Assignment - V

MSA & Phylogeny

Name: Gowlapalli Rohit

Roll No: 2021101113

(1) MSA = Multiple Sequence Alignment

MSA is a powerful tool used in bioinformatics to compare and analyze multiple sequences simultaneously

→ Phylogenetic Analysis: Multiple alignment can be used to infer evolutionary relationships b/w sequences & construct a phylogenetic tree

→ Functional Analysis: MSA can help identify conserved regions within a set of sequences, which are likely to be important for their function

→ Structure prediction: MSA can be used to predict the structure of a protein or RNA molecule based on the conserved regions observed in the alignment

→ Annotation of genome sequences: MSA can be used to annotate functional regions of genomes such as promoters, exons & introns

* In DNA sequences MSA is used in

i) Genome sequences assembly - Shotgun sequencing

ii) Discovering new regulatory elements

iii) Inferring evolutionary relationships

iv) DNA barcoding

v) SNP identification

vi) Developing primers & probes - use conserved regions to develop
   - Primers for PCR
   - Probes for DNA microarray

* In protein sequences, MSA is used in

i) Homology modeling of proteins

ii) Building phylogenetic tree

iii) Constructing scoring matrices - PAM, BLOSUM

iv) Predicting secondary & tertiary structures of new sequences

v) Identifying conserved regions, motifs, blocks in protein sequences - to characterize protein families

vi) Identify related proteins in Database searches

→ Also helps in finding Regions rich in insertions / deletions, building gene / protein families and in secondary structure prediction

→ (Yes), A multiple alignment carries more information than mere pair-wise alignment By comparing multiple sequences, a multiple alignment can reveal additional information such as the degree of conservation of certain residues across different sequences, and the presence of insertions & deletions relative to a common ancestor. This additional information can be valuable for inferring functional & evolutionary relationships b/w the sequences

(2) Sum of pairs (SP) is a simple way to score

Evaluate a multiple alignment is to evaluate the cost column by column

$$\text{Sum of pairs (SP)} = \sum_{i<j} D(S_i, S_j)$$

↓

using unit cost: mismatch costs 1, match 0 and indel costs 1

For example:

$$\text{Column cost} \begin{bmatrix} L \\ L \\ A \\ P \\ G \\ S \\ - \\ G \end{bmatrix} = 6+6+5+4+3+2 = 26$$

→ Summing the scores of all possible combinations of AA pairs in a column of MSA

→ Assumes a model for evolutionary change in which any of the sequence could be the ancestor of others

→ The main drawback of the SP score is that it does not take into account the phylogenetic relationships between the sequences being aligned. This means that the SP score may not accurately reflect the true evolutionary relationships b/w the sequences

→ There are problems with SP scoring system as illustrated in the example:

Sequence	Col-A	Col-B	Col-C
1	...N...	...N...	...N...
2	...N...	...N...	...N...
3	...N...	...N...	...N...
4	...N...	...N...	...N...
5	...N...	...N...	...N...
score	60	24	9

(Using Blosum62): → score for N=10 seq

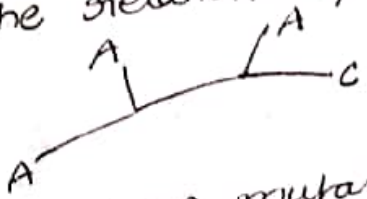
N-N: 6
N-C: 3
C-C: 9

→ The sum of the pairs scoring scheme tends to overweight the contribution of differences, from many very similar sequences

Consider the column $\begin{bmatrix} A \\ A \\ A \\ C \end{bmatrix}$

The sum of pairs scores for the column
$= 3 - 3 = 0$

However, the relationship b/w the sequences is



Then, a single A→C mutation can explain the data and thus SP tends to overcount mutations for related sequences, it ignores the assumption of a shared ancestor. As a result, the SP approach for evaluating MSA's is inherently problematic from an evolutionary standpoint

→ An alternative scoring system is the maximum-likelihood (ML) score, which takes into account the phylogenetic relationships b/w the sequences The ML score uses a probabilistic model of sequence evolution to calculate the likelihood of observing the observed multiple alignment given a particular tree topology & substitution model. The tree topology & substitution model are optimized to maximize the likelihood of the observed data, which provides a more accurate measure of the evolutionary relationships b/w the sequences

(3) Progressive Approach

↳ Align each sequence to every other pair-wise

→ Compute distances b/w each aligned pair (e.g: no of mismatches)

→ Construct a phylogenetic tree

→ Cluster closely related sequences

→ Align closely related sequences first

→ Gaps inserted in closely related sequences are propogated

Progressive Alignment - involves constructing a succession of pairwise alignments

Drawbacks:

→ Not globally optimal

→ Error made at any stage is propogated throughout the final result

→ Time complexity is high in the worst-case scenario. It happens when the sequences in the set are unrelated

→ The $O(N^3)$ time complexity of Neighbour-Joining makes

⇒ The Related Neighbor Joining technique, which lowers the constraints for linking tree nodes, can lessen the shortcomings of the Progressive alignment approach. As a result, the time complexity is reduced to $O(N^2 \log N)$

→ We could also use Bayesian methods, iterative refinement methods (IRA, IRMSD) that help reduce alignment errors & improve accuracy of phylogenetic reference

④ Given several sequences that are L=50 residues long.

Given that alignment of 4 sequences takes

$(2L)^{N-2} = 10^{2N-4} = 10^4$ seconds

Alignment of N sequences takes time

$= (2L)^{N-2} = 10^{2N-4}$ seconds

Given Time = 5 billion years $= 5 \times 10^9$ years

$= 5 \times 10^9 \times 365$ days

$= 5 \times 10^9 \times 365 \times 86400$ seconds

$= 5 \times 365 \times 864 \times 10^{11}$

$= 15768 \times 10^{13}$ seconds

$10^{2N-4} = 15768 \times 10^{13}$

Applying $\log_{10}$ on both sides,

$2N-4 = 13 + \log_{10} 15768 = 13 + 4.197$

$2N-4 = 17.197$

$2N = 21.197$

$\boxed{N = 10.5985}$

This implies that our computer can align 10 sequences in 5 billion years

⑥ MSA using progressive Approach

S_1S_2:

		G	A	T	T	C	A
	0	-1	-2	-3	-4	-5	-6
G	-1	1	0	-1	-2	-3	-4
T	-2	0	0	1	0	-1	-2
C	-3	-1	-1	0	0	1	0
T	-4	-2	-2	0	1	0	0
G	-5	-3	-3	-1	0	0	-1
A	-6	-4	-2	-2	-1	-1	1

Alignment → GAT_TCA
6_TCTGA

Score = 1
Distance = 3

S_1S_3:

		G	A	T	T	C	A
	0	-1	-2	-3	-4	-5	-6
G	-1	1	0	-1	-2	-3	-4
A	-2	0	2	1	0	-1	-2
T	-3	-1	1	3	2	1	0
A	-4	-2	0	2	2	1	2
T	-5	-3	-1	1	3	2	1
T	-6	-4	-2	0	2	2	1

Two alignments are possible → GAT_TCA
GATAT_T

↳ GAT_TCA
GATATT_

Score = 1

Distance = 3

$S_2 S_3$:

		G	T	C	T	G	A
	0	-1	-2	-3	-4	-5	-6
G	-1	1	0	-1	-2	-3	-4
A	-2	0	0	-1	-2	-3	-2
T	-3	-1	1	0	0	-1	-2
A	-4	-2	0	0	-1	-1	0
T	-5	-3	-1	-1	1	0	-1
T	-6	-4	-2	-2	0	0	-1

Two Alignments are possible:

G-TCTGA | G_TCTGA
GATATT_ | GATAT_T

Score = -1, mismatch = Distance = 4

$S_2 S_4$:

		G	T	C	T	G	A
	0	-1	-2	-3	-4	-5	-6
G	-1	1	0	-1	-2	-3	-4
T	-2	0	2	1	0	-1	-2
C	-3	-1	1	3	2	1	0
A	-4	-2	0	2	2	1	2
G	-5	-3	-1	1	1	3	2
C	-6	-4	-2	0	0	2	2

Alignment: GTCTGA
GTCAGC

Score = 2
mismatch = distance = 2

$S_1 S_4$:

		G	A	T	T	C	A
	0	-1	-2	-3	-4	-5	-6
G	-1	1	0	-1	-2	-3	-4
T	-2	0	0	1	0	-1	-2
C	-3	-1	-1	0	0	1	0
A	-4	-2	0	-1	-1	0	2
G	-5	-3	-1	-1	-2	-1	1
C	-6	-4	-2	-2	-2	-1	0

Two Alignments are possible:

GATTCA__
G--TCAGC

GATTCA__
G-T-CAGC

Score = 0, mismatch = distance = 4

$S_3 S_4$:

		G	A	T	A	T	T
	0	-1	-2	-3	-4	-5	-6
G	-1	1	0	-1	-2	-3	-4
T	-2	0	0	1	0	-1	-2
C	-3	-1	-1	0	0	-1	-2
A	-4	-2	0	-1	1	0	-1
G	-5	-3	-1	-1	0	0	-1
C	-6	-4	-2	-2	-1	-1	-1

Alignment: GAT_ATT
G_TCAGC

Score = -1
mismatch
= distance = 4

Distance matrix

	S_1	S_2	S_3	S_4
S_1		3	3	4
S_2			4	2
S_3				4
S_4				

$D(S_1, S_2) = 3$

$D(S_1, S_4) = 4$

$D(S_1, S_2\text{-}S_4) = 4/2$
$= 3.5$

Grouping S_2 & S_4 as they have lowest mismatches

	S_1	S_2-S_4	S_3
S_1		3.5	3
S_2-S_4			4
S_3			

Grouping S_1 and S_3 as they have lowest number of mismatches

	S_1-S_3	S_2-S_4
S_1-S_3		3.75
S_2-S_4		

For S_1-S_3 and S_2-S_4

		G	A	T	–	T	C	A
		G	A	T	A	T	–	T
	0	-1	-2	-3	-4	-5	-6	-7
G	-1	1	0	-1	-2	-3	-4	-5
G	-1	1	0	-1	-2	-3	-4	-5
T	-2	0	0	1	0	-1	-2	-3
T	-2	0	0	1	0	-1	-2	-3
C	-3	-1	-1	0	0	-1	-1	-2
C	-3	-1	-1	0	0	-1	-1	-2
T	-4	-2	-1	-1	-1/2	0	-1	-1
A	-4	-2	-1	-1	-1/2	0	-1	-1
G	-5	-3	-2	-2	-3/2	-1	-1	-2
G	-5	-3	-2	-2	-3/2	-1	-1	-2
A	-6	-4	-3	-3	-5/2	-2	-3/2	-3/2
C	-6	-4	-3	-3	-5/2	-2	-3/2	-3/2

Final MSA

S_2	G	–	T	C	T	G	A
S_4	G	–	T	C	A	G	C
S_1	G	A	T	–	T	C	A
S_3	G	A	T	A	T	–	T

Final score:

Score of column-1 = 1+1+1+1+1+1
= 6

Score of column-2 = -1-1-1-1-1+1 = -4

Score of column-3 = 1+1+1+1+1+1
= 6

Score of column-4 = 1-1-1-1-1-1 = -4

Score of column-5 = -1+1+1-1-1+1 = 0

Score of column-6 = 1-1-1-1-1-1 = -4

Score of column-7 = -1+1-1-1-1-1 = -4

Final score = 6-4+6-4+0-4-4
= -4

∴ Score of alignment using sum-of-pairs method = -4

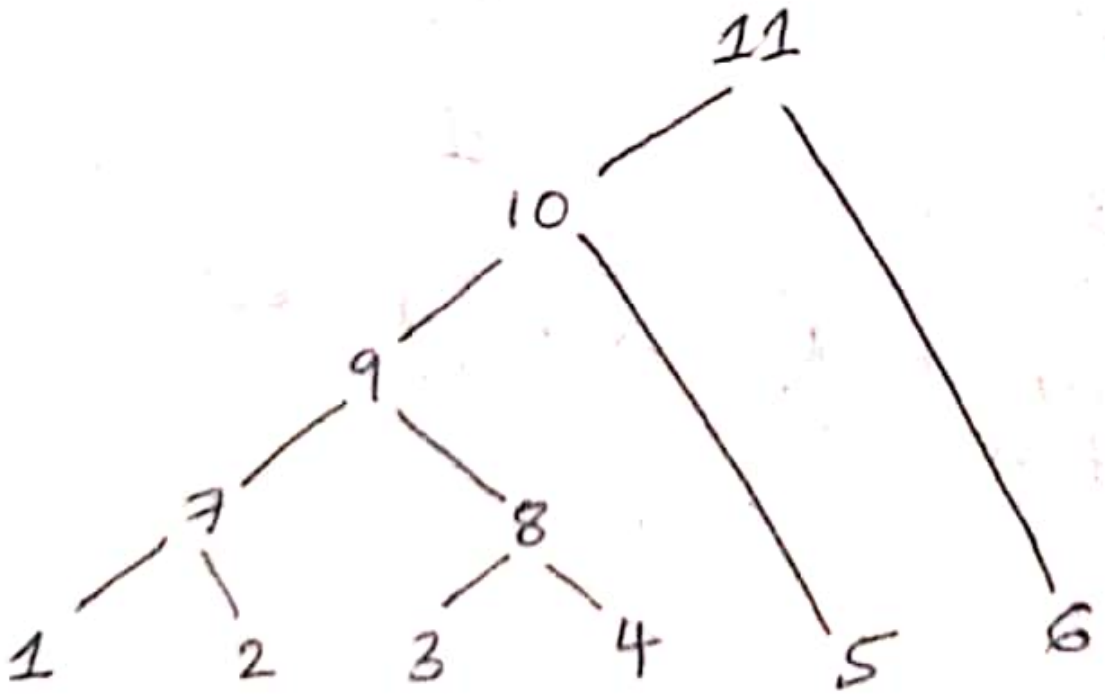
∴ Best Alignment using MSA

```
G - T C T G A   — S2
G - T C A G C   — S4
G A T - T C A   — S1
G A T A T - T   — S3
```

(7)

Species	Site 1	2	3	4
1	T	C	A	A
2	G	C	A	T
3	T	T	T	T
4	G	A	T	A
5	G	A	A	C
6	A	T	A	G

Tree → (((((1,2),(3,4)),5),6)



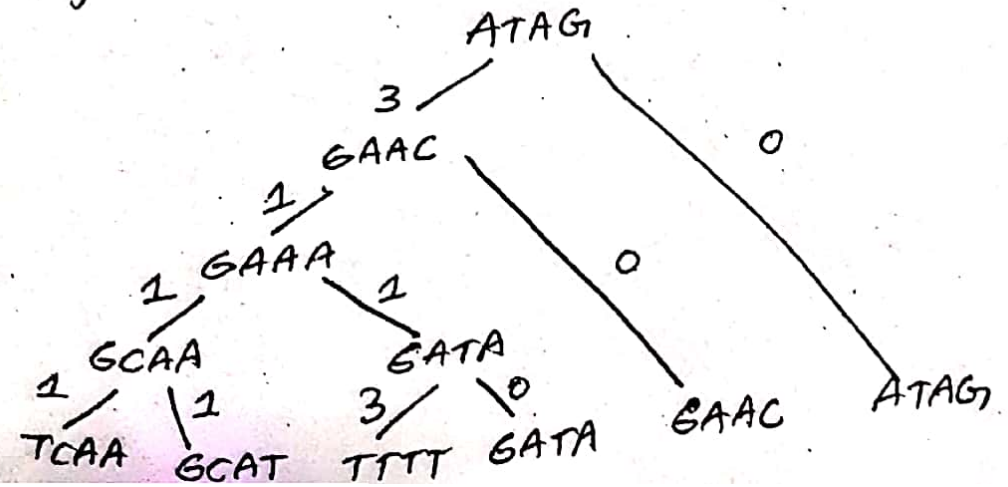
Let X_{iu} denote set of possible alignments at node i and u^{th} position & X_i denotes possible sequences at node i

for $i = 1, \dots, n$ (Leaf nodes)
  for $u = 1, \dots, K$ (K – length sequence)
    X_i = given sequence
    X_{iu} = value of u^{th} position of X_i
    $L_i = 0$ (No. of substitutions upto node i)

for $i = n+1, \dots, 2n-1$
  No_of_changes = 0 (Let X_m, X_n be children of X_i)
  for $u = 1, \dots, K$
    if $X_{mu} \cap X_{nu} == \emptyset$: {
      $X_{iu} = X_{mu} \cup X_{nu}$
      No_of_changes += 1
    }
    else :
      $X_{iu} = X_{mu} \cap X_{nu}$

$L_i = L_m + L_n +$ No_of_changes

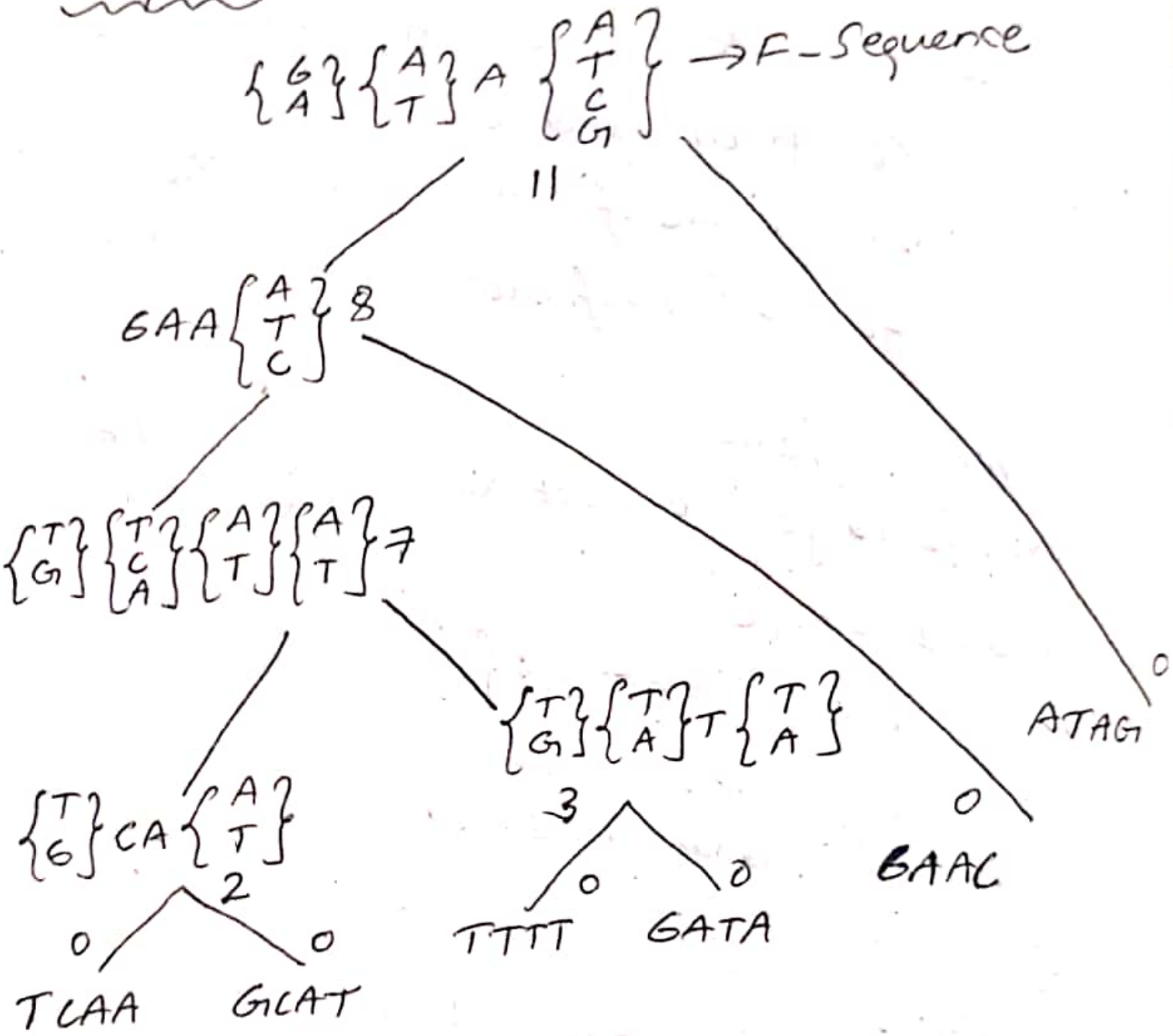
⇒ By following the above method, A possible alignment is as follows



Total mutations = 1 + 1 + 3 + 1 + 1 + 1 + 3

= 11

F-sequence



Parsimony score = Total number of mutations from root

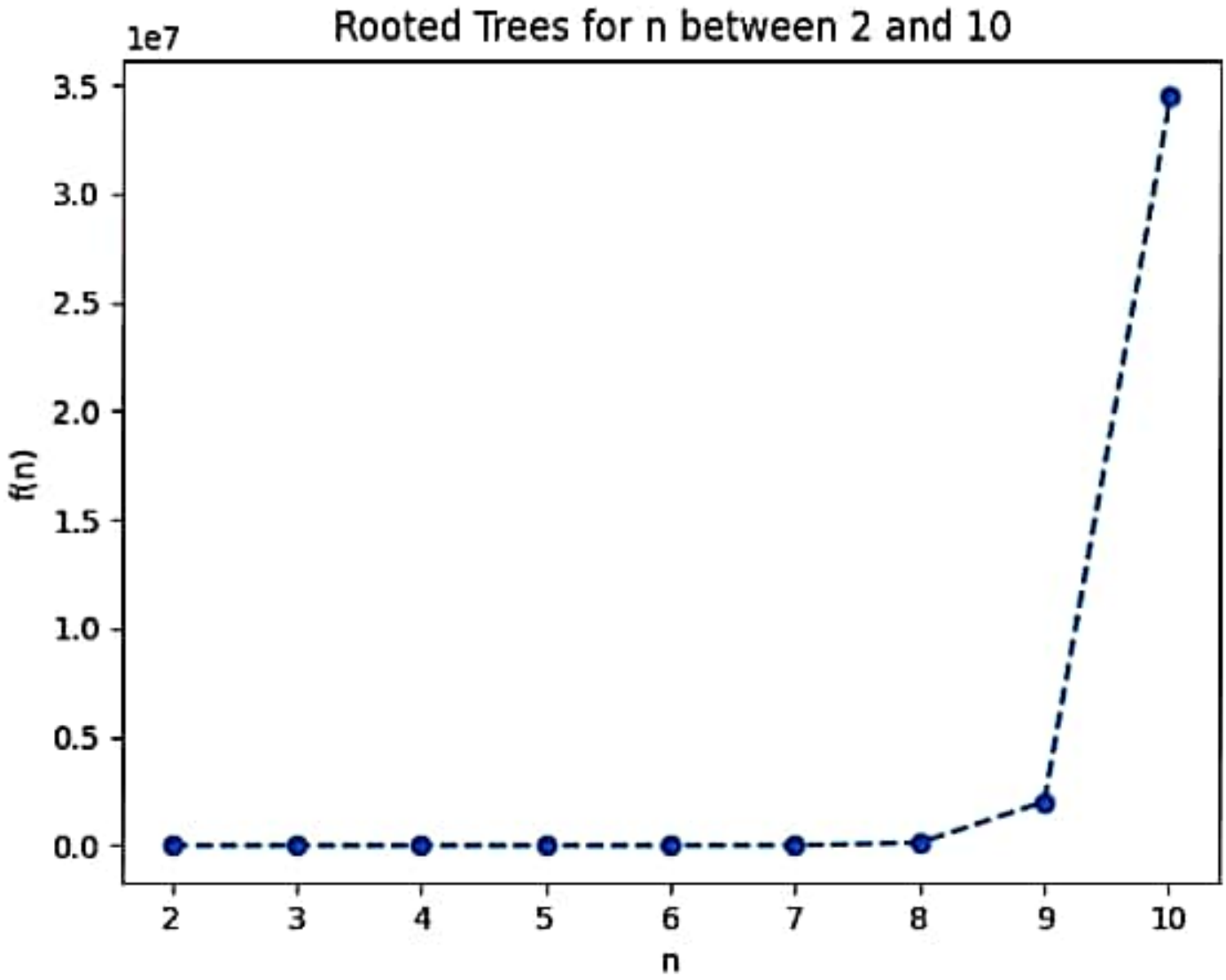
= 11

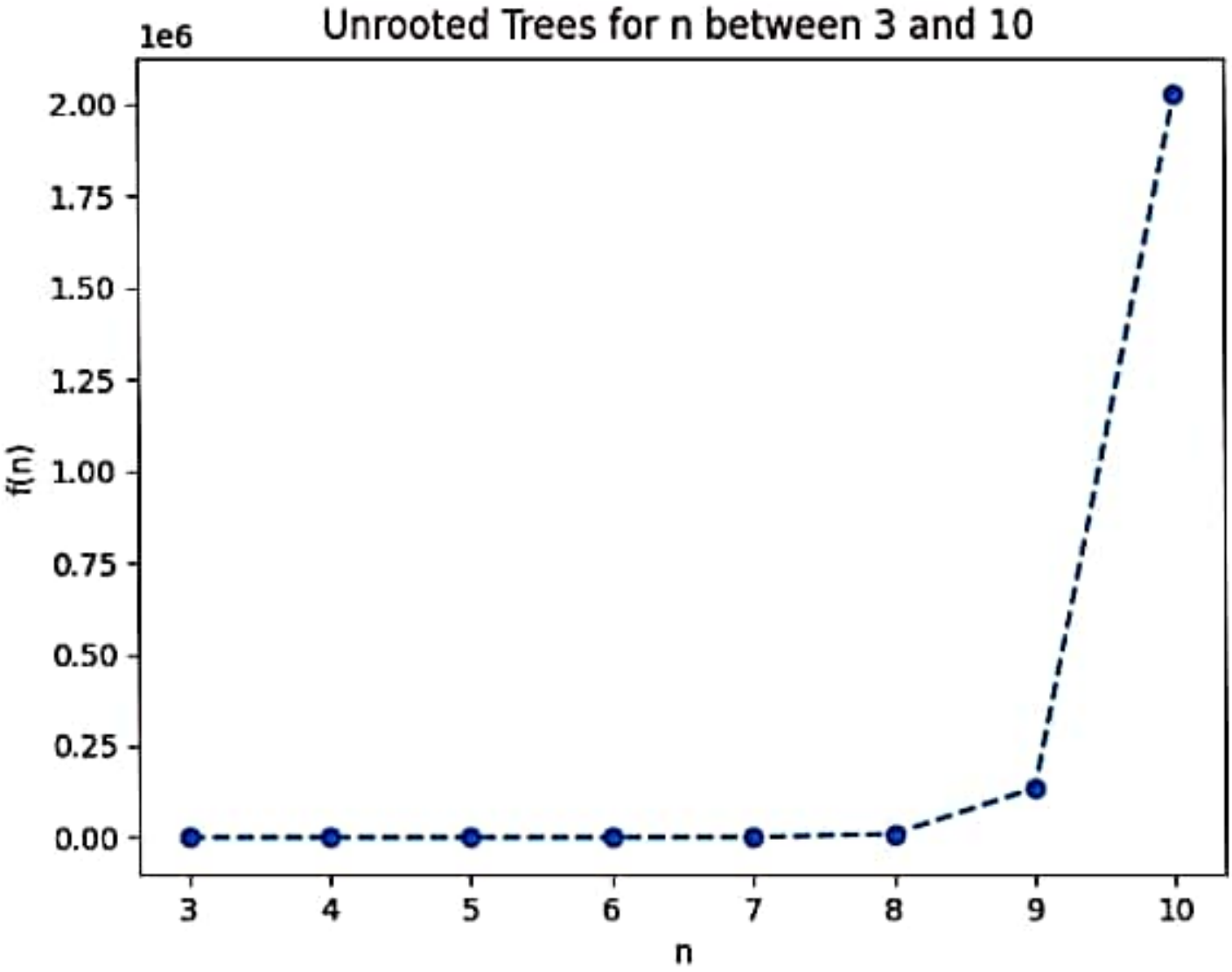
(8) For n terminal taxa,

number of unrooted trees $= \frac{(2n-5)!}{2^{n-3}[(n-3)!]}$

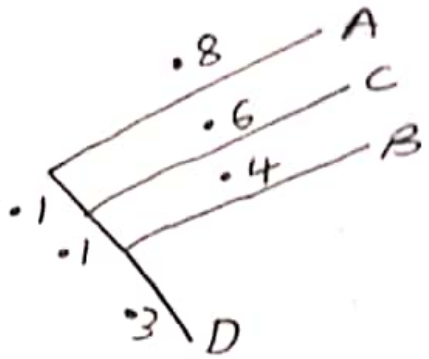
number of rooted trees $= \frac{(2n-3)!}{2^{n-2}(n-2)!}$

Taxa (n)	No. of unrooted trees	No. of rooted trees
1	1	1
2	1	1
3	1	3
4	3	15
5	15	105
6	105	945
7	945	10395
8	10395	135135
9	135135	2027025
10	2027025	34459425

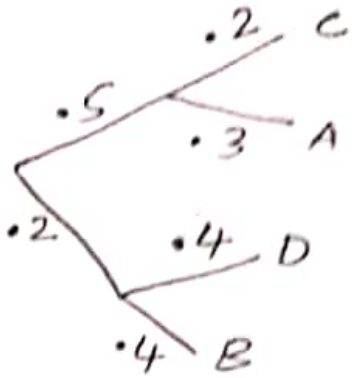




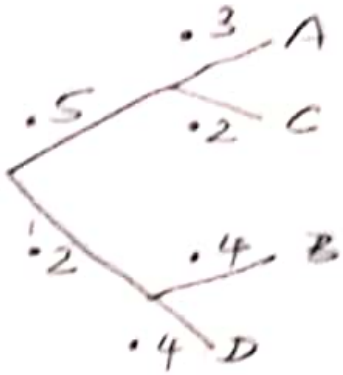
(9)



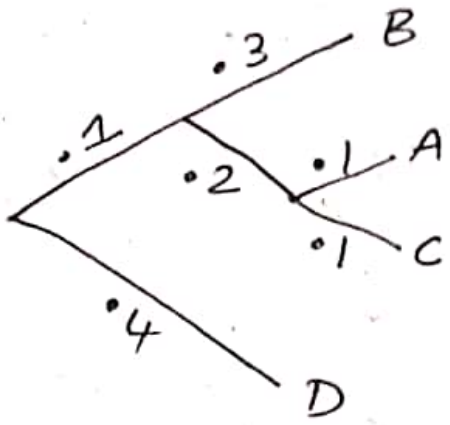
T_1



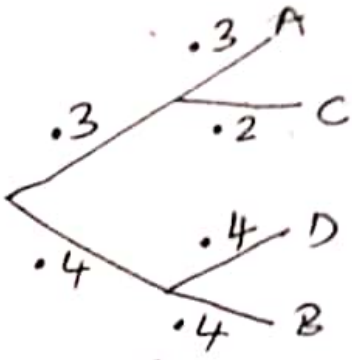
T_2



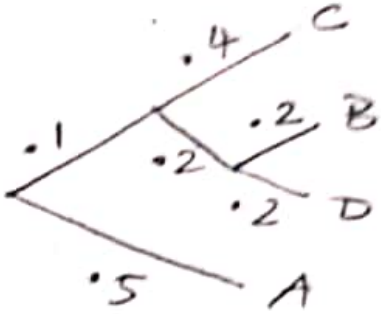
T_3



T_4



T_5



T_6

(9)

D_{xy} denotes distance b/w node x and node y

(a) Consider T_2,

$D_{root-A} = 0.5+0.3 = 0.8$

$D_{root-B} = 0.2+0.4 = 0.6$

$D_{root-C} = 0.5+0.2 = 0.7$

$D_{root-D} = 0.2+0.4 = 0.6$

Consider T_3,

$D_{root-A} = 0.5+0.3 = 0.8$

$D_{root-B} = 0.2+0.4 = 0.6$

$D_{root-C} = 0.5+0.2 = 0.7$

$D_{root-D} = 0.2+0.4 = 0.6$

As we can observe the corresponding distances b/w roots & nodes are same in both Trees.

→ Hence T_2 and T_3 are the same, as rooted metric trees

	T_1	T_2	T_3	T_4	T_5	T_6
A	0.8	0.8	0.8	0.4	0.6	0.5
B	0.6	0.6	0.6	0.4	0.8	0.5
C	0.7	0.7	0.7	0.4	0.5	0.5
D	0.5	0.6	0.6	0.4	0.8	0.5

(b) Distances b/w nodes from their Nearest common predecessor (upto any level) are considered for unrooted metric trees

Distance b/w A and C in

$T_1(D,A,C) = 1.5$

$T_2(D,A,C) = 0.5$

$T_3(D,A,C) = 0.5$

$T_4(D,A,C) = 0.2$

$T_5(D,A,C) = 0.5$

$T_6(D,A,C) = 1$

} T_1, T_4, T_6 don't match with others and hence need not be checked for other combinations

---

$T_2(D_A,D) = T_2(D_A,B) = 1.4$

$T_3(D_A,D) = T_3(D_A,B) = 1.4$

$T_5(D_A,D) = T_5(D_A,B) = 1.4$

$T_2(D_B,C) = T_3(D_B,C) = T_5(D_B,C) = 1.3$

$T_2(D_B,D) = T_3(D_B,D) = T_5(D_B,D) = 0.8$

$T_2(D_C,D) = T_3(D_C,D) = T_5(D_C,D) = 1.3$

⇒ Hence T_2, T_3, T_5 are the same, as unrooted metric trees

(c)

	T_1	T_2	T_3	T_4	T_5	T_6
A	1	2	2	3	2	1
B	3	2	2	2	2	3
C	2	2	2	3	2	2
D	3	2	2	1	2	3

Table representing Levels of A, B, C, D in various trees [It is assumed that root is at level 0

From the table,

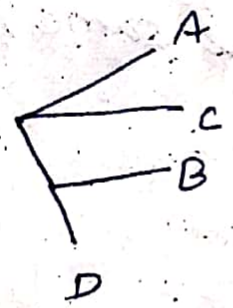
→ T_2 and T_6 have their corresponding nodes A, B, C, D at same levels

→ T_2, T_3, T_5 have their corresponding nodes A, B, C, D at same levels

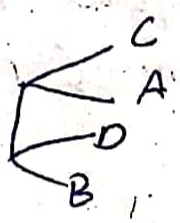
(T_1, T_6) and (T_2, T_3, T_5) are the same, as rooted Topological trees

(d) Two trees are said to be unrooted topological if they result in the same skeleton after their ancestor is removed and they are connected with an edge

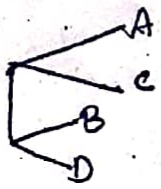
$(T_1)_{modify}$ =



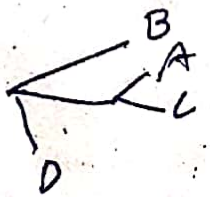
$(T_2)_{modified}$ =



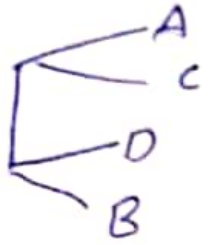
$(T_3)_{modify}$ =



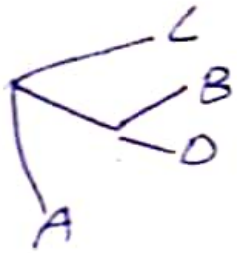
$(T_4)_{modify}$ =



$(T_5)_{modify}$ =



$(T_6)_{modify}$ =



From the above modified structures, we can conclude that they are all the same skeletons

⇒ $(T_1, T_2, T_3, T_4, T_5, T_6)$ are the same, as unrooted Topological trees

(e) For a molecular clock to operate, distance b/w nodes & roots should be same for all nodes

	T_1	T_2	T_3	T_4	T_5	T_6
A	0.8	0.8	0.8	0.4	0.6	0.5
B	0.6	0.6	0.6	0.4	0.8	0.5
C	0.7	0.7	0.7	0.4	0.5	0.5
D	0.5	0.6	0.6	0.4	0.8	0.5

From the above Table, we can conclude that

$(T_D, X)_4 = (T_D, Y)_4$ where X, Y are any nodes in T_4, T_4

$(T_D, X)_6 = (T_D, Y)_6$ where X, Y are any nodes in T_6

→ Hence for (T_4, T_6) a molecular clock appears to be operating