Final exam, Genome Informatics 20221228 *Write your name on each sheet.*

Name & student ID: _____

	B	**C**	**D**
A	9	6	4
B	x	13	9
C	x	x	8

Problem 1.
Use UPGMA to infer a possible tree for leaves **A**,**B**,**C**,**D** (topology & edge lengths) from the distance matrix shown. Discuss if the inferred tree is reasonable.

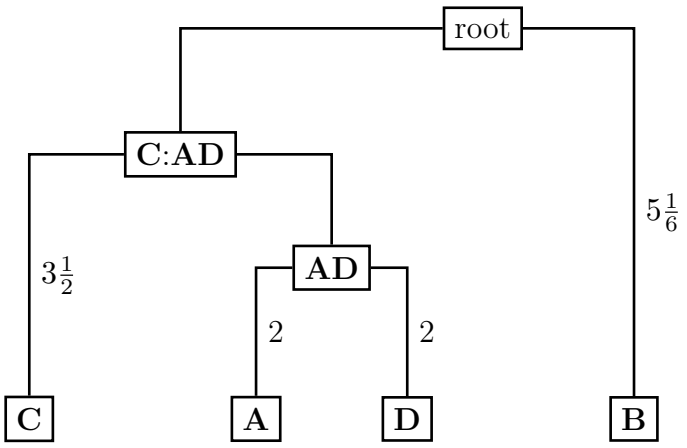
Solution: We start with each leaf as its own cluster, positioned at a "height" of zero from the bottom of the tree. First we the nearest cluster (leaf) pair, which is $\texttt{dist}(\mathbf{A},\mathbf{D}) = 4$. Let's call that node **AD**, and note that its "height" from the bottom of the tree should be $4/2 = 2$.
Now we compute distances from the internal node **AD**, to the leaves **C** and **D**.

$$\texttt{dist}(\mathbf{B},\mathbf{AD}) = \texttt{avg}\{\texttt{dist}(\mathbf{B},\mathbf{A}),\texttt{dist}(\mathbf{B},\mathbf{D})\} = \texttt{avg}\{9,9\} = 9$$
$$\texttt{dist}(\mathbf{C},\mathbf{AD}) = \texttt{avg}\{\texttt{dist}(\mathbf{C},\mathbf{A}),\texttt{dist}(\mathbf{C},\mathbf{D})\} = \texttt{avg}\{6,8\} = 7$$

The other distance is $\texttt{dist}(\mathbf{B},\mathbf{C}) = 13$. $\texttt{dist}(\mathbf{C},\mathbf{AD}) = 7$ is the smallest of the three, so we merge it to form **C**:**AD** with a height of $7/2 = 3.5$. Distance between **B** and **C**:**AD** is the average distance between **B** and any sequence in {**A**,**C**,**D**}.

$$\texttt{dist}(\mathbf{B},\mathbf{C:AD}) = \texttt{avg}\{\texttt{dist}(\mathbf{B},\mathbf{C}),\texttt{dist}(\mathbf{B},\mathbf{A}),\texttt{dist}(\mathbf{B},\mathbf{D})\} = \texttt{avg}\{13,9,9\} = 10\tfrac{1}{3}$$

Finally the root node must be the direct ancestor of **B** and **C**:**AD**, and should be placed at a height of $10\frac{1}{3}/2 = 5\frac{1}{6}$.



Is the tree reasonable? Not really, because the distances implied by the tree are different from the given distance matrix:

	B	**C**	**D**
A	9	6	4
B	x	13	9
C	x	x	8

Given Distances

	B	**C**	**D**
A	7	9	4
B	x	9	7
C	x	x	9

Distances implied by tree.

We could have predicted this in advance by noting that the distances given in the distance matrix are not ultrametric. For example the distances between **A**, **B**, and **C** are all different {9,6,13}, but to be an ultrametric two of those distances would have to be the same and the third one smaller (or equal).

Name & student ID: _____

	B	**C**	**D**
A	9	6	4
B	x	13	9
C	x	x	8

Problem 2.
Use Saitou Nei Neighbor Joining (NJ) to infer a possible tree for leaves **A**,**B**,**C**,**D** (topology & edge lengths) from the distance matrix shown.

Solution: For a set of leaves L of size n, with distances $\forall_{i \neq j} \mathtt{D}(i,j) > 0$, $\forall_i \mathtt{D}(i,i) \equiv 0$; NJ merges according to minimizing a kind of normalized distance known as the neighbor joining function:
In mathematical notation, merge the two leaves x and y which minimize:

$$\mathtt{NJ}(x,y) \stackrel{\text{def}}{=} (n-2)\mathtt{D}(x,y) - (\sigma_x + \sigma_y)$$

Where we define $\sigma_i \stackrel{\text{def}}{=} \sum_{i \in T} \mathtt{D}(i,x)$ as the total distance from leaf x to any other leaf. (the book Biological Sequence Analysis uses notation $r_x \stackrel{\text{def}}{=} \frac{\sigma_x}{n-2}$)

$$\sigma_{\mathbf{A}} = 9 + 6 + 4 = 19 \qquad \sigma_{\mathbf{C}} = 6 + 13 + 8 = 27$$
$$\sigma_{\mathbf{B}} = 9 + 13 + 9 = 31 \qquad \sigma_{\mathbf{D}} = 4 + 9 + 8 = 21$$

	B	**C**	**D**
A	(2)(9) - (19 + 31) = -32	(2)(6) - (19 + 27) = -34	(2)(4) - (19 + 21) = -32
B	x	(2)(13) - (31 + 27) = -32	(2)(9) - (31 + 21) = **-34**
C	x	x	(2)(8) - (27 + 21) = -32

NJ function value of each pair of leaves

The minimal (most negative) value is -34, so leaves **B**,**D** should be assigned a parent which we denote as **BD**. **BD** should be added to the tree with a distance of:

$$\mathtt{dist}(\mathbf{B},\mathbf{BD}) = (\tfrac{1}{2})\Big(\mathtt{dist}(\mathbf{B},\mathbf{D}) - \frac{\sigma_{\mathbf{D}} - \sigma_{\mathbf{B}}}{n-2}\Big) = (\tfrac{1}{2})\Big(9 - \frac{21-31}{4-2}\Big) = 7$$
$$\mathtt{dist}(\mathbf{D},\mathbf{BD}) = (\tfrac{1}{2})\Big(\mathtt{dist}(\mathbf{B},\mathbf{D}) - \frac{\sigma_{\mathbf{B}} - \sigma_{\mathbf{D}}}{n-2}\Big) = (\tfrac{1}{2})\Big(9 - \frac{31-21}{4-2}\Big) = 2$$

Distances from other leaves to **BD** are the average distance from the leaf to **B** and **D**, minus the average distance from **BD** to **B** and **D**.

$$\mathtt{D}(\mathbf{A},\mathbf{BD}) = (\tfrac{1}{2})\Big(\mathtt{D}(\mathbf{A},\mathbf{B}) + \mathtt{D}(\mathbf{A},\mathbf{D}) - \mathtt{D}(\mathbf{B},\mathbf{D})\Big) = (\tfrac{1}{2})(\ 9 + 4 - 9) = 2$$

$$\mathtt{D}(\mathbf{C},\mathbf{BD}) = (\tfrac{1}{2})\Big(\mathtt{D}(\mathbf{C},\mathbf{B}) + \mathtt{D}(\mathbf{C},\mathbf{D}) - \mathtt{D}(\mathbf{B},\mathbf{D})\Big) = (\tfrac{1}{2})(13 + 8 - 9) = 6$$

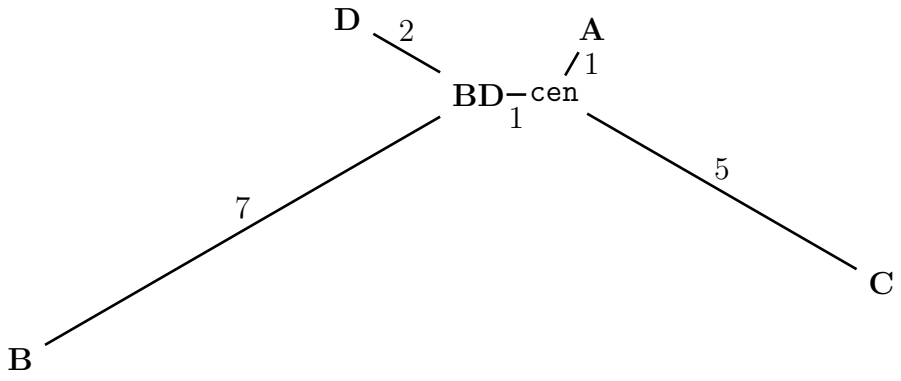
The distance matrix of the tree after merging **B** and **D** is:

	BD	**C**
A	2	6
BD	x	6

NJ infers unrooted trees, and there is only one possible topology (a kind of star topology) for an unrooted tree with three leaves, so we are essentially done. Let `cen` denote the center of the "star".
We can arbitrarily choose a pair to merge and the results should be equivalent. For example if we merge **A**, **BD** to form `cen`, the distance from the remaining leaf **C** to `cen` should be one half `dist`$(\mathbf{C},\mathbf{A})$ + `dist`$(\mathbf{C},\mathbf{BD})$ - `dist`$(\mathbf{A},\mathbf{BD})$.

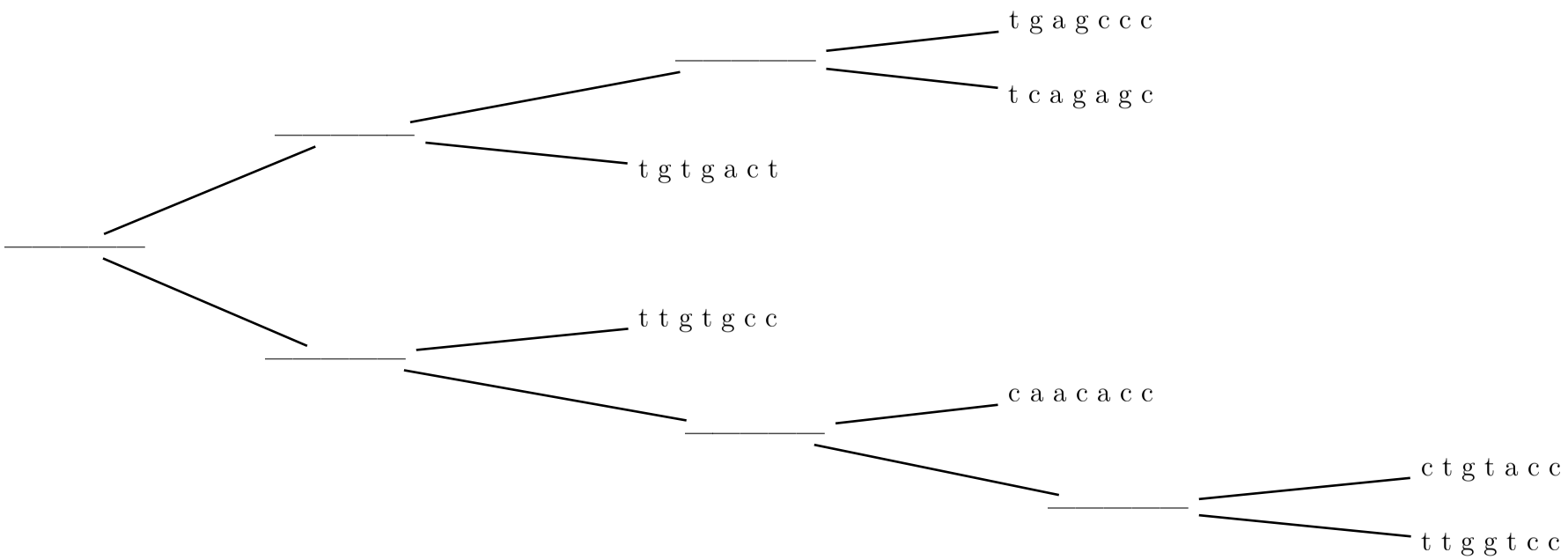
$$\mathtt{D}(\mathbf{C},\mathtt{cen}) = \tfrac{1}{2}(6 + 6 - 2) = 5$$

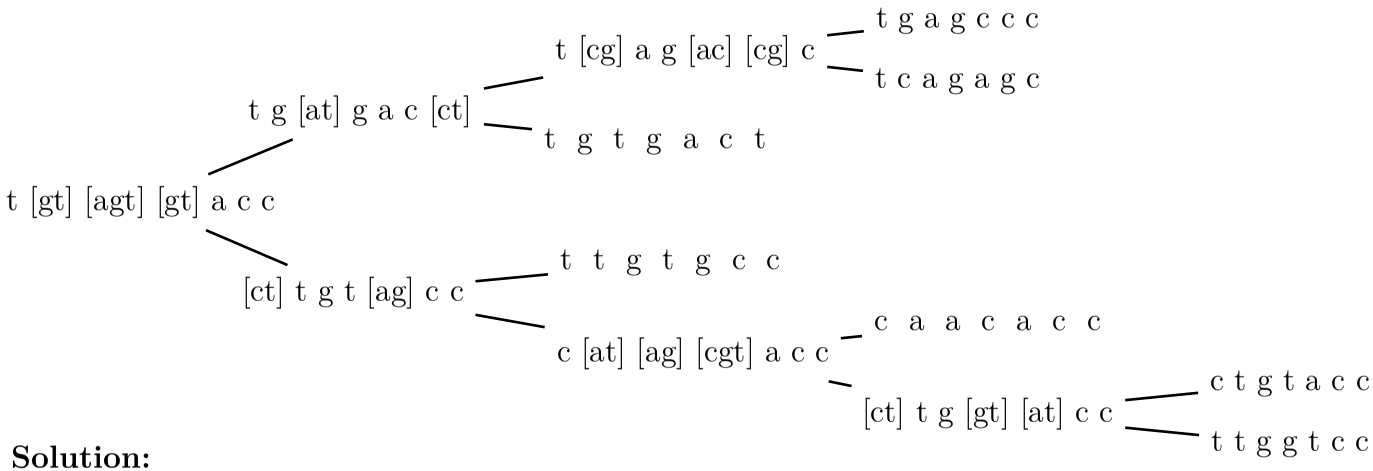
Therefore $\mathtt{D}(\mathbf{A},\mathtt{cen}) = \mathtt{D}(\mathbf{BD},\mathtt{cen}) = 6 - 5 = 1$



Name & student ID: _____

Problem 3.
Use Maximum Parsimony to reconstruct possible sequences of the ancestor nodes for the tree on the following page. Use notation like [ac] to represent sets. For example "t [ac]" denotes t followed by either a or c.


t g a g c c c
t c a g a g c
t g t g a c t
t t g t g c c
c a a c a c c
c t g t a c c
t t g g t c c


t g a g c c c
t [cg] a g [ac] [cg] c
t c a g a g c
t g [at] g a c [ct]
t g t g a c t
t [gt] [agt] [gt] a c c
t t g t g c c
[ct] t g t [ag] c c
c a a c a c c
c [at] [ag] [cgt] a c c
c t g t a c c
[ct] t g [gt] [at] c c
t t g g t c c

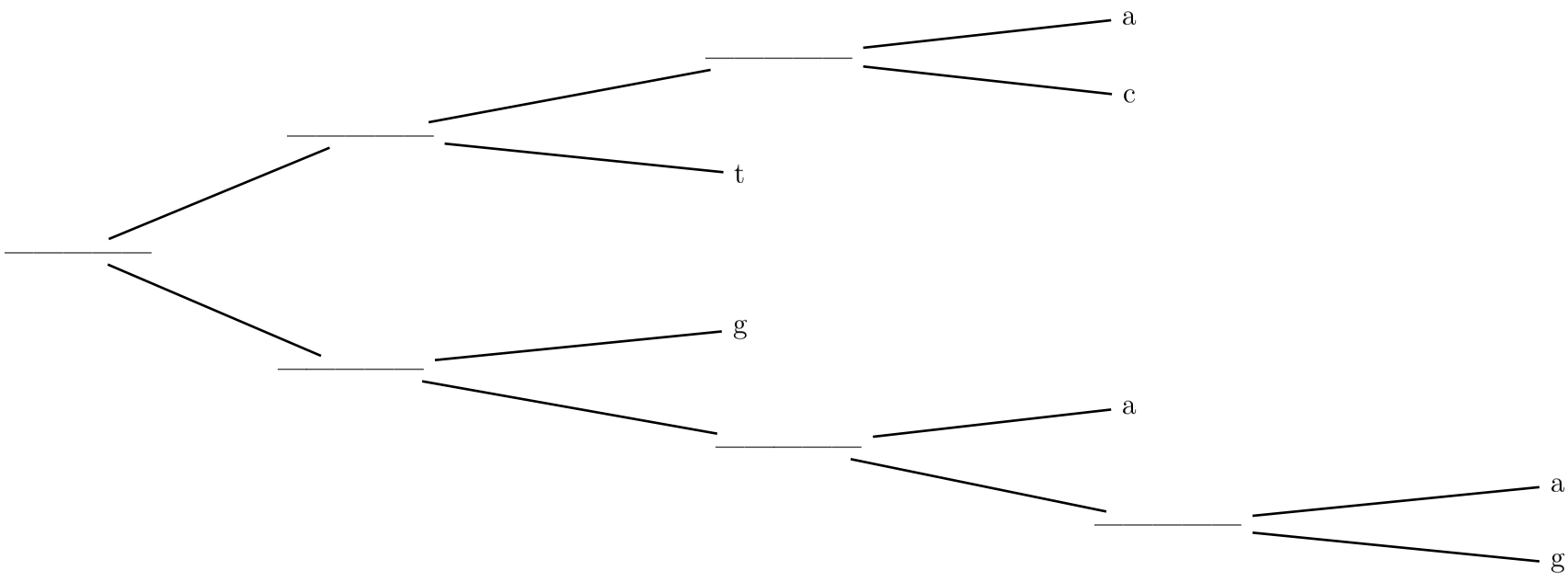
Solution:

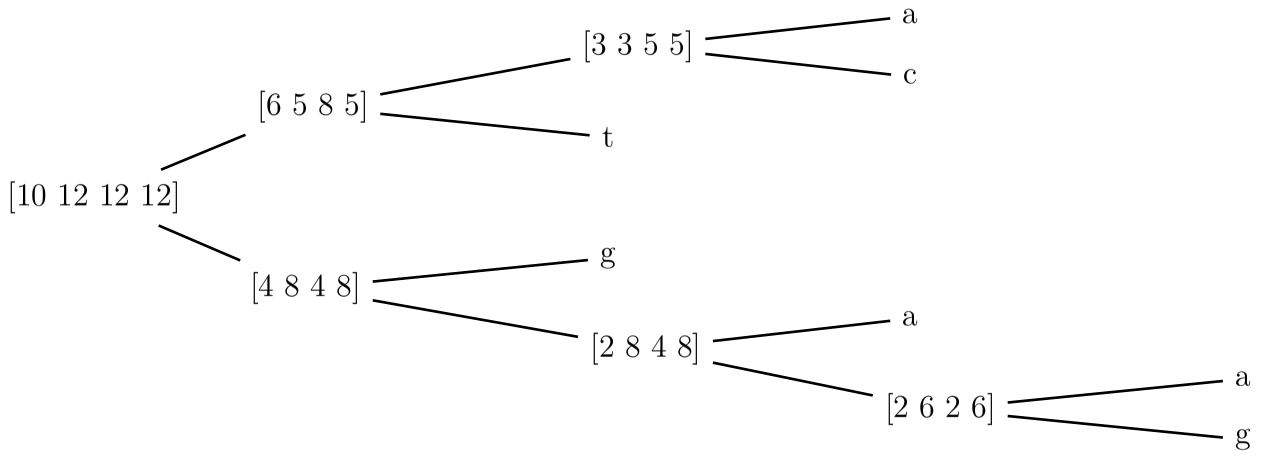
Name & student ID: _____

Problem 4.
Assuming the following symmetric substitution costs:

	c	g	t
a	3	2	3
c	x	3	2
g	x	x	3

Use weighted weighted parsimony to compute the minimum cost of each possible nucleotide for each blank in the tree on the following page.
For notation use vectors in the order [a c g t], for example [6 3 2 5] denotes a minimum cost of 6, 3, 2 or 5; conditioned on the base in that sequence being a,c,g, or t, respectively.





Solution: