ABIN Assignment 3
Dr. Debarka Sengupta
26/10/23

Guidelines

1. Deadline: 2nd November 2023, 11:59 pm (Midnight).
2. Late submission: 3rd November 2023, 11:59 pm (Midnight) (50% of the obtained marks will be deducted).
3. Further late submissions after the 3rd November will be awarded zero.
4. All coding assignments must be submitted as .ipynb files with proper comments.
5. For every question you should write a function with appropriate name and define the test case outside the function. The test case should be passed in the function as argument, and the function should return the result. If any group does not follow it, they will be awarded zero without further evaluation.

For example in Q1:

```
def Q1(<test case arguments>):

…code block……………………………
………………………..


        return <output>
output = Q1(sample test case)
```

6. Standard IIIT-D plagiarism policy applies.
7. The assignments have to be submitted in the following manner:

(a) Create a **single** Jupyter notebook with proper demarcation of questions and responses (text,
code, command, explanation, output, graphs). The accepted language is Python.

(b) A PDF file comprising the entire Jupyter Notebook and a separate Jupyter Notebook file needs to be uploaded (zip format) at the assignment link on Google Classroom before the deadline.

(c) One group should submit only from the registered submission ID during group formation. Rest of the members must turn in the assignments with private comments mentioning the name, roll, and submission id. Any violation of above will disqualify your submission from evaluation.

(d) The name of the PDF and jupyter notebook file should be a combination of the group number
and assignment number. For example, group1 1.pdf, where 'group1' is the group number and '1' is the assignment number.

8. No shift of the deadline is allowed.

Question 1 (5 Points)

Download coronavirus sequences of three different strains(Ex. Alpha, Omicron, etc). Perform Multiple Sequence Alignment (MSA) using tools such as ClustalW. Download the generated MSA file and write a python script to identify mutations at each position of the three sequences. Also calculate percent identity between the three sequences and comment on the pathogenesis of the mutations(if any).

Question 2 (5 Points)

Find K-mer of length 5 and implement De-Bruijn Graph Assembly for the following sequence with minimum overlap of 2:
TGTAGAAAGTACCCAGTGCTCAGTATAG

Question 3 (5 Points)

Take a nucleotide sequence as input and construct a Burrows Wheeler Transform (BWT) of the sequence.
Given - A string Text
Return – BWT(Text)

Sample Input –
GCGTGCCTGGTCA$

Sample Output –
ACTGGCT$TGCGGC