Indraprastha Institute of Information Technology Delhi (IIITD)
Department of Computational Biotechnology

BIO213 – Introduction to Quantitative Biology

MID-SEM EXAM (March 03, 2024)

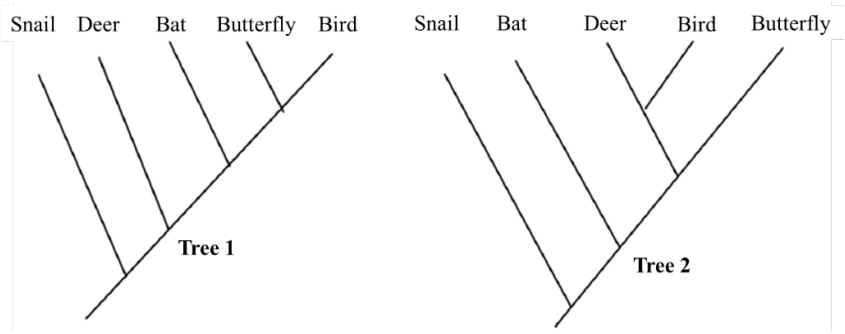
Time duration: 1 hour **Total marks:** 60

Question 1. Differentiate between **any 5** of the following: **(10 marks)**

i. BLOSUM and PAM substitution matrices
ii. blastx and tblastx
iii. Cladogram and Phylogram
iv. Linear and Affine gap penalty
v. Bootstrap and Jackknife methods for tree evaluation
vi. Global and local alignment

You can evaluate this question on your own. Make sure that the major differences have been mentioned. 2 marks each.

Question 2. Applying the principle of parsimony, the ability to fly is most likely to have gone according to which of the following evolutionary trees? Justify your answer. **(3 marks)**



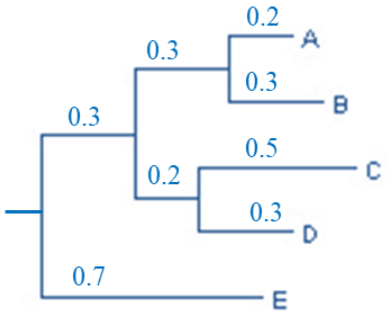
Answer: Tree 1, according to principle of parsimony simplest explanation that fits for best results, Tree 1 explains the the most simplest path that can be followed.

2 marks for correct answer + 1 mark for justification

OR

Question 2. Draw the phylogenetic tree that corresponds to (((A:0.2, B:0.3):0.3,(C:0.5, D:0.3):0.2):0.3, E:0.7):1.0. Is this a rooted or an unrooted tree?

Answer:



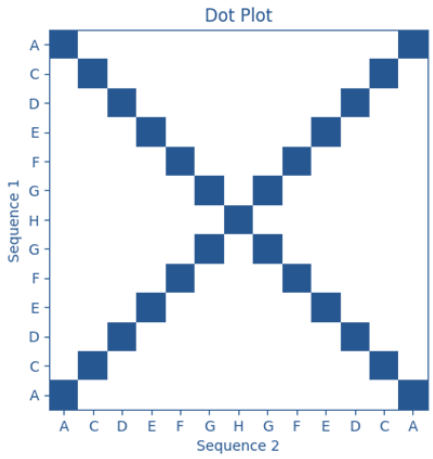
2 marks for correct tree + 1 mark for correct branch lengths

Question 3. i) Construct a dot plot for the following sequences, and comment on the type of similarity you observe in the sequences.

seq1 = ″ACDEFGHGFEDCA″ seq2 = ″ACDEFGHGFEDCA″

ii) What are the limitations of dot plots? **(4+2 marks)**

Answer:



2.5 marks

Type of similarity: Sequences are palindrome in nature 1.5 marks

Limitation: (Any 2) - 1 mark each

- Rely on visual analysis
- Difficult to find optimal alignments
- Difficult to estimate significance of alignments
- Insensitive to conserved substitutions (e.g. L→ I or S → T)
- Compares only two sequences (*vs*. multiple alignment)
- Time consuming (1,000 bp *vs*. 1,000 bp = 10^6 operations, 1,000,000 bp *vs*. 1,000,000 bp = 10^{12} operations)
- Any other reasonable answer (like DOT plot can be noisy for bigger sequence)

Question 4. Align the given two reads to make a contiguous sequence (contig) using dynamic programming approach. **(10 marks)**

R1: ATTCGCGAG *R2:* GGCGAGCTCA

Scoring scheme: Match= +2, Mismatch= -1, Gap= -2

		A	T	T	C	G	C	G	A	G
	0	0	0	**0**	0	0	0	0	0	0
G	0	-1	-1	-1	**-1**	2	0	2	0	2
G	0	-1	-2	-2	-2	**1**	1	2	1	2
C	0	-1	-2	-3	0	-1	**3**	1	1	0
G	0	-1	-2	-2	-2	2	1	**5**	3	3
A	0	2	-2	-3	-4	0	1	3	**7**	5
G	0	0	1	-1	-3	-2	-1	3	5	**9**
C	0	-1	-1	0	1	-1	0	1	3	7
T	0	-1	1	1	-1	0	-2	-1	1	5
C	0	-1	-1	0	3	1	2	0	-1	3
A	0	2	0	-2	1	2	0	1	2	2

ATTCGCGAG---

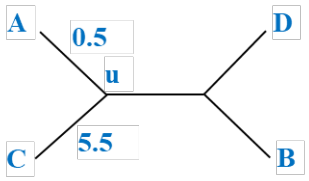
---GGCGAGCTCA

Correct matrix = 8 marks, up to 5 mistakes 7 marks

Correct traceback and alignment shown = 2 marks

Question 5. Assume that the first step of the neighbor-joining algorithm joins taxa A and C, that the edge to A has a length 0.5, the edge to C has length 5.5, and that the new distance matrix relating the remaining groups is as follows. Find the neighbor-joining tree. **(10 marks)**

	A/C = u	B	D
A/C = u	0	8	2
B	8	0	8
D	2	8	0



r(u) = 0+8+2 = 10

r(B) = 8+8 = 16

r(D) = 2+8 = 10 (if correct up till this part - 2 marks)

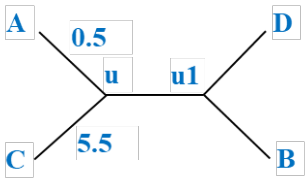
New distance matrix:

M(U,B) = 8-[(10+16)/(3-2)] = -18

M(U,D) = 2- [10+10] = -18

M(B,D) = 8- [16+10] = -18 (joining B & D together with common ancestor u1)

(if correct up till this part - 6 marks)

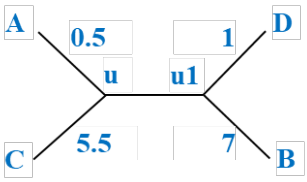


Distance of B & D from u1:

S(B,u1) = d(BD)/2 + [(r(B) - r(D))/2(N-2)] = 8/2 + [(16-10)/2(3-2)] = 4+3 = 7

S(D,u1) = d(BD) - s(Bu1) = 8-7 = 1

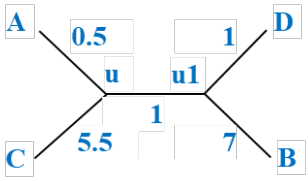
(if correct up till this part - 8 marks)



Distance between u and u1:

d(u,u1) = [d(B,u) + d(D,u) - S(B,D)]/2 = (8+2-8)/2 = 1

(if correct up till this part - 10 marks)



Question 6. Answer the following. **(5 marks)**

1. Log odds ratios express the probability of a substitution event occurring without knowledge of ancestral sequences. State TRUE or FALSE? TRUE
2. Which of the following PAM substitution matrices is comparable to BLOSUM80?
   a. PAM 1 b. PAM 250
3. In genetic algorithm, producing two new offspring by combining selected bits of strings from the two parents is called mutation. State TRUE or FALSE? FALSE
4. BLOSUM matrices are numbered directly proportional to the percentage identity of sequences in the BLOCKS database. State TRUE or FALSE? TRUE
5. Which of the following will contribute in increasing alignment with non-homologous sequences:
   a. Increasing match score
   b. Increasing mismatch score
   c. Decreasing the gap penalty

Question 7. A researcher is working with the following sequences of flagellin proteins from six different bacteria of similar lineage to find the phylogenetic relationship between.

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
K	L	S	A	A	D	H	K	R	G	V	R	T	N	G	V	L	M	G	E
K	I	S	A	A	D	H	K	R	G	V	R	C	Q	G	V	L	M	G	E
K	I	S	A	A	E	H	K	R	A	V	H	C	N	G	V	I	M	G	E
K	I	S	A	A	E	H	G	R	K	V	H	C	Q	G	V	L	M	A	E
K	I	S	A	A	E	H	G	R	I	V	S	C	N	G	V	L	M	G	E
K	L	S	A	A	D	H	G	R	V	V	S	C	Q	G	M	L	M	G	E

(i) What sites in the above alignment layout would be informative for a parsimony analysis? **(2 marks)**
Informative sites: 2, 6, 8, 12, 14
Deducted 1 mark if one incorrect site, if more, no marks.

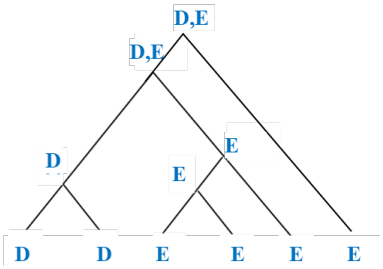
(ii) How many rooted trees can be constructed to describe the possible evolutionary relationships between these 6 taxa? **(2 marks)**

For n = 6 (number of taxa),

Number of rooted trees = $(2n-3)! / (2^{n-2})\,(n-2)! = (9)! / (2^{4})(4)! = 945$

Answer in factorial is acceptable

(iii) Draw any one possible rooted tree for these six taxa using amino acids at the position under consideration as D, D, E, E, E, E. Label each of the internal nodes with the most likely candidate for the inferred ancestral sequence. What is the minimum number of substitutions required by your tree topology? **(6 marks)**



Minimum number of substitutions will be 1 (with E at the root node).

This question can have many answers, any possible tree topology with correct labelling of nodes and minimum mutations mentioned will get full marks.

Question 8. Which of the following is NOT a correct bootstrap sequence sample of the given original sequence sample? Justify your answer **(6 marks)**

Original sample:

```
AGGAGGTCCAGA
AGAAGGTCCAGA
AGGGGGTCCAGA
AAAGGATCCAGA
```

Bootstrap samples:

a.
```
GGAGATCCAGAC
GGAAATCCAGAC
GGGGATCCAGAC
AAGAATCCAGAC
```

b.
```
AAGGAGGTCAGA
AAGAACGTCAAA
AAGGGGGTCAGG
AAAAGAGTCAAG
```

c.
```
GGAAGGTCCAG
AAAAGGTCCAG
GGAGGGTCCAG
AAAGGATCCAG
```

b. GCGA is not an original column

c. sequence length is shorter than the original layout

3 marks each (1.5 marks for correct identification, 1.5 marks for explanation)