**Programming Assignment – II**

**Deadline: 7th Sept**

Using C/C++/Java/Python:

1. Write a program to change format of a DNA sequence: from Fasta format to Flat file format and vice versa. (File: take any gene sequence)
2. Write a script to parse the features table from a GenBank file. (File: FJ959345; NM\_175762)
3. It is assumed that A = T and G = C in a genome. Compute G+C content, (G+C)/(A+T+G+C), deviation from A = T, (A-T)/(A+T) and deviation from G = C, (G-C)/(G+C) to confirm. Use a non-overlapping sliding window of size 50Kb. [Hint use some bacterial genome]
4. Write a program to find ‘most frequent’ 9-mers in *Vibrio Cholerae* genome (Chr1: NC\_002505.1, Chr2: NC\_002506.1): (i) exact match, (ii) allowing ‘d’ mismatches. [Hint: count for reverse complements also). Are the ‘most frequent’ 9-mers the same in Salmonella enterica (NC\_003198.1) as identified in *Vibrio Cholerae* genome? Plot the GC skew and find which ‘most frequent’ 9-mer is in the vicinity of the region where GC skew changes sign.
5. Using the k-mer detection program in Q-4 above, carry out comparison of the SARS-CoV-2 genome with bat coronavirus genome using k=6 (exact matches). (i) Based on your analysis can you infer that SARS-CoV-2 coronavirus is related to bat coronavirus? (ii) Based on the GC content of SARS-CoV-2 genome, what can you say about the stability of this virus at high temperatures.