CSE 5370: Bioinformatics
Homework 2

1.1 Reasoning About Time Complexity [5 points]

In your write up, explain which specific parts of the brute force solution in the section 1.1 code give the solution an undesirable time complexity and state what that time complexity is. State what the time complexity of your improved solution implemented in the"scsfast.py" file is.

- We got undesirable outcome at permutations part. Inside itertools.permutations. That makes the time complexity as O(n!).
- The timecomplexity of improved solution is O(n^3*m).

2 De Bruijn Graphs

You want to *assemble* a set of DNA sequencing reads. In this question, you will be reasoning about using a De Bruijn graph to do this.

2.1 Generating Your Own Unique Data [10 points]

You are provided with a python script named datasetGenerator hw2.py. This program will take in your UTA student ID as an argument and generates a unique artificial dataset of genome sequencing reads. To run the code, simply run:

>> python3 datasetGenerator_hw2.py --ID **1001960046**

Running the program will create a file named **1001960046**.txt in the same directory this program is located in. The file will have one simulated sequencing read per line.

2.2 Generating K-mers [20 points]

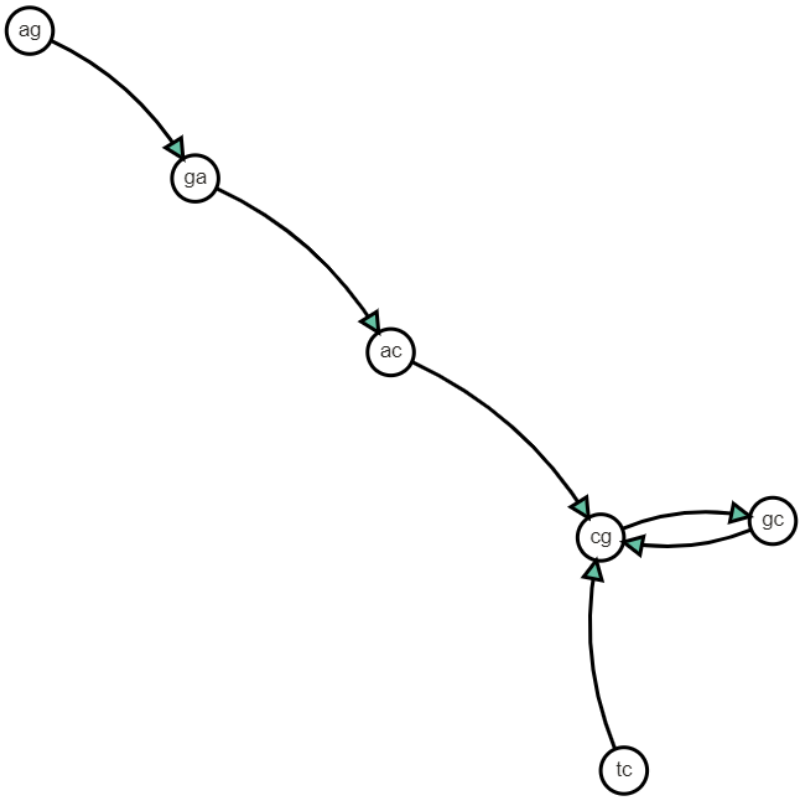
Assuming that k=2, calculate all possible K-mers from the set of reads in your unique dataset. Create a file called "kmers.txt" with one calculated K-mer per line and include this in your submission.

- **Kmers.txt file is included in the Zip file.**

2.3 Generating a De Bruijn Graph [20 points]

From the K-mers that you calculated in section 2.2, generate and plot a De Bruijn graph. Include your plotting code in your submission. Your plotting code should read in your"kmers.txt" file. Include a plot of your De Bruijn graph in your write up.

You should generate the De Bruijn graph from your calculated K-mers from section 2.2 with k=2.



2.4 Eulerian Cycles [10 points]

Discuss in your write-up whether or not your De Bruijn graph has an Eulerian cycle. The example De Bruijn graph in Figure 1 has an Eulerian cycle. The example De Bruijn graph in Figure 2 below does not have an Eulerian Cycle.

- **The De Bruijn graph for the generated data has no Eulerian cycle because in the graph their no path that passes through every node.**

2.5 Generating an Assembly [10 points]

If your De Bruijn has an Eulerian Cycle, include the assembled genome sequence in your write up. If your graph does not have an Eulerian Cycle, state what the minimum set of reads (of the same length as those generated for this problem) that could add a Eulerian Cycle to your De Bruijn graph are, and then state what the assembled genome sequence would be with those reads added.

Solution:

- By adding **cga,gag** we can make it Eulerian.
- Genome after adding reading is **tcgcgagacg.**

Another solution is:

- By adding **cgt**, **gtc** we can make it eulerian
- Genome : **agacgcgtcg**

We don't have a gtc reading in the dataset generator so first solution is in line with the readings of datagenerator.

3 Difficulty Adjustment [5 points]

Your answers to this section will be used to adjust the difficulty of future assignments in the class.

- How long did this assignment take you to complete?

  It took me around 2 days to complete the task

- If the assignment took you longer than the 15 hours, which parts were overly difficult?

  The first part of the assignment took me more time as I felt it was difficult.