Science – II Final Exam

May 2023

Duration: 2hrs **Total Marks: 60**

Roll No.:..... **Program:.....**

Room No.:..... **Seat No.:.....** **Date:.....**

Invigilator's Signature:.....

1. (a) If the length of sequence read is 100bases, and 10x coverage is desired, give the number of reads that will be required to sequence a human genome. [2]
   (b) If the following DNA sequence is used as template for RNA synthesis, give the RNA sequence that will be synthesized in the correct orientation. [2]
   5' CATTGCCAGT 3'

Soln: (a) No. of Reads required = GenomeLength * DesiredCoverage / ReadLength
$= 3 \times 10^9 \times 10/100 = 3 \times 10^8 = 300$Million reads

(b) First write its complement:
5' CATTGCCAGT 3'
3' GUAACGGUCA 5'
Then the synthesized RNA sequence in 5' to 3' orientation is:
5' ACUGGCAAUG 3'
- i.e., synthesized RNA sequence is basically the complement of the template DNA sequence with T replaced by U, when read in the 5' to 3' orientation

2. (a) Give the expected frequency of observing the pattern TATAAT (TATA box) in a genome of length 10MB and A = 0.3?
   (b) Give the basic steps (or a pseudocode) to identify restriction recognition sites in a genome. [2,3]

Soln: (a) $(0.3)^6 \times 10^7 = 7290$
(b) use the property that restriction recognition sites are palindromic sequences. For a restriction recognition site of size k, take a sliding window of size k and check if i=1 and i=k are complementary bases, if yes, check for i=2 and i=k-1, and so on.

3. (a) Which of the following two alignments is likely to evolutionarily more plausible? Why?

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

   (b) Find the number of times letter T occurs in a sequence of length N = 1000 if the probability of occurrence of T, p(T) = 0.15. [2,4]

Soln: (a). Alignment (i) - since in this case 2 adjacent deletions is likely to be a single deletion event compared to 2 independent deletions in close vicinity in (ii).

(b) Expected value = n x p(T) = .15 x 1000 = 150,
Variance = n x p(T) (1 - p(T)) = 1000 x .15 x .85 = 127.5,
Std. deviation = 11.29

No. of times T would occur in a sequence of length 1000, if p(T) = 0.15 will lie between 150 ± 11.29, i.e., **139 – 161**

4. Which variant of the dynamic programming algorithm would you use if
   (i) the two sequences are closely related
   (ii) the two sequences are distantly related
   (iii) you have to find the order of fragments in genome assembly
   Give reasons for your choice. **[2,2,2]**

Soln: (i) Global alignment, because you would expect similarity over the full length of the sequence from closely related species.

(ii) Local alignment / Suboptimal local alignment – in sequences from distantly related species, you would more of mutations such as substitutions and indels, which would lead to only functionally/structurally important regions conserved. Thus one can only expect short conserved regions in such a case.

(iii)Overlap alignment/Semi-global alignment – In this case we expect similarity only at the end of the sequences (3' end of one sequence with 5' end of another or vice versa) and we do not want to penalize the over-hanging ends.

5. (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? **[3,3]**

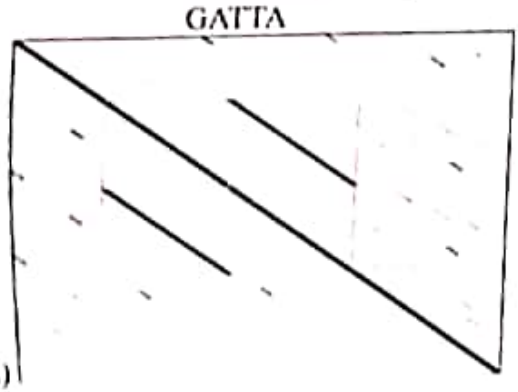
Soln: (a) There is higher probability of finding similar instead of exact matches between sequences that are evolutionary distant. This helps in identifying distantly related sequences in the database and also help in using larger k-mer, thereby increasing the computational efficiency of the algorithm.

(b) (i) using PSI-BLAST, or, (ii) by repeated searches using hits from the first BLAST search as query.

6. (a) Draw the self dotplot of a sequence containing a tandem repeat region.
   (b) Obtain local alignment between the following two sequences using the scoring scheme: Match = +1, Mismatch = -1, INDEL = -2. **[2,5]**

GCGGTTGACG
GATTA

Soln: (a)



g a -1 -1 a

g
c
g
g
t
t
g
a
c
g

(b) Fill the F(i,j) matrix to show the best local alignment and score

7. (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): [2,3]

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

Soln: (a) MSA can be informative in

- inferring evolutionary relationships – building phylogenetic trees
- improve pairwise alignment
- constructing scoring matrices – PAM, BLOSUM
- predicting secondary and tertiary structures of new sequences
- identifying conserved motifs, patterns, and blocks – to characterize protein families
- homology modelling of proteins

(b) 0.4 x 0.2 x 6 + 0.4 x 0.3 x 6 + 0.1 x 0.2 x (-3) + 0.1 x 0.3 x (-3) = 0.48 + 0.72 – 0.06 – 0.09 = 1.20 – 0.15 = 1.05

8. (a) For N=4 sequences, give the possible number of (i) unrooted, and (ii) rooted trees.
(b) For the given distance matrix relating four taxa, construct the phylogenetic tree using UPGMA method. [2,5]

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

Soln: (a) No. of unrooted trees = 3, No. of rooted trees - 15
(b) In the table giving the distances between taxa, pick the two closest taxa, S1 and S3. Because they are .27 apart, we draw the edges with each edge 0.27/2 = 0.135 long.

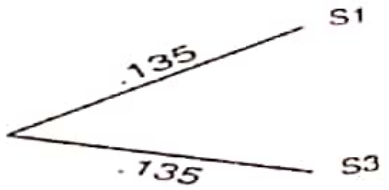


Figure 5.6. UPGMA: step 1

We then combine S1 and S3 into a group, and average the distances of S1 & S3 to each different taxon to get the distance from the group. S1-S3 to that taxon. For e.g., the distance between S1-S3 and S2 is (.45 + .40)/2 = .425, and the distances between S1-S3 and S4 is (.53 + .62)/2 = .575. Our table thus collapses to:

Table: Distances between Groups: UPGMA Step-1

	S1-S3	S2	S4
S1-S3		.425	.575
S2			.50

Now, we simple repeat the process, using the distance in the collapsed table. Because the closest taxa and/or groups in the new table are S1-S3 and S2, which are .425 apart, we draw the figure:

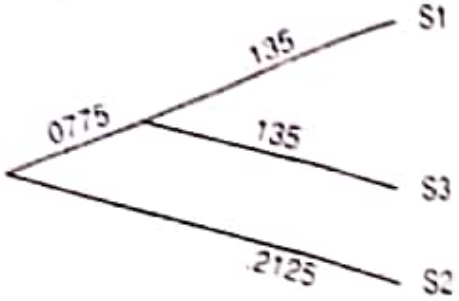


Figure 5.7. UPGMA: step 2

The edge to S2 must have length .425/2 = .2125, while the other new edge must have length (.425/2) - .135 = .0775, because we already have the edges of length .135 to account for some of the distance between S2 and the other taxa.

Again combining taxa, we form a group S1-S2-S3, and compute its distance from S4 by averaging the original distances from S4 to each of S1, S2, and S3. This gives us (.53 + .5 + .62)/3 = .55. (Note this is not the same as averaging the distance from S4 to each of S1-S3 and to S2. The new collapsed table would have this as its only entry.

	S1-S2-S3
S4	.55

We draw the final tree, estimating that S4 is .55/2 = .275 from the root. The final edge has length .0625, since that places the other taxa .275 from the root as well.

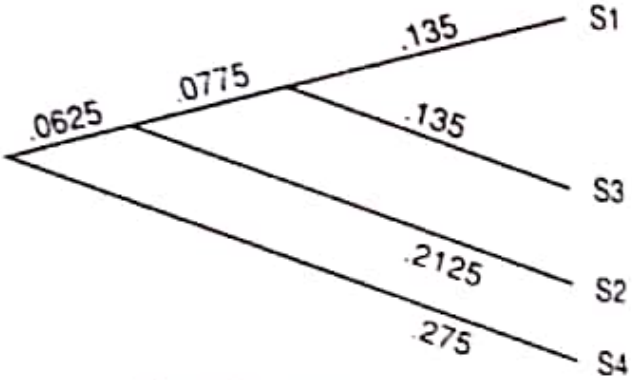
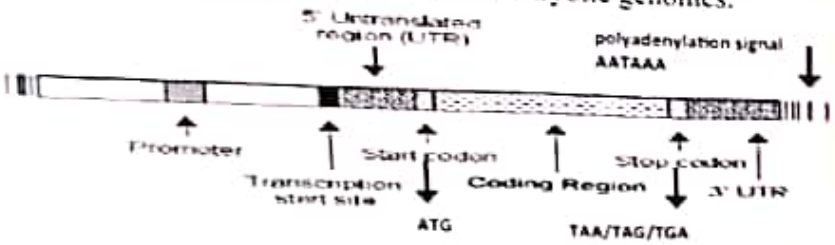
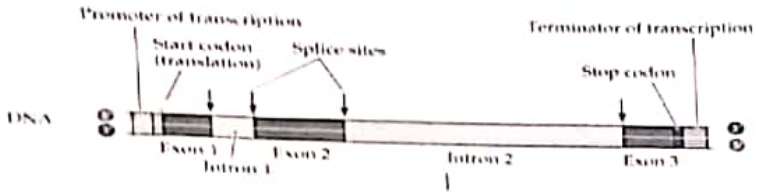


Figure 5.8. UPGMA: step 3.

9. (a) Show schematically gene structure in prokaryotes and eukaryotes.
(b) Briefly discuss any method for predicting genes in a genome. Discuss it advantages or limitations for prediction in prokaryotic and eukaryotic genomes. [2,5]

Soln:





(b) Using ORF, Codon usage, Codon prototype, Markov model, position asymmetry, Fourier spectra – check the paper uploaded for details of any of these methods.

ORF – good for finding coding regions in prokaryotes, but would fail in eukaryotes because of non-contiguous gene structure, a stop codon may occur in the intronic region

Codon usage – works fine for both prokaryotic and eukaryotic gene structure – however may miss out small exons

HMM - works better than codon usage for identifying both prokaryotic and eukaryotic gene structure – however may miss out small exons

Position asymmetry/Fourier spectrum – useful when no training data is available, but the signal is weak and small genes/exons can be missed.

10. (a) What is the space and time complexity of dynamic programming algorithm for pairwise sequence alignment?
(b) Which according to you is a major issue? Give reasons to support your answer and provide a possible solution. **[2,5]**

Soln: (a) O(n,m) for both space and time complexity

(b) Space – when comparing large genomes/chromosomes using DP

Time – when searching large databases, e.g., GenBank using DP

Solution – For Time

- use heuristics to find short conserved regions first, and if only many such conserved ungapped regions can be identified, extend them, and if they exceed a pre-defined threshold, perform DP only in the band around the diagonal.
- Split the datasets and perform search in parallel

Solution – For space – use linear space alignment algorithm.