Bioinformatics: Phylogeny Assignment

1. i.) **Case 1:** Ancestral sequence is S_2

We know, $f_b = f_{bA} + f_{bT} + f_{bG} + f_{bC}$

where $f_b \rightarrow$ frequency of base 'b' in the ancestral sequence

$\Rightarrow f_A = 150$, $f_G = 250$, $f_C = 350$, $f_T = 250$

Thus, $P_0 = (P_A \quad P_G \quad P_C \quad P_T) = (0.15 \quad 0.25 \quad 0.35 \quad 0.25)$

& $M = \begin{bmatrix} P_{A|A} & P_{A|G} & P_{A|C} & P_{A|T} \\ P_{G|A} & P_{G|G} & P_{G|C} & P_{G|T} \\ P_{C|A} & P_{C|G} & P_{C|C} & P_{C|T} \\ P_{T|A} & P_{T|G} & P_{T|C} & P_{T|T} \end{bmatrix} = \begin{bmatrix} 0.7 & 0.1 & 0.1 & 0.1 \\ 0.1 & 0.7 & 0.1 & 0.1 \\ 0.1 & 0.1 & 0.7 & 0.1 \\ 0.1 & 0.1 & 0.1 & 0.7 \end{bmatrix}$

Here, M is a Jukes-Cantor matrix with $\boxed{\alpha = 0.3}$

However, $Mp_0 \neq P_0 \Rightarrow P_0$ is not a equilibrium distribution for M

ii.) Case 2: Ancestral sequence is S_β

$$\Rightarrow f_A = 190 \;,\; f_G = 250,\; f_C = 310 \;,\; f_T = 250$$

Thus, $P_0 = (P_A \;\; P_G \;\; P_C \;\; P_T) = (0.19 \;\; 0.25 \;\; 0.31 \;\; 0.25)$

& $M = \begin{bmatrix} P_{A|A} & P_{A|G} & P_{A|C} & P_{A|T} \\ P_{G|A} & P_{G|G} & P_{G|C} & P_{G|T} \\ P_{C|A} & P_{C|G} & P_{C|C} & P_{C|T} \\ P_{T|A} & P_{T|G} & P_{T|C} & P_{T|T} \end{bmatrix} = \begin{bmatrix} 0.552 & 0.06 & 0.04 & 0.06 \\ 0.131 & 0.7 & 0.08 & 0.1 \\ 0.184 & 0.14 & 0.79 & 0.14 \\ 0.131 & 0.1 & 0.08 & 0.7 \end{bmatrix}$

Here, M is not a Jukes-Cantor matrix

Further, $Mp_0 \neq P_0 \Rightarrow P_0$ is not a equilibrium distribution for M

2. 1.) UPGMA :

Step 1: Merging sequences S_1 & S_2

	S_1/S_2	S_3	S_4	S_5
S_1/S_2	0	1.005	0.72	0.965
S_3	—	0	0.62	0.42
S_4	—	—	0	0.37

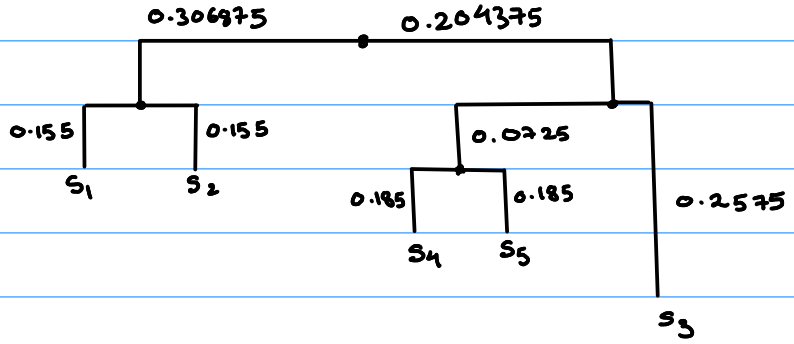
Step 2: Merging sequences S_4 & S_5

	S_1/S_2	S_3	S_4/S_5
S_1/S_2	0	1.005	0.8425
S_3	—	0	0.515

Step 3: Merging sequences S_3 & S_4/S_5

	S_1/S_2	$S_3/S_4/S_5$
S_1/S_2	0	0.92375

Tree:

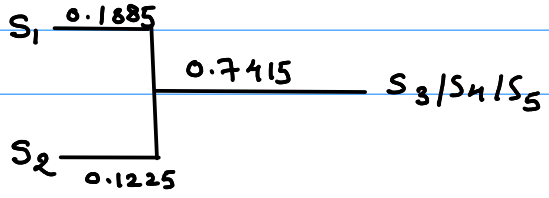


2.) FM Method:

Step 1: Calculating distance of S_1 & S_2, by clubbing S_3, S_4 & S_5

	S_1	S_2	$S_3/S_4/S_5$
S_1	0	0.31	0.93
S_2		0	0.863

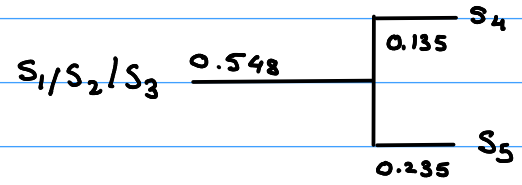
Tree:



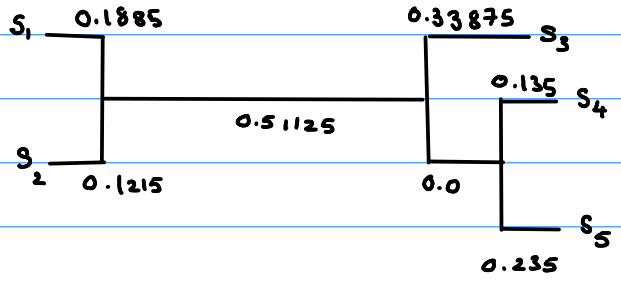
Step 2: Calculating distance of S_4 & S_5, by clubbing S_1, S_2 & S_3

	$S_1/S_2/S_3$	S_4	S_5
$S_1/S_2/S_3$	0	0.683	0.783
S_4		0	0.37

Tree:



Final Tree:



3.) **NJ Method:**

Step 1: Merging sequences S_1 & S_2

	S_1/S_2	S_3	S_4	S_5
S_1/S_2	0	1.005	0.72	0.965
S_3	—	0	0.62	0.42
S_4	—	—	0	0.37

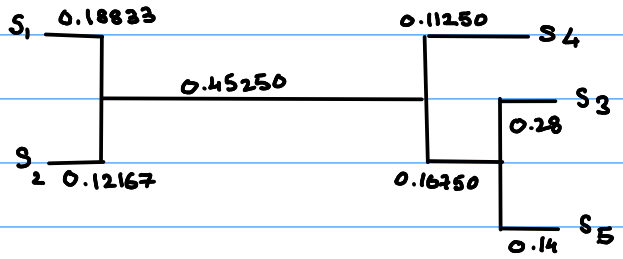
Step 2: Merging sequences S_3 & S_5

	S_1/S_2	S_4	S_3/S_5
S_1/S_2	0	0.72	0.6328
S_4	—	0	0.379

Step 3: Merging sequences S_4 & S_3/S_5

	S_1/S_2	$S_4/S_3/S_5$
S_1/S_2	0	0.92375

Final Tree:

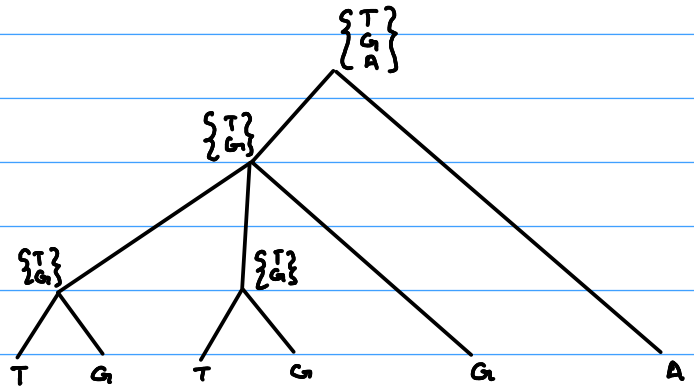


4.) We get topologically same Tree by UPGMA & FM, but not through NJ

5) ∘ FM & UPGMA are distance-based methods.
- FM does not enforce the Molecular Clock Assumption, unlike UPGMA.
- Thus, FM provides much more accurate results, albeit with slower speeds.

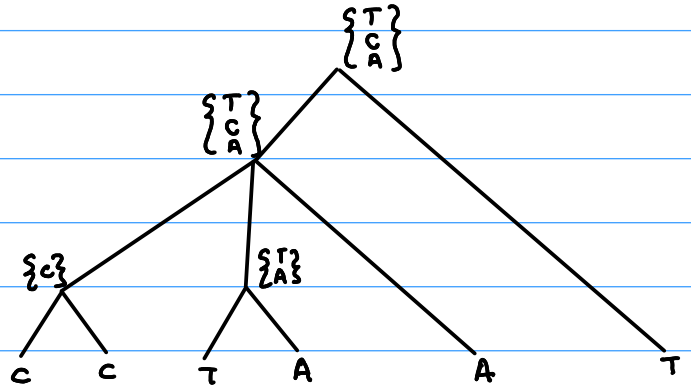
6.) ∘ NJ method is comparitively rapid & gives more accurate results than UPGMA
- It allows for unequal branch lengths, which provides a more accurate picture.

3. For i=0



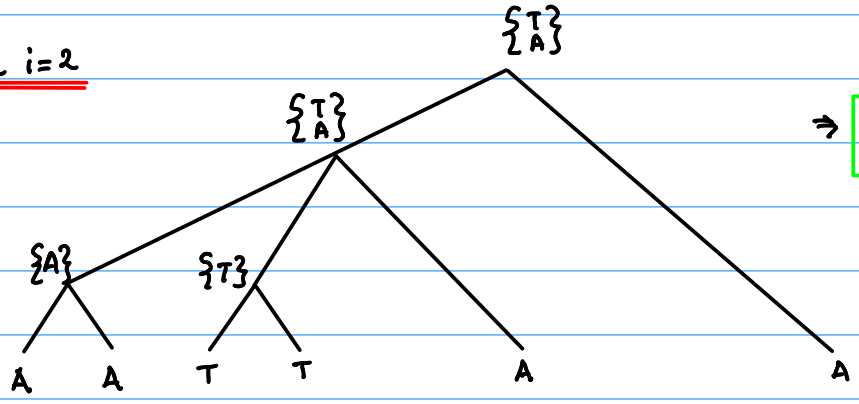
⇒ Parsimony = 4

For i=1



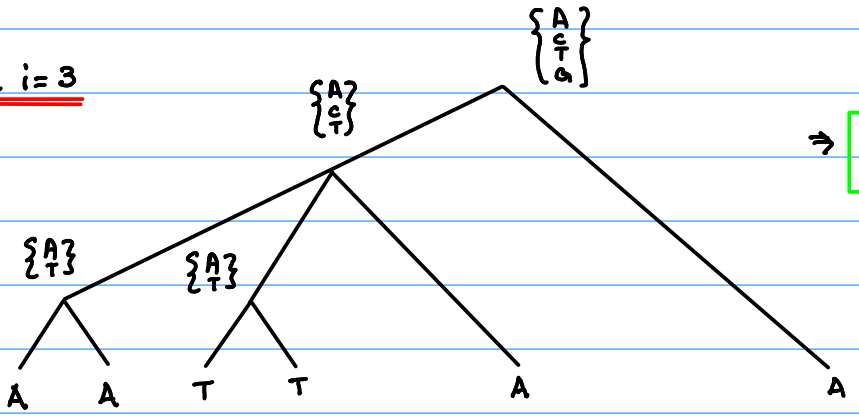
⇒ Parsimony = 3

For i=2



⇒ Parsimony = 2

For i=3



⇒ Parsimony = 4

Thus, Net Parsimony = 13 & F sequence = {T, G, A}, {T, C, A}, {T, A}, {A, C, T, G}

4. ∘ The program (code) is provided in the file `Q5.py`.

∘ The parameters are stored in the file `parameters.py`.

∘ The results generated is provided as `Q4.txt`. Additionally, the results are generated during run-time of the program.

5. ∘ The program (code) is provided in the file `Q5.py`.

∘ The plot generated is provided as `Q5.png`. Additionally, the plot is generated during run-time of the program.

6. ∘ The program (code) is provided in the file `Q6.py`.

∘ The results generated is provided as `output.txt`. Additionally, the results are generated during run-time of the program, and are rewritten to `output.txt`.

∘ Distance matrix from Q2 has been used (since there was no distance matrix for Q3), which is sent as input through `input.txt`, which is a csv file with delimiter '|'.