Course Project

Submission Format- In .ipynb jupyter notebook, write the answers of the subjective questions in markdown cells. The video presentation should be in .mp4 format. You'll be uploading both the jupyter notebook and the video on your google drive. We will share a google form on **30th july** where you will submit this assignment as a google drive link.

Q1. It's possible for there to be multiple different shortest common superstrings for the same set of input strings. Consider the input strings ABC,BCA,CAB. One shortest common superstring is ABCAB but another is BCABC.

1) What is the length of the shortest common superstring of the following strings?
   CCT,CTT,TGC,TGG,GAT,ATT
   How many different shortest common superstrings are there?
   Hint: you can modify the scs function to keep track of this.
2) Make the overlap graph and De Bruijn graph of above 3-mers. (this can be generated via code)

Q2. Download this FASTQ file containing synthetic sequencing reads from a mystery virus: https://d28rh4a8wq0iu5.cloudfront.net/ads1/data/ads1_week4_reads.fq
All the reads are the same length (100 bases) and are exact copies of substrings from the forward strand of the virus genome. You don't have to worry about sequencing errors, ploidy, or reads coming from the reverse strand. Assemble these reads using the greedy algorithm.
How many As,Ts,Cs and Gs are there in the full assembled genome? Report as percentage.
Hint: the length of the virus genome is 15894 bases

Q3. Download the following file from NCBI using the GenBank code U01317.1 (Download file in GenBank format) Now answer the following questions regarding this sequence:

1) How many times does CTTAGAACGGAAATCTTAGT or its reverse complement occur in the above sequence? Eg. if AGGT occurs 10 times and its reverse complement ACCT occurs 12 times, report the final answer as 22.
2) Use Naive exact matching, Boyer-Moore and k-mer indexing each for the above problem. Compare their performances ie. time taken to execute the code and no. of character comparisons performed by each.

Q4. Use gene ID 65262 and 11946 to download the gene ATP synthase subunit in rats and mice. Find the edit distance between these two gene sequences such that

1) Transversions are given score of 2, transitions of 2, gaps are given score of 2
2) Transversions are given score of 2, transitions of 4, gaps are given score of 4
3) Transversions are given score of 4, transitions of 2, gaps are given score of 8
4) Find the hamming distance between the two (splice the larger string to take the first n elements, where n is the length of the shorter string)

Which of the above four methods is the best way to find the edit distance?

Comment on the potential uses of such a program. (answer in markdown)

Q5. Use the code NC_045512.2 to download the coronavirus genome (cds file). Now do the following:

1) Print the protein sequence and reverse complement of the helicase enzyme
2) Take the ORF7a protein sequence and divide it into 8,9,10,11,12-mers each. Assemble the sequence back with each of these k-mers. Compare the results, are they the same?
3) By using the “random” library of python (random.randint method), randomly generate 20 15-mers (ie. 20 substrings of length 15) of the surface glycoprotein gene. Create a naive exact matching function that allows for two mismatches and returns the location of mismatches (AATG in ACTAACG will return (0,3) as matches) use this function to find matches of these 20 substrings in RNA dependent RNA polymerase gene of sars-cov-2. Note- use random.seed so that your results are replicable.

Q6. Create a short video presentation (about 1 to 2 min long) describing the read alignment problem, de novo assembly problem and the various algorithms we have learnt to tackle these problems. Go into detail about the specific challenges faced while solving these problems (don’t describe each algorithm separately, focus more on the difficulties and how to solve them for example- polyploidy)