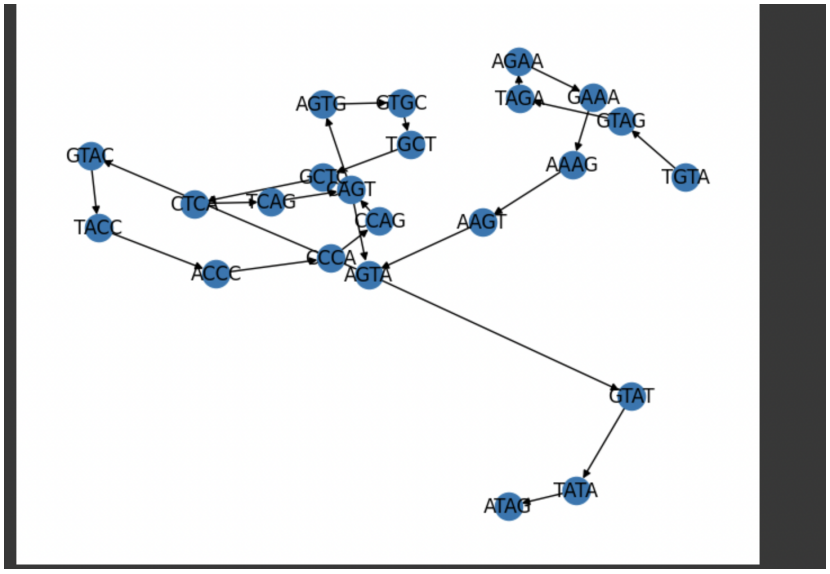
Abin assignment -3

ACTGGCT$TGCGGC

Q3.
This is the following BWT output for the input test case ‘GCGTGCCTGGTCA$’ The dollar sign is usually used as a special character to represent the end of the string it is important to keep the dollar sign as it helps differentiate between string with common prefixes and helps such strings be properly sorted in the transformation process

The Burrows-Wheeler Transform (BWT) is a reversible data transformation technique commonly used in data compression and data indexing. It is an essential component of several compression algorithms and plays a crucial role in bioinformatics, particularly in DNA sequence compression and sequence alignment

Q2.
De Bruijn assembly, also known as the de Bruijn graph-based assembly, is a computational method used in bioinformatics for reconstructing the genome or DNA sequence of an organism from short DNA sequence fragments, typically generated by high-throughput sequencing technologies such as next-generation sequencing (NGS)



Assembled sequence

```
1: Find K-mers and Assemble Sequence
2: Plot De Bruijn Graph
Enter the number of the option you want to execute: 1
Assembled Sequence: TGTAGAAAGTACCCAGTGCTCAGTATAG
```

Q1.
Percent identity, in the context of sequence alignment, represents the degree of similarity or resemblance between two biological sequences, such as DNA, RNA, or protein sequences. It is usually expressed as a percentage and quantifies the proportion of identical positions (matching residues or bases) between the two sequences when they are aligned.

A higher percent identity means greater degree of similarity, and we have tried to find out the percentage of similarity between the three strains

We have performed multiple sequence alignments of three strains of COVID, which are:

(i)Alpha
(ii)Beta
(iii) Gamma

```
Position 6856: Counter({'T': 2, 'N': 1})
Position 6857: Counter({'C': 2, 'N': 1})
Position 6858: Counter({'T': 2, 'N': 1})
Position 6859: Counter({'T': 2, 'N': 1})
Position 6860: Counter({'A': 2, 'N': 1})
Position 6861: Counter({'G': 2, 'N': 1})
Position 6862: Counter({'G': 2, 'N': 1})
Position 6863: Counter({'T': 2, 'N': 1})
Position 6864: Counter({'T': 2, 'N': 1})
Position 6865: Counter({'A': 2, 'N': 1})
Position 6866: Counter({'A': 2, 'N': 1})
Position 6867: Counter({'A': 2, 'N': 1})
Position 6868: Counter({'A': 2, 'N': 1})
Position 6869: Counter({'A': 2, 'N': 1})
Position 6870: Counter({'A': 2, 'N': 1})
Percent Identity: 78.19%
```

Pathogenesis: The interpretation of the MSA file is given below

A 78% sequence identity between two genetic sequences suggests a moderate level of similarity. This similarity can provide insights into the evolution, virulence, and drug resistance of pathogens. It's useful for designing diagnostic tests and identifying conserved regions for vaccine development. However, pathogenesis involves many other factors, including host responses and environmental conditions, making sequence identity just one piece of the puzzle in understanding how diseases develop and progress.