Genome Informatics 2021 Fall Final exam.

Problem 1.
Assume an organism evolves with constant substitution rates such that after $t = 1000$ generations the probabilities of the form $\mathrm{P}[\mathrm{a}|\mathrm{a}, t = 1000]$, $\mathrm{P}[\mathrm{a}|\mathrm{c}, t = 1000]$, etc. are as shown.

$$S(t = 1000) \;=\; \left[\begin{array}{ccccc} & \mathrm{c} & \mathrm{t} & \mathrm{a} & \mathrm{g} \\ \mathrm{c} & 0.95 & 0.03 & 0.01 & 0.01 \\ \mathrm{t} & 0.03 & 0.95 & 0.01 & 0.01 \\ \mathrm{a} & 0.01 & 0.01 & 0.95 & 0.03 \\ \mathrm{g} & 0.01 & 0.01 & 0.03 & 0.95 \end{array}\right]$$

Task: fill in the numbers for $S(t = 2000)$

	c	t	a	g
c	0.9036	0.0572	0.0196	0.0196
t	0.0572	0.9036	0.0196	0.0196
a	0.0196	0.0196	0.9036	0.0572
g	0.0196	0.0196	0.0572	0.9036

Solution:

$$S(2000) = S(1000) \times S(1000)$$

So it suffices to square the matrix $S(1000)$.
For example, if the base starts out as g. After the first 1000 generations its probability mass should be distributed as: g:0.95, a:0.03, c:0.01, t:0.01. So the probability it is, say a, after another 1000 generations is:

g⟶g⟶a   g⟶a⟶a   g⟶c⟶a   g⟶t⟶a
(0.95)(0.03) + (0.03)(0.95) + (0.01)(0.01) + (0.01)(0.01) = 0.0572

Problem 2.

1. Describe precisely the set of DNA strings accepted by the following context free grammar and the size of that set (in other words, how many distinct strings are accepted by that grammar).
2. Give an example of one of the strings, sketch it as a stem-loop structure and show its derivation from the grammar.

$$\begin{aligned}
S &\longrightarrow cW_1g \mid gW_1c\\
W_1 &\longrightarrow aW_2u \mid uW_2a\\
W_2 &\longrightarrow aW_3u \mid uW_3a\\
W_3 &\longrightarrow cW_4g \mid gW_4c\\
W_4 &\longrightarrow aW_5 \mid cW_5 \mid gW_5 \mid uW_5\\
W_5 &\longrightarrow aW_6 \mid cW_6 \mid gW_6 \mid uW_6\\
W_6 &\longrightarrow aW_7 \mid cW_7 \mid gW_7 \mid uW_7\\
W_7 &\longrightarrow a \mid c \mid g \mid u
\end{aligned}$$

Solution: Any string of length 8 with 4 residues forming a stem loop such that the inner two pairs are a:u or u:a, and the outer two pairs are c:g or g:c. There are $(2^4)(4^4) = 4096$. Example sequence and derivation: cuag cgag cuag, where the loop is cgag, and the stem is:

```
5'   3'
 c ≡ g
 u = a
 a = u
 g ≡ c
c     g
  g  a
```

And the derivation is:

$$S \longrightarrow cW_1g \longrightarrow cuW_2ag \longrightarrow cuaW_3uag \longrightarrow cuagW_4cuag \longrightarrow cuagW_4cuag \longrightarrow cuagcW_5cuag$$

$$cuagcgagcuag \longleftarrow cuagcgaW_7cuag \longleftarrow cuagcgW_6cuag \longleftarrow cuagcW_5cuag$$

Problem 3.

A

	w	x	y
x	8		
y	10	6	
z	7	11	13

B

	w	x	y
x	8		
y	8	4	
z	6	8	8

C

	w	x	y
x	8		
y	13	9	
z	6	9	16

The three matrices above indicate distances between 4 species $\{w, x, y, z\}$. Two of the three matrices hold additive distances and one of those is also ultrametric.

Tasks:
1. Determine which of the three matrices {**A**,**B**, **C**} are additive and which one is ultrametric (and show the reason).
2. Use UPGMA to infer a phylogenetic tree (with branch lengths) of the one that is ultrametric. Show your calculation.
3. Use Neighbor-Joining to infer a phylogenetic tree (with branch lengths) of the one which is additive but not ultrametric. Show at least the first iteration of the neighbor-joining algorithm.

Solution: First we test for the ultrametric condition.

triples	Distances	Mat. **A**	Mat. **B**	Mat. **C**
w,x,y	(w,x):(w,y):(x,y)	8:10:6	8:8:4✓	8:13:9
w,x,z	(w,x):(x,z):(w,z)		8:8:6✓	
x,y,z	(x,y):(x,z):(y,z)		4:8:8✓	

Where ✓ marks distances forming an acute isoceles triangle. Note it is sufficient to find one counter-example. Only **B** is ultrametric. Since **B** is ultrametric, it is also additive. Next we test **A** and **C** for additivity. The additivity test requires that following distances form an obtuse isoceles triangle.

$$d_{xy} + d_{yz} : d_{wy} + d_{xz} : d_{wz} + d_{xy}$$

Let's try this for matrices **A** and **C**.

Matrix	$d_{xy} + d_{yz} : d_{wy} + d_{xz} : d_{wz} + d_{xy}$	
A	$8 + 13 : 10 + 11 : 7 + 6$	$= 21 : 21 : 11$✓
C	$9 + 16 : 13 + 9 : 6 + 9$	$= 25 : 22 : 15$

So **A** is additive but **C** is not.

UPGMA on Matrix B

Recall the distance matrix **B** is:

$$\mathbf{B} = \begin{array}{c|ccc} & w & x & y \\ \hline x & 8 & & \\ y & 8 & 4 & \\ z & 6 & 8 & 8 \\ \hline \end{array}$$

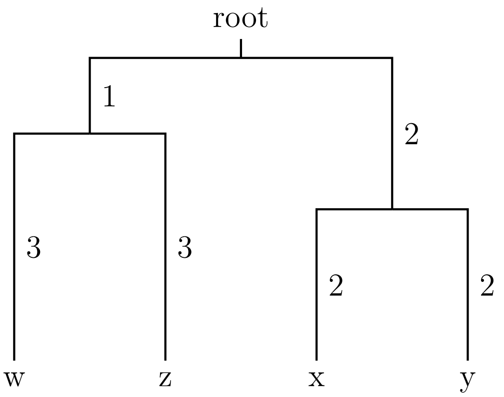
The first step of UPGMA would be to merge the minimum distance (x, y) into a cluster (parent node). Let's call the merged node $C_{xy} : \{x, y\}$. The height of C_{xy} should be $\frac{1}{2}\, d(x, y) = 2$. Distances from w to C_{xy} are calculated according to:

$$d(w, C_{xy}) = \text{avg}\big(d(w, x),\, d(w, y)\big) = \text{avg}(8, 8) = 8$$
$$d(z,\, C_{xy}) = \text{avg}\big(d(z,\, x),\, d(,\, zy)\big) = \text{avg}(8, 8) = 8$$

These are both greater than $d(w, z) = 6$, so in the next step we would create a cluster. C_{wz} at height $\frac{1}{2}\, d(w, z) = 3$.
Finally to compute the height of the root, we can compute $d(C_{xy}, C_{wz})$.

$$d(C_{xy}, C_{wz}) = \text{avg}_{i \in \{x,y\}, j \in \{w,z\}} d(i, j) = \text{avg}(8, 8, 8, 8) = 8$$

So the root is at height $(\frac{1}{2})(8) = 4$.



Rooted tree inferred from distances in matrix **B** by UPGMA.

Neighbor-Joining on Matrix A

For convenience, I copy the distance matrix **A** here:

$$\mathbf{A} = \begin{array}{c|ccc} & w & x & y \\ \hline x & 8 & & \\ y & 10 & 6 & \\ z & 7 & 11 & 13 \\ \hline \end{array}$$

The first step is to compute r_i (a kind of average distance involving taxon i) for each taxon. The general formula is:

$$r_i = \frac{\Sigma_{j \neq i} d(i,j)}{(n-2)}$$

where n is the number of taxa, in this problem $n = 4$.

$$\begin{aligned}
r_w &= \tfrac{1}{2}\big(d(w,x) + d(w,y) + d(w,z)\big) = \tfrac{1}{2}(\ 8 + 10 + 7) = \tfrac{1}{2}(25) = 12.5 \\
r_x &= \tfrac{1}{2}\big(d(x,w) + d(x,y) + d(x,z)\big) = \tfrac{1}{2}(\ 8 + \ 6 + 11) = \tfrac{1}{2}(25) = 12.5 \\
r_y &= \tfrac{1}{2}\big(d(y,w) + d(y,x) + d(y,z)\big) = \tfrac{1}{2}(10 + \ 6 + 13) = \tfrac{1}{2}(29) = 14.5 \\
r_z &= \tfrac{1}{2}\big(d(z,w) + d(z,x) + d(z,y)\big) = \tfrac{1}{2}(\ 7 + 11 + 13) = \tfrac{1}{2}(31) = 15.5
\end{aligned}$$

Next compute D_{ij} according to:

$$D_{ij} = d(i,j) - r_i - r_j$$

$$\mathbf{D} = \begin{array}{c|ccc} & w & x & y \\ \hline x & d(w,x) - r_w - r_x & & \\ y & d(w,y) - r_w - r_y & d(x,y) - r_x - r_y & \\ z & d(w,z) - r_w - r_z & d(x,z) - r_x - r_z & d(y,z) - r_y - r_z \\ \hline \end{array}$$

$$\mathbf{D} = \begin{array}{c|ccc} & w & x & y \\ \hline x & 8 - 12.5 - 12.5 & & \\ y & 10 - 12.5 - 14.5 & 6 - 12.5 - 14.5 & \\ z & 7 - 12.5 - 15.5 & 11 - 12.5 - 15.5 & 13 - 14.5 - 15.5 \\ \hline \end{array}$$

$$\mathbf{D} = \begin{array}{c|ccc} & w & x & y \\ \hline x & -17.0 & & \\ y & -17.0 & -21.0 & \\ z & -21.0 & -17.0 & -17.0 \\ \hline \end{array}$$

The smallest of these are D_{xy} and D_{wz}. Here we choose D_{xy}, introducing a parent node of x and y (let's name it a) with edge lengths from x and y defined by:

$$\begin{aligned} d(x,a) &= \tfrac{1}{2}(d_{xy} + r_x - r_y) = \tfrac{1}{2}(6 + 12.5 - 14.5) = \tfrac{1}{2}(6-2) = 2 \\ d(y,a) &= \tfrac{1}{2}(d_{xy} + r_y - r_x) = \tfrac{1}{2}(6 + 14.5 - 12.5) = \tfrac{1}{2}(6+2) = 4 \end{aligned}$$

Then compute the distances between the new parent node a and the other tree nodes, according to:

$$d(a,k) = \tfrac{1}{2}(d(x,k) + d(y,k) - d(x,y))$$

Plugging in the numbers gives:

$$\begin{aligned} d(a,w) &= \tfrac{1}{2}(d(x,w) + d(y,w) - d(x,y)) = \tfrac{1}{2}(\ 8 + 10 - 6) = \tfrac{1}{2}(12) = 6 \\ d(a,z) &= \tfrac{1}{2}(d(x,z) + d(y,z) - d(x,y)) = \tfrac{1}{2}(11 + 13 - 6) = \tfrac{1}{2}(18) = 9 \end{aligned}$$

In the new tree distances become:

	w	a
a	6	
z	7	9

The new or updated r's are:

$$\begin{aligned} r_w &= d_{wa} + d_{wz} = 6 + 7 = 13 \\ r_z &= d_{zw} + d_{za} = 7 + 9 = 16 \\ r_a &= d_{aw} + d_{az} = 6 + 9 = 15 \end{aligned}$$

$$\mathbf{D} = \begin{array}{c|cc} & w & a \\ \hline a & d(w,a) - r_w - r_a & \\ z & d(w,z) - r_w - r_z & d(a,z) - r_a - r_z \\ \hline \end{array}$$

$$= \begin{array}{c|cc} & w & a \\ \hline a & 6 - 13 - 15 = -22 & \\ z & 7 - 13 - 16 = -22 & 9 - 15 - 16 = -22 \\ \hline \end{array}$$

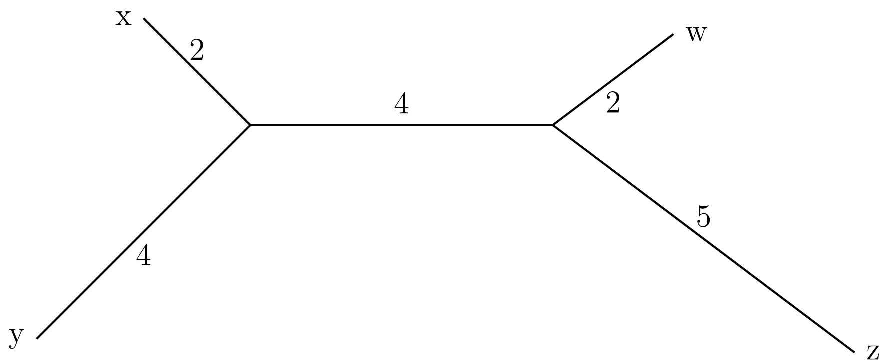
One of the minimum corrected distance is $D(w,z) = -22$, so w and z should be joined at a node b. The edge lengths from w and z to b can be calculated with:

$$\begin{aligned} d(w,b) &= \tfrac{1}{2}(d_{wz} + r_w - r_z) = \tfrac{1}{2}(7 + 13 - 16) = \tfrac{1}{2}(7-3) = 2 \\ d(z,b) &= \tfrac{1}{2}(d_{wz} + r_z - r_w) = \tfrac{1}{2}(7 + 16 - 13) = \tfrac{1}{2}(7+3) = 5 \end{aligned}$$

And distance to a as computed by:

$$d_{ab} = \tfrac{1}{2}(d(aw) + d(az) - d(wz)) = \tfrac{1}{2}(6 + 9 - 7) = \tfrac{1}{2}(8) = 4$$

At this point we know the tree topology ((xy)(wz)), and all the edge lengths.✓



Unrooted tree inferred from distances in matrix **A** by neighbor-joining.