Université de Rouen Normandie
Master de Bioinformatique 1re année

Projet de Langages de scripts

2022 – 2023

Prédictions des cibles des miARN

Les micro-ARN (ou miARN) sont de courtes séquences d'acides ribonucléiques (ARN) simple-brin propres aux cellules eucaryotes. Ils possèdent en moyenne 22 nucléotides (en général de 21 à 24). Les miARN sont des régulateurs post-transcriptionnels. Ils sont dérivés d'un précurseur de miARN qui est une structure en épingle à cheveux. Les deux côtés de la tige peuvent produire un miARN mature. À l'origine, à ces miARN matures ont été assignés des noms basés sur l'abondance relative du miARN. Le miARN majoritairement exprimé avait pour nom, par exemple, miR-76a, et celui issu de la branche opposée avait pour nom miR-76a*. Au cours du temps de plus en plus de données de séquençage ayant été générées, ce format a changé de telle sorte que le niveau d'expression n'intervient plus dans les noms des miARN. À partir de la version 19 de miRBase, qui recense tous les miARN, toutes les séquences matures sont désormais nommées comme miR-76a-5p ou miR-76a-3p pour représenter le bras du précurseur dont ils sont issus. De manière à unifier l'ancienne et la nouvelle nomenclature chaque miARN s'est vu attribuer un identifiant MIMAT unique. Les miARN régulent l'expression des gènes en se fixant sur les ARN messagers de manière à empêcher leur traduction. Prédire la fixation d'un miARN sur un messager est une tâche compliquée. Il existe différentes méthodes et de nombreux outils dédiés à cette opération.

Le but de ce projet est d'écrire un programme Python qui lit trois fichiers de données :

— un fichier qui associe des noms de miARN à leur MIMAT (`aliases-rno.tsv`) ;

— un fichier qui associe des identifiants de transcripts à leur identifiant de gène (`ncbi-refseq-rno.tsv`) ;
— un fichier contenant des scores de prédictions de fixation entre un transcript et un miARN (`predictions-rno.tsv`).

Le programme devra alors créer un fichier par MIMAT (avec comme nom le MIMAT avec l'extension `.tsv`) qui contiendra les identifiants des gènes ciblés par le miARN avec le score de fixation associé. Ces lignes devront être triées dans l'ordre décroissant des scores.

Extrait de `aliases-rno.tsv` :

```
MIMAT0003125 rno-miR-1;rno-miR-1-3p;
MIMAT0003126 rno-miR-133b;rno-miR-133b-3p;
MIMAT0017085 rno-let-7a-1*;rno-let-7a-1-3p;
MIMAT0017086 rno-let-7a-2*;rno-let-7a-2-3p;
MIMAT0017286 rno-miR-466b-2*;rno-miR-466b-2-3p;
```

Extrait de `ncbi-refseq-rno.tsv` :

```
100360880 NM_001256509
108348108 NM_001329896
112400 NM_001271125
```

Extrait de `predictions-rno.tsv` :

```
rno-miR-466b-2-3p NM_001271125 54.2776586455
rno-miR-1-3p NM_001256509 64.0254
rno-miR-466b-2-3p NM_001256509 94.05034
```

Avec les valeurs précédentes on doit obtenir deux fichiers :

Dans `MIMAT0003125.tsv` :

```
100360880 64.0254
```

Dans `MIMAT0017286.tsv` :

```
100360880 94.05034
112400 54.2776586455
```

De plus, il est possible que dans le fichier de prédictions on trouve des miARN sans MIMAT ou des transcripts sans gène. Vous devez donc aussi produire le cas échéant deux fichiers nommés `missing-mimat.txt` et `missing-gene.txt` contenant respectivement les miARN sans MIMAT et les transcripts sans gène.

Vous devez déposer une archive compressée contenant votre programme sur Universitice pour le 20 décembre 2022.