(Write your answers as clearly and precisely as possible within the space provided in the ANSWER sheet that is given separately)

Part – A – MCQ (Q1 – 6)

Each question (Q1-6) carries 1 mark
For every wrong answer, 0.5 marks will be deducted

Q1) The number of possible rooted and unrooted trees that can describe the possible relationships among four family members are respectively:

a) 15 and 15
b) 15 and 3
c) 3 and 15
d) None of these

Q2). Which of the statements given below is correct ?

a) Similarity is a measurable quantity (x% similar), but one cannot say the two sequences are x% homologous
b) Homology is a measurable quantity (x% homologous), but one cannot say the two sequences are x% orthologous
c) One can say the two sequences are x% similar and y% homologous
d) None of these

Q3). An alignment of four homologous sequences are given below. In Parsimony method used in phylogenetic tree, one has to find the informative and uninformative site. Which sites (the nucleotide positions) are informative?

GGGGGG
GGGAGT
GGATAG
GATCAT

a) 3rd and 4th
b) 4th and 5th
c) 5th and 6th
d) All sites

Q4). Two species are found to share a cluster of 8 genes, but the genes are in different orders in the two species. The orders are represented by signed permutations:

Species X 1, 2, 3, 4, 5, 6, 7, 8
Species Y 1, 2, -5, -4, -3, 8, 6, 7

The transformation between the two gene orders.....

a) cannot be achieved by inversions alone
b) can be achieved by one translocation and one inversion
c) can be achieved by three inversions
d) requires six separate genome rearrangement events

Q5). Which one of the following expressions is likely to retrieve more matches in a database search?

a) D-A-V-I-D
b) [DE]-A-V-I-[DE]
c) [DE]-[AVILM]-X-E
d) D-A-V-E

Q6). Your BLAST search against a database returned 100 hits. Of these, 17 were false positives. However, you know that there were 165 sequences in the database that should have been returned a hit with your query sequence. The sensitivity and selectivity of BLAST respectively are:

b) 83 and 82
b) 82 and 83
c) 165 and 83
d) 0.50 and 0.83

Part – B

Q7). Describe these terms very important while performing annotation (a) ORF (b) UTR (c) mRNA processing

Q8). What is PSSM? How is it derived? Name a tool - where & how is it used. What is its utility?
(2 + 2 + 1 + 1 marks)

Q9). My research team (DAILAB) in our department is actively conducting research on integrated drug screening for stress, aging and cancer intervention with prime focus on elucidation of functional mechanisms of natural drugs. We are studying a gene called "Gene X" that is involved in the regulation of secondary metabolite biosynthesis in an Indian medicinal plant called *Ashwagandha*. The protein encoded by Gene X appears to be a good candidate for a possible "switch" protein that determines whether a metabolite A or B is formed in the metabolic pathway. Now, we have managed to obtain a partial cDNA sequence (3000 nucleotides long) as well as a clone containing the corresponding genomic DNA sequence (4000 bp long). The genomic clone includes exactly 1000 additional nucleotides upstream (on the 5' side) of the 5' end of the cDNA clone sequence. When the Gene X sequence was BLASTed against the existing database, it was found that Gene X has ~ 90% sequence identity with part of a larger protein (Ash-P) encoded by a gene previously cloned and sequenced in *Ashwagandha*. The function of the *Ashwagandha* protein (Ash-P) is not known, but mutations in the Ash-P gene result in the metabolite A formation rather than metabolite B. The Gene X appears to have at least two exons (each about 500 bp). There is a candidate for a translational start signal (ATG) within one exon (at position 1500 on the genomic clone) and a cluster of 3 potential translational STOP codons within the other exon (beginning at position 3000 on the genomic clone).

Question 9-A **(2 marks)**

i) Draw a line to represent the map of the cDNA clone
ii) Label the 5' end of the cDNA
iii) Label the positions of the translational START and STOP codons
iv) Draw two boxes to represent the two exons mentioned above (at their approximate locations)

Question 9-B **(1 mark)**

Would you use any of the gene prediction methods to predict the exact intron/exon boundaries in the Gene X? Explain.

Question 9-C **(1 mark)**

You are now asked to check the genomic DNA sequence between 1-1000 for potential additional exon(s) not present in the cDNA clone. You find a region in the genomic DNA sequence (between positions 500 and 800) that is strongly predicted to contain another ORF. What are the criteria you would assume/choose to make such a prediction?

Q10). Imagine you are working with an unusual protein. It has very weak similarity to anything in the public databases. To ensure that you keep up-to-date with the information available, you "BLAST" your protein against the nr database at NCBI now and then. When you did your last search, a few months ago, the top hit had E value 0.1. Now, doing the same search again, the very same top hit has E value 0.2, even though the score is the same. How come? **(1 mark)**

Q11). What are contigs? How can different contigs be connected together in a genome assembly? **(1 mark)**

Q12). In what linear order would these five sequences assemble into a contig? All sequences are on the same strand. Answer just with the linear order of sequence reads from 5' to 3'. Use at least 7 base pairs minimal overlap.
(1 mark)

1. CCGACTCCAGCCTCCACTGCCTCGAGCCCCCC
2. GCTCTCCCAGTTCTCCGACTCCAG
3. CAGAGAGCACCAGCTCCACAAG
4. CTCCACAAGGGACCTGCTCTCCC
5. GCCCCCCTGTACGAAGTGGACACTCTC
