Lab 1: Modeling a Library for the Manipulation of Ribonucleic Acids (RNAs)

The goal of this series of labs is to build a library that allows easy manipulation and study of RNA sequences. This course is around object-oriented programming principles. Therefore, besides complying with the use case specificities, particular attention will be paid to the quality of the proposed model in terms of *extensibility*, i.e., the ability of a system to be effortlessly expanded or adjusted with new features without needing major alterations to the current codebase. Indeed, you have to conceive your models while keeping in mind that the following labs will impact your current modeling. In other words, change is inevitable, but you have to minimize the amount of it.

Instructions

For this first lab, we ask you to provide:

1. a UML class diagram
2. an object diagram
3. a Python implementation
4. Python code examples to show how the library can be practically used by the end-user. The RNA sequences provided as examples in the following sections can be used to test your implementation.

All the work will be done via GitHub. The repositories will be private. Do not share your models nor implementation with your classmates.
Access your personal repository and clone it to your machine. Make a first change to the main `readme.md` file: replace `Firstname [Lastname] (@github_username)` with your information. The last name is optional.

Please follow this structure for each lab:
`lab1/`
  `src/` ← here goes your Python implementation
  `model/` ← here goes your UML class and object diagrams
  `readme.md` ← put here the diagrams and add some comments about your design choices

To draw your UML class diagrams, we suggest you use draw.io.
Please, put your class (and object) diagram in the folder `/lab1/model/` with a naming convention like `class-diagram.png` (and `object-diagram.png`).

1. RNA sequences and their spatial conformations

In this first part, we model RNA sequences and their spatial conformations (or 3-dimensional structures).

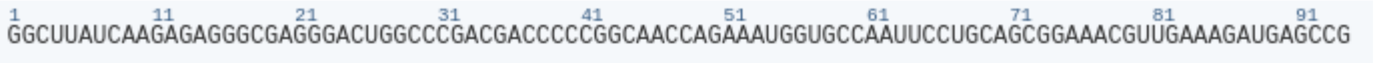
Introduction into RNA

RNA is a nucleic acid akin to DNA, yet it has significant structural and functional distinctions. It consists of a single strand of nucleotides, each made up of a ribose sugar, a phosphate group, and one of four nitrogenous bases: adenine (A), uracil (U), cytosine (C), and guanine (G). In contrast to DNA, RNA features uracil (U) in place of thymine (T) and typically exists as a single strand. Figure 1 illustrates an example of a 3D RNA structure.

Figure 1: 3D structure of SAM-I riboswitch with the Actinomyces-1 k-turn (source:https://www.rcsb.org/3d-view/7EAF/1)

Residue

A residue is a single building block of RNA. When RNA forms a strand, individual nucleotides are linked together by phosphodiester bonds. Once incorporated into the RNA chain, each nucleotide is often referred to as a nucleotide residue[1]. Figure 2 illustrates the sequence of nucleotides corresponding to the RNA 3D structure depicted above.



```
1         11        21        31        41        51        61        71        81        91
GGCUUAUCAAGAGAGGGCGAGGGACUGGCCCGACGACCCCCGGCAACCAGAAAUGGUGCCAAUUCCUGCAGCGGAAACGUUGAAAGAUGAGCCG
```

Figure 2: Sequence of nucleotides of the RNA reference sequence 7EAF (source:https://www.rcsb.org/3d-view/7EAF/1)

Atom

Atoms represent the tiniest unit in an RNA molecule. Each nucleotide contains several atoms: a phosphate group (P, O atoms), ribose sugar (C, O atoms), and a nitrogenous base (A, U, G, C, which includes C, N, O atoms[2]).

From an RNA sequence to its 3D structure

To get the 3D structure of an RNA sequence, we need to know the coordinates of its residues, which in turn are determined by the 3D coordinates of the atoms that compose the individual residues. This 3D structure

[1] because part of the original nucleotide (such as a phosphate group) is lost in the bonding process.

[2] The atom depends on the type of the base: purines (A and G) and pyrimidines (U and C).

can be determined experimentally (e.g., X-ray crystallography or NMR spectroscopy). For example, the following are the atomic coordinates of a series of atoms (the first six atoms) that can be found in the RNA sequence 7EAF depicted in Figures 1 and 2.

```
[...]
OP3   -9.698   3.426 -31.854 O
P     -8.782   4.433 -32.440  P
OP1   -8.042   5.072 -31.321  O
OP2   -8.000   3.751 -33.502  O
O5'   -9.668   5.577 -33.115  O
C5'   -11.071   5.663 -32.885  C
[...]
```

You can notice the atom name followed by its atomic coordinates (x, y, z).

Chain

The level above a residue is the *chain*. Discontinuities in the sequence of nucleotides in an RNA sequence can occur. In these situations, the different continuous strands separated by these discontinuities form chains. A single RNA molecule can have one or multiple chains.

Model

Depending on the experimental technology used to determine the 3D structure of RNA sequences, multiple distinct structures can be obtained. A given structure is then referred to as a Model. In a structure derived from an X-ray crystallography experiment typically includes only one model, with some exceptions. NMR structures usually comprise many different models. These models represent a possible conformation of the RNA structure.

2. Families of RNA

In this second part, we turn to modeling families of RNA sequences. The study of families of RNA sequences is an essential topic including for better understanding the functions of RNAs. *Rfam* is one of the databases that reference RNA families identified in the living as of today. The current version of the Rfam database (version 15.0)[3] features no less than 4,178 RNA families.

Families

A RNA family (in the Rfam database) consists of evolutionarily related RNA sequences sharing common sequence similarity, secondary structure, and often function. Some of the families featured in the Rfam database include *rRNA* (Ribosomal RNA); *tRNA* (Transfer RNA), which is essential for protein translation; *miRNA* (MicroRNA), which are regulatory RNAs that inhibit gene expression; and Other *ncRNAs* (Non-coding RNAs). The RNA structure 7EAF described earlier belongs to the *SAM riboswitch* Rfam family. More information about this family can be found here: https://rfam.org/family/SAM.

[3] https://rfam.org/

Clans

RNA families can further be grouped into clans, which help resolve cases where closely related sequences belong to different functional families. For example, the SAM riboswitch Rfam family is a member of the clan CL00012, which contains the following three members: *SAM*, *SAM-I-IV*, and *SAM-IV*.

Species

Species refer to the organisms (bacteria, animals, plants, etc.) that contain RNA sequences belonging to a particular RNA family. Figure 3 illustrates the distribution of the SAM riboswitch Rfam family across species.

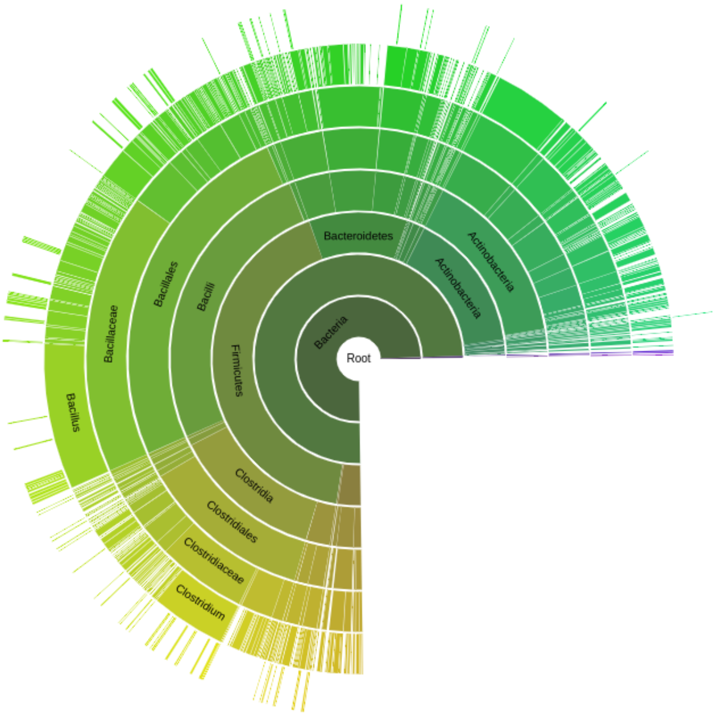


Figure 3: Distribution of the SAM riboswitch Rfam family across species (source:https://rfam.org/family/SAM#tabview=tab4)

Phylogenetic tree

A phylogenetic tree represents the evolutionary relationships between RNA sequences in a given RNA family. It is constructed using multiple sequence alignments and computational phylogenetics to illustrate how different RNA sequences are related through common ancestry. These trees help in understanding RNA structure conservation, functional similarities, and evolutionary divergence within RNA families. You can take a look at the phylogenetic tree of the SAM riboswitch Rfam family, which includes the RNA structure 7EAF discussed earlier, here: https://rfam.org/family/SAM#tabview=tab5. A much smaller phylogenetic tree is that of the *2dG-I* (RF01510) family and is depicted in Figure 4.

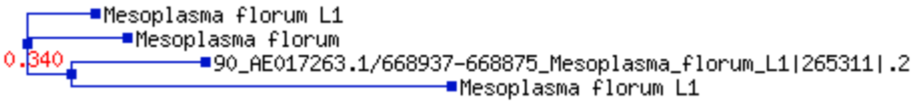


Figure 4: Phylogenetic tree of the 2dG-I family (source:https://rfam.org/family/RF01510#tabview=tab5)