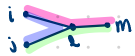
NEIGHBOR-JOINING ALGORITHM → constructs an unrooted phylogenetic tree, unlike UPGMA

- additivity: weaker concept than universal clock ~ says we only care that we can add edges and get accurate distances.
  ↪ defn: 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
- if additivity holds, but universality of the molecular clock fails, we can still reconstruct the tree by the neighbor-joining algorithm
- Main idea of the algorithm: the true tree we are trying to reconstruct

  given a [theoretical] tree w/ additive edge lengths, we will reconstruct a tree T from pairwise distances between its leaves as follows:

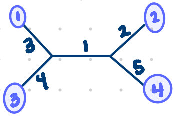
  1. Find a pair of neighboring leaves i and j i.e. 2 leaves with the same parent ℓ
  2. Remove i and j from the list of leaves and add ℓ to the current list of nodes
- define the distance of node ℓ to leaf m by this formula:

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



$$= \frac{1}{2}(\cancel{d_{i\ell}} + d_{\ell m} + \cancel{d_{j\ell}} + d_{\ell m} - \cancel{d_{i\ell}} - \cancel{d_{j\ell}})$$
$$= \frac{1}{2}(d_{\ell m} + d_{\ell m}) = \frac{1}{2}(2 d_{\ell m})$$
$$= d_{\ell m}$$

caveat: picking 2 closest leaves is not good enough, i.e. picking i, j with minimal d_{ij} is not good enough



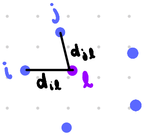
$d_{12} = 6$ ← smallest distance in the tree, but these nodes are not neighbors
$d_{13} = 7$
$d_{34} = 10$
$d_{42} = 7$

define: $D_{ij} = d_{ij} - (a_i + a_j)$ → now it can be proved that, for a pair of leaves for which D_{ij} is minimum, they are neighboring leaves (have a common parent)

$$a_i = \frac{1}{|L|-2} \sum_{\ell \in L} d_{i\ell}$$

L = set of leaves

neighbor-joining alg



input: set of n sequences

pairwise distance matrix d_{ij}

initialize:

T = set of leaf nodes, one for each input sequence

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

iteration:

pick a pair i, j from L for which D_{ij} is minimal

define a new node ℓ and set $d_{\ell m} = \frac{1}{2}(d_{im} + d_{jm} - d_{ij})$

add ℓ to T with edge length $d_{i\ell} = \frac{1}{2}(d_{ij} + a_i - a_j)$ and $d_{j\ell} = d_{ij} - d_{i\ell}$

Remove i and j from L and add ℓ to L # recalculate D_{ij} each iteration to account for the new collection of nodes L

Termination:

when L consists of 2 leaves $\overset{(i,j)}{;}$ add the remaining edge between i and j with length d_{ij}

* Look @ the phylogeny slides to see this alg in action *