Science -II (Biology)
Final Exam April 2022

Roll No.: **Seat No.:**

Date of Exam: **Investigators Signature:**

Duration: 3hrs

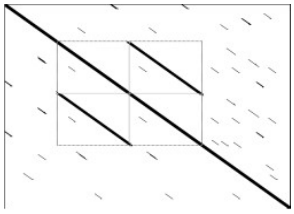
Total Marks: 90

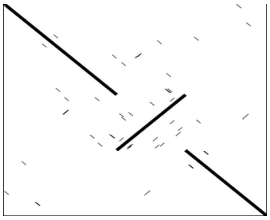
<u>Note:</u> Answer PART-A in the space provided, and use additional sheet for PART-B.

<u>PART-A</u>: (Marks: 32)

1. Show how a self-dot plot looks like if the sequence contains a (i) tandem repeat and (ii) an inverted repeat. **[2]**

Soln:





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

Soln: Echolation is an ancient characteristic in sperm whale. Analysis of the two mitochondrial ribosomal gene suggested that sperm whales are more closely related to baleen whales than to other toothed whales. This implies echolocation is present in both branches of whales, which suggests that the common ancestor had this characteristic which was subsequently lost in the baleen whales as a result of adaptation to their lifestyle. Thus echolation is an ancient characteristic in sperm whale.

3. Scores of ungapped local alignment of random sequences follow ---------- distribution. **[1]**

Soln: extreme-value

4. (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:

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

Soln: (a)All the residues in a column represent the ancestral residues as well as recent evolutionary changes. Since any one of the residues could have been the ancestor residue, the computation of sum of pairs in MSA by scoring every residue with every other residue in each column is carried assuming each one as the ancestor residue.

(b) -1-2+1-2-1-2 = -6

5. Is e-value in the BLAST program dependent on the size of database being searched? If yes, will it increase/decrease with increase in the size of the database? **[2]**

Soln: Yes. It will decrease.

6. How would you randomize a nucleotide sequence of length L = 100 without changing the composition of the nucleic acids? **[2]**

Soln: (a) In a loop over a very large N (~ 10 times L), call two random numbers between (1 to L) and swap the positions of the two amino acids. R.

Else, Compute the frequencies of each of the nucleic acids in the given sequence. Break the interval of 0 – 1 into parts corresponding to these frequencies. Call a random no. between 0 – 1. Depending on the range in which the random no. falls, add that particular nucleic acid. Repeat this 100 times.

7. Give one example of biological relevance for analyzing the number and distribution of k-tuples: (i) k=1, (ii) k=2, (iii)k=3? **[3]**

Soln: (i) k=1: composition of the sequence, origin of replication, melting point of DNA.

(ii) k=2: distribution of dinucleotide GC – helps in identifying protein-coding regions, set of di-nucleotide relative abundance values constitutes a "genomic signature" of an organism, Useful in identifying genomic islands (laterally transferred regions) devoid of coding, GC skew - to predict locations of replication origins and termini in prokaryotes

(iii) k=3: codon usage – for identifying protein-coding regions, frequency of stop codons to identify open reading frames, Compute codon adaptation index (CAI) that provides an indication of gene expression level - it compares the distribution of codons actually used in a particular protein with the preferred codons for highly expressed genes,

8. In BLAST database search algorithm, the match/mismatch ratio for comparing nucleotide sequences is chosen to be large for highly conserved sequences, while it is small for divergent sequences. Give reasons, why? **[2]**

Soln: When you are looking for very conserved sequences, you expect to see very low %age of mismatch, hence you want to penalize it heavily in this case. While in the case of divergent sequences, we expect to see more positions with mismatch, and hence we would want to penalize it not so heavily in this case.

9. Which of the two scoring matrices, BLOSUM80 and BLOSUM45 would you use for comparing closely related sequences and evolutionarily distant sequences? **[1]**

Soln: BLOSUM80 - closely related sequences and BLOSUM45 - evolutionarily distant sequences.

10. 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]

Soln: 5^2 = 25, more than 20. So the minimum codon length required in this case is 2.

11. Gene finding is easier in prokaryotes compared to eukaryotes. Give two reasons to support this statement. **[2]**

Soln: (i) contiguous coding region in case of prokaryotic gene compared to eukaryotic genes.
(ii) high gene density in prokaryotes.
(iii) exons are not multiple of 3, leading to frame-shift errors.
(iv) stop codons may be present in the intronic regions.

12. You are given a set of ESTs belonging to the same gene locus. Suggest a simple approach to identify the correct order of the ESTs to build the mRNA sequence? **[2]**

Soln: By concatenating all the ESTs and obtaining a self-dot plot, the correct order of ESTs can be identified.
Or, using a program such as polydot in EMBOSS, to obtain an all-against-all dotplots of a set of sequences.

13. Suppose you have extracted some mRNA and do not know what gene it corresponds to, or the genome of the organism is not available. How would you design primers for this sequence?

[2]

Soln: First obtaining cDNA, then inserting it into a bacterial plasmid of known sequence (cloning) using restriction enzymes. Using primers for the plasmid vector DNA, sequence the unknown DNA and then create primers for the sequence of interest.

14. A template strand of DNA is 3′ TACCGATTGCA 5′. Which of the following DNA strand that is created from this template during replication? **[1]**
    A. 5′ TGCAATGCCTA 3′
    B. 5′ TAGGCATTGCA 3′
    C. 5′ AUGGCUAACGU 3′
    D. 5′ ATGGCTAACGT 3′

Soln: D

15. Presence of which of the following cell structures indicates that it is a eukaryotic cell: **[1]**
    A. DNA
    B. Nucleus
    C. Cytoplasm
    D. Ribosomes

Soln: B

16. Which of the following two alignments is preferable? Give reason. **[2]**

A.

ACAAT

A -A – T

B.

ACAAT

A - - AT

Soln: B. because in alignment two indels are represented as a single mutational event.

17. Unequal usage of codons in the coding regions appears to be a universal feature of the genomes across the phylogenetic spectra. What is the likely cause for this bias? **[2]**

Soln: (i) the uneven usage of the amino acids in the existing proteins and (ii) the uneven usage of synonymous codons.

18. In the Jukes Cantor model, a C→T transition is more likely than a C→G transversion. **[1]**

Soln: False

19. Fourier transform of a coding sequence would result in a peak at 3. **[1]**

Soln: False. It's at p = 1/3.

<u>PART – B</u>: Answer in the additional answer sheets. (Marks: 58)

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]

Soln: (i) global alignment – when the sequences are closely related and we can expect an end-to-end alignment.

(ii) local alignment – when the sequences are distantly related, during database search to identify homologs, identify conserved regions/motifs, identify common domains shared by two sequences, etc.

(iii) suboptimal matches – to identify multiple-copies of a domain/motif, more than one conserved region between a pair of sequences.

(iv) overlap matches – in sequence assembly of genomic sequences, in gene assembly, in EST clustering, using ESTs for identifying genes in a genome.

(v) Linear space alignment algorithm – while comparing large genomic sequences, to reduce the space complexity.

2. (a) Give the (i) boundary conditions and (ii) recursive relations in the dynamic program variant for genomic sequence assembly.
   (b) What changes in the above conditions will lead to global alignment?
   (c) Why the conditions for obtaining global alignment not suitable for sequence assembly?
   (d) Which is more important in sequence alignment: the choice of alignment (global or local) or the scoring scheme? Give your reasons. **[6:2,1,2,1]**

Soln: (a) $F(0,0) = 0$, $F(i,0) = 0 = F(0,j)$, $F(i,j) = \max\{F(i,j)+sij, F(i-1,j)-d, F(i,j-1)-d\}$

(b) change in the boundary conditions: $F(i,0) = -id$, $F(0,j) = -jd$

(c) We do not want to penalize the over-hanging ends.

(d) The most reasonable approach is to use the program based on the appropriate algorithm for the analysis at hand, and then to choose the scoring system carefully. Small changes in the scoring system can abruptly change an alignment from a local to a global one.

3. (a) What is the major difference between the approach of FASTA and BLAST in identifying the seed for extending matches?
   (b) How is the significance of an alignment evaluated in the BLAST program?
   (c) When would you use PSI-BLAST ? **[6:2,2,2]**

Soln: (a) BLAST is based on the motivation that - Good alignments should contain many <u>close</u> matches while FASTA is based on the assumption that - Good alignments should contain many <u>exact</u> matches. Thus, in BLAST each word in the database sequence, a list of "neighbourhood" words, scoring above a threshold T are considered as a 'match'.

(b) If the score of the alignment observed is no better than might be expected from a random permutation of the sequence, then it is likely to have arisen by chance. For computing the statistics one needs random sequences. Align these random sequences using the Smith-Waterman algorithm and compute the score for the optimal alignment. Scores of optimal alignment follow an extreme value distribution. Measure the mean and standard deviation of the scores of the alignments of randomized sequences. If the randomized sequences score as well as the original one, the alignment is unlikely to be significant.

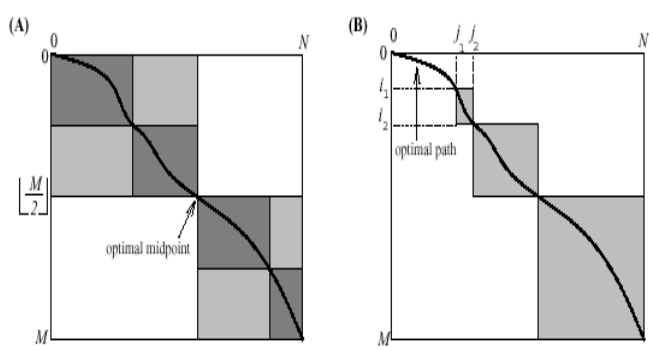
(d) Position Specific Iterated (PSI) BLAST has been designed for finding remote homologues. When sequence identities dip below 25%, BLAST normally fails to recognize the similarity. PSI-BLAST is able to work in the range 15%-25% sequence identity levels by using a scoring matrix tailor-made to find sequences similar to the query from the initial BLAST search. The process is iterated until you get no more new hits and the algorithm has converged.

4. (a) There are two major problems with the progressive approach for multiple sequence alignment: (i) the local minimum problem, and (ii) the choice of alignment parameters. How can these two problems be addressed?
   (b) In ClustalW, what is the justification of retaining gaps introduced in the earlier alignments?
   (c) Why are sequences weighted in ClustalW? **[8: 4,2,2]**

Soln: (a) (i) The local minimum problem - dependence on the initial pairwise alignment. The more distantly related these sequences are, the more errors will get propagated through the alignment. To correct for errors introduced by initial alignment, iterative methods may be used. Iterative methods attempt to avoid this by repeatedly aligning subgroups of sequences.
(ii) The choice of alignment parameters: when adding more distantly related sequences to the msa, we are dealing with different evolutionary distances so that the substitution matrix and the gap penalties must be different at every stage of merging the clusters.
(b) The positions of the gaps that were introduced during the early alignments of the closely related sequences are not changed as new sequences are added. This is justified because the placement of gaps in alignments between closely related sequences is much more accurate than between distantly related ones.
(c) Sequences are weighted to correct for unequal sampling across all evolutionary distances in the data set. This down-weights sequences that are very similar to other sequences in the data set and up-weights the most divergent ones.

5. Write short notes on (i) discuss Hirschberg's algorithm for aligning two sequences, and (ii) Importance of bootstrapping in phylogenetic analysis. **[6]**

Soln: (i) Hirschberg's alignment algorithm uses the principle of divide and conquer for aligning two sequences.



Steps: Let the two sequences be of size m and n respectively.
Let u = m/2 (integer part)
Identify a v such that the cell (u, v) is on the optimal alignment, i.e., v is the column where the alignment crosses the i = u row of the matrix

- this splits the dynamic programming problem into two subproblems: from (0, 0) to (u, v), & from (u, v) to (m, n)
- the full alignment will be the concatenation of the optimal alignments for these two separate submatrices

This is done recursively, by successively halving each region, until sequences of zero length are being aligned, or alternatively, sequences are short enough and standard O(nm) alignment and traceback method can be used.

But how do we find v?

By combining the results of "forward" and "backward" dynamic programming passes at row u, for each point along the middle row, i.e.,

- the optimal score from (0, 0) to each point in row u and the optimal score from that point to (m,n)

Adding these numbers gives the optimal score over all paths from (0, 0) to (m, n) that pass through all the points in row u. A sweep along the middle row, checking these sums, determines a point (m/2, v) where an optimal path crosses the middle row. This is then done recursively.

(ii) Bootstrapping:

Once a tree has been chosen by some method, it would be desirable to quantify how confident one is of it. This is given by the statistical technique - bootstrapping.

In this procedure, the true data sequences are used to create a set of new pseudo-replicate sequences of the same length. Bases at a particular site in the new sequences are chosen to be the bases appearing in a randomly chosen site in the original sequences. Considering pseudo-replicates is similar to randomizing a given sequence; to construct a random sequence, we may call randomly any character at each position to construct a new random sequence, here we call any random column to construct a new MSA and a new tree corresponding to this MSA.

In bootstrap analysis:

- All sites are considered independent as if freely available in a 'hat' to pick
- Available sites are picked up randomly to reconstruct a new alignment of the <u>original size</u> and a new phylogeny

Process is repeated many times to ascertain the strength of clustering. New "alignment" may contain several sites multiple times while some other sites may be absent - sampling with replacement - as a result same column may occur a number of times in the new 'random' alignment.

Phylogenies are compared to calculate values [Bootstrap value] that signify the number of times a given branch/cluster occurred in the Multiple bootstrap trees

Higher the value - higher the confidence of phylogenetic inference

In general values < 50% provide very poor support

6. (a) Define an open reading frame. Discuss issues with using open reading frame (ORF) approach for gene identification.
   (b) In computational gene prediction, give the advantage and limitation of homology-based approach.
   (c) Briefly describe a model-independent approach for finding genes. **[10:3,3,4]**

Soln: (a) A long sequence between two stop codons devoid of stop codons in-between is called an ORF.

Addition or deletion of one or two bases will cause all the codons scanned to be different - sensitive to frame shift errors

- Fails to identify very small coding regions
- In general, the largest ORF is the one that codes for proteins – need not be always true.
- Fails to identify the occurrence of overlapping long ORFs on opposite DNA strands (genes and 'shadow genes')
- Overlapping genes on the same strand observed in bacterial genomes – an overlap of 2-3 bases in an operon.

In the case of eukaryotes,

- due to the existence of interweaving exons and introns – stop codons may exist in intronic regions making it difficult to identify the correct ORF
- a gene region may encode many proteins – due to alternative splicing/alternative translation initiation
- Exon length need not be multiple of three – resulting in frameshift between exons
- Gene may be intron-less (single-exon genes)
- Relatively low gene density - only 2 - 3% of the human genome codes for proteins

(b) The genes whose homologs are present in the database can be readily identified by homology-based database search approach – with the database sizes increasing exponentially, this is now routine first step in gene prediction and nearly ~50% of genes can be thus identified. Limitation - only homologs of known genes can be identified, novel genes, or genes specific to an organism cannot be identified by this approach, and apparent homology can be sometimes misleading.

(c) Refer to ppt

7. (a) The recognition site for Sau3A I is GATC and is contained in the recognition site of BamH I, GGATCC. Will the two REs give the same results? If not, which one will give larger number of fragments?

   (b) The enzymes BamH I and Bal II recognise different sequences but leave the same sticky ends:

   BamH I: ----------G|G A T C C ------

   Bal II: ----------A|G A T C T ------

   Will the two enzymes result in the same number of fragments in a random DNA sequence? Give reasons.

   (c) Construct a restriction map of a linear fragment of DNA, using the following data. Give the fragment lengths and their positions on a linear scale. **[6:2,2,2]**

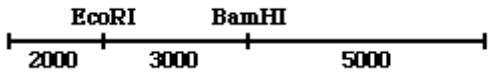
DNA	**Sizes of Fragments (bp)**
uncut DNA	10,000
DNA cut with EcoRI	8000, 2000
DNA cut with BamHI	5000, 5000
DNA cut with EcoRI + BamHI	5000, 3000, 2000

Soln: (a) No. The cut made by BamHI is between GG and CC, while by Sau3A I is between GA and TC. The one with smaller restriction site will make larger number of fragments, since a pattern of four will occur more often than a pattern of size six.

(b) Yes, because the size of the pattern is the same, and all 6-mers will occur with the same frequency in a random DNA sequence.

(c) F1: L=2000, Pos: 1 – 2000,

F2: L=3000, Pos: 2001 – 5000
F3: L=5000, Pos: 50001 – 10,000



The evolutionary distance given by Kimura 2-parameter model accurate as it considers different substitution rates for transitions and transversions, while Jukes-Cantor model assumes them to occur at the same rate.

8. (a) Use the distance table below and apply the UPGMA method to compute distances and build a tree.

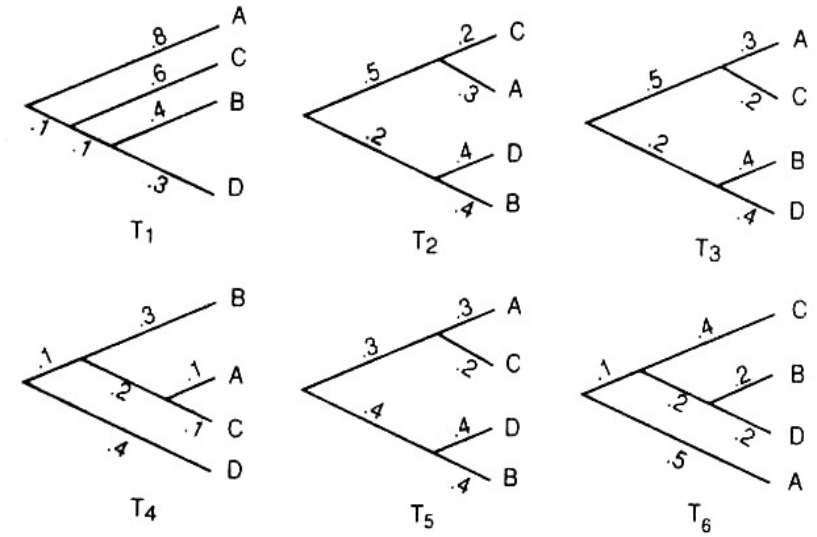
	S1	**S2**	**S3**	**S4**
S1		**.45**	**.27**	**.53**
S2			**.40**	**.50**
S3				**.62**

(b) Give all possible unrooted trees for any four taxa A, B, C, D. **[6:3,3]**

Soln: Refer to ppt.

9. Consider the trees in the figure below: **[5]**

(a) Which of them are the same, as rooted metric trees?
(b) Which of them are the same, as unrooted metric trees?
(c) Which of them are the same, as rooted topological trees?
(d) Which of them are the same, as unrooted topological trees?
(e) For which trees does a molecular clock appear to be operating?



Soln: (a) Same rooted metric trees: T2, T3 [Hint: compute the distance matrix between all taxa, trees exhibiting same distances between taxa are similar rooted trees, if there rooted topology exhibits same distances between internal nodes also]

(b) Same unrooted metric trees: T2, T3, T5 [Hint: compute the distance matrix between all taxa, trees exhibiting same distances between taxa are similar rooted trees, if there topology is also the same]

(c) Same rooted topological trees: T2, T3, T5 and T1, T6 [Hint: same rooted branching pattern]

(d) Same unrooted topological trees: all [Hint: same branching pattern, exhibiting same neighbours]

(e) Trees for which molecular clock appear to be operating: T4 and T6 [Hint: compute the distance of each taxa from the root, if all the taxa are placed at equidistant from the root, molecular clock is operating]