Outline: Feb 28

- Continue Neighbor-Joining (NJ)
- Theory of the Q-criteria
- Consistency of NJ

**Notes:**
- Office hours TODAY 1-3pm
- Create “cheat-sheet” for midterm
- Choose partners for Lab 5
Lab 4 Runtime plot examples
Hannah and Melissa
Angelina and Rye

Read Mapping Runtime, length of reference = 168859

- built-in index
- FM-Index

runtime (s)

number of reads (n)
Charlotte and Emily
Lesia and Linda
Continue Neighbor-Joining (NJ)
NJ initialization

Input

We are given a set of samples $\mathcal{X}$ and a dissimilarity map $\delta$ on $\mathcal{X}$.

Initialization

- Create a star tree with center vertex $c$ and an edge $(c, u)$ between $c$ and all samples $u \in \mathcal{X}$.
- Let $N_c$ be the set of neighbors of $c$ and $n = |N_c|$ (cardinality of $N_c$). Set $d$ equal to $\delta$.

$$N_c = \{b, e, f, g, h\}, \quad |N_c| = 5$$
NJ Iterative step (part a)

(a) Find vertices $f, g$ that minimize the $Q$-criteria. Note that UPGMA would only use the first term in this formula, $d(i, j)$. The remaining terms represent how far $i$ and $j$ are from the other vertices.

$$Q(i, j) = (n - 2) \cdot d(i, j) - S_i - S_j,$$

where

$$S_i = \sum_{k \in N_i} d(i, k)$$
NJ Iterative step (part a)

(a) Find vertices $f, g$ that minimize the $Q$-criteria. Note that UPGMA would only use the first term in this formula, $d(i, j)$. The remaining terms represent how far $i$ and $j$ are from the other vertices.

$$Q(i, j) = (n - 2) \ d(i, j) - S_i - S_j,$$

where

$$S_i = \sum_{k \in N_c} d(i, k)$$

UPGMA
NJ Iterative step (part a)

(a) Find vertices \( f, g \) that minimize the \( Q \)-criteria. Note that UPGMA would only use the first term in this formula, \( d(i, j) \). The remaining terms represent how far \( i \) and \( j \) are from the other vertices.

\[
Q(i, j) = (n - 2) \left( d(i, j) - S_i - S_j \right),
\]

where

\[
S_i = \sum_{k \in N_i} d(i, k)
\]

UPGMA

How far away \( i \) and \( j \) are from all the other vertices (further away means we’ll join them earlier)
NJ Iterative step (part b)

(b) Join $f$ and $g$ at internal vertex $v$. Now $N_c$ contains $v$ but not $f$ and $g$. Compute the new edges weights:

\[
d(f, v) = \frac{1}{2}d(f, g) + \frac{1}{2(n-2)}[S_f - S_g] \\
d(g, v) = \frac{1}{2}d(f, g) + \frac{1}{2(n-2)}[S_g - S_f]
\]
(b) Join $f$ and $g$ at internal vertex $v$. Now $N_c$ contains $v$ but not $f$ and $g$. Compute the new edges weights:

\[
\begin{align*}
d(f, v) &= \frac{1}{2} d(f, g) + \frac{1}{2(n-2)} [S_f - S_g] \\
d(g, v) &= \frac{1}{2} d(f, g) + \frac{1}{2(n-2)} [S_g - S_f]
\end{align*}
\]
(b) Join $f$ and $g$ at internal vertex $v$. Now $N_c$ contains $v$ but not $f$ and $g$. Compute the new edges weights:

\[ d(f, v) = \frac{1}{2} d(f, g) + \frac{1}{2(n-2)} [S_f - S_g] \]

\[ d(g, v) = \frac{1}{2} d(f, g) + \frac{1}{2(n-2)} [S_g - S_f] \]
(b) Join \( f \) and \( g \) at internal vertex \( v \). Now \( N_c \) contains \( v \) but not \( f \) and \( g \). Compute the new edges weights:

\[
\begin{align*}
d(f, v) &= \frac{1}{2} d(f, g) + \frac{1}{2(n-2)} [S_f - S_g] \\
d(g, v) &= \frac{1}{2} d(f, g) + \frac{1}{2(n-2)} [S_g - S_f]
\end{align*}
\]
NJ Iterative step (part b)

(b) Join $f$ and $g$ at internal vertex $v$. Now $N_c$ contains $v$ but not $f$ and $g$. Compute the new edges weights:

$$d(f, v) = \frac{1}{2} d(f, g) + \frac{1}{2(n-2)} [S_f - S_g]$$

$$d(g, v) = \frac{1}{2} d(f, g) + \frac{1}{2(n-2)} [S_g - S_f]$$

The difference between how far $f$ and $g$ are from other vertices. In this example $g$ is on average further from other vertices, so $d(g, v) > d(f, v)$.
(b) Join \( f \) and \( g \) at internal vertex \( v \). Now \( N_c \) contains \( v \) but not \( f \) and \( g \). Compute the new edges weights:

\[
\begin{align*}
  d(f, v) &= \frac{1}{2} d(f, g) + \frac{1}{2(n-2)} [S_f - S_g] \\
  d(g, v) &= \frac{1}{2} d(f, g) + \frac{1}{2(n-2)} [S_g - S_f]
\end{align*}
\]

The difference between how far \( f \) and \( g \) are from other vertices. In this example \( g \) is on average further from other vertices, so \( d(g, v) > d(f, v) \)

\[ N_c = \{ b, e, h, v \}, \quad |N_c| = 4 \]
NJ Iterative step (part c)

(c) Compute the distances from $v$ to all remaining vertices $i \in N_c$:

\[ d(i, v) = \frac{1}{2} [d(f, i) - d(f, v)] + \frac{1}{2} [d(g, i) - d(g, v)] \]
NJ Iterative step (part c)

(c) Compute the distances from $v$ to all remaining vertices $i \in N_c$:

\[
d(i, v) = \frac{1}{2}[d(f, i) - d(f, v)] + \frac{1}{2}[d(g, i) - d(g, v)]
\]

Another way to write this:

\[
d(i, v) = \frac{1}{2}[d(f, i) + d(g, i) - d(f, g)]
\]
NJ Termination

Termination

When $n = 3$, the tree topology does not change since we have obtained a binary tree. We still need to run the last iteration though to determine the 3 remaining edge weights. The output is then the tree topology and all edge weights.

$N_c = \{e, v, w\}, \quad |N_c| = 3$
Termination

When $n = 3$, the tree topology does not change since we have obtained a binary tree. We still need to run the last iteration though to determine the 3 remaining edge weights. The output is then the tree topology and all edge weights.

We could “merge” e and w at c, then we would find $d(e,c)$ and $d(w,c)$ in step (b) and find $d(v,c)$ in step (c)

$$N_c = \{e, v, w\}, \quad |N_c| = 3$$
Handout 13 Solution
\( n = 5 \)

\[ \begin{array}{ccccc}
A & B & C & D & E \\
1 & 6 & 13 & 15 & 18 \\
\end{array} \]

\[ \begin{array}{ccccc}
0 & 1 & 3 & 5 & 0 \\
0 & 2 & 5 & 5 & 0 \\
0 & 5 & 5 & 0 & 0 \\
\end{array} \]

\[ \begin{array}{ccccc}
A & B & C & D & E \\
-26 & -22 & -16 & -16 & \\
\end{array} \]

\[ \begin{array}{ccccc}
A & B & C & D & E \\
-26 & -22 & -16 & -16 & \\
\end{array} \]

\[ d(D, v) = 1 = \frac{1}{2} d(D, E) = \frac{1}{2} \sqrt{(50 - 50)} \]

\[ d(E, v) = 1 = \frac{1}{2} d(E, C) = \frac{1}{2} \sqrt{(50 - 50)} \]

\[ N = 3 \]

\[ N = 3\]
Q-criteria theory and consistency
NJ: is consistent if \( S \) is a tree metric, then \( S' \) (induced tree metric from NJ) is equal to \( S \)
length of tree walk 
= 12 + 5 + 9 + 7 
= 33

tree walk (Hautain 15)
Q-criteria

Finding f&g that minimize the average total "tree length"

want the tree that minimizes the total "amount of evolution"