Part-I. Answer in the space provided. (Marks: 15)

1. If the length of sequence read is 100bases. and 30x coverage is desired, give the number of reads that will be required to sequence a SARS-CoV2 genome (L_G = 30Kb). [2]

2. If the following DNA sequence is used as template for RNA synthesis, give the RNA sequence that will be synthesized in the correct orientation. [1]

5′ CATTGCCAGT 3′

3. Give the expected frequency of observing the pattern GCCGCC (GC box) in a genome of length 10MB and G = 0.3? [2]

4. Which of the following two alignments is likely to evolutionarily more plausible? Why?

(i)	GCGGC	(ii)	GCGGC
	G - - GC		G- G – C

[2]

5. Draw the self dotplot of a sequence containing a tandem repeat region. [2]

6. Echolation in sperm whale is an ancient or derived characteristic? How was it determined? [2]

7. In the four-nucleotide DNA code, the 20 amino acids are encoded in codons of length three. Suppose Martians have same 20 amino acids but have a five-nucleotide code (A, C, G, T, Z). What is the minimum codon length required to encode Martian proteins? Justify your answer. [2]

8. (a) What's the underlying assumption in the computation of sum of pairs scoring method in MSA? [2]

(b) Using the scoring scheme (2, -1, -2), compute the score for the column:

$$\begin{pmatrix} G \\ A \\ - \\ G \end{pmatrix}$$

Part-II: Answer in additional answer booklet. (Marks: 35)

1. Under what situations would you use the following variants of dynamic programming algorithm (give examples): (i) global alignment, (ii) local alignment, (iii) suboptimal matches, (iv) overlap matches, (v) Linear space alignment algorithm. **[5]**

2. (a) BLAST program uses 'neighbourhood' words instead of 'exact' words as seeds to look for similarity in sequence database search. What is the motivation behind this?
(b) BLAST program misses some good biological homologies below the accepted statistical cut off value. How would you identify these distant homologies?
(c) How is the significance of an alignment evaluated in the BLAST program? **[8:3,3,2]**

3 (a) What information can be obtained by performing multiple sequence alignment of protein sequences?
(b) Compute the score of the alignment of S1-S2 with the alignment of S3-S4, given the weight of four sequences, S1, S2, S3, and S4 are 0.4, 0.1, 0.2, 0.3, respectively. [Use BLOSUM62 scoring scheme (N-N: 6, N-C: -3, C-C: 9):

S1: ---N--- S3: ---N---
S2: ---C--- with S4: ---N---

(c) In ClustalW, what is the justification of retaining gaps introduced in the earlier alignments? **[6:2,2,2]**

4. (a) Show schematically gene structure in prokaryotes and eukaryotes.
(b) Define an open reading frame. Give at least one issue with using open reading frame (ORF) approach for gene identification in prokaryotes and eukaryotes. **[5:2,3]**

5. (a) For N=4 sequences, give the possible number of (i) unrooted, and (ii) rooted trees.
(b) Draw all possible unrooted trees for four taxa: A, B, C, D. **[5:2,3]**

6. (a) What is the space and time complexity of dynamic programming algorithm for pairwise sequence alignment?
(b) How would you address these issues for tasks such as database search, or whole genome comparisons. **[6: 2,4]**