Indraprastha Institute of Information Technology Delhi (IIITD)

Department of Computational Biotechnology

BIO213 – Introduction to Quantitative Biology

MID-SEM EXAM (March 07, 2022)

Time duration: 1 hour **Total marks:** 50

Instructions:

1. Answers need to handwritten. Typed documents will NOT be accepted.
2. Scan all the answer sheets and compile them into a single PDF file for submission.
3. You will get additional 10 minutes to prepare the document for submission. The excuse of slow/unstable internet connection will strictly NOT be entertained.

Question 1. Differentiate between **any 3** of the following: (2 X 3 = 6 marks)

1) Primary and secondary databases
2) BLOSUM and PAM substitution matrices
3) tblastx and tblastn
4) Cladogram and Phylogram

You can evaluate this question on your own. Make sure that the major differences have been mentioned.

Question 2. What are the common types of errors that add to the complexity of the problem of DNA fragment assembly? (4 marks)

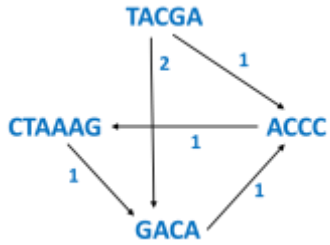
Should have proper explanation about base call errors, chimera and contamination.

Question 3. Find the overlap between the given two sequences using dynamic programming approach. (8 marks)
S: VAEGEWGVAA **T:** REWVAEA **Scoring scheme:** Match= +4, Mismatch= -1, Gap= -5

		V	A	E	G	E	W	G	V	A	A
	0	0	0	0	0	0	0	0	0	0	0
R	0	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
E	0	-1	-2	3	-2	3	-2	-2	-2	-2	-2
W	0	-1	-2	-2	2	-2	7	2	-3	-3	-3
V	0	4	-1	-3	-3	1	2	6	6	1	-4
A	0	-1	8	3	-2	-4	0	1	5	10	5
E	0	-1	3	12	7	2	-3	-1	0	5	9
A	0	-1	3	7	11	6	1	-4	-2	4	9

```
---VAEGEWGVAA
REWVAEA------
```

Question 4. Construct an overlap graph for F= (TACGA, ACCC, GACA, CTAAAG). Find a shortest common superstring for this collection. (5 marks)




```
TACGA
   GACA
      ACCC
         CTAAAG
_______________
TACGACACCCTAAAG
```

Give 3 marks for the graph and 2 marks for the fragment assembly. The alignment layout given should result in the shortest string.

Question 5. What is the smallest value of ε such that the layout below is valid under the Reconstruction model? (4 marks)

F= (ACCGT, CGTGC, TTAC, TGCCGT)

```
--ACCGT--
----CGTGC
TTAC-----
-TGCCGT--
---------
TTACCGTGC
```

There exists one error between the last fragment and the consensus sequence.
So, d_s(TGCCGT, TTACCGTGC) = 1

Now, we know that d_s(TGCCGT, TTACCGTGC) $\leq \varepsilon$ |TGCCGT|

Therefore, $1 \leq \varepsilon$ 6.

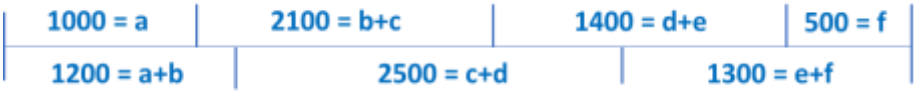
So, the smallest value for ε = 1/6

Or

Find a solution given the following results of a double digest experiment.
Enzyme A: 1000, 2100, 1400, 500 (4 marks)
Enzyme B: 1200, 2500, 1300
Enzyme A+B: 1000, 200, 1900, 600, 800, 500

A+B: 1000 200 1900 600 800 500
a b c d e f

A: 1000= a, 2100= b+c, 1400= d+e, 500=f
B: 1200= a+b, 2500=c+d, 1300= e+f

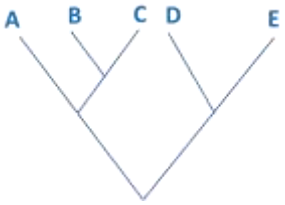


Question 6. Draw the phylogenetic trees that correspond to the following Newick format. Comment whether the trees obtained are the same or different. (5 marks)

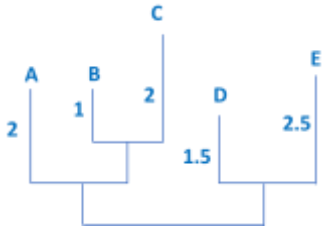
i. (((B,C)A)(D,E))

ii. ((B:1,C:2)A:2)(D:1.5,E:2.5))

i.



ii.



The (ii) tree gives information about the branch length as well. Therefore, we get evolutionary distance/relationship between the different Taxa.

2marks for each tree and 1 mark for the comment (please use your own judgement).

Question 7. Use UPGMA to reconstruct a phylogenetic tree using the following distance matrix: (6 marks)

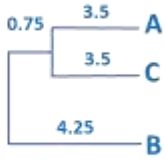
Species	A	B	C	D
A	0	-	-	-
B	8	0	-	-
C	7	9	0	-
D	12	14	11	0

I.

	A	B	C	D
A	0	-	-	-
B	8	0	-	-
C	7	9	0	-
D	12	14	11	0

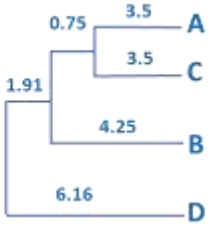
II. d(AC,B)=[d(AB)+d(BC)]/2 = [8+9]/2 = 8.5
d(AC,D)=[d(AD)+d(CD)]/2 = [12+11]/2 = 11.5

	AC	B	D
AC	0	-	-
B	8.5	0	-
D	11.5	14	0



III. Unweighted:
d(ACB,D)=[d(AD)+d(CD)+d(BD)]/3 = [12+11+14]/3 = 12.33

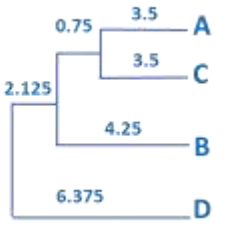
	ACB	D
ACB	0	-
D	12.33	0



OR

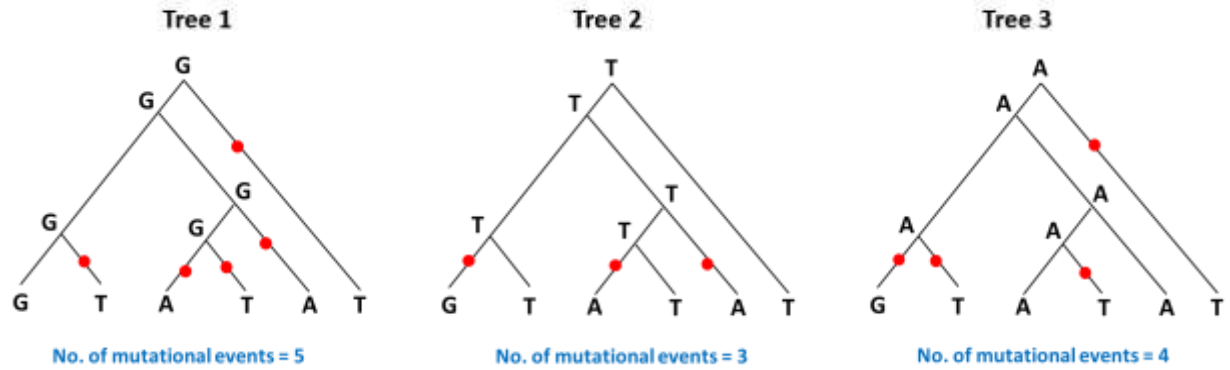
Weighted:
d(ACB,D)=[d(AC,D)+d(B,D)]/2 = [11.5+14]/2 = 12.75

	ACB	D
ACB	0	-
D	12.75	0



2 marks if only step I, 4 marks up to step II, 6 marks if all correct.
4 marks if trees are correct but branch lengths are wrong.

Question 8. Which of the following trees is the best solution as per the principle of maximum parsimony? Justify your answer (6 marks)



No. of mutational events/evolutionary changes are the least in tree 2, therefore it will be considered as the best solution by maximum parsimony.

Question 9. Match the following. (6 marks)

1. Protein data bank	i. Protein clustering on sequence similarity
2. GenBank	ii. Dot plot
3. CODIS	iii. Mutations and crossover
4. BLOCKS	iv. Biomolecular structural database
5. Genetic algorithm	v. Short tandem repeats
6. FASTA	vi. Nucleotide database

1-iv, 2-vi, 3-v, 4-I, 5-iii, 6-ii
1 mark for each of the correct answers