UPGMA – Rewriting the distance equation

Johannes Falk, October 25, 2023

The distance between cluster is in the UPGMA algorithm calculated via the equation:

$$d_{ij} = \frac{1}{|C_i||C_j|} \sum_{p \in C_i; q \in C_j} d_{pq} \tag{1}$$

This equation requires that one computes the sequence distances of all possible sequence pairs between the two clusters C_i and C_j. Towards the end of tree generation, this can lead to many combinatorial possibilities. However, there is a possibility to rewrite the equation in such a way that already calculated distances between clusters can be further used in the next step.
Let us assume a cluster C_k that was created by the union of the two clusters C_i and C_j; $C_k = C_i \cup C_j$. Using Eq. 1 we then get for the distance between this cluster C_k and any other cluster C_l:

$$d_{kl} = \frac{1}{|C_k||C_l|} \sum_{p \in C_k; q \in C_l} d_{pq} = d_{kl} = \frac{1}{|C_k||C_l|} \sum_{p \in C_i \cup C_j; q \in C_l} d_{pq} \tag{2}$$

We now split the sum into two parts. The first part contains all pairwise distances that contain a sequence from C_i, the second part contains all sequences that contain a sequence from C_j:

$$d_{kl} = \frac{1}{|C_k||C_l|} \sum_{p \in C_i \cup C_j; q \in C_l} d_{pq} = \frac{1}{|C_k||C_l|} \left(\sum_{p \in \mathbf{C_i}; q \in C_l} d_{pq} + \sum_{p \in \mathbf{C_j}; q \in C_l} d_{pq} \right) \tag{3}$$

By comparison with Eq. 1 we can now identify substitutions for the two sums:

$$\sum_{p \in \mathbf{C_i}; q \in C_l} d_{pq} = d_{il}|C_i||C_l| \tag{4}$$

$$\sum_{p \in \mathbf{C_j}; q \in C_l} d_{pq} = d_{jl}|C_j||C_l| \tag{5}$$

If we insert these substitutions into Eq. 3, we obtain:

$$d_{kl} = \frac{d_{il}|C_i| + d_{jl}|C_j|}{|C_k|} = \frac{d_{il}|C_i| + d_{jl}|C_j|}{|C_i| + |C_j|}, \tag{6}$$

where we used in the last step that $|C_k| = |C_i \cup C_j| = |C_i| + |C_j|$