Exercise sheet 10: UPGMA

A phylogenetic tree (also phylogeny or evolutionary tree [1]) is a branching diagram or a tree showing the evolutionary relationships among various biological species or other entities based upon similarities and differences in their physical or genetic characteristics. All life on Earth is part of a single phylogenetic tree, indicating common ancestry.

https://en.wikipedia.org/wiki/Phylogenetic_tree

UPGMA (Unweighted Pair Group Method with Arithmetic Mean) is a simple agglomerative or hierarchical clustering method used in bioinformatics for the creation of phylogenetic trees. UPGMA assumes a constant rate of evolution (molecular clock hypothesis), and is not a well-regarded method for inferring phylogenetic trees unless this assumption has been tested and justified for the data set being used.

https://en.wikipedia.org/wiki/UPGMA

Exercise 1 - WPGMA

Note

Distances for a merged cluster e, where $e = c \cup d$:

$$WPGMA : dist(x, e) = \frac{dist(x, c) + dist(x, d)}{2}$$

In the following steps we calculate the evolutionary tree using WPGMA and the pairwise distances in the following distance matrix.

1a)

Which leaves should be selected first?

Hide

Hint

- ☐ c and d
- ☐ a and b
- ☐ d and e

[1] Felsenstein, Joseph, and Joseph Felenstein. Inferring phylogenies. Vol. 2. Sunderland, MA: Sinauer associates, 2004.

Solution

- ☐ c and d
- ☒ a and b
- ☐ d and e

1b)

Calculate the corresponding distance for the set of leaves from *a)*.

Hide

Solution 1.5

1c)

Fill in the distance matrix with the correct distances form the set of leaves (aka. internal node) from *a)* to all other leaves.

Hide

Solution

1d)

Which nodes are joined next given the correct distance matrix from *c)*?

Hide

Hint

- ☐ c and d
- ☐ {a,b} and e
- ☐ {c, d} and e
- ☐ e and a

Solution

- ☒ c and d
- ☐ {a,b} and e
- ☐ {c, d} and e
- ☐ e and a

1e)

Fill in a distance matrix with the remaining nodes and leaves.

Hide

Solution

1f)

What does the **subpart** of the tree look like in Newick format after selecting and joining your answer from *e)*

Note

The following answers will be given in Newick format. Feel free to inspect them using an online tool.

Hide

Hint

- ☐ `((c : 3, d : 3) : 3.5, e : 3.5);`
- ☐ `((c : 3, d : 3) : 0.5, e : 3.5);`
- ☐ `((a : 1.5, b : 1.5) : 2.75, e : 4.25);`

Solution

☐ `((c : 3, d : 3) : 3.5, e : 3.5);`
☒ `((c : 3, d : 3) : 0.5, e : 3.5);`
☐ `((a : 1.5, b : 1.5) : 2.75, e : 4.25);`

1g)

Following the approach from the previous exercises, what does the **whole tree** look like.

Hide

Hint

☐ `((a : 1.5, b : 1.5) : 4, ((c : 3, d : 3) : 0.5, e : 3.5) : 2);`
☐ `((a : 1.5, b : 1.5) : 4.25, ((c : 3, d : 3) : 0.5, e : 3.5) : 2.25);`
☐ `(((c : 3, d : 3) : 3.5, e : 3.5): 4, (a : 1.5, b : 1.5) : 2);`

Solution

☒ `((a : 1.5, b : 1.5) : 4, ((c : 3, d : 3) : 0.5, e : 3.5) : 2);`
☐ `((a : 1.5, b : 1.5) : 4.25, ((c : 3, d : 3) : 0.5, e : 3.5) : 2.25);`
☐ `(((c : 3, d : 3) : 3.5, e : 3.5): 4, (a : 1.5, b : 1.5) : 2);`

Exercise 2 - UPGMA

2a)

Imagine using UPGMA instead of WPGMA for construction of a tree. Which of the following statements is True?

Statements

☐ There will only be a difference in edge lengths. Overall topology will stay the same.
☐ The tree in Exercise 1 will not change
☐ UPGMA is equal to WPGMA if the number of leaves in the two clusters (|c| and |d|) is the same.
☐ UPGMA can end up with wrong topologies when using non-ultrametric distances.

Hint: Formula

$$UPGMA : dist(x, e) = \frac{|c|dist(x, c) + |d|dist(x, d)}{|c| + |d|}$$

Solution

☐ There will only be a difference in edge lengths. Overall topology will stay the same.
☐ The tree in Exercise 1 will not change
☒ UPGMA is equal to WPGMA if the number of leaves in the two clusters (|c| and |d|) is the same.
☒ UPGMA can end up with wrong topologies when using non-ultrametric distances.

Exercise 3 - Ultrametric

3a)

Which of the following distance matrices are ultrametric?

1)

2)

3)

4)

Hide

Hint Note

Definition Ultra-Metric:

$$w(x, y) = 0 \leftrightarrow x = y \quad \text{(identity)} \quad (1)$$
$$w(x, y) = w(y, x) \quad \text{(symmetric)} \quad (2)$$
$$w(x, z) \leq w(x, y) + w(y, z) \quad \text{(triangle inequality)} \quad (3)$$
$$w(x, z) \leq max\{w(x, y), w(y, z)\} \quad \text{(strong triangle inequality)} \quad (4)$$

Solution 2)

Exercise 4 - Programming assignment

Programming assignments are available via Github Classroom and contain automatic tests.

We recommend doing these assignments since they will help you to further understand this topic.

Access the Github Classroom link: Programming Assignment: Sheet 10.