Projet 1: structure secondaire des ARN

Les ARN sont des molécules simple brin dont la fonction est liée à plusieurs niveaux d'information: la séquence formée de l'enchainement des nucléotides A, C, G et U, la structure secondaire qui capture les appariements entre nucléotides, l'organisation de la structure dans l'espace, et enfin les interactions avec les autres molécules présentes dans la cellule. Parmi ces différents niveaux, la structure secondaire est déterminante, car elle constitue un échaffaudage pour le repliement final.

Les appariements de la structure secondaire sont de trois sortes:

- C avec G, impliquant 3 liaisons hydrogène
- A avec U, impliquant 2 liaisons hydrogène
- G avec U, impliquant 1 liaison hydrogène

Ces appariements ont également la propriété de ne pas se croiser, et de ne pas se former entre nucléotides dont la distance sur la séquence est inférieure à trois. Toute position est impliquée dans *au plus* une paire de bases. Voir l'exemple de la Figure 1.

L'objectif du projet est de créer une librairie qui permet de manipuler des structures secondaires d'ARN. Un exemple de librairie est représentée au format UML Figure 4.

Vous implémenterez cette librairie dans un script python, afin de répondre aux différentes questions du projet.

Partie 1: formats de description

Il existe deux formats principaux pour décrire les structures secondaires: le format CT (connect) et le format parenthésé.

Le format CT est un format tabulé constitué de trois colonnes, avec n lignes où n est la longueur de la séquence d'ARN. La première colonne indique la position du nucléotide dans la séquence (de 1 à n), la deuxième colonne indique la valeur du nucléotide à cette position (A, C, G ou U) et la troisième colonne indique avec quelle position le nucléotide est apparié. Dans le cas où le nucléotide n'est pas apparié (il est libre), on indique 0 par convention. Les appariements sont donc représentés par des couples de positions. Un exemple est donné en Figure 2.

Le format parenthésé ressemble au format Fasta, avec une ligne supplémentaire pour les appariements. La première ligne est donc une ligne d'entête commençant par un chevron >. La deuxième ligne contient la séquence nucléotidique. La troisième ligne indique les appariements deux à deux sous forme de parenthèses: la parenthèse ouvrante est le premier nucléotide de l'appariement et la parenthèse fermante correspondante le second nucléotide de l'appariement. Les positions libres sont indiquées avec des points. Un exemple est donné en Figure 3.

Question 1 *Parser une structure d'ARN au format Connect, vérifier sa validité et calculer le nombre total de liaisons hydrogènes de la structure.*

Question 2 *Parser une structure d'ARN au format parenthésé, vérifier sa validité et calculer le nombre total de liaisons hydrogènes de la structure.*

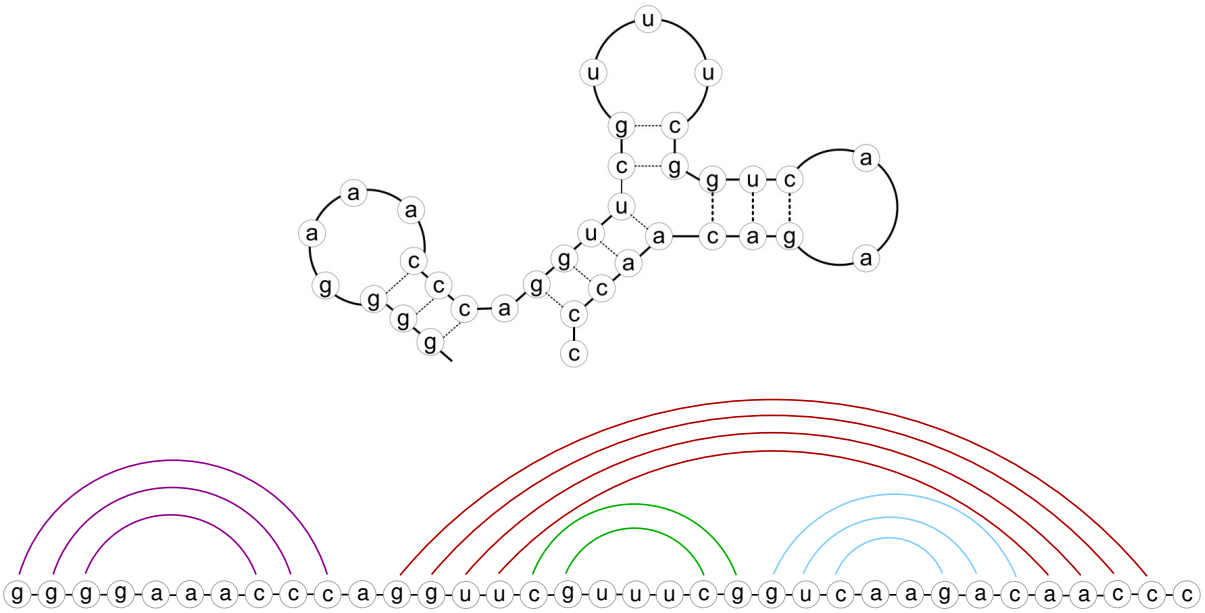

Figure 1: Exemple de structure secondaire

Partie 2: équivalence de deux structures secondaires

Une structure secondaire peut être décomposée en *tiges*: une tige est une succession d'appariements consécutifs. Par exemple, la structure de la Figure 1 contient quatre tiges, en violet, rouge, vert et bleu. On dit que deux structures sont *équivalentes* si elles contiennent la même organisation en tiges. tiges.

Question 3 *Ecrire une fonction qui détermine si deux structures sont équivalentes*

Partie 3: prédiction de structure

La prédiction de structure consiste à déterminer l'ensemble des appariements à partir de la séquence nucléoditique. Pour cela, on calcule la structure qui maximise le nombre de liaisons hydrogène, ce qui se fait par programmation dynamique.

Question 4 *Ecrire une fonction qui prend en entrée une séquence d'ARN et calcule le nombre de liaisons hydrogènes maximal.*

Question 5 *Compléter la question précédente en reconstruisant la structure optimale (ou une structure optimale s'il en existe plusieurs).*

Question 6 *(falcultative) Ecrire une fonction qui calcule toutes les structures optimales possibles.*

```
1  g 10
2  g 9
3  g 8
4  g 0
5  a 0
6  a 0
7  a 0
8  c 3
9  c 2
10 c 1
11 a 0
12 g 34
13 g 33
14 u 32
15 u 31
16 c 22
17 g 21
18 u 0
19 u 0
20 u 0
21 c 17
22 g 16
23 g 30
24 u 29
25 c 28
26 a 0
27 a 0
28 g 25
29 a 24
30 c 23
31 a 15
32 a 14
33 c 13
34 c 12
35 c 0
```

Figure 2: Représentation au format CT de la structure de la Figure 1

```
>sequence test
ggggaaacccagguucguuucggucaagacaaccc
(((....))).((((((...))(((..)))))))).
```

Figure 3: Représentation au format parenthésée de la structure de la Figure 1

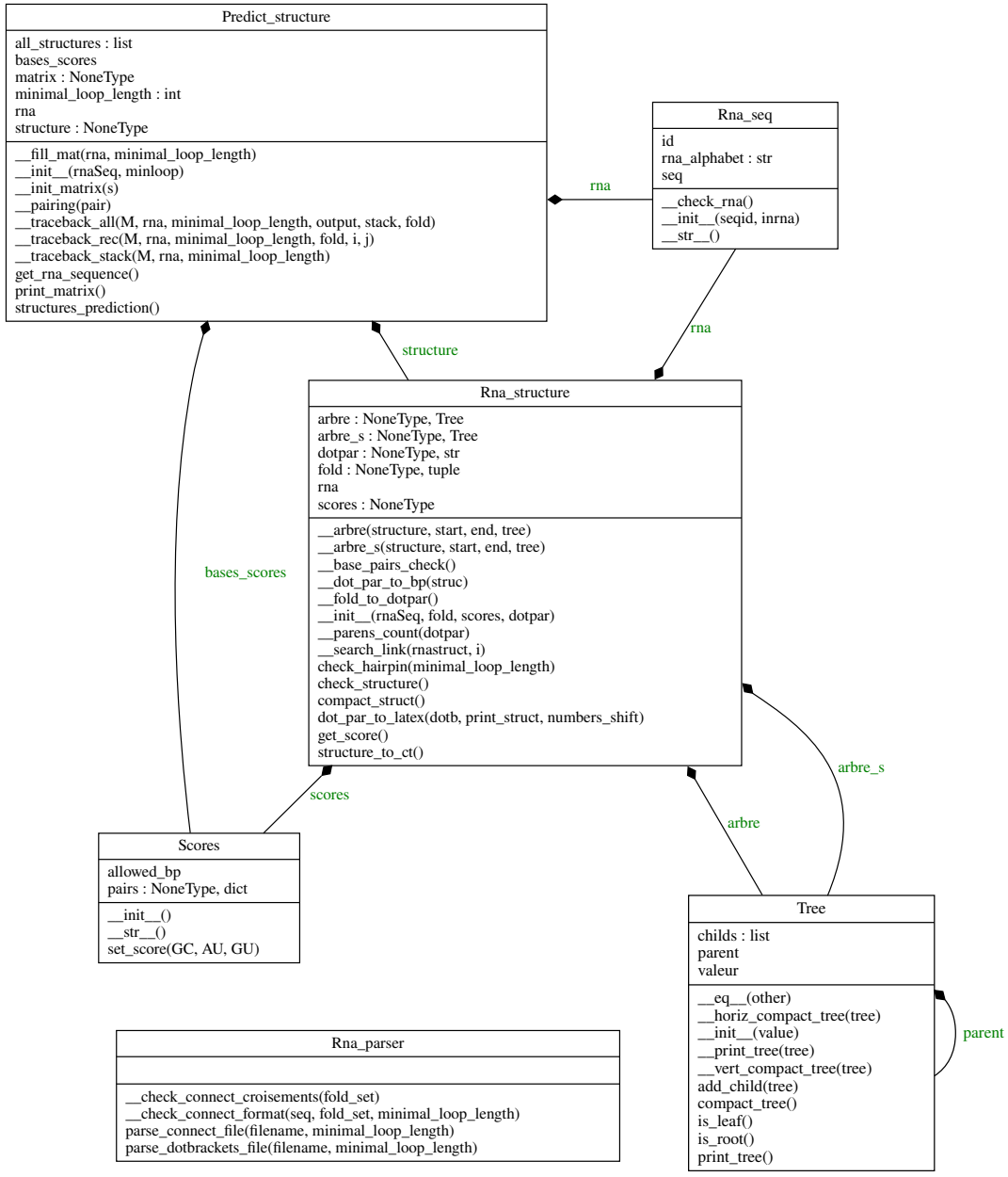

Figure 4: Exemple d'un diagramme UML pour la librairie