P. Clote

Clustering
\[(f,\gamma)A = ([f],\{g\})_{\text{inst}}\]

1. Initialize set \( \mathcal{C} \) to consist of \( n \) initial singleton clusters

\[\{u\}, \ldots, \{1\}\]

2. Initialize function \( \text{dist}(p, \gamma) \) on \( \gamma \) by defining for all \( (f,\gamma) \in \mathcal{C} \)

Input: \( D \) distance matrix

Algorithm

UPGMA
\[
\mathcal{Z} \left( f', \mathcal{P}_{\mathrm{dist}} + (f', \mathcal{C})_{\mathrm{dist}} \right) = (f', \mathcal{C})_{\mathrm{dist}} = (f', \mathcal{C})_{\mathrm{dist}}
\]

\[\varepsilon \neq f \text{ define for all } \varepsilon \in f \text{ with } \mathcal{C} \]

\[
\varepsilon \in f \text{ at } \varepsilon.
\]

\[\text{distance to its leaves } \overline{\text{leaves}} \text{ of the subtree rooted}
\]

\[\text{at } \varepsilon.
\]

\[\text{define a node with label } \varepsilon \text{ and daughters } \mathcal{C}, \mathcal{D} \text{ where the } e \text{ has}
\]

\[\{\varepsilon\} \cap \{p', \varepsilon\} = \mathcal{C}
\]

\[\text{define new cluster } \varepsilon, \mathcal{P}, \mathcal{C}; \text{ define}
\]

\[\mathcal{P}(p', \mathcal{C})_{\mathrm{dist}} = \mathcal{P}_{\mathrm{dist}}
\]

\[\text{minimal, define}
\]

\[\text{determine pair } \mathcal{P}, \mathcal{C} \text{ of clusters in such that } \mathcal{A}
\]

\[\text{is}\]

\[\text{repeat } n - 1 \text{ times}
\]
In implementation, can choose label of cluster $c$ to be smallest integer $1 \leq i \leq n$ belonging to $c$, and hence one can update distance matrix $D$ directly. Suppose the labels of the two closest clusters are $i_0, j_0$ and that $i_0 < j_0$. Then
erase \( j \)-th row and column of \( D \):

\[
0 = [0^j] y D = [y][0^j] D
\]

for \( u > j, \quad 0 = y \)

for

// copy \( E \) into \( D \) and symmetry

\[
[y][0^j] E = [0^j][y] D = [y][0^j] D
\]

for \( u > j, \quad 0 = y \)

for

else

\[
\frac{2}{[y][0^j] D + [y][0^j] D} = [y][0^j] E
\]

\( (0 \neq [y][0^j] D) \hat{\wedge} \hat{\wedge} (0 \neq y) \hat{\wedge} \hat{\wedge} (0 \neq y)) \) if

\[
(\quad u > j, \quad 0 = y \quad ) \quad \text{for}
\]
\[(l', r') D = (\{l\}, \{r\})_{dist}\]

1. Initialize set \(C\) to consist of \(n\) initial singleton clusters
2. Initialize function \(dist(c', p)\) on \(C\) by defining for all \(c\) in \(C\) and \(l', r'\) in \(n\) the distance matrix \(D\)

\text{WPGMA Algorithm}
\[
\frac{|p| + |c|}{(f,p)_{dist} |p| + (f,c)_{dist} |c|} = (f, e)_{dist} = (f, e)
\]

\( e \neq f \) for all \( e \in f \) with \( e \) is leaf.

distance to its leaves.

(e) define a node with label \( e \) and daughters \( c, d \), where the edge has

\[
\{e\} \cap \{p, c\} - c = c
\]

(p) define new cluster \( e \); define

\[
(p, c)_{dist} = \min p
\]

minimal; define

(a) determine pair \( c, d \) of clusters in such that \( \text{dist}(c, d) \) is

3. Repeat \( n-1 \) times
\[
\frac{|p| + |c|}{(1, r) A \sum_{f \in \mathcal{F} \cap p} \frac{|f|}{1} + (1, r) A \sum_{f \in \mathcal{F} \cap c} \frac{|f|}{1}} = (f, \text{dist}(\varepsilon))
\]

Ind. Hyp.

\[
\frac{|p| + |c|}{} = (f, \text{dist}(\varepsilon))
\]

Then

\[
p \cap c = \varepsilon
\]

Proof. By induction. Suppose inducivity that \( e = p \).

\[
(1, r) A \sum_{f \in \mathcal{F} \cap \varepsilon} \frac{|f|}{1} = (f, \text{dist}(\varepsilon))
\]

WP cram for clusters \( e \) and \( f \). Then

\[
\text{Proposition. Let } \text{dist}(\varepsilon) \text{ be the distance defined by} \]

cluster without using a recursive definition, as follows.

In the case of WP cram, we can define the distance function on
\[(l',\gamma)D \sum_{f \in \mathcal{H}} \frac{|f||\varnothing|}{I} = \frac{|p| + |\varnothing|}{(l',\gamma)D \sum_{f \in \mathcal{H}} |f|} = \]
the topology.

FACT. If tree is ultrametric, then UPGMA correctly computes

\[
\begin{array}{c}
A & p & c & q & a \\
\circ & \circ & \circ & \circ & \circ \\
\downarrow & \downarrow & \downarrow & \downarrow & \downarrow \\
3 & 3 & 3 & 3 & 3 \\
\end{array}
\]
UPGMV correctly computes the topology.

Fact. If tree is additive, then Pars transformed distance additive tree.
matrix generates the correct topology of \( \mathcal{L} \).

Then \( \cup \mathcal{C} \cup \mathcal{M} \) applied to the transformed distance between \( \mathcal{L} \) and the leaves. The average distance between \( \mathcal{L} \) and

\[
\frac{1}{u} \sum_{i=1}^{u} \left( \sum_{j=1}^{u} \frac{\sum_{k=1}^{u} \left( \left( \frac{1}{u} \right)^{j-1} \left( 1 - \frac{1}{u} \right)^{i-1} \right) \left( 1 - \frac{1}{u} \right) \sum_{k=1}^{u} \left( \left( \frac{1}{u} \right)^{j-1} \left( 1 - \frac{1}{u} \right)^{i-1} \right) \right)}{u - i + 1} \right) = \delta'_{u} \]

Define the transformed distance

\[
\delta'_{u} \]

\( \delta'_{u} \) is the path length of the unique path connecting nodes \( \delta'_{u} \) of \( \mathcal{L} \) is the path length of the unique path connecting taxa \( \delta'_{u} \). Then suppose that the distance \( \delta'_{u} \) between

Let be a phylogenetic tree with ancestor \( \mathcal{L} \) and leaves \( \mathcal{L} \).

**Theorem.** (Parts transformed distance method)
\[(f,g)_{\text{root}} \cdot p - \frac{\pi}{p} = \frac{\pi}{\pi - \pi - \pi} \]

so

\[(f,g)_{\text{root}} \cdot \pi_0 - \frac{\pi}{\pi} + \pi_0 = \frac{\pi}{\pi} \]

Since distance is path length between nodes in \( L \) we have

\[(f,g)_{\text{root}} \cdot p - \frac{\pi}{p} = \frac{\pi}{\pi} \]

Leaves \( f, g \). Claim that

Proof. Let \( \text{lead}(f, g) \) denote the least common ancestor of
Thus we have the same distance from the root to any leaf, and
ancestor of $\tau$ and $\tau'$ is $\tau$, so $d_p = 0$, which is the least common
multiple: the distance from the root $\tau$ to any leaf $\tau'$ is an
integer. It is immediate that the matrix $(\delta_{i,j})$ is an
the correct topology.

Applying $\Upsilon_{CMA}$ to the transformed distance matrix yields
has the values as follows.

\[
(j', l')_0\nu \nu_0 - \nu_0 = \nu_0 + \frac{\mathcal{Z}}{(j', l')_0 \nu - \nu_0} = (j', l')_0
\]

where the distance matrix \((j', l')_0\) and so the transformed \(\mathcal{Z}\) is computed that produces, for additive, non-ultrametric example:

\[
\begin{pmatrix}
0 & 19 & 11 & 11 & 18 & 11

19 & 0 & 12 & 14 & 21 & 14

11 & 12 & 0 & 13 & 18 & 13

18 & 21 & 13 & 0 & 6 & 13

11 & 6 & 14 & 6 & 0 & 14
\end{pmatrix}
\]
\[
\begin{pmatrix}
0 & 7.2 & 6.2 & 7.2 \\
7.2 & 0 & 6.2 & 7.2 \\
6.2 & 7.2 & 0 & 7.2 \\
7.2 & 7.2 & 7.2 & 0
\end{pmatrix}
\]

Transformed distance matrix $\mathcal{F} \in \mathbb{R}^{n \times n}$
Applying UPGMA, get