Final Exam

Introduction

The exam is based on work we covered in the course. To do all the tasks listed below you can modify existing code or commands from the GitHub repositories we used in the course (for example, `shell_stuff` and `little_things`)

You can work in teams, but if you do this, you need to say who your team members are.

Note: based on the labwork, everyone is passing the course. This exam will help to determine your final grade.

The goal of the work is to filter a dataset of miRNA-mRNA target prediction data and examine the data in a biological context. The raw data is named `miRAW_raw_data.tsv.zip`

The dataset is a list of target prediction results generated by the miRAW software package. Each line contains one prediction result between a 3'UTR (identified by its `ENSG` ID) and a miRNA (identified by its `MIMATID`).

For example, line 2 shows the predicted target event between the **TSC1** gene, and miRNA **miR-4422**).

The provided code (`target_prediction_parsing.py`) shows you how to load the file and filter the predictions by *energy* and *probability*.

How to submit your work

The code you use to generate the results, together with a report written in Word, OpenOffice, or Markdown, should be pushed to a new **private GitHub** repository created on your own GitHub account. You can then invite me to access your repository. If you are working in a team, you can also invite team members to access the repository.

The tasks

There are seven tasks:

Task 1: Reviewing the data

Question 1.1: how many predictions are there in the raw data file?
Question 1.2: How many unique genes are there in the list?
Question 1.3: How many unique miRNAs are there in the list?

Task 2: Filtering the data

Filter the data using the target_prediction_parsing.py to only keep predictions with an energy > 10 and probability > 0.999.

Question 2: How many predictions are left after filtering?

Task 3: Targeting types

There are two types of targeting events that occur in miRNA targeting. These are called *canonical* and *non-canonical* targeting events. Canonical binding occurs when the seed region of the miRNA is involved when the miRNA binds to the target. Column N reports whether a target event is canonical or non-canonical.

Question 3.1: What percentage of the reported targets are non-canonical?

Question 3.2: Is this consistent with what is reported in the literature?

Task 4: 3'UTR Lengths

Line 1 shows the predicted targeting between the 3'UTR of **TSC1** and **miR-4422**. The full descriptor for **TSC1** in this entry is

```
ENSG00000165699___TSC1_HUMAN___9-132891349-132896234n
```

This means that the 3'UTR for **TSC1** is located on Chromosome `9`, from nucleotides `132891349` to `132896234`.

Question 4.1: What are the lengths of all the 3'UTRs?
Plot a histogram of the length distribution. (Hint: we did something similar in the course when we plotted the length distribution of the miRNA and pre-miRNAs in the **GFF3** file we downloaded from **miRBase**)

Task 5: miRNA binding along the 3'UTR

Column D in the dataset shows the start position of the target in the 3'UTR. For example, for the first entry, the miRNA binding occurs at position `2156` on the 3'UTR.

Question 5.1: Plot the distribution of locations of the start positions for all 3'UTRs.

Question 5.2: Are there any biases in the location of where the miRNAs bind on the 3'UTR? (for example, are there more towards the 5' or 3' end of the sequence?)

Question 5.3: Are there any differences between where the canonical and non-canonical targeting events bind?

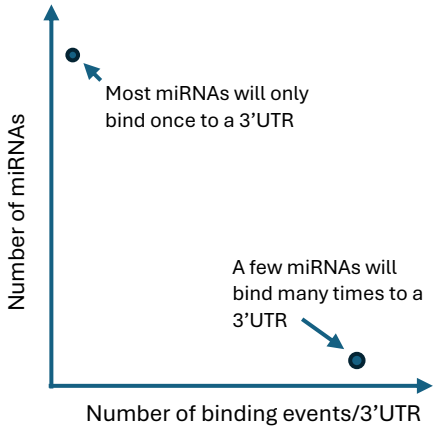
Question 5.4: Display these two types of binding on the plot?

Task 6: multiple miRNA binding events

If you look at the targeting events on lines 3, 4 & 5 in the dataset, you can see that **miR-4429** binds in three different locations on the same 3'UTR (**TSC1**)

Question 6.1: Generate another histogram to show how many times one miRNA binds to the same 3'UTR for <u>all</u> miRNA and all 3'UTRs.

For example

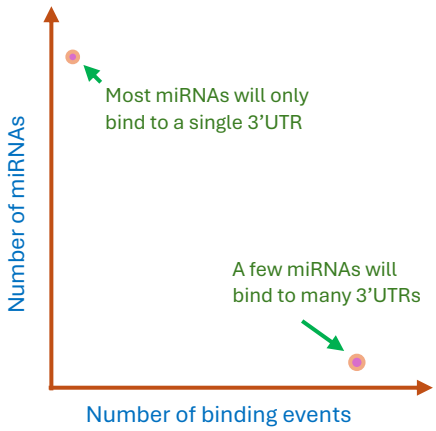


Hint: You can generate the data for this plot using **Panda** Summation functions.

Task 7: generate a miRNA--3'UTR targeting network

As we discussed in the course, one miRNA targets many 3'UTRs, and one 3'UTR is targeted by many 3'UTRs. So, what we have is a miRNA-3'UTR targeting network. One way to characterise this network is to look at the *connectivity* of the network.

Question 7.1: Similar to the plot above, generate a Connectivity Plot.



This time however, the plot shows how many times a miRNA targets a 3'UTR.
Once again, you can generate the data using the Panda summing functions by selecting the rows in the dataframe for each miRNA, and then counting how many unique 3'UTRs occur for each miRNA.

Question 7.2: Which miRNAs target the most 3'UTRs?
Question 7.3: What is the highest number of 3'UTRs targeted by a single miRNA?
Question 7.4: What are the roles in disease of the three miRNAs that have the most 3'UTR targets?