14.1 The UPGMA algorithm

In Chapter 3, we will learn about algorithms to construct phylogenetic trees. In the phylogeny problem we are given

- a set of DNA or protein sequences, \( \{s_1, s_2, \ldots, s_n\} \)
- a distance matrix \( d = [d_{ij}] \) where \( d_{ij} \) is the (evolutionary) distance between sequence \( i \) and sequence \( j \).

The goal is to build a tree that has the sequences \( s_1, \ldots, s_n \) as its leaves and for which the arrangement of the internal nodes is representative of the shared evolutionary history of the sequences. More specifically, each internal node should represent the most recent common ancestor (MRCA) of all its descendant leaves; in particular, the root node should be the most recent common ancestor of all \( n \) sequences.

How do we compute the distance matrix \( d \)? Before the rise of sequencing technologies, distances between species were defined in terms of physical and morphological characteristics. As you can imagine, the choice of such characteristics is fairly arbitrary, and the resulting distances were often not reflective of the true evolutionary past. On your midterm, you saw how one might use alignment algorithms to compute distance measures between two sequences. Using molecular sequence data gives us a more principled way in which to infer phylogenies.

14.1.1 Working with distances

Here we introduce the Unweighted Pair Group Method using Arithmetic averages algorithm for clustering sequences. The inputs are a set of sequences \( s_1, \ldots, s_n \) and a matrix \( d \) that gives the distance \( d_{\alpha\beta} \) between any pair of sequences \( \alpha \) and \( \beta \). Let \( S_i \) and \( S_j \) be two disjoint subsets of sequences from the input set of sequences.

**Definition 14.1** The distance between two clusters \( S_i \) and \( S_j \) is

\[
d_{ij} = \frac{1}{|S_i||S_j|} \sum_{s \in S_i, s' \in S_j} d_{s,s'}
\]

In words, \( d_{ij} \) is the average distance between pairs of sequences from each cluster.

We create a new cluster by taking the union of existing clusters, i.e. \( S_l = S_i \cup S_j \).
Theorem 14.2 The distance between the new cluster \( l \) and the previous cluster \( m \) is
\[
d_{lm} = \frac{d_{im}|S_i| + d_{jm}|S_j|}{|S_i| + |S_j|}.
\]

Proof: By our above definition,
\[
d_{lm} = \frac{1}{|S_i \cup S_j||S_m|} \sum_{s \in S_i \cup S_j, s' \in S_m} d_{s,s'}
\]
\[
= \frac{1}{(|S_i| + |S_j|) \cdot |S_m|} \sum_{s \in S_i, s' \in S_m} d_{s,s'} + \frac{1}{(|S_i| + |S_j|) \cdot |S_m|} \sum_{s \in S_j, s' \in S_m} d_{s,s'}
\]
\[
= \frac{|S_i|}{|S_i| + |S_j|} \cdot \frac{1}{|S_m|} \sum_{s \in S_i, s' \in S_m} d_{s,s'} + \frac{|S_j|}{|S_i| + |S_j|} \cdot \frac{1}{|S_m|} \sum_{s \in S_j, s' \in S_m} d_{s,s'}
\]
\[
= \frac{d_{im}|S_i| + d_{jm}|S_j|}{|S_i| + |S_j|}.
\]

14.1.2 The algorithm

Input: A set of sequences \( \{s_1, \ldots, s_n\} \) and a distance matrix \( d_{n \times n} \).

Initialization: For \( k = 1, \ldots, n \), set \( S_k = \{s_k\} \). In other words, every sequence starts off as its own cluster. These clusters will be the leaves of the tree. Since they are at the bottom of the tree, they are at height (time) 0.

Iteration: Find two clusters \( S_i \) and \( S_j \) for which \( d_{ij} \) is minimal. Using these two clusters, create the new cluster \( S_l = S_i \cup S_j \). Compute the distance \( d_{lm} \) for every cluster \( S_m \) other than \( S_i \) and \( S_j \). Define a new node \( l \) with daughter nodes \( i \) and \( j \) and place it at height \( \frac{1}{2}d_{ij} \). Add node \( l \) to the current list of clusters and remove nodes \( i \) and \( j \).

Termination: When only two clusters \( S_i \) and \( S_j \) remain, place the root of the entire tree at height \( \frac{1}{2}d_{ij} \).

The UPGMA algorithm makes the “common molecular clock” assumption, which says that all sequences evolve (i.e. mutate) at the same rate. This implies that the distances in the matrix \( d \) are additive, a term that we will elaborate on later in the context of the neighbor joining algorithm.