Phylogenetic trees

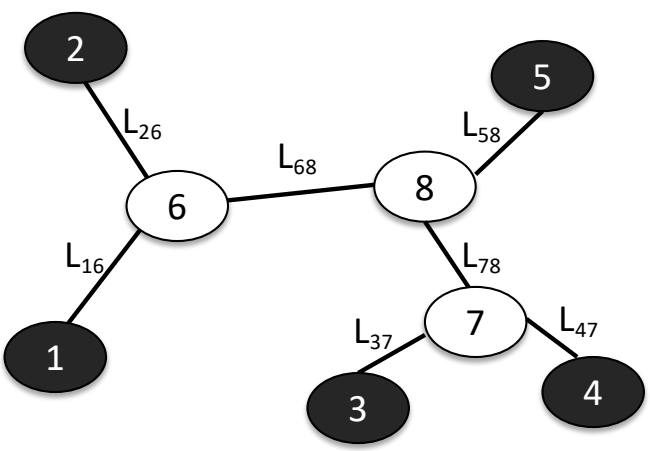
Distance-based methods and UPGMA

Outline

- Distance-based methods for phylogenetic tree estimation
- Computing distances between sequences
- The UPGMA distance-based method
- The “molecular clock assumption”

Basic idea of distance-based methods

- Suppose we can compute a "distance", d_{ij}, between each pair of taxa based on some data (e.g., sequences)
- Can we come up with a tree structure (with lengths assigned to branches) that accurately reflect the pairwise distances?
  - i.e., is there a tree such that d_{ij} is equal to the length of the path between i and j in the tree, for all pairs of taxa, i and j?



$$d_{15} \stackrel{?}{=} L_{16} + L_{68} + L_{58}$$

Distance-based methods for phylogenetic tree reconstruction

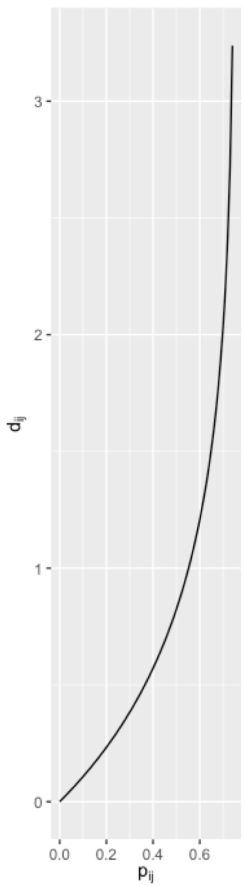
- Given $n \times n$ distance matrix for n entities (e.g., taxa), construct the tree for these n entities
- Algorithms
  - UPGMA
  - Neighbor joining
- Assume additivity and sometimes a "molecular clock"
- Additivity means we can add up the branch lengths of the tree connecting two nodes and get their distances
  - In other words, "additivity" of the distances means that there exists some tree that perfectly explains these distances
- In practice, distances will only be approximately additive

Defining distance between sequences

- Fractional alignment mismatch for two sequences *i* and *j*
  - $p_{ij} = m_{ij}/L_{ij}$
    - Gives an estimate of changes per site
  - m_{ij}: Number of mismatches between sequences *i* and *j*
  - L_{ij}: Number of aligned positions between sequences *i* and *j*
  - Assumes that changes have happened only once
    - Underestimates the distance between sequences
- Jukes Cantor distance
  - Removes assumption above
  - The most likely evolutionary distance d_{ij} between sequences *i* and *j*, where p_{ij} is the fractional mismatch defined above

$$d_{ij} = -\frac{3}{4}\ln(1 - \frac{4}{3}p_{ij})$$

Jukes Cantor Distance



$$d_{ij} = -\frac{3}{4}\ln(1 - \frac{4}{3}p_{ij})$$

UPGMA algorithm for phylogenetic tree reconstruction

- UPGMA: Unweighted pair group method using arithmetic averages
- Start with each taxon as its own disconnected node
- At each step, merge two closest nodes to create a new node in the tree
  - Set new node at height determined by nodes being merged
  - Recompute distance between new node and all other nodes
- Intermediate nodes will correspond to a set of taxa
- We will call taxa associated with an intermediate node i cluster C_i
- Need to compute
  - Distance between two clusters
  - Height

Computing distance between clusters

- Let i and j be two nodes
- Let C_i be the cluster of taxa for node i
- Let C_j be the cluster of taxa for node j
- lC_jl: Number of taxa in C_j
- Distance between nodes i and j

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

Computing distance from a new node

- Let k be a new node to be created from merging i and j
- Let C_i be the cluster of taxa for node i
- Let C_j be the cluster of taxa for node j
- Distance d_{kl} between nodes k and l, $l \neq i$ and $l \neq j$

$$d_{kl} = \frac{1}{|C_k||C_l|} \sum_{p \in C_k, q \in C_l} d_{pq}$$

- This is equal to

$$d_{kl} = \frac{d_{il}|C_i| + d_{jl}|C_j|}{|C_i| + |C_j|}$$

Derivation of the efficient distance calculation

$$
\begin{aligned}
d_{kl} &= \frac{1}{|C_k||C_l|} \sum_{p \in C_k, q \in C_l} d_{pq} \\
&= \frac{1}{|C_k||C_l|} \left(\sum_{p \in C_i, q \in C_l} d_{pq} + \sum_{p \in C_j, q \in C_l} d_{pq} \right) \\
&= \frac{1}{|C_k||C_l|} \left(|C_i||C_l| d_{il} + \left|C_j\right||C_l| d_{jl} \right) \\
&= \frac{|C_i| d_{il} + \left|C_j\right| d_{jl}}{|C_k|} \\
&= \frac{|C_i| d_{il} + \left|C_j\right| d_{jl}}{|C_i| + \left|C_j\right|}
\end{aligned}
$$

UPGMA algorithm

- Input
  - n taxa
  - Distance matrix for all pairs of *n* taxa, d_{ij}
- Output
  - Tree T
- Initialization
  - Assign each taxon i to its own cluster C_i
  - Define one leaf of T for each taxon
- Iterate until only two clusters remain
  - Find two nodes C_i and C_j that have the smallest d_{ij}
  - Define new cluster $C_k = C_i \cup C_j$
  - Define daughters of k as i and j, place at height $d_{ij}/2$
  - Add k to cluster set. Remove i and j from the set of clusters
- Terminate
  - When only two clusters C_i and C_j remain, place root at $d_{ij}/2$

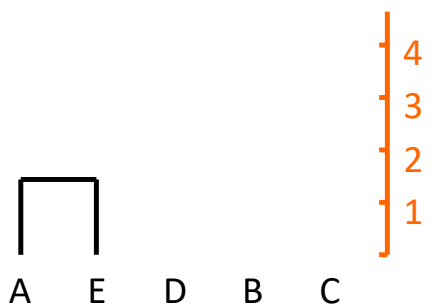
UPGMA example

initial state

	A	B	C	D	E
A	0	8	8	5	3
B		0	3	8	8
C			0	8	8
D				0	5
E					0

after one merge

	AE	B	C	D
AE	0	8	8	5
B		0	3	8
C			0	8
D				0



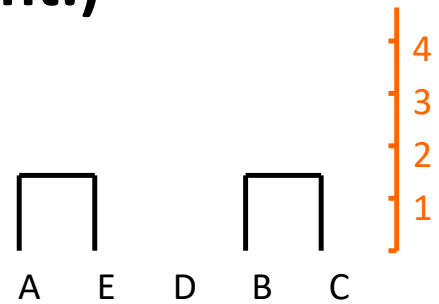
Example calculation

$$d_{(A,E)B} = \frac{d_{AB} + d_{EB}}{1+1} = \frac{16}{2}$$

UPGMA example (cont.)

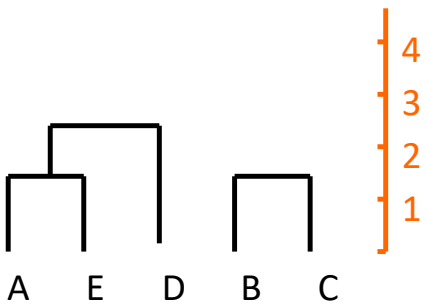
after two merges

	AE	BC	D
AE	0	8	5
BC		0	8
D			0



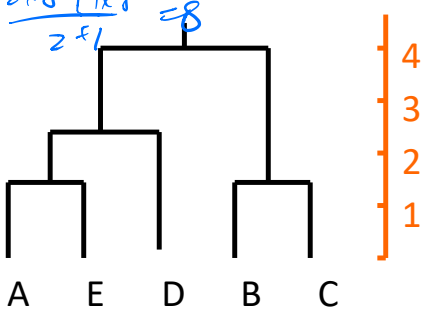
after three merges

	AED	BC
AED	0	8
BC		0



$d_{AED,BC} = \frac{2 \times d_{AE,BC} + 1 \times d_{D,BC}}{2+1} = \frac{2\times 8 + 1\times 8}{2+1} = 8$

final state



UPGMA relies on the molecular clock assumption

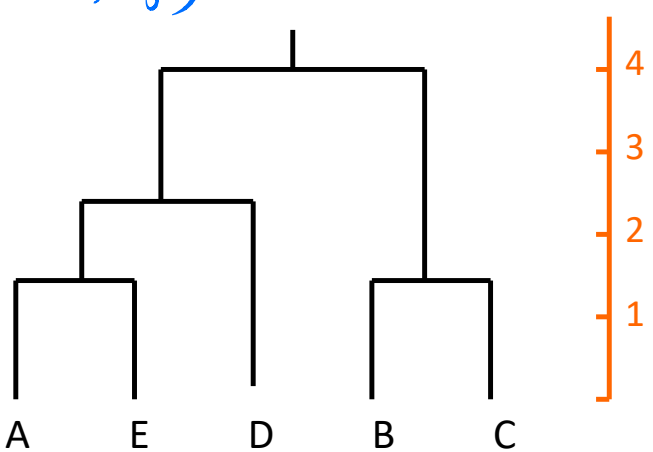
- Sequences diverge at the same rate at all points in the phylogeny
- Distance from any leaf to root is the same.
- If this is true the distances are said to have an "ultrametric" property
- This assumption is rarely true in practice

The molecular clock assumption & ultrametric data

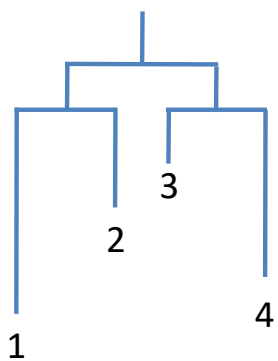
- Ultrametric data: for all triplets of taxa, i, j, k, the pairwise distances between them are either all equal, or two are equal and the remaining one is smaller

$d_{ij} \leq \max(d_{ik}, d_{jk})$

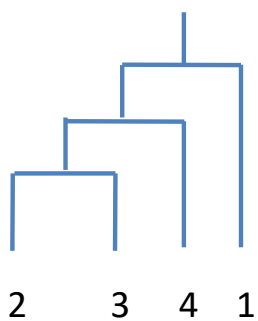
	A	B	C	D	E
A	0	8	8	5	3
B		0	3	8	8
C			0	8	8
D				0	5
E					0



Problem with UPGMA when the molecular clock assumption does not hold



Actual tree

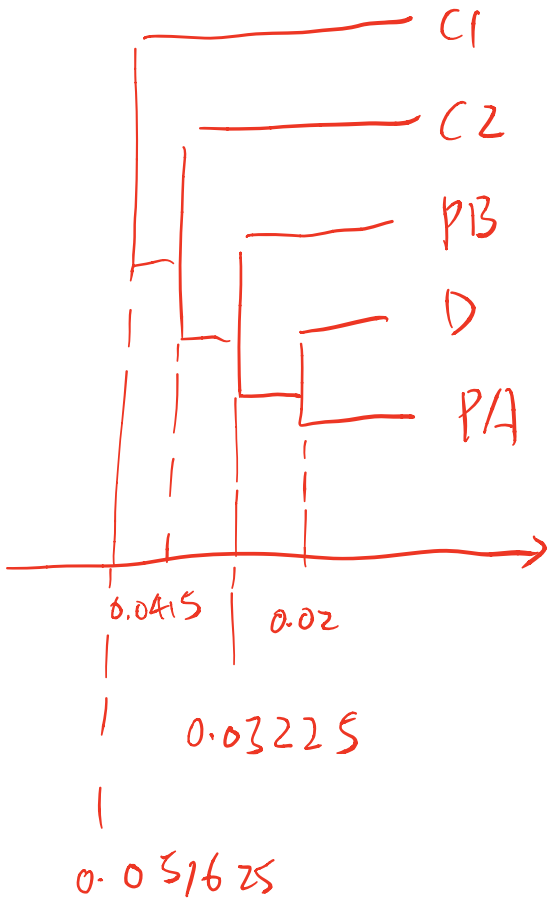


Constructed by UPGMA

Summary

- Distance-based methods construct trees that attempt to fit pairwise distance data
- Pairwise distances may be computed from pairwise alignments using the appropriate corrections
- UPGMA is a distance-based method that is successful if the data are ultrametric or perhaps approximately ultrametric

	C1	C2	D	PA	PB
C1	0	0.09	0.098	0.105	0.12
C2		0	0.072	0.076	0.101
D			0	0.04	0.061
PA				0	0.068
PB					0



	D, PA	C1	C2	PB
D, PA	0	0.1015	0.074	0.0645
C1		0	0.09	0.12
C2			0	0.101
PB				0

	D, PA, PB	C1	C2
D, PA, PB	0	0.107667	0.083
C1		0	0.09
C2			0

	D, PA, PB, C2	C1
D, PA, PB, C2	0	0.16325
C1		0