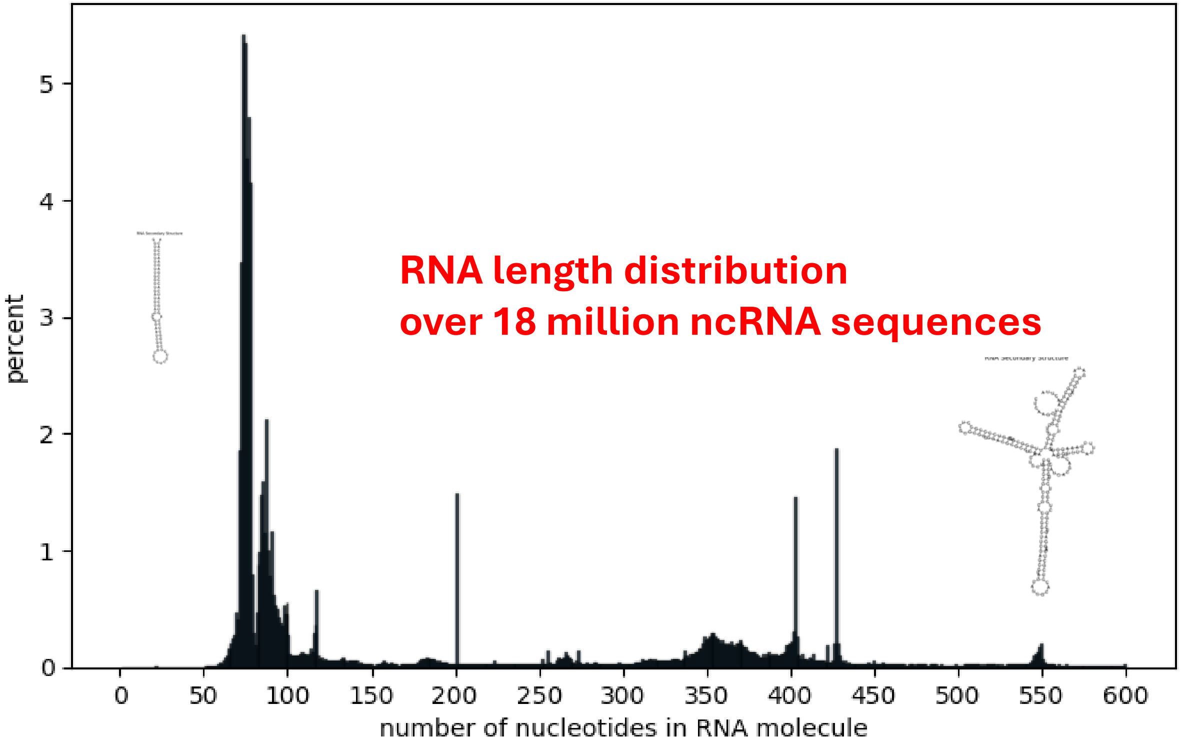
```
{ "url": "http://rnacentral.org/api/v1/rna/URS0002915621",
"rnacentral_id": "URS0002915621",
"md5": "fee3fe68dbd91ee898bffd9d4b89b2e9",
"sequence": "AUGGAUGGUUGAUCAGAGAACGUACAUUUUAUAAAUGGUGUAUGUCAAUUGAUCCACAGUCCCU",
"length": 64,
"xrefs": "http://rnacentral.org/api/v1/rna/URS0002915621/xrefs",
"publications": "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 Secondary Structure





lncRNA
antisense_RNA
miRNA
ncRNA
ribozyme
hammerhead_ribozyme
other
snoRNA
rRNA
rna_type
SRP_RNA
tmRNA
snRNA
tRNA
misc_RNA
scaRNA
RNase_MRP_RNA
pre_miRNA
sRNA
RNase_P_RNA
telomerase_RNA
circRNA



RNA length distribution
over 18 million ncRNA sequences

OUTLINE

- **Dataset**
  - What is RNA molecule
  - RNAcental database

- **Tools description**
  - RNA alignment Tool
  - RNA 3D structure parameters Tool
  - The structure-guided feedback Tool

- **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

- **Example use case**
  - Single point mutation in tRNA associated with MELAS syndrome

- **Conclusions**
- **What’s next ?**

RNA alignment Tool: RNA Prompt analysis using Large Language Model

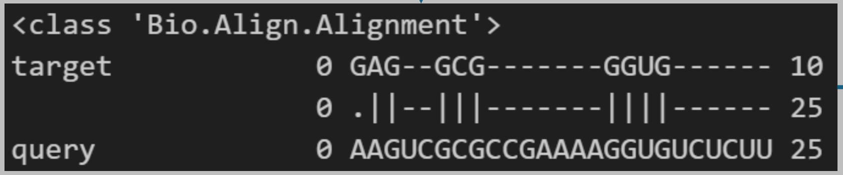
The answer is wrong!

================================ Human Message =================================
Current RNA sequence (input): GUUAAGAUGGCAGAGCCCGGUAAUCGCAUAAAACUUAAAACUUUACAGUCAGAGGUUCAAUUCCUCUUCUUAACACCA
How many nucleotides in this sequence?
================================== Ai Message ==================================
I'll count the nucleotides in the given RNA sequence.
Let me analyze the sequence:
GUUAAGAUGGCAGAGCCCGGUAAUCGCAUAAAACUUAAAACUUUACAGUCAGAGGUUCAAUUCCUCUUCUUAACACCA
Counting each nucleotide character:
G-U-U-A-A-G-A-U-G-G-C-A-G-A-G-C-C-C-G-G-U-A-A-U-C-G-C-A-U-A-A-A-A-C-U-U-A-A-A-A-C-U-U-U-A-C-A-G-U-C-A-G-A-G-G-U-U-C-A-A-U-U-C-C-U-C-U-U-C-U-U-A-A-C-A-C-C-A
~~**Total nucleotides: 76**~~
~~The RNA sequence contains 76 nucleotides, consisting of:~~
~~- Guanine (G): 19 nucleotides~~
~~- Uracil (U): 25 nucleotides~~
~~- Adenine (A): 20 nucleotides~~
~~- Cytosine (C): 12 nucleotides~~

The Correct answer:
Total nucleotides: 78
- Guanine (G): 13 nucleotides
- Uracil (U): 21 nucleotides
- Adenine (A): 26 nucleotides
- Cytosine (C): 18 nucleotides

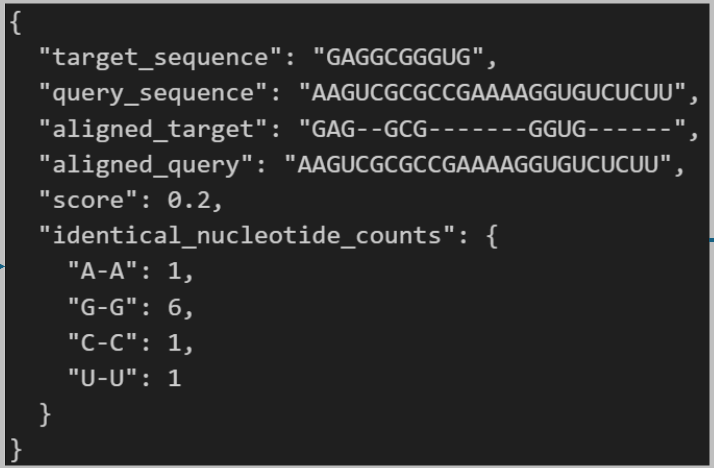
Including additional context would help make the answer clearer and more specific

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
  }
}
```

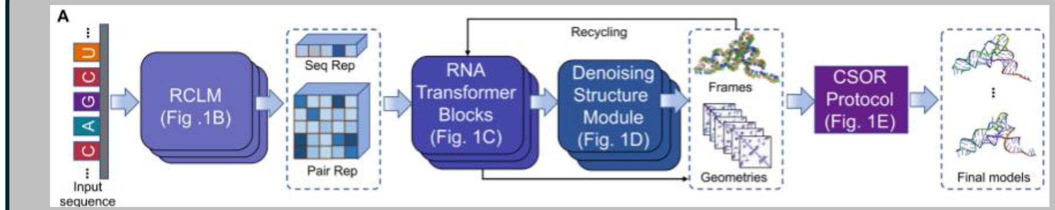
dictionaries **are better understood** by LLMs

USE for
LLM
prompt

RNA 3D structure parameters Tool:

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

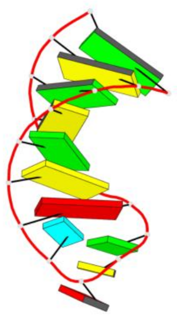
DRfold2 enables the prediction of RNA 3D structures (in PDB format) from nucleotide sequences provided in FASTA format



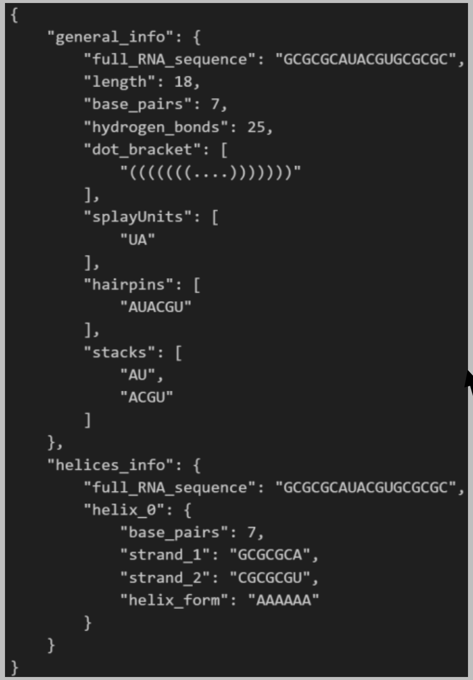
DRfold2 pipeline for end-to-end RNA structure prediction.

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**)

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**)

Input

Output for LLM input

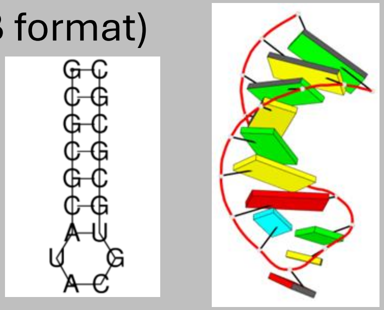
From RNA sequence to a parametric description of its 3D structure

The Structure-Guided Feedback Tool: the inverse folding problem

Current RNA sequence
GCGCGCAUACGUGCGCGC

DRfold2

current RNA 3D structure
(PDB format)

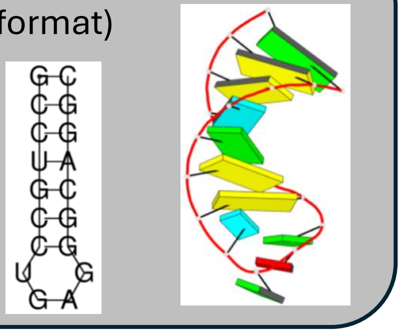


RiboDiffusion

The backfolded RNA sequence
GCCUGCCUGAGGGCAGGC

DRfold2

backfolded RNA 3D structure
(PDB format)



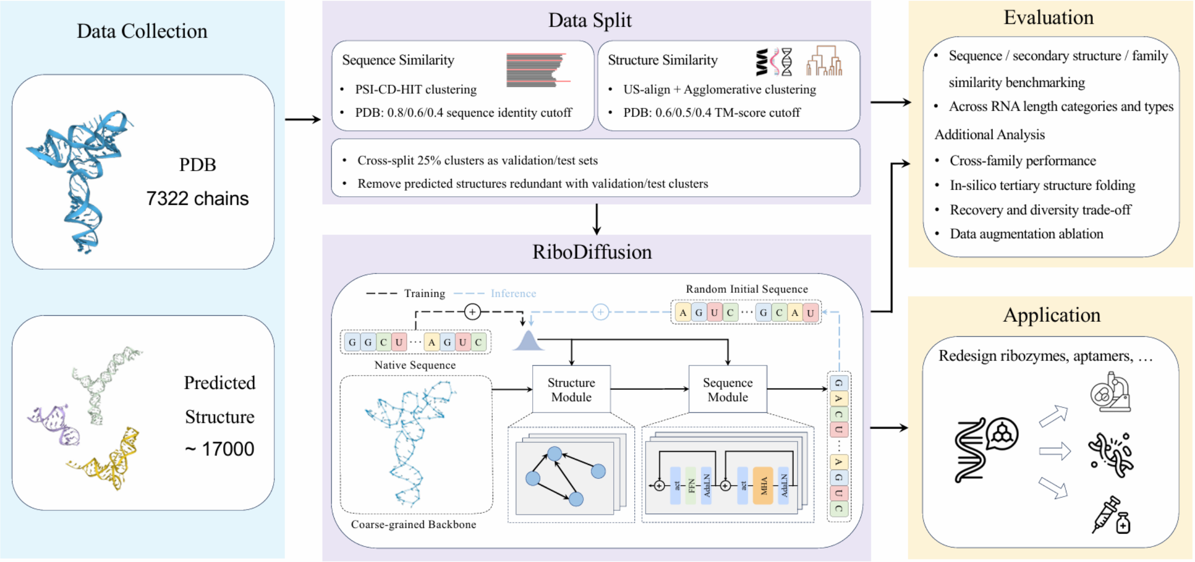

Data Collection
PDB
7322 chains
Predicted Structure ~ 17000
Data Split
Sequence Similarity
PSI-CD-HIT clustering
PDB: 0.8/0.6/0.4 sequence identity cutoff
Structure Similarity
US-align + Agglomerative clustering
PDB: 0.6/0.5/0.4 TM-score cutoff
Cross-split 25% clusters as validation/test sets
Remove predicted structures redundant with validation/test clusters
RiboDiffusion
Training
Inference
Random Initial Sequence
Native Sequence
Structure Module
Sequence Module
Coarse-grained Backbone
Evaluation
Sequence / secondary structure / family similarity benchmarking
Across RNA length categories and types
Additional Analysis
Cross-family performance
In-silico tertiary structure folding
Recovery and diversity trade-off
Data augmentation ablation
Application
Redesign ribozymes, aptamers, …

Fig. 1: Overview of RiboDiffusion for tertiary structure-based RNA inverse folding. We construct a dataset with experimentally determined RNA structures from PDB, supplemented with additional structures predicted by an RNA structure prediction model. We cluster RNA with different cut-offs for sequence or structure similarity and make cross-split to evaluate models. RiboDiffusion trains a neural network with a structure module and a sequence module to recover the original sequence from a noisy sequence and a coarse-grained RNA backbone extracted from the tertiary structure. RiboDiffusion then uses the trained network to iteratively refine random initial sequences until they match the target structure. We present a comprehensive evaluation and analysis of the proposed method.

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

I believe that incorporating the backfolded sequence allows the language model to learn key base-pairing dependencies underlying RNA tertiary structure.

OUTLINE

- **Dataset**
  - What is RNA molecule
  - RNAcental database

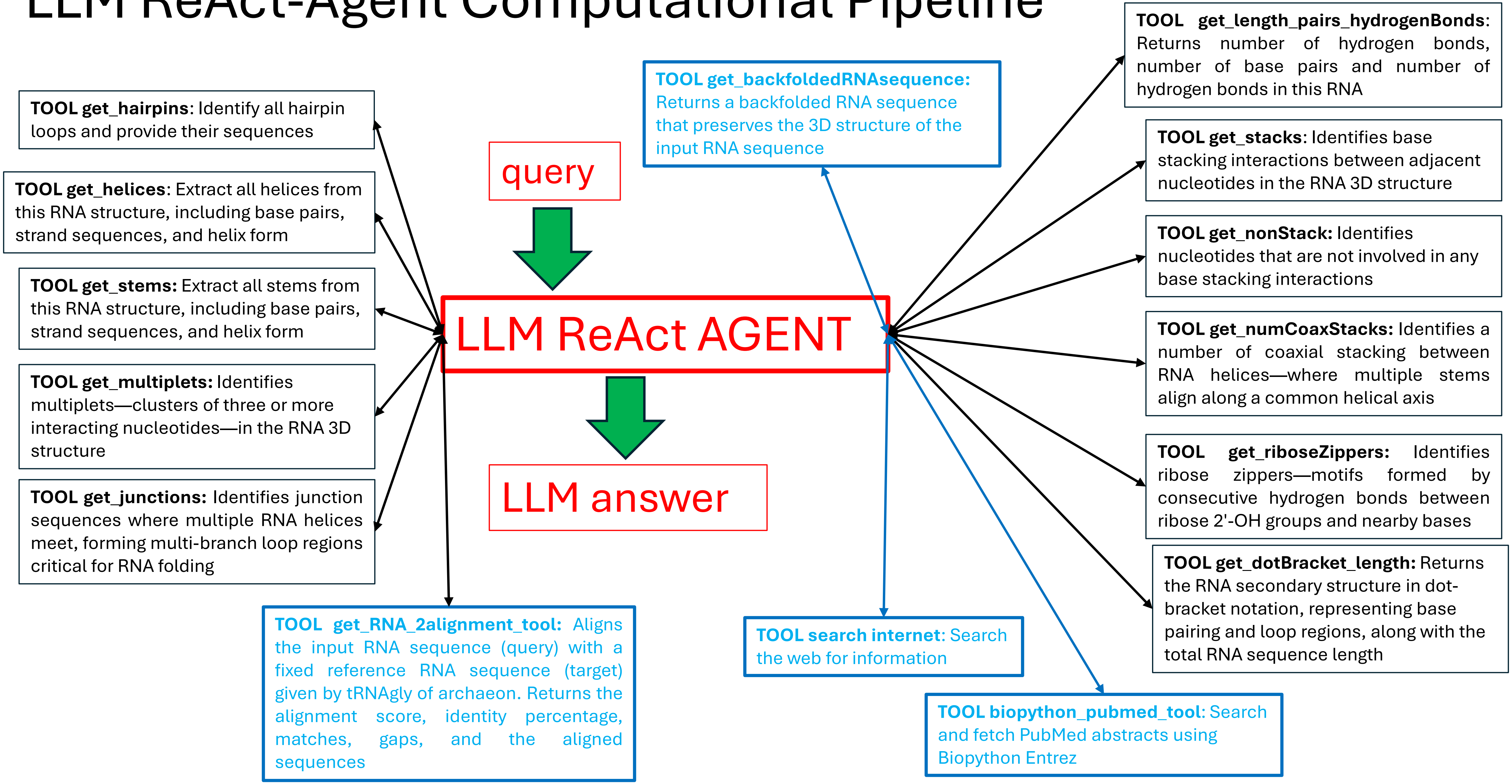
- **Tools description**
  - RNA alignment Tool
  - RNA 3D structure parameters Tool
  - The structure-guided feedback Tool

- **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

- **Example use case**
  - Single point mutation in tRNA associated with MELAS syndrome

- **Conclusions**
- **What’s next ?**

LLM ReAct-Agent Computational Pipeline



query

LLM ReAct AGENT

LLM answer

TOOL get_hairpins: Identify all hairpin loops and provide their sequences

TOOL get_helices: Extract all helices from this RNA structure, including base pairs, strand sequences, and helix form

TOOL get_stems: Extract all stems from this RNA structure, including base pairs, strand sequences, and helix form

TOOL get_multiplets: Identifies multiplets—clusters of three or more interacting nucleotides—in the RNA 3D structure

TOOL get_junctions: Identifies junction sequences where multiple RNA helices meet, forming multi-branch loop regions critical for RNA folding

TOOL get_RNA_2alignment_tool: Aligns the input RNA sequence (query) with a fixed reference RNA sequence (target) given by tRNAgly of archaeon. Returns the alignment score, identity percentage, matches, gaps, and the aligned sequences

TOOL get_backfoldedRNAsequence: Returns a backfolded RNA sequence that preserves the 3D structure of the input RNA sequence

TOOL search internet: Search the web for information

TOOL biopython_pubmed_tool: Search and fetch PubMed abstracts using Biopython Entrez

TOOL get_length_pairs_hydrogenBonds: Returns number of hydrogen bonds, number of base pairs and number of hydrogen bonds in this RNA

TOOL get_stacks: Identifies base stacking interactions between adjacent nucleotides in the RNA 3D structure

TOOL get_nonStack: Identifies nucleotides that are not involved in any base stacking interactions

TOOL get_numCoaxStacks: Identifies a number of coaxial stacking between RNA helices—where multiple stems align along a common helical axis

TOOL get_riboseZippers: Identifies ribose zippers—motifs formed by consecutive hydrogen bonds between ribose 2'-OH groups and nearby bases

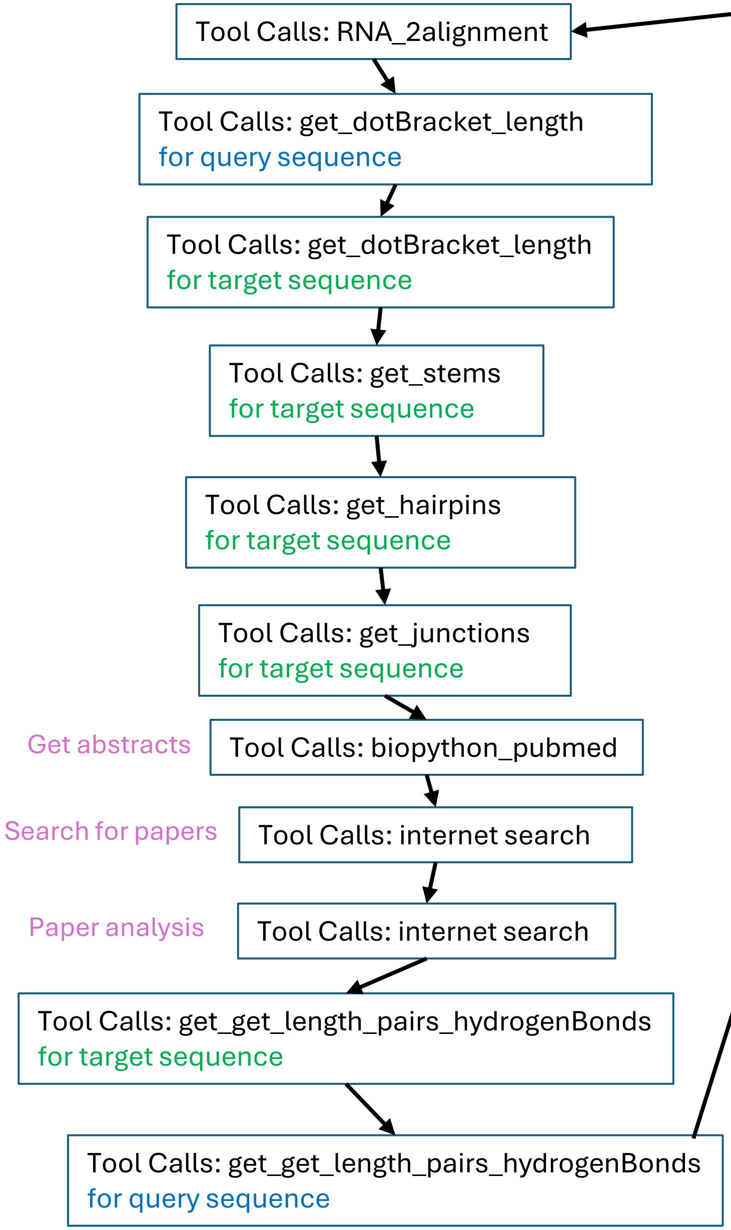
TOOL get_dotBracket_length: Returns the RNA secondary structure in dot-bracket notation, representing base pairing and loop regions, along with the total RNA sequence length

LLM ReAct Agent: Comparison of Two RNA Sequences

Provide a detailed comparison of the following two RNA molecules: query AGAGAGAGAGAGAGAGAGAG 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: AGAGAGAGAGAGAGAGAGAG (20 nucleotides)
Target: GCACCGAUAGUCUAGUGGUAGGACUUAUCCCUGCCUAGGAUAGAGCCCGGGUUCAAAUCCCGGUCGGUGCACCA (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 (AGAGAGAGAGAGAGAGAGAG):
- **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

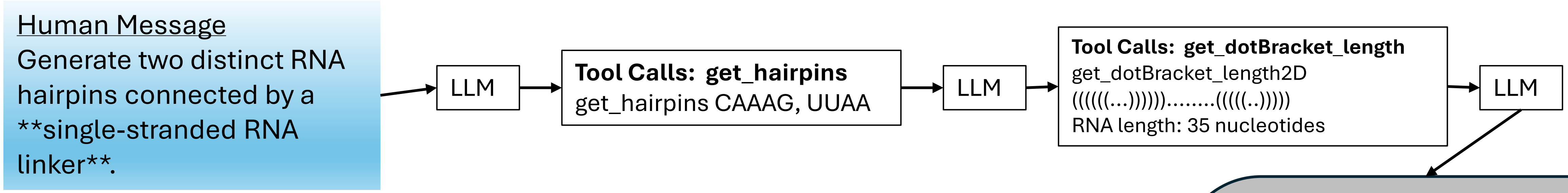
## 4. Similarities and Differences

### 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**.

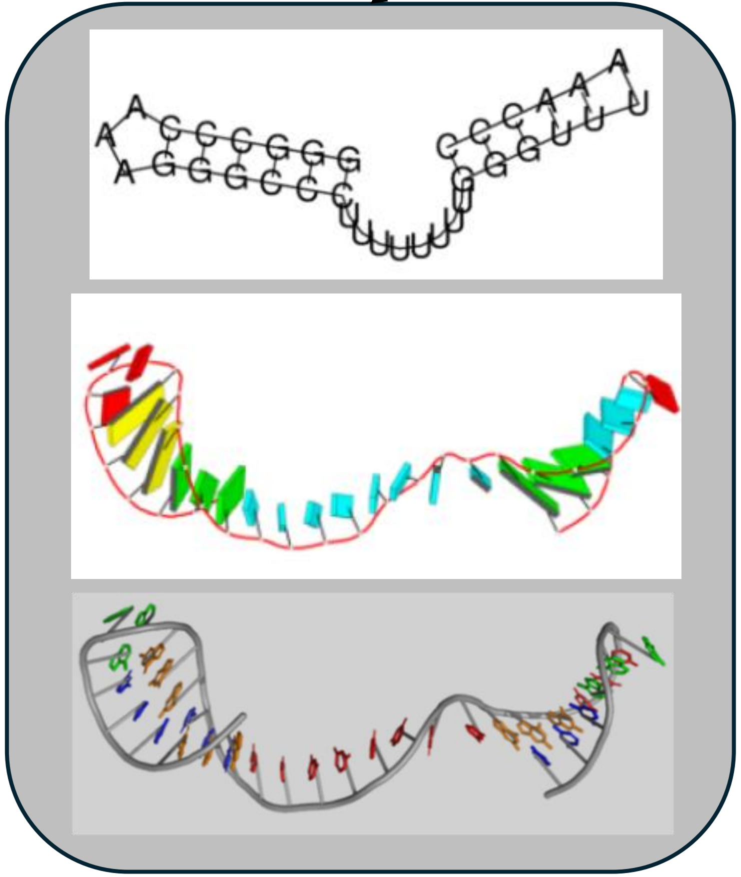
LLM

Tool Calls: get_hairpins
get_hairpins CAAAG, UUAA

LLM

Tool Calls: get_dotBracket_length
get_dotBracket_length2D
((((((...))))))........(((((..)))))
RNA length: 35 nucleotides

LLM



```
================================[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```\nGGGCCCAAAGGGCCCUUUUUUUUGGGUUUAAACCC\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\nNow let me analyze this structure to verify the hairpins:", 'type': 'text'}, {'id': 'toolu_01Q645amkxCPhDjA4qxeQbVH', 'input': {'__arg1': 'GGGCCCAAAGGGCCCUUUUUUUUGGGUUUAAACCC'}, 'name': 'get_hairpins', 'type': 'tool_use'}]Tool Calls: get_hairpins (toolu_01Q645amkxCPhDjA4qxeQbVH) Call ID: toolu_01Q645amkxCPhDjA4qxeQbVH Args: __arg1: GGGCCCAAAGGGCCCUUUUUUUUGGGUUUAAACCCc:\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': 'GGGCCCAAAGGGCCCUUUUUUUUGGGUUUAAACCC'}, 'name': 'get_dotBracket_length', 'type': 'tool_use'}]Tool Calls: get_dotBracket_length (toolu_0149fG8yTvHaMatfqfVq44nJ) Call ID: toolu_0149fG8yTvHaMatfqfVq44nJ Args: __arg1: GGGCCCAAAGGGCCCUUUUUUUUGGGUUUAAACCC================================[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:** `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 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:
AAAUUUAAAGGGCCCUUUUGGGCCCAAAUUUGGGCCCUUUUGGGCCC

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

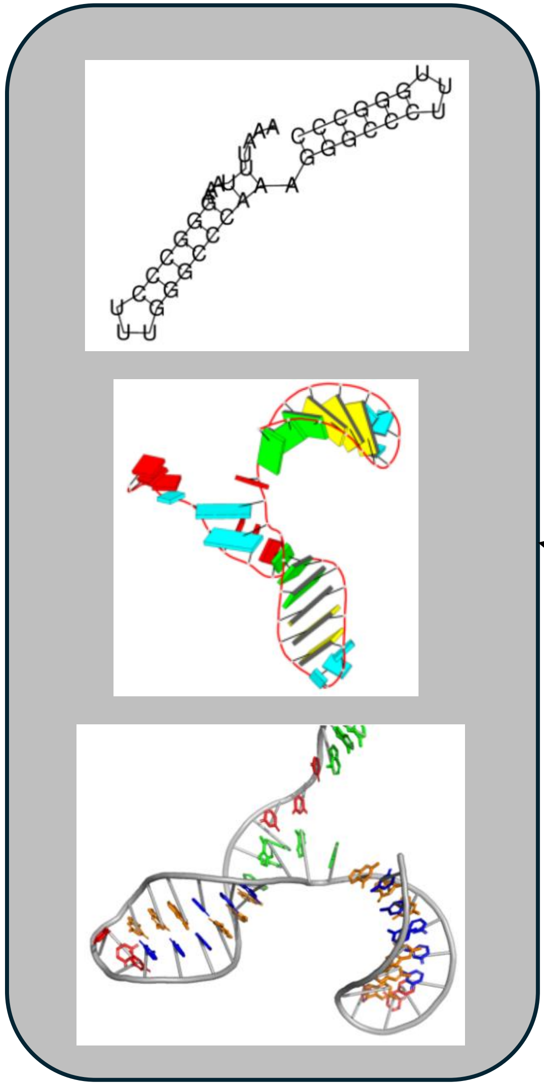
Let me create a shorter sequence that meets your ~40 nucleotide
AAAUUUAAAGGGCCCUUUGGGCCCAAAGGGCCCUUUGGGCCC

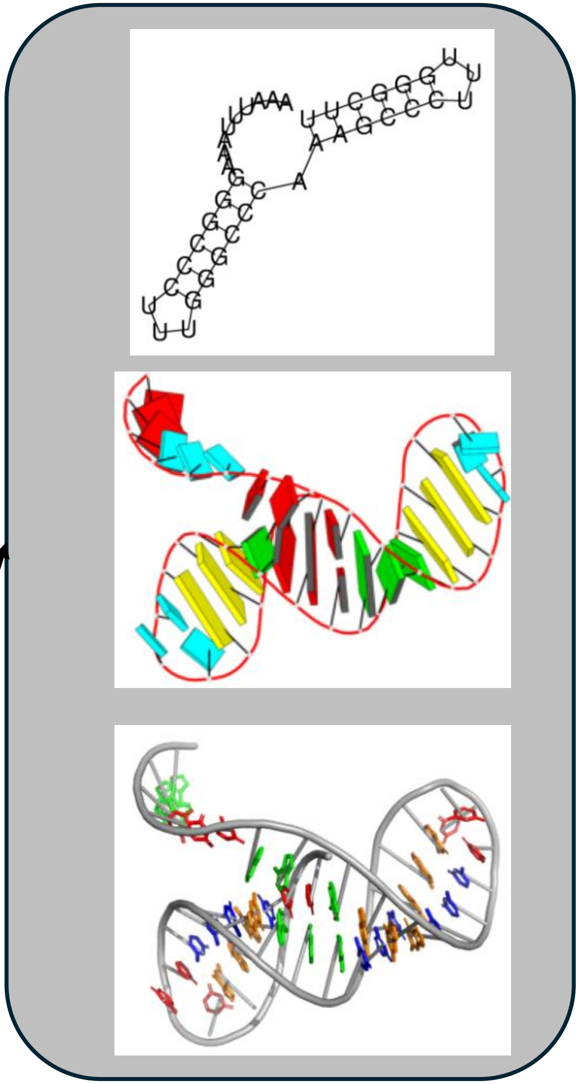
What is ‘overlapping’ in LLM internal raisoning ??

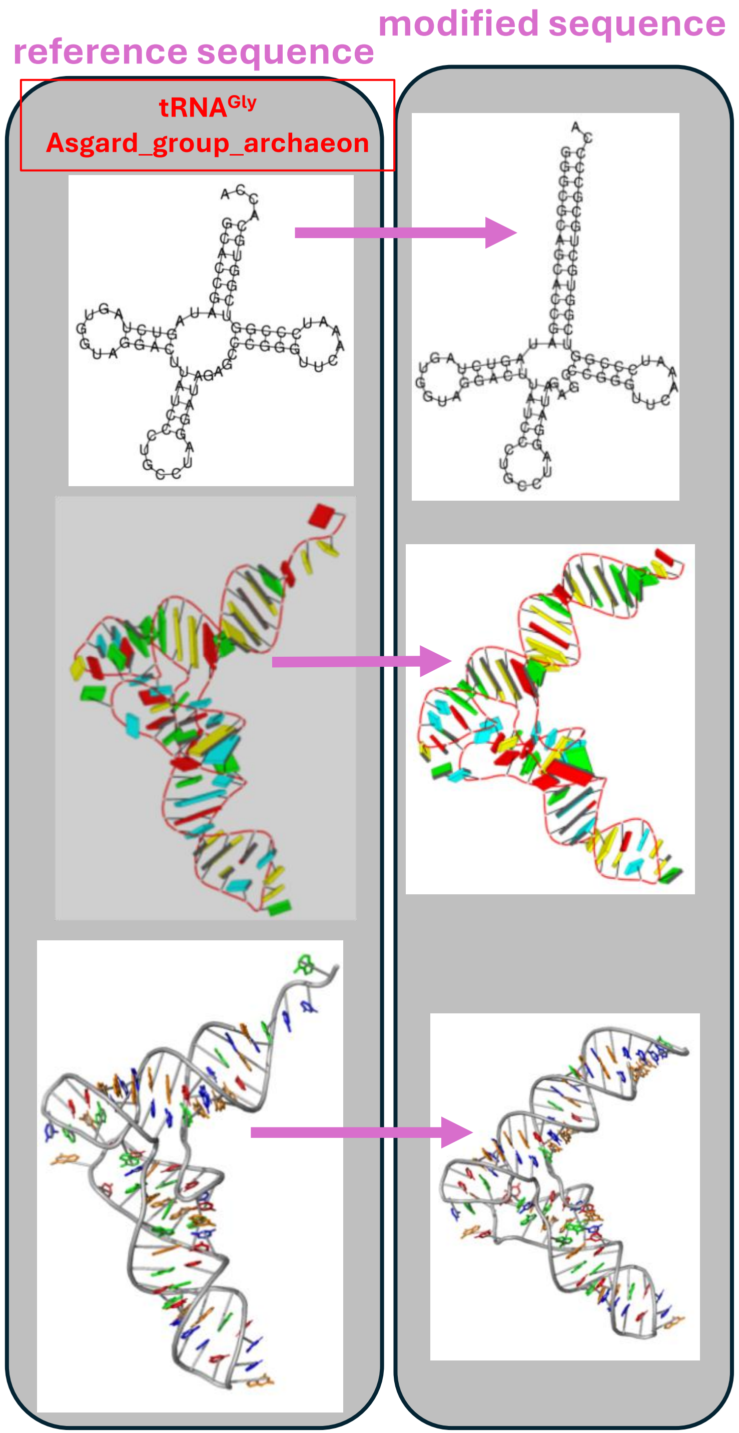
'This structure shows some overlapping, let me try a cleaner design:',
'AAAUUUAAAGGGCCCUUUGGGCCCAAAGCCCUUUGGGCUU’
Tool Calls: get_dotBracket_length

'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

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...

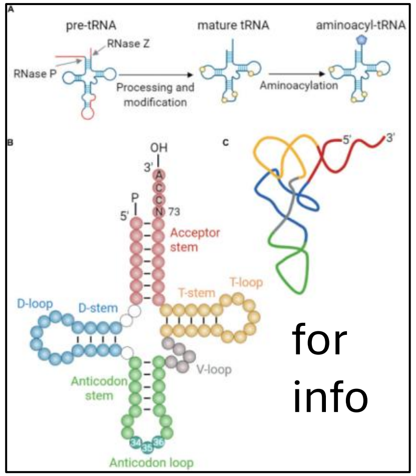


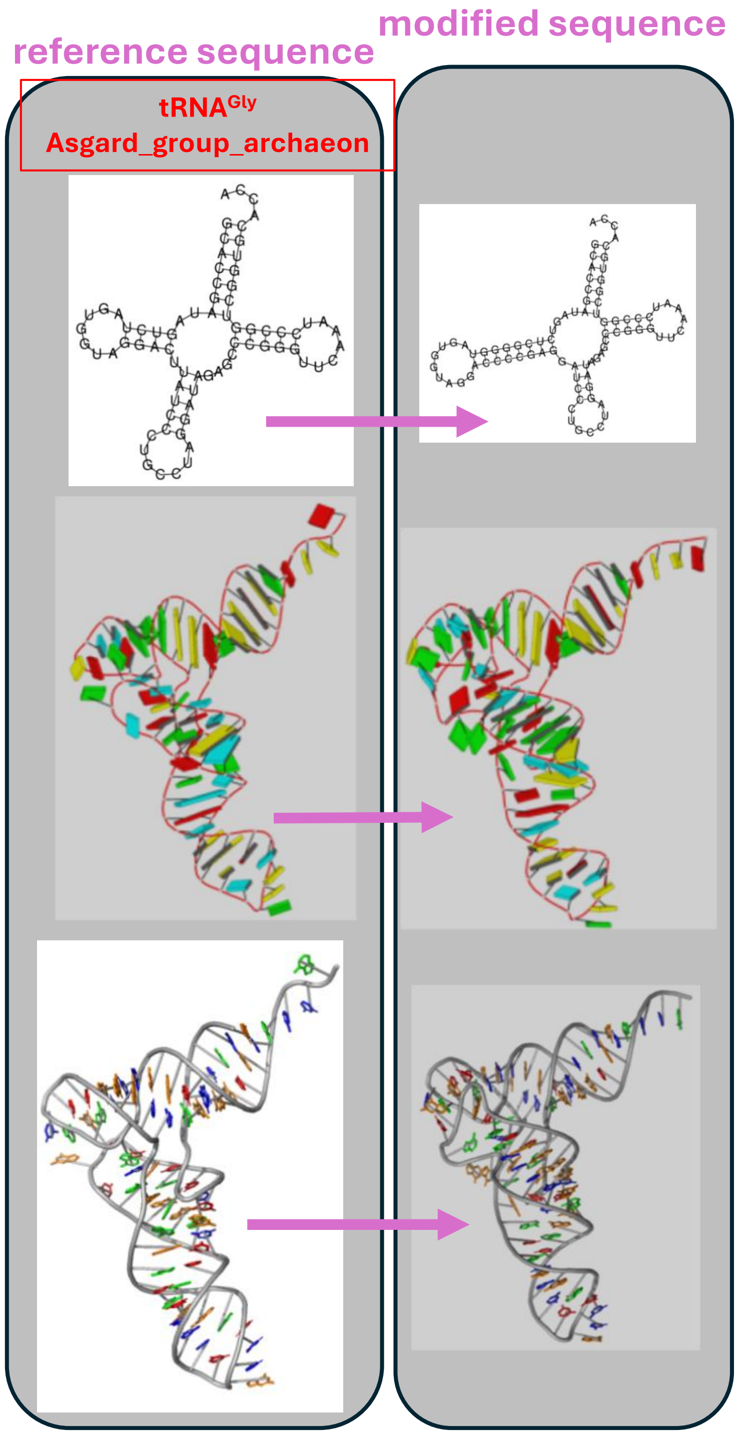
Figure adapted from Frontiers in Molecular Biosciences (Irem Avcilar-Kucukgoze et al., 2020), CC BY 4.0. DOI: 10.3389/fmolb.2020.610617

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



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."

It does not work without backfolded 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.

OUTLINE

- **Dataset**
  - What is RNA molecule
  - RNAcental database

- **Tools description**
  - RNA alignment Tool
  - RNA 3D structure parameters Tool
  - The structure-guided feedback Tool

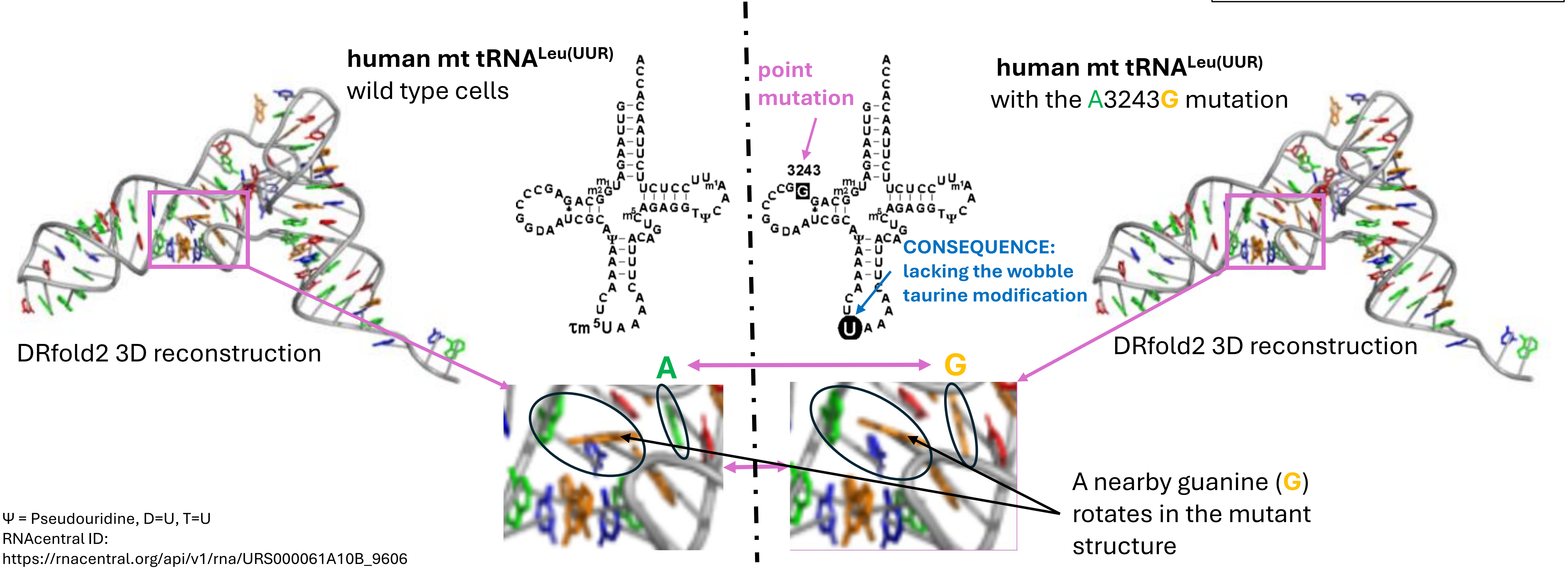
- **LLM ReAct-Agent Computational Pipeline**
- **Results for LLM ReAct-Agent**
  - Comparison of two RNAs
  - Generate two connected RNA hairpins without RNA template
  - Generate two connected RNA hairpins from a given RNA template
  - Targeted modification of a specific region of an RNA molecule

- **Example use case**
  - Single point mutation in tRNA associated with MELAS syndrome

- **Conclusions**
- **What's next ?**

A Single Mutation, Systemic Consequences: Single Point Mutation in tRNA Associated with MELAS Syndrome

point mutations (A3243G or T3271C) → mt tRNALeu(UUR) molecule **lacking the taurine modification** → mt tRNALeu(UUR) **lacking the wobble taurine modification** has a codon-specific **deficit in translational ability**. → **M**itochondrial myopathy, **E**ncephalopathy, **L**actic **A**cidosis, and **S**troke-like episodes (**MELAS**)



human mt $tRNA^{Leu(UUR)}$
wild type cells

point mutation

human mt $tRNA^{Leu(UUR)}$
with the A3243G mutation

CONSEQUENCE: lacking the wobble taurine modification

DRfold2 3D reconstruction

DRfold2 3D reconstruction

A nearby guanine (G) rotates in the mutant structure

Ψ = Pseudouridine, D=U, T=U
RNAcentral ID:
https://rnacentral.org/api/v1/rna/URS000061A10B_9606

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."
```

First LLM query on the MELAS syndrome topic

human mt tRNA$^{\text{Leu(UUR)}}$
with the A3243G mutation

human mt tRNA$^{\text{Leu(UUR)}}$
wild type cells

Let me start by examining their structures and then align them to identify the specific mutation

Tool Calls: get_dotBracket_length (for query RNA)
Tool Calls: get_dotBracket_length (for target RNA)

Tool Calls: RNA_2alignment

Now let me analyze the structural elements of both RNA molecules to understand how the mutation affects their structure:

Tool Calls: get_stems (for query RNA)
Tool Calls: get_stems (for target RNA)
Tool Calls: get_hairpins (for query RNA)
Tool Calls: get_hairpins (for target RNA)
Tool Calls: get_length_pairs_hydrogenBonds (for query RNA)
Tool Calls: get_length_pairs_hydrogenBonds (for target RNA)

LLM Analysis Report — see next slide

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**GCCCGGUAAUCGCAUAAAACUUAAAACUUUACAGUCAGAGGUUCAAUUCCUCUUCUUAACACCA
- **Target sequence:**
GUUAAGAUGGCAG**A**GCCCGGUAAUCGCAUAAAACUUAAAACUUUACAGUCAGAGGUUCAAUUCCUCUUCUUAACACCA

## **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
- **Hydrogen 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)

Second LLM query on the MELAS syndrome topic

```
query = (
  f"The query sequence is:\n{tRNA_A2G}\n\n"
  f"The target (healthy) sequence is:\n{tRNA_healthy}\n\n"
  "Explain why the mutation in the query sequence leads to the absence of the wobble taurine modification.\n"
  "Support your explanation by analyzing how the relevant RNA properties differ between the target and query.\n\n"
  "Present the output in clearly labeled sections.")
```

I'll analyze the mutation between the query and target sequences to explain how it leads to the absence of the wobble taurine modification. Let me start by aligning the sequences and then examine their structural properties.
Tool Calls: RNA_2alignment

Now let me analyze the structural properties of both sequences:
Tool Calls: get_dotBracket_length (for target sequence)
Tool Calls: get_dotBracket_length (query sequence)

Let me examine the stems and helices in both sequences:
Tool Calls: get_stems (for target sequence)
Tool Calls: get_stems (for query sequence)
Tool Calls: get_hairpins (for target sequence)
Tool Calls: get_hairpins (for query sequence)
Tool Calls: get_stacks (for target sequence)
Tool Calls: get_stacks (for query sequence)
Tool Calls: get_multiplets (for target sequence)
Tool Calls: get_multiplets (for query sequence)

Tool Calls: biopython_pubmed
(__arg1: wobble taurine modification tRNA mitochondrial mutation)

1. 1. Nucleic Acids Res. 2021 Mar 18;49(5):2816-2834. doi: 10.1093/nar/gkab104.
2. The human tRNA taurine modification enzyme GTPBP3 is an active GTPase linked to mitochondrial diseases.
3. Peng GX(1)(2), Zhang Y(1), Wang QQ(2), Li QR(3), Xu H(4), Wang ED(1)(2), Zhou XL(1).

3. Hum Mol Genet. 2015 Jan 1;24(1):167-84. doi: 10.1093/hmg/ddu427. Epub 2014 Aug 22.
16. The ROS-sensitive microRNA-9/9* controls the expression of mitochondrial tRNA-modifying enzymes and is involved in the molecular mechanism of MELAS syndrome.
17. Meseguer S(1), Martínez-Zamora A(2), García-Arumí E(3), Andreu AL(3), Armengod ME(4).

2. Brain Nerve. 2017 Feb;69(2):111-117. doi: 10.11477/mf.1416200650.
9. [MELAS: Mitochondrial Encephalomyopathy, Lactic Acidosis and Stroke-Like Episodes].
10. [Article in Japanese]
11. Murakami H(1), Ono K.

4. Adv Exp Med Biol. 2013;776:153-65. doi: 10.1007/978-1-4614-6093-0_16.
23. Taurine deficiency and MELAS are closely related syndromes.
24. Schaffer SW(1), Jong CJ, Warner D, Ito T, Azuma J.

Now let me search for more specific information about the A to G mutation at position 3243:
Tool Calls: search
(__arg1: tRNA leucine A3243G mutation wobble position taurine modification MELAS)

Although at least 30 distinct mitochondrial DNA mutations have been associated with MELAS , about 80% of patients have the m.3243A > G tRNALeu(UUR ... de S Maharjan · 2024 · Cité 2 fois — Mitochondrial haplotype mutation alleviates respiratory defect of MELAS by restoring taurine modification in tRNA with 3243A > G mutation . Nucleic Acids Res ... Feature papers are submitted upon individual invitation or recommendation by the scientific editors and must receive positive feedback from the ... Not surprisingly, mutations in enzymes that modify cytoplasmic and mitochondrial tRNAs are also linked to a diverse range of clinical phenotypes. ... spectrometry data show that, while m1G9, m2G10, and m1A58 are all fully modified in the MELAS tRNA , the taurine modification at the normally wobble ...

LLM Analysis Report — see next slide

LLM (ReAct agent) Analysis Report (see previous slide)

# Analysis Report: Mutation Impact on Wobble Taurine Modification

## 1. SEQUENCE ALIGNMENT ANALYSIS
The alignment analysis reveals a **single nucleotide mutation** between the query and target sequences:
- **Position 14**: A → G mutation (A3243G in mitochondrial nomenclature)
- **Sequence context**: CAGAGCC (target) → CAGGGCC (query)
- **Overall similarity**: 98.72% identity (77/78 matches)

## 2. STRUCTURAL COMPARISON
### Secondary Structure Analysis
Both sequences maintain the same overall secondary structure:
- **Dot-bracket notation**: Both show identical folding patterns with the same stem-loop architecture
- **RNA length**: 78 nucleotides for both sequences
- **Base pairing**: The mutation does not disrupt the major stem regions
### Hairpin Loop Differences
The mutation affects the first hairpin loop:
- **Target (healthy)**: CAGAGCCCGGUAAUCG
- **Query (mutated)**: CAGGGCCCGGUAAUCG
### Stacking Interaction Changes
The mutation alters local stacking interactions:
- **Target**: Contains UC and UA stacking patterns
- **Query**: Shows altered stacking with different GG-rich regions
### Multiplet Formation
Significant changes in three-nucleotide interactions:
- **Target**: UAA, GAC, GGUAA multiplets
- **Query**: UGA, GAC, GCU multiplets

## 3. MOLECULAR MECHANISM EXPLANATION
### Why the A→G Mutation Prevents Taurine Modification
1. **Wobble Position Impact**: The A3243G mutation occurs in the mitochondrial tRNA^Leu(UUR) and affects the structural context required for taurine modification at the wobble position (position 34).
2. **Enzyme Recognition Disruption**:
  - The mutation alters the local RNA structure and base stacking patterns
  - GTPBP3 and MTO1 enzymes, which cooperatively catalyze 5-taurinomethyluridine (τm5U) biosynthesis, require specific structural motifs for recognition
  - The changed multiplet patterns (UAA→UGA) disrupt the enzyme binding interface
3. **Structural Constraint Changes**:
  - The hairpin loop alteration (CAGAGCC→CAGGGCC) changes the local electrostatic environment
  - Modified stacking interactions affect the accessibility of the wobble position
  - The GG-rich region in the mutant may create stronger internal base pairing that prevents enzyme access

## 4. FUNCTIONAL CONSEQUENCES
### Translation Defects
- **Codon Recognition**: The absence of τm5U modification specifically affects UUA codon decoding
- **Aminoacylation**: Reduced tRNA^Leu(UUR) aminoacylation capacity
- **Mitochondrial Protein Synthesis**: Decreased translation of mitochondria-encoded proteins
### Disease Pathogenesis
- **MELAS Syndrome**: This mutation is found in ~80% of MELAS patients
- **Oxidative Stress**: Increased ROS production due to respiratory chain deficiency
- **Energy Metabolism**: Impaired ATP synthesis leading to lactic acidosis

## 5. CONCLUSION
The A→G mutation at position 14 (corresponding to m.3243A>G) prevents wobble taurine modification by:
1. **Disrupting enzyme recognition sites** through altered multiplet formation patterns
2. **Changing local RNA structure** in the critical hairpin loop region
3. **Modifying base stacking interactions** that are essential for GTPBP3/MTO1 enzyme binding
4. **Creating unfavorable structural constraints** that prevent access to the wobble position
This structural analysis explains why this single nucleotide change has such profound consequences for mitochondrial tRNA function and ultimately leads to the MELAS phenotype through defective protein synthesis.

OUTLINE

- **Dataset**
  - What is RNA molecule
  - RNAcental database

- **Tools description**
  - RNA alignment Tool
  - RNA 3D structure parameters Tool
  - The structure-guided feedback Tool

- **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

- **Example use case**
  - Single point mutation in tRNA associated with MELAS syndrome

- **Conclusions**
- **What's next ?**

CONCLUSIONS

- Although an LLM alone **cannot reliably count** nucleotides without external tools, a simple LLM agent is already capable of effectively answering questions related to non-coding RNA sequences.

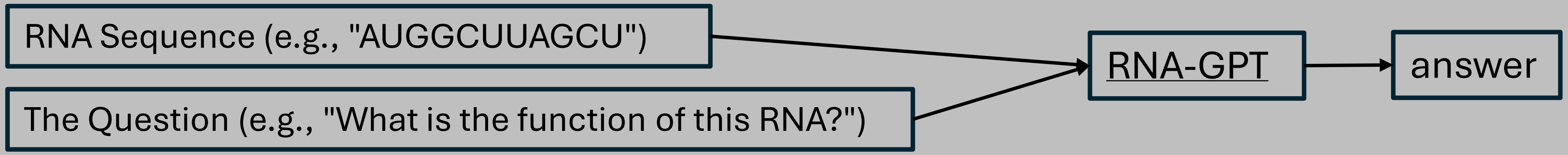
- LLM REAct agent can be used not only for de novo ncRNA sequence **generation** but also for **modifying** existing sequences, providing a potential advantage over conditional diffusion models in guided ncRNA design tasks.

- 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.

1. What’s next ?

1. **Search for mistakes in the answers and correct the prompt accordingly**
2. **Comprehensive literature research**
3. **More tools, leading toward a multi-agent pipeline**

4. Use RNA-GPT to construct a new tool based on answering questions about given RNA sequence.


RNA Sequence (e.g., "AUGGCUUAGCU")
The Question (e.g., "What is the function of this RNA?")
RNA-GPT
answer

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**

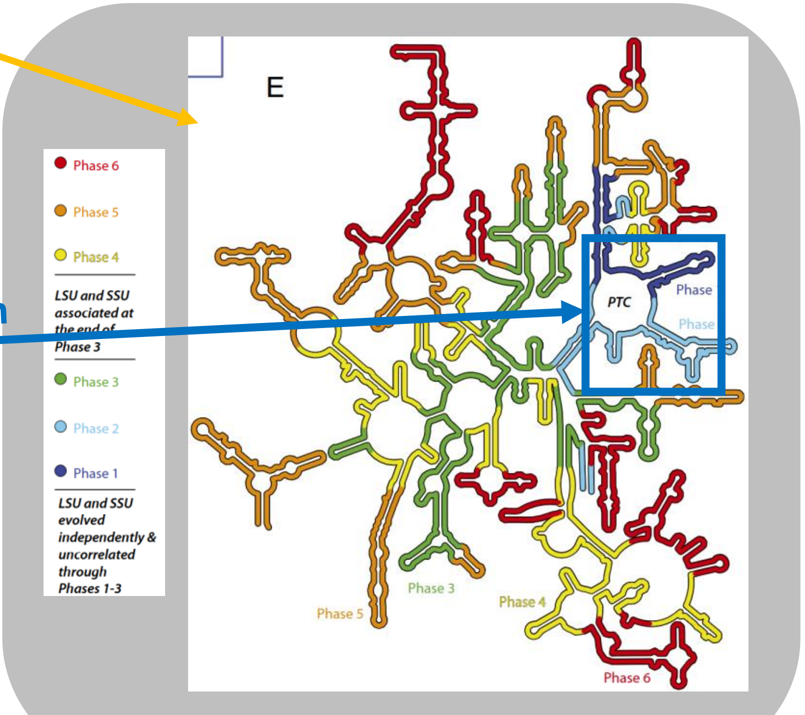
5. Fundamental Problems (see next slide) ?

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

2. What is the oldest know ncRNA ?

rRNA evolution: Age Variability Across Different Regions

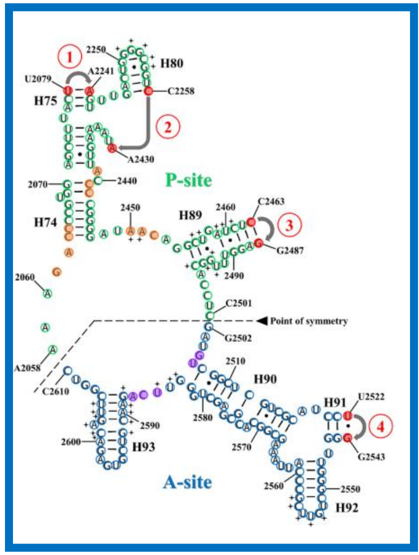
large ribosomal subunit (LSU) rRNA


E
Phase 6
Phase 5
Phase 4
LSU and SSU associated at the end of Phase 3
Phase 3
Phase 2
Phase 1
LSU and SSU evolved independently & uncorrelated through Phases 1-3
PTC
Phase
Phase
Phase 5
Phase 3
Phase 4
Phase 6

LSU of tRNA (E. coli) :

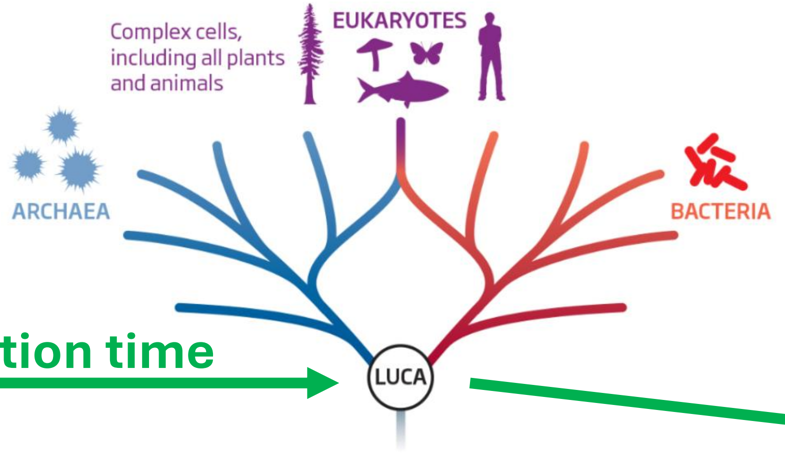
Blue part is the oldest one

the oldest functional part, conserved throughout evolution



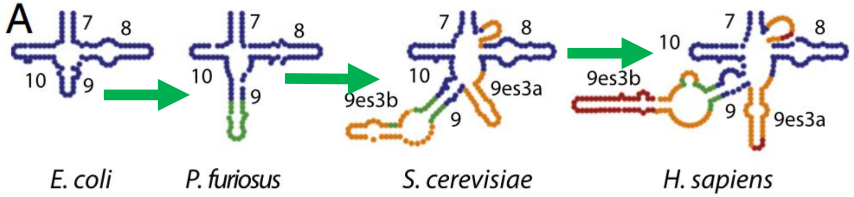
The pseudosymmetrical region of rRNA (**SymR**; *Agmon et al., 2005*), derived from the LSU secondary structure of Thermus thermophilus (*Petrov et al., 2013*).
(*Madhan R. Tirumalai et al., 2021*)

evolution time



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)

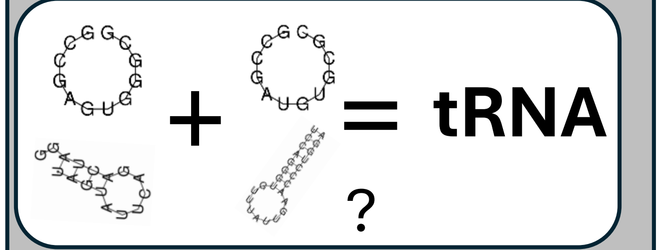
Peptidyl Transferase Center (PTC) is the oldest part of ribosomes and contains no proteins

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. Peptidyl Transferase Center (PTC) Sequences

the idea

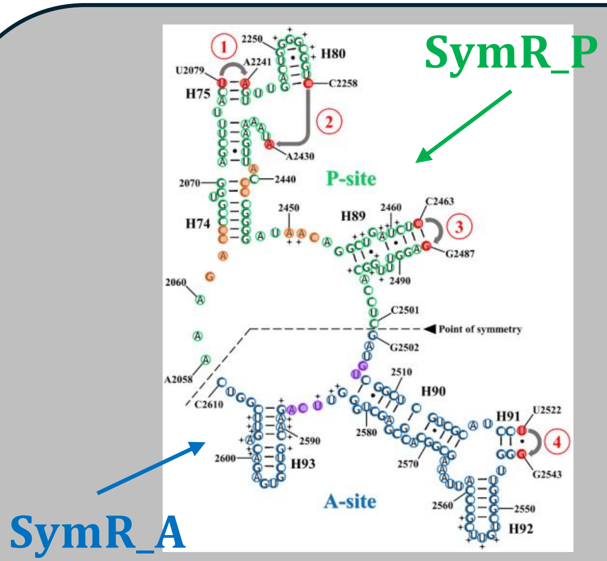
the dimerization of two similar RNA structures



"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

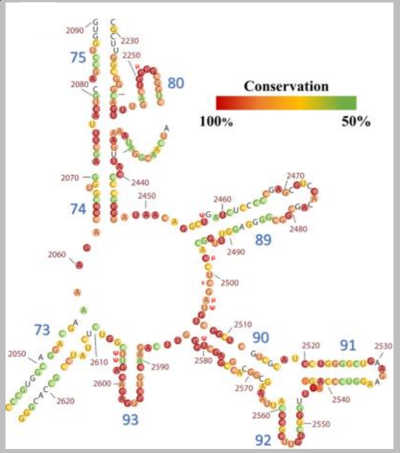


SymR_PA = SymR_P + SymR_A

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



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)