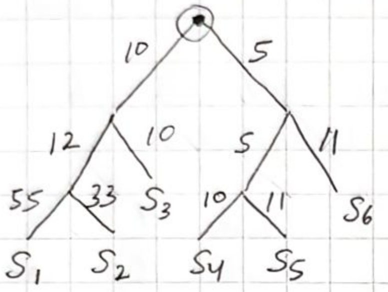
Ch3. Phylogenetic Trees Algorithms

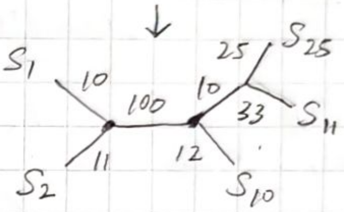
11/11/2025

3.1. Methods for Constructing Trees:
Distance, Parsimony, Max Likelihood (Stat Methods)
3.2. Distance based: UPGMA
3.3. Distance based: Neighbor-Joining
3.4. Parsimony Methods
3.5. Max Likelihood methods



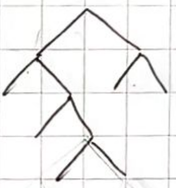
$S = \{S_1, \dots, S_6\}$
with a common ancestor

← rooted and unrooted trees
↓



branching order
evolutionary time

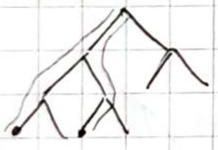
direction of evolution
(rooted tree)



parsimony method
branching order
no evolutionary time inference

morphological vs molecular

UPGMA



universal molecular clock

Additivity & Neighbor-Joining

No universal molecular clock assumption

Additivity: Given a tree, its edge lengths are additive if the distance between any pair of leaves is the sum of the lengths of the edges on the unique path connecting them.

In a tree (graph theoretic) between any 2 nodes, there is a unique path connecting them.

UPGMA trees satisfy this condition. If additivity holds, but universal molecular clock fails then the Neighbor-Joining Algorithm reconstructs the phylogenetic tree of the INPUT species.

Given: a tree T with additive edge lengths,

We will reconstruct T from pairwise distances between its leaves, as follows.

We find a pair of neighboring leaves, say i and j, i.e. the two leaves have the same parent ℓ

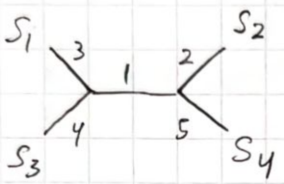
Remove i, j from the list of leaves and add node ℓ to the current list of nodes, defining the distance of node ℓ to a generic leaf m by

$$d_{\ell m} = \frac{1}{2}(d_{im} + d_{jm} - d_{ij})$$

bc of additive

Caveat: picking 2 closest leaves is not good enough because they may not be neighbors

i, j
$d_{ij} = \min$



$\boxed{d_{12} = 6}$
$d_{13} = 7$
$d_{14} = 9$
$d_{23} = 7$
$d_{24} = 7$
$d_{34} = 10$

To avoid the caveat, we subtract the average distances to all leaves.

Define $D_{ij} = d_{ij} - (a_i + a_j)$

L = list of leaves at that time in the alg

$$a_i = \frac{1}{|L| - 2} \sum_{r \in L} d_{ir} \qquad |L| = \text{size of } L$$

We can prove that $D_{ij} = \min$ then i, j are neighbors

NEIGHBOR-JOINING ALGORITHM for tree T

Initialization: INPUT: set of sequences $S = \{s_1, \dots, s_n\}$
pairwise distances d_{ij}

T = the set of leaf nodes, one for each input species (sequence, molecular data)

$L = T$; $a_i = \frac{1}{|L| - 2} \sum_{r \in L} d_{ir}$ $D_{ij} = d_{ij} - (a_i + a_j)$

Iteration: pick a pair i, j from L for which $D_{ij} = \min$
Define new node ℓ and set $d_{\ell m} = \frac{1}{2}(d_{im} + d_{jm} - d_{ij})$ for all $m \in L$
Add node ℓ to T with edge length
$d_{i\ell} = \frac{1}{2}(d_{ij} - a_i - a_j)$
$d_{j\ell} = d_{ij} - d_{i\ell}$
Remove i, j from L and add ℓ to L

joining

Termination: when L consists of 2 leaves only i and j, add the remaining edge between i and j with length d_{ij}