UPGMA/WPGMA

(Un)weighted Pair Group Method with Arithmetic Mean

Julian Löffler

12.11.2018

Motivation for UPGMA / WPGMA

- Consider a set of sequences:

A: TCAACTAC
B: ACTGCAAA
C: GGCTGTAA
D: AGTTGCAA
E: TTTGAACT

- The aim is grouping the most similar sequences, regardless of their evolutionary rate or phylogenetic affinities.

Motivation for UPGMA / WPGMA

- Unweighted Pair Group Method with Arithmetic Mean
- Weighted Pair Group Method with Arithmetic Mean
- Used for the creation of guide trees.

UPGMA Algorithm

1. Compute the distance between each pair of sequences.
2. Treat each sequence as a cluster C by itself.
3. Merge the two closest clusters. The distance between two clusters is the average distance between all their sequences:

$$d(C_i, C_j) = \frac{1}{|C_i||C_j|} \sum_{r \in C_i, s \in C_j} d(r, s)$$

4. Repeat 2. and 3. until only one cluster remains.vllt

Example

- Let A, B, C, D, E be sequences.
- Say we have calculated a distance matrix D:

	A	B	C	D	E
A	0	8	4	6	8
B	8	0	8	8	4
C	4	8	0	6	8
D	6	8	6	0	8
E	8	4	8	8	0

Example

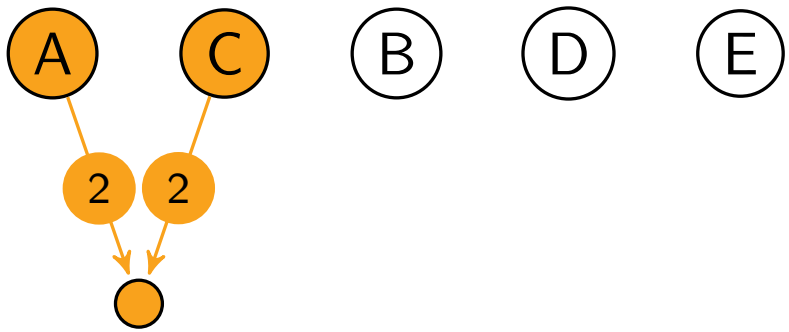
- Create guide tree for the sequences

	A	B	C	D	E
A	0	8	4	6	8
B	8	0	8	8	4
C	4	8	0	6	8
D	6	8	6	0	8
E	8	4	8	8	0

Example I

- $\{A\}$ and $\{C\}$ are the closest clusters
  $\Rightarrow$ merge $\{A\}$ and $\{C\}$ into $\{A,C\}$.
- Each branch gets a total distance of $\frac{d(\{A\},\{C\})}{2} = \frac{4}{2}$.

	A	B	C	D	E
A	0	8	4	6	8
B	8	0	8	8	4
C	4	8	0	6	8
D	6	8	6	0	8
E	8	4	8	8	0



Example I

- The distance between $\{A, C\}$ and the other clusters is the average distance between all their sequences.
- Example: the new distance between $\{A, C\}$ and $\{B\}$ is:

$$d(\{A, C\}, \{B\}) = \frac{1}{|\{A, C\}||\{B\}|}(d(\{A\}, \{B\}) + d(\{C\}, \{B\}))$$
$$= \frac{1}{2}(8 + 8) = 8$$

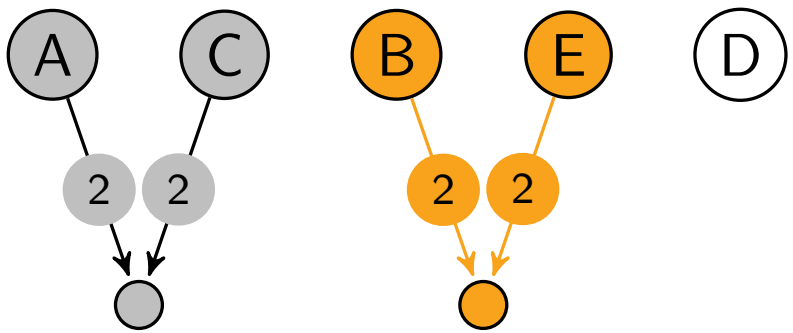
	A	B	C	D	E
A	0	8	4	6	8
B	8	0	8	8	4
C	4	8	0	6	8
D	6	8	6	0	8
E	8	4	8	8	0

	A,C	B	D	E
A,C	0	8	6	8
B	8	0	8	4
D	6	8	0	8
E	8	4	8	0

Example II

- {B} and {E} are the closest clusters
  $\Rightarrow$ merge {B} and {E} into {B,E}.
- Assign $\frac{d(\{B\},\{E\})}{2} = \frac{4}{2} = 2$ to each branch.

	A,C	B	D	E
A,C	0	8	6	8
B	8	0	8	4
D	6	8	0	8
E	8	4	8	0



Example II

- The distance between $\{B, E\}$ and the other clusters is the average distance between all their sequences.

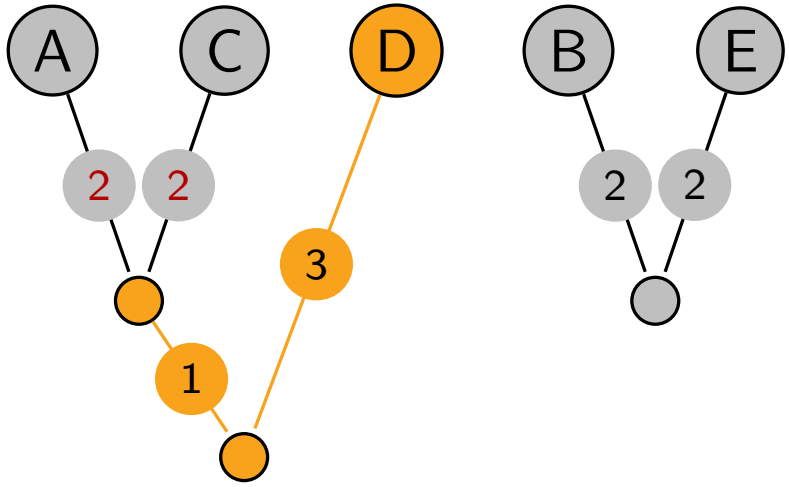
	A,C	B	D	E
A,C	0	8	6	8
B	8	0	8	4
D	6	8	0	8
E	8	4	8	0

	A,C	B,E	D
A,C	0	8	6
B,E	8	0	8
D	6	8	0

Example III

- {A,C} and {D} are the closest clusters
  $\Rightarrow$ merge {A,C} and {D} into {A,C,D}.
- Assign $\frac{d(\{A,C\},\{D\})}{2} = \frac{6}{2} = 3$ total length to each branch.
- Observe, that the 1 is obtained calculating $3 - 2 = 1$.

	A,C	B,E	D
A,C	0	8	6
B,E	8	0	8
D	6	8	0



Example III

- The distance between $\{A, C, D\}$ and $\{B, E\}$ is the average distance between all their sequences.

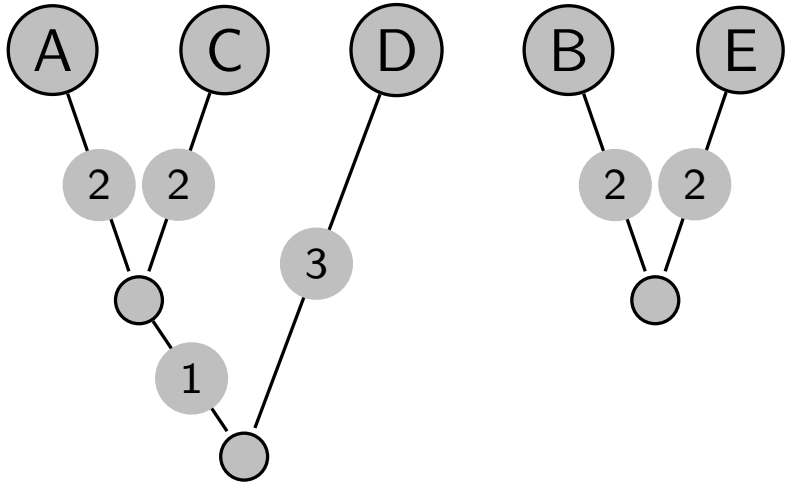
	A,C	B,E	D
A,C	0	8	6
B,E	8	0	8
D	6	8	0

	A,C,D	B,E
A,C,D	0	8
B,E	8	0

Example IV

- {A,C, D} and {B,E} are remaining.
  $\Rightarrow$ merge {A,C,D} and {B,E} into {A,C,D, B, E}.

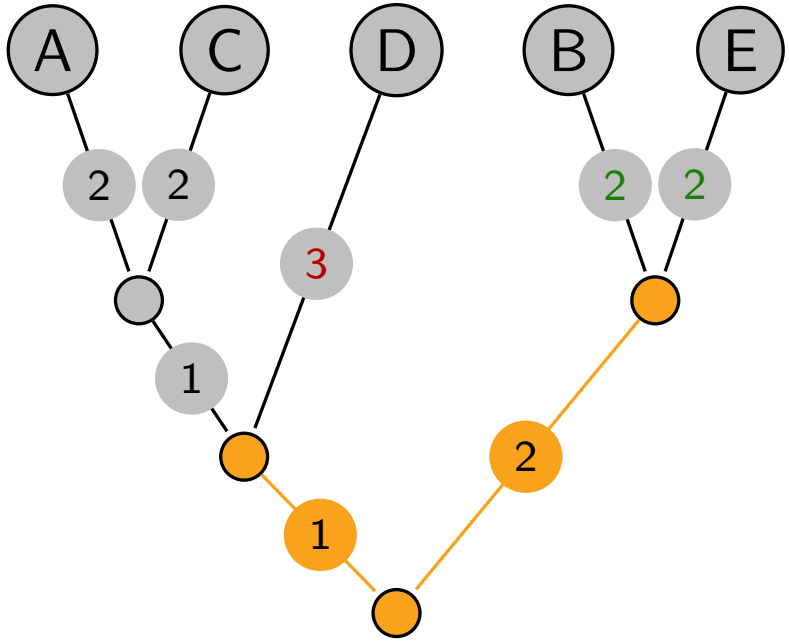
	A,C,D	B,E
A,C,D	0	8
B,E	8	0



Example IV

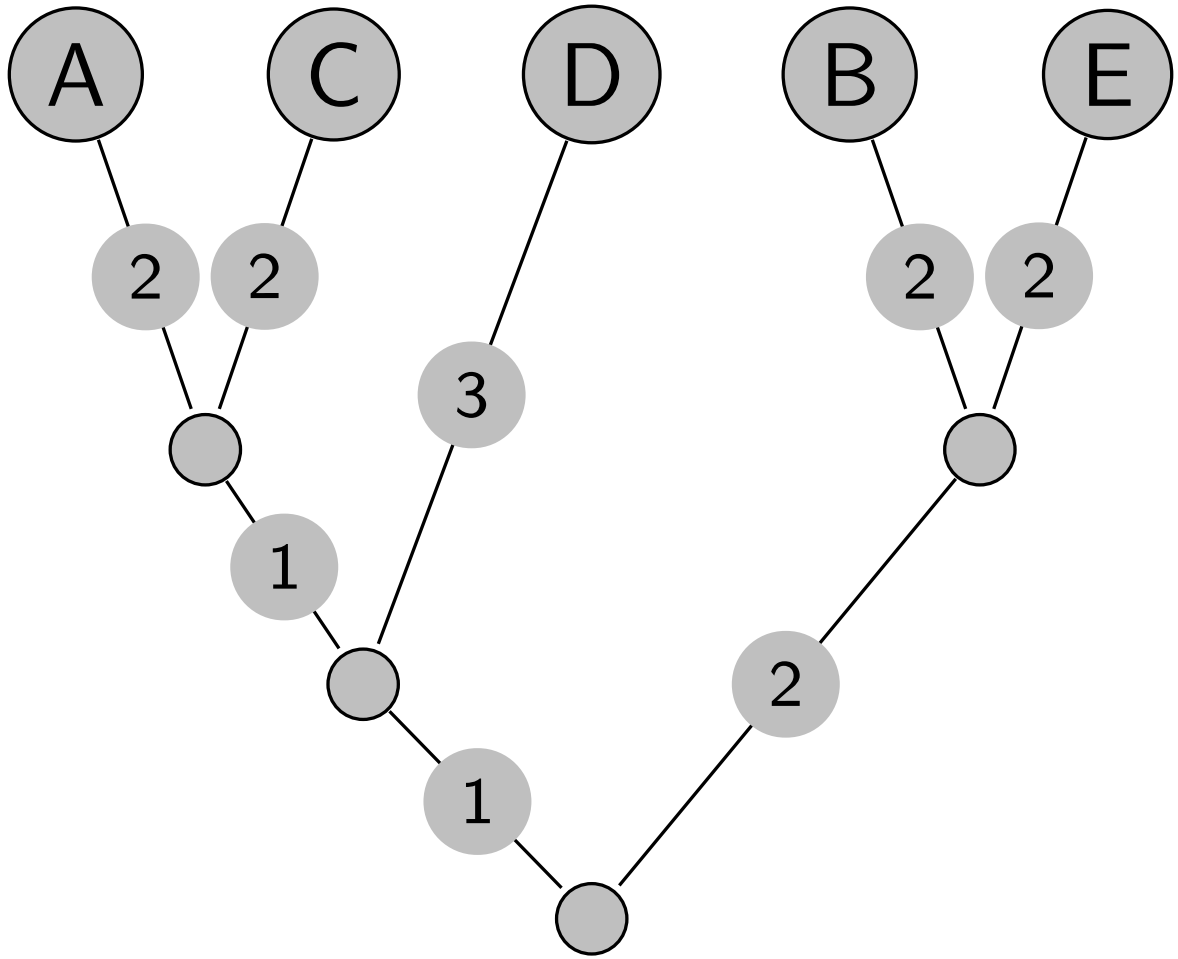
- Assign $\frac{d(\{A,C,D\},\{B,E\})}{2} = \frac{8}{2} = 4$ total length to each branch.
- Observe, that the 1 is obtained calculating $4 - 3 = 1$.
- Observe, that the 2 is obtained calculating $4 - 2 = 2$.

	A,C,D,B,E
A,C,D,B,E	0



Example V

- Our beautiful tree:



WPGMA

- In WPGMA the distance between clusters is calculated as a simple average:

$$d(C_i \cup C_j, C_k) = \frac{d(C_i, C_k) + d(C_j, C_k)}{2}$$

- Computationally easier than UPGMA.
- Unequal numbers of taxa in the clusters cause problems $\Rightarrow$ the distances in the original matrix do not contribute equally to the intermediate calculations.
- The branches do not preserve the original distances.
- Final result is therefore said to be weighted.

Final notes

- Clustering works only if the data are ultrametric.
- Ultrametric distances are defined by the satisfaction of the three-point condition:
  - For any three taxa it holds:

$$d(A, C) \leq max(d(A, B), d(B, C))$$

- So we assume that all taxa evolve with the same constant rate
- $O(n^3)$ for the trivial approach.
- $O(n^2 log(n))$, when using a heap for each cluster
- $O(k3^k n^2)/O(n^2)$ implementations for special cases. (by Fionn Murtagh, by Day and Edelsbrunner)

References

The content of this set of slides, is based on:

- Construction of a distance tree using clustering with the Unweighted Pair Group Method with Arithmatic Mean (UPGMA): `https://www.icp.ucl.ac.be/~opperd/private/upgma.html`
- UPGMA Wikipedia article: `https://en.wikipedia.org/wiki/UPGMA`
- WPGMA: `http://www.wikiwand.com/en/WPGMA`