The Neighbor-Joining Method

Input is a matrix of pairwise distances.

"Molecular clocks" & the ULTRAMETRIC Properties of distances.

UPGMA algorithm constructs phylogenetic trees where the edge lengths (estimation of evolutionary distance)
molecular clock has constant rate of evolution on all branches. "uniform".

What this means is that the sum of the times down the path from the root to the leaves is the same.

we add up the edge costs along a path in the tree.
Now we consider a weaker assumption about the molecule choice:

**Additivity**

Cost of path is sum of cost of the edges of the path.

True for UPGMA

**The Ultrametric Condition**

A set of distances \( D = \{d_{ij}\} \) is said to be ultrametric if for any triplet of sequences \( i,j,k \), the distances \( d_{ij}, d_{jk}, d_{ik} \) are either all equal or two are equal and the third one is smaller.
If additivity holds but uniform molecular clock fails the
"neighbor joining" algorithm can construct a phylogenetic
tree from the input distance matrix.

Intuition

Edge given a tree $T$ with additive
lengths, we will reconstruct $T$ from
pairwise distances between its
leaves as follows:

Find a pair of neighboring
leaves, say $i, j$, i.e.
leaves that have the same
parent \( e \) in \( T \) between neighboring leaves of \( T \)

- Remove \( ij \) from the list of leaves and add \( e \) to the current list of nodes and define the distance

\[
    d_{im} = \frac{1}{2} (d_{im} + d_{jm} - d_{ij})
\]

for all the other nodes \( m \) as follows:
By additivity the distances $d_{i,j}$ are exactly what they should be.

In this way we can eliminate leaves reducing the number of leaves by 1, until we end up with just a pair of leaves. That is the last clustering step.

---

Caveat

Picking the two closest leaves (smallest distance) is not good enough!!!
picking \( i, j \) with \( d_{ij} \) minimal may not work because they are not adjacent leaves in the reconstructed tree.

Which are the pair of leaves at min distance?

But they are not adjacent leaves in \( T \).

The algorithm avoids the caveat as follows:

Construct a new distance matrix

\[
D = \frac{1}{2} dist^* \text{ is the input matrix of pairwise distances}
\]
$M = \sum M_{ij}^2$ a new matrix of pairwise distances

There is a correction step for for every species

$$a_i = \frac{1}{|\mathcal{L}| - 2} \sum_{n \neq i, n \in \mathcal{L}} d_{in}$$

Species i

$L = \text{set of leaves}

|L| = \text{size of L}$

$a_1, a_2, \ldots$

Now we can prove that for a pair of leaves $i, j$ for which

$M_{ij}$ is minimum

it follows they are neighboring leaves in the tree.

$M_{ij} = \delta_{ij} - (a_i + a_j)$
THE NEIGHBOR JOINING ALG

Initialization:

INPUT: Set of sequences/species and their matrix of pairwise distances

\( T = \) the set of leaf nodes, one for each sequence

\( L = T \)

\[ q_i = \frac{1}{11-2} \sum_{n \in L} d_{in} \]

\[ M_{ij} = s_{ij} - (q_i + q_j) \]

Iteration

- Pick a pair \( i, j \) from \( L \) for which \( M_{ij} \) is minimal

- Define a new node \( \ell \) and set
\[ d_{mn} = \frac{1}{2} (d_{im} + d_{jm} - d_{ij}) \]

for every \( m \in L \)

Add \( e \) to \( T \) with edge length

\[ d_{ie} = \frac{n}{2} \left( d_{ij} + (a_i + a_j) \right) \]

\[ d_{je} = d_{ij} - d_{ie} \]

join \( e \) to \( i \) and \( j \)

\[ T = \{ . , \hat{e} , \ldots \} \]

Remove \( i \) and \( j \) from \( L \)

Add \( e \) to \( L \)
TERMINATION

When L contains only two leaves i and j, then add an edge between them with edge length \( d_{ij} \). 

---

Two other Types of phylogeny algorithms.

**PARSimony** = the shortest based phylogeny explanation.

---

Margaret Dayhoff

A A
B C
A B
A C

PAM

BLOSUM
DAYhoff

"molecular clock"

Smallest site tree

substitution along the edges of the most parsimonious tree

1PAM = 1 in 100 amino acid changes in protein

PAM 50
PAM 200

Probabilistic Phylogenetic Trees

Maximum Likelihood Solution