CS482 PS2 report

Arvin Asgharian Rezaee (a29asgha)

March 2025

1 Preliminary

The assignment was done using the following setup for python.

```
> python --version
Python 3.12.8
```

The given files use Numpy and Biopython to implement their solution and are needed to run them. To install them, run the following

```
> pip install numpy
> pip install biopython
```

all of the functions have ”–help” argument built in, and running it can showcase what arguments are for each file

2 kmer_composition.py

To find the K-mer frequency, I first filtered out any ambiguous bases in the input with ”N”, so we don’t record them in our results when calculating the frequency vector. I did not remove the N base since, there can be cases where extra frequencies can be read.
Then I created a HashMap of each k-mer to its count, iterated the sequence while updating the table, and then printed the result to the output file. Below is an example of how to run the python file:

```
> python .\kmer_composition.py --k 3 --input .\candidate.fas
```

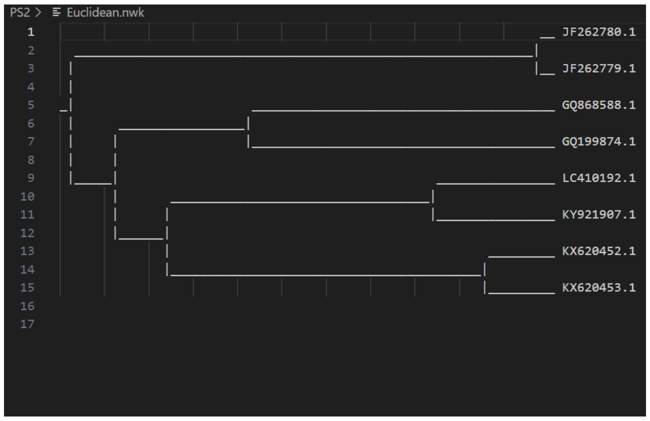
3 de_bruign_graph.py

To construct the graph, I used the algorithm provided in the course lectures. The only difference being, I saved the edges inside of a set(), to avoid duplicates being saved. Furthermore whenever going over every k-mer, I also calculate the reverse complement of the sequence, and then added the appropriate graph edges. below is an example of how to run the python file:

```
> python .\de_bruign_graph.py --input reads.txt
```

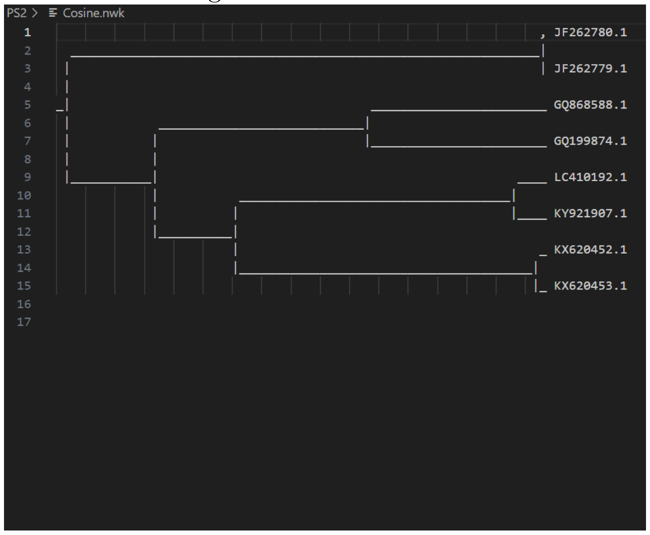
4 phylogeny_tree.py

Downloading and processing the sequences using biopython and code written on Q1 is pretty straightforward. The only difference is we normalize the frequency vectors by the sum of all vectors, which limits the values between 0 and 1. Now we use each distance function to create the distance matrices and use Biopython implementation of UPMGA to construct and print out the trees. Below is an initial diagram of each tree.


PS2 > Euclidean.nwk

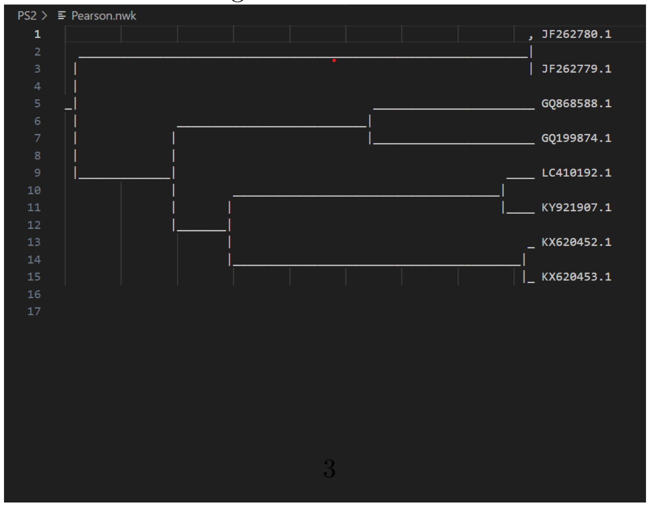
```
                                                                __ JF262780.1
  _____________________________________________________________|
 |                                                             |__ JF262779.1
 |
_|                     ___________________________________________ GQ868588.1
 |     _______________|
 |    |               |___________________________________________ GQ199874.1
 |    |
 |____|                                   _______________________ LC410192.1
      |      ____________________________|
      |     |                            |_______________________ KY921907.1
      |_____|
            |                                    _________ KX620452.1
            |___________________________________|
                                                |_________ KX620453.1
```

Figure 1: Euclidean tree


PS2 > Cosine.nwk

```
                                                                , JF262780.1
  _____________________________________________________________|
 |                                                             | JF262779.1
 |
_|                                        _________________________ GQ868588.1
 |           ____________________________|
 |          |                            |_________________________ GQ199874.1
 |          |
 |__________|                                                  ____ LC410192.1
            |          ___________________________________________|
            |         |                                           |____ KY921907.1
            |_________|
                      |                                              _ KX620452.1
                      |_____________________________________________|
                                                                    |_ KX620453.1
```

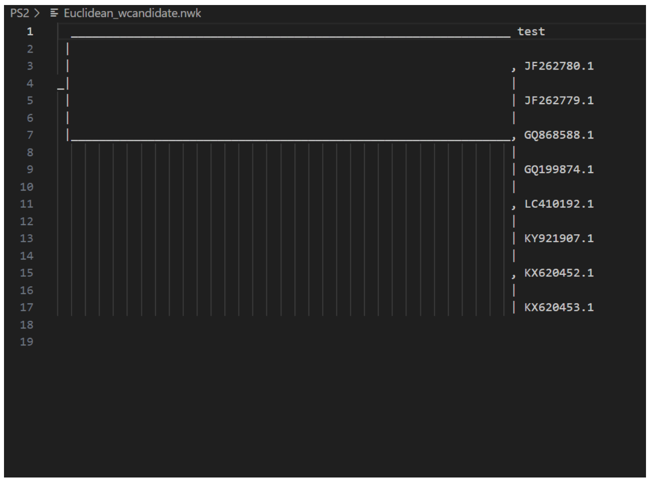
Figure 2: Cosine tree


PS2 > Pearson.nwk

```
                                                                , JF262780.1
  _____________________________________________________________|
 |                                                             | JF262779.1
 |
_|                                        _______________________ GQ868588.1
 |            ___________________________|
 |           |                           |_______________________ GQ199874.1
 |           |
 |___________|                                                ____ LC410192.1
             |       ____________________________________________|
             |      |                                            |____ KY921907.1
             |______|
                    |                                                _ KX620452.1
                    |_______________________________________________|
                                                                    |_ KX620453.1
```

Figure 3: Pearson tree

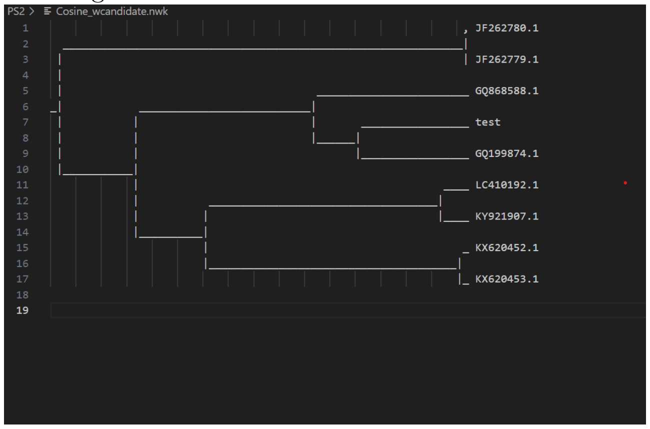
Now we add the candidate sequence to and rerun the analysis:



```
PS2 > Euclidean_wcandidate.nwk

  ______________________________________________ test
 |
 |                                              , JF262780.1
_|                                              |
 |                                              | JF262779.1
 |                                              |
 |______________________________________________, GQ868588.1
                                                |
                                                | GQ199874.1
                                                |
                                                , LC410192.1
                                                |
                                                | KY921907.1
                                                |
                                                , KX620452.1
                                                |
                                                | KX620453.1
```

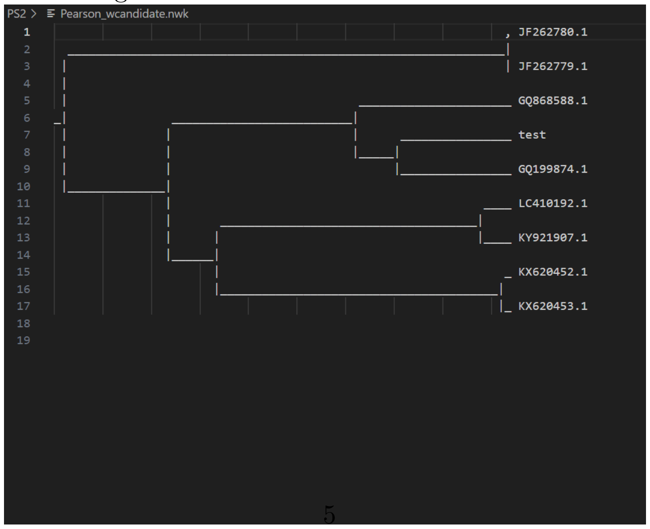
Figure 4: Euclidean tree with test candidate



```
PS2 > Cosine_wcandidate.nwk

                                                 , JF262780.1
  _______________________________________________|
 |                                               | JF262779.1
 |
 |                              __________________ GQ868588.1
_|                             |
 |         ____________________|
 |        |                    |      _____________ test
 |        |                    |_____|
 |        |                          |_____________ GQ199874.1
 |________|
          |                                     ____ LC410192.1
          |         ___________________________|
          |        |                           |____ KY921907.1
          |________|
                   |                                _ KX620452.1
                   |_______________________________|
                                                   |_ KX620453.1
```

Figure 5: Cosine tree with test candidate



```
PS2 > Pearson_wcandidate.nwk

                                                 , JF262780.1
  _______________________________________________|
 |                                               | JF262779.1
 |
 |                              __________________ GQ868588.1
_|                             |
 |         ____________________|
 |        |                    |      _____________ test
 |        |                    |_____|
 |        |                          |_____________ GQ199874.1
 |________|
          |                                     ____ LC410192.1
          |         ___________________________|
          |        |                           |____ KY921907.1
          |________|
                   |                                _ KX620452.1
                   |_______________________________|
                                                   |_ KX620453.1
```

Figure 6: Pearson tree with test candidate

1. **How well are subtypes clustered in the initial tree?** The tree layout given the fact that the same subspecies are clustered together seems to be correct. The Euclidean tree has higher distances for the same subtypes than Cosine and Pearson. Cosine and Pearson have very similar structure.

2. **How does the test sequence affect the tree?** In the Euclidean tree, the test candidate fails to be clustered with any of the subtypes. The distance between the test and each other subtype is too much. Meanwhile, the Pearson and the Cosine tree manage to put the candidate in a very similar and fairly reasonable place. The placing does not "break the logic" of the initial tree

3. **What is the predicted subtype of the test sequence?** subtype 2

4. **Which metric is preferable for alignment-free analysis?** Pearson or cosine.