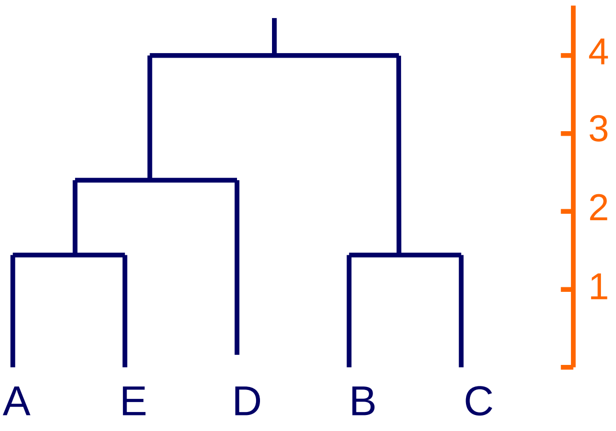
Distance-Based Approaches to Inferring Phylogenetic Trees

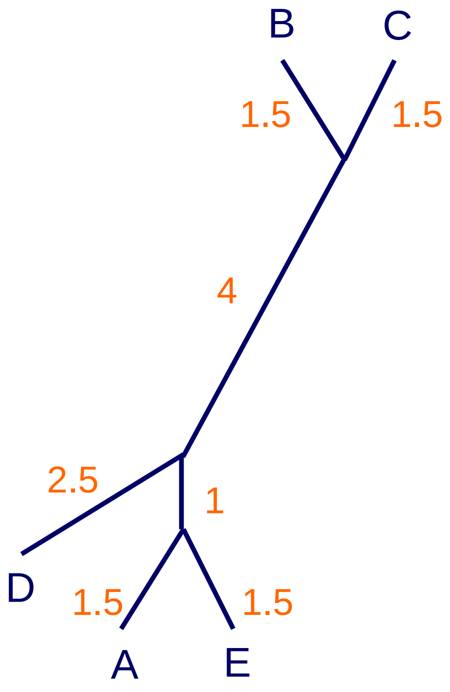
BMI/CS 576
www.biostat.wisc.edu/bmi576.html
Mark Craven
craven@biostat.wisc.edu

Representing distances in rooted and unrooted trees

dist(A,C) = 8
dist(A,D) = 5



distances represented by summed height of edges to reach common ancestor

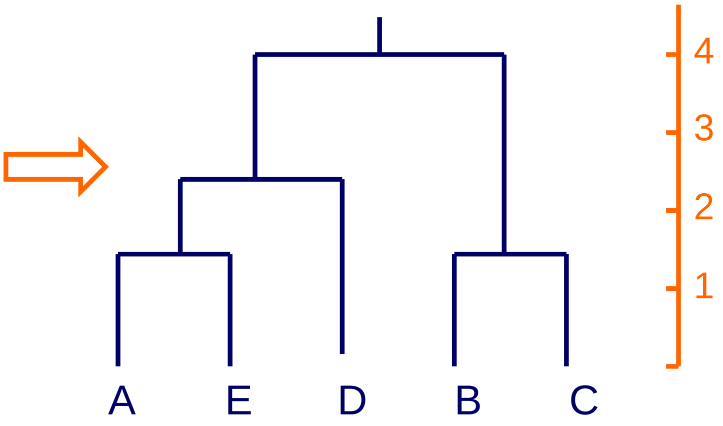


distances represented by summed length of edges to reach common ancestor

Distance-based approaches

- **given**: an $n \times n$ matrix M where M_{ij} is the distance between taxa i and j
- **do**: build an edge-weighted tree such that the distances between leaves i and j correspond to M_{ij}

	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



Where do we get distances?

- commonly obtained from sequence alignments

$$f_{ij} = \frac{\#\text{mismatches}}{\#\text{matches} + \#\text{mismatches}}$$

in alignment of sequence i with sequence j

$$\text{dist}(i, j) = f_{ij}$$

- to correct for multiple substitutions at a single position:

$$\text{dist}_{\text{Jukes-Cantor}}(i, j) = -\frac{3}{4}\ln\left(1 - \frac{4}{3} f_{ij}\right)$$

Distance metrics

- properties of a distance metric

$$\text{dist}(x_i, x_j) \geq 0$$

$$\text{dist}(x_i, x_i) = 0$$

$$\text{dist}(x_i, x_j) = \text{dist}(x_j, x_i)$$

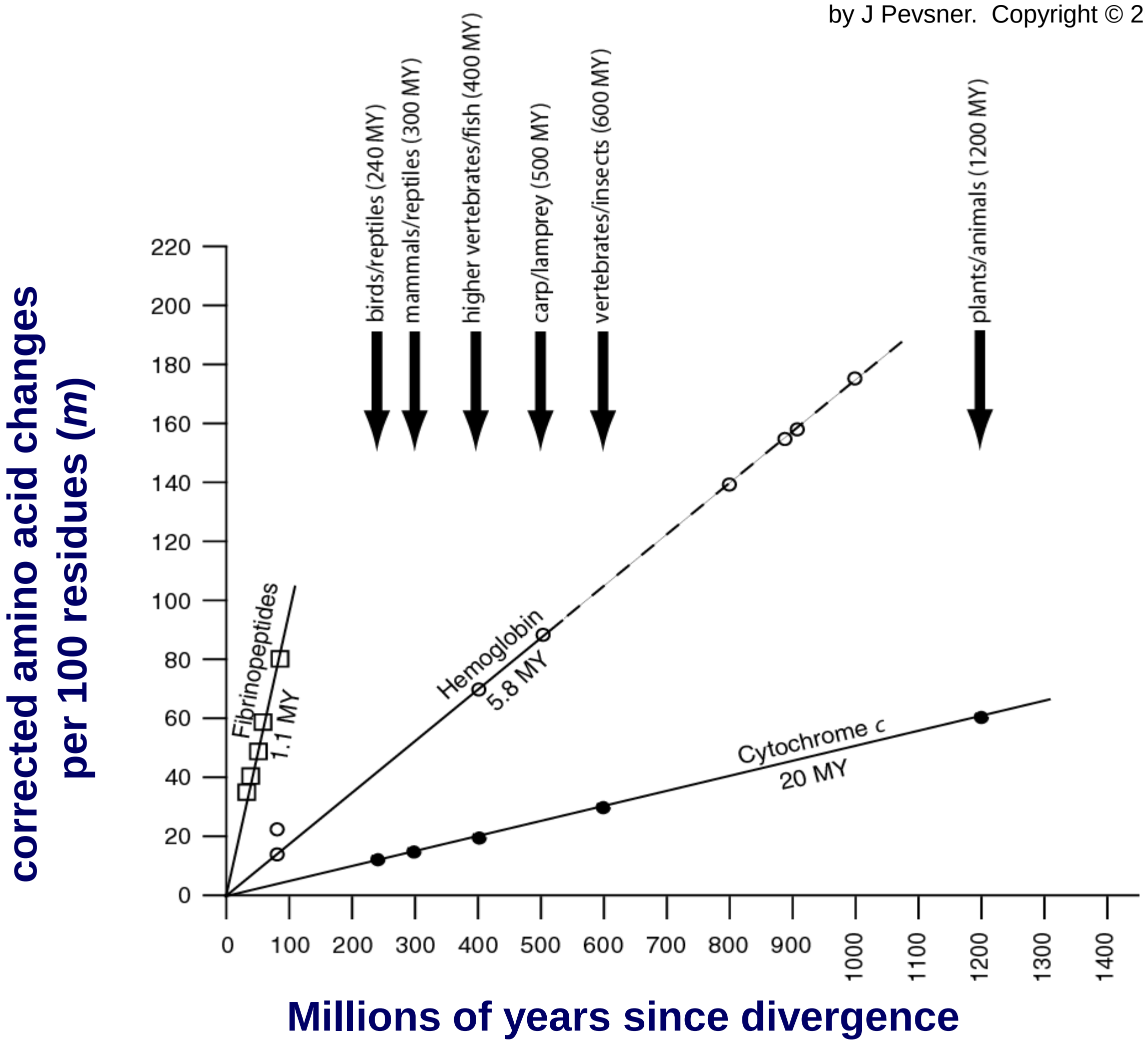
$$\text{dist}(x_i, x_j) \leq \text{dist}(x_i, x_k) + \text{dist}(x_k, x_j)$$

The molecular clock hypothesis

- In the 1960s, sequence data were accumulated for small, abundant proteins such as globins, cytochromes c, and fibrinopeptides. Some proteins appeared to evolve slowly, while others evolved rapidly.

- Linus Pauling, Emanuel Margoliash and others proposed the hypothesis of a molecular clock: *For every given protein, the rate of molecular evolution is approximately constant in all evolutionary lineages*

Figure from *Bioinformatics and Functional Genomics* by J Pevsner. Copyright © 2009 by Wiley.



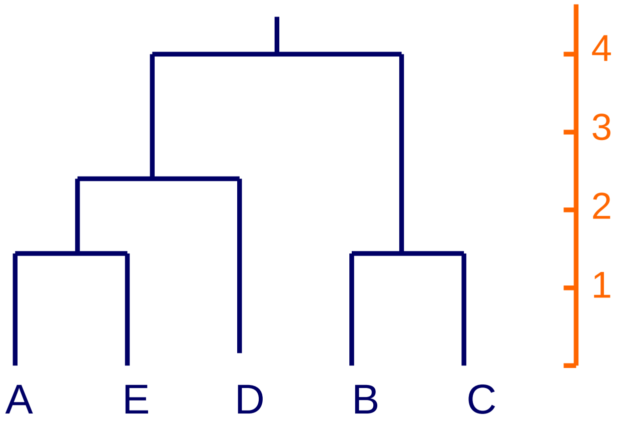
The molecular clock assumption & ultrametric data

- the molecular clock assumption is not generally true: selection pressures vary across time periods, organisms, genes within an organism, regions within a gene
- if it does hold, then the data is said to be *ultrametric*

The molecular clock assumption & ultrametric data

- ultrametric data: for any triplet of sequences, i, j, k, the distances are either all equal, or two are equal and the remaining one is smaller

	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



The UPGMA method

(Unweighted Pair Group Method using Arithmetic Averages)

- given ultrametric data, UPGMA will reconstruct the tree *T* that is consistent with the data
- basic idea:
  - iteratively pick two taxa/clusters and merge them
  - create new node in tree for merged cluster
- distance d_{ij} between clusters C_i and C_j of taxa is defined as

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

(avg. distance between pairs of taxa from each cluster)

UPGMA algorithm

```
assign each taxon to its own cluster
define one leaf for each taxon; place it at height 0
while more than two clusters
    determine two clusters $i, j$ with smallest $d_{ij}$
    define a new cluster $C_k = C_i \cup C_j$
    define a node $k$ with children $i$ and $j$; place it at height $d_{ij} / 2$
    replace clusters $i$ and $j$ with $k$
    compute distance between $k$ and other clusters
join last two clusters, $i$ and $j$, by root at height $d_{ij} / 2$
```

UPGMA

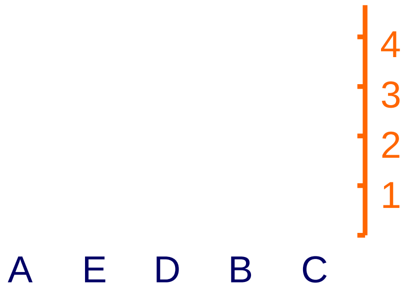
- given a new cluster C_k formed by merging C_i and C_j
- we can calculate the distance between C_k and any other cluster C_l as follows

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

UPGMA example

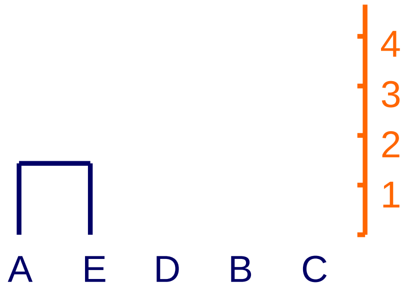
initial state

	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

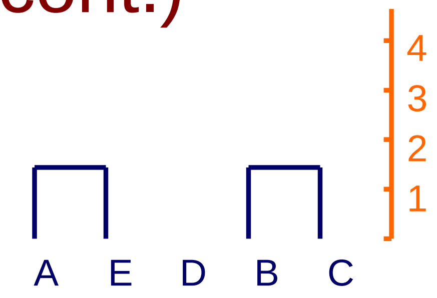
	AE	B	C	D
AE	0	8	8	5
B		0	3	8
C			0	8
D				0



UPGMA example (cont.)

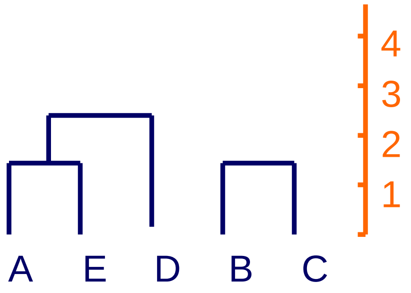
after two merges

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

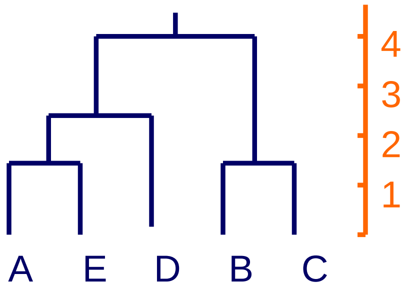


after three merges

	AED	BC
AED	0	8
BC		0



final state

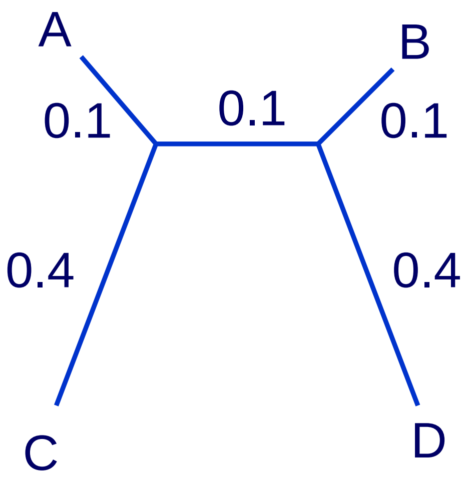


Neighbor joining

- unlike UPGMA
  - doesn't make molecular clock assumption
  - produces unrooted trees
- does assume *additivity*: distance between pair of leaves is sum of lengths of edges connecting them
- like UPGMA, constructs a tree by iteratively joining subtrees
- two key differences
  - how pair of subtrees to be merged is selected on each iteration
  - how distances are updated after each merge

Picking pairs of nodes to join in NJ

- at each step, we pick a pair of nodes to join; should we pick a pair with minimal d_{ij} ?
- suppose the real tree looks like this and we're picking the first pair of nodes to join?



$$\boxed{d_{AB} = 0.3}$$

$$d_{AC} = 0.5$$

- wrong decision to join A and B: need to consider distance of pair to other leaves

Picking pairs of nodes to join in NJ

- to avoid this, pick pair to join based on D_{ij} [Saitou & Nei ’87; Studier & Keppler ’88]

$$D_{ij} = d_{ij} - (r_i + r_j)$$

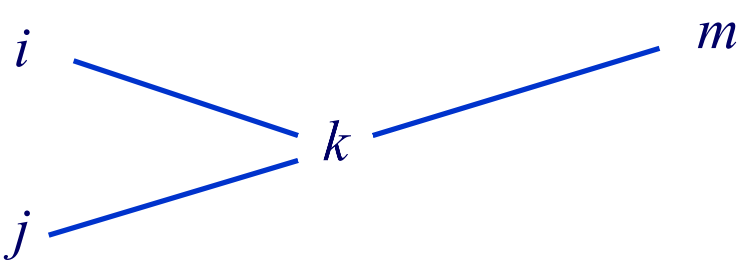
$$r_i = \frac{1}{|L|-2} \sum_{k \in L} d_{ik}$$

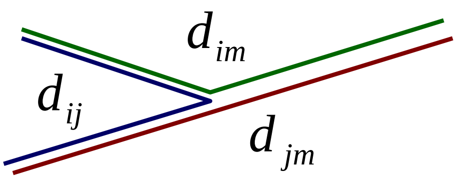
where L is the set of leaves

Updating distances in neighbor joining

- given a new internal node k, the distance to another node m is given by:

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

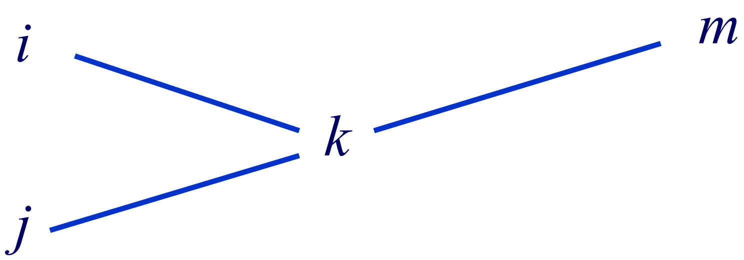




Updating distances in neighbor joining

- can calculate the distance from a leaf to its parent node in the same way

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



$$d_{jk} = d_{ij} - d_{ik}$$

Updating distances in neighbor joining

- we can generalize this so that we take into account the distance to all other leaves

$$d_{ik} = \frac{1}{2}(d_{ij} + r_i - r_j)$$

where

$$r_i = \frac{1}{|L|-2} \sum_{m \in L} d_{im}$$

and L is the set of leaves

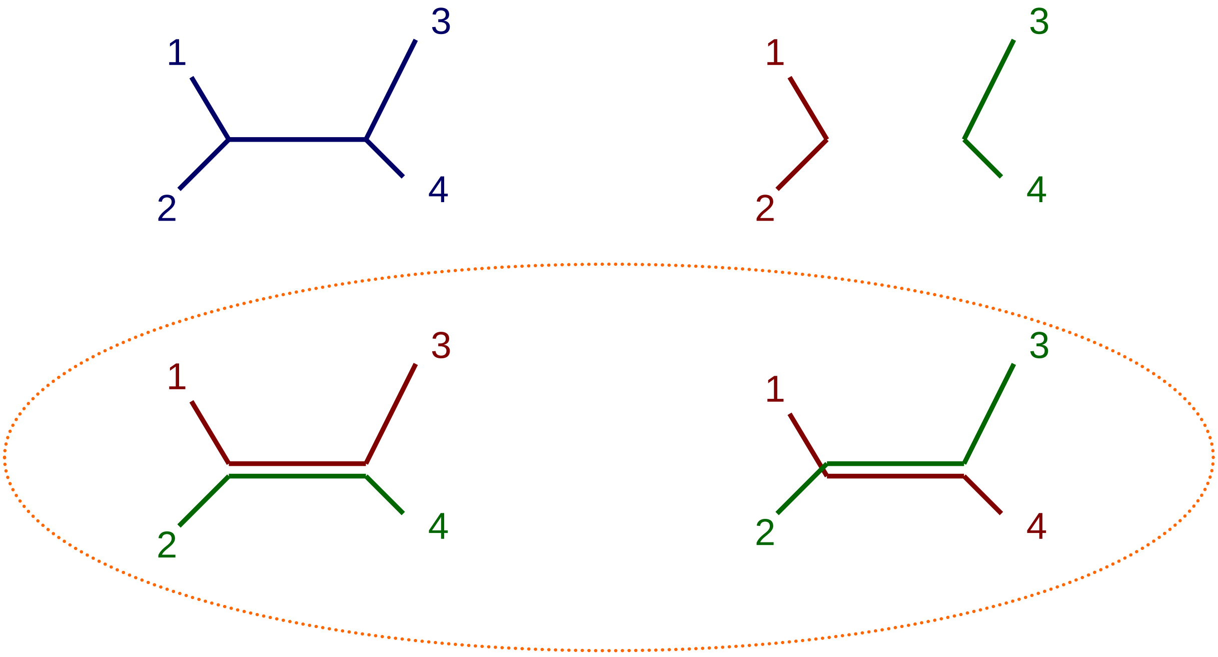
- this is more robust if data aren’t strictly additive

Neighbor joining algorithm

```
define the tree $T$ = set of leaf nodes
$L = T$
while more than two subtrees in $T$
    pick the pair $i, j$ in $L$ with minimal $D_{ij}$
    add to $T$ a new node $k$ joining $i$ and $j$
    determine new distances
        $d_{ik} = \frac{1}{2}\left(d_{ij} + r_i - r_j\right)$
        $d_{jk} = d_{ij} - d_{ik}$
        $d_{km} = \frac{1}{2}\left(d_{im} + d_{jm} - d_{ij}\right)$ for all other $m$ in $L$
    remove $i$ and $j$ from $L$ and insert $k$ (treat it like a leaf)
join two remaining subtrees, $i$ and $j$ with edge of length $d_{ij}$
```

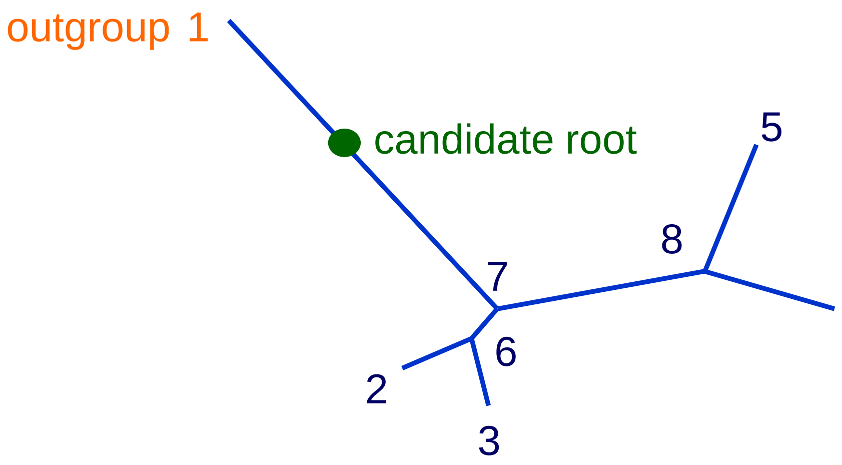
Testing for additivity

- for every set of four leaves, i, j, k, and l, two of the distances $d_{ij} + d_{kl}$, $d_{ik} + d_{jl}$ and $d_{il} + d_{jk}$ must be equal and not less than the third



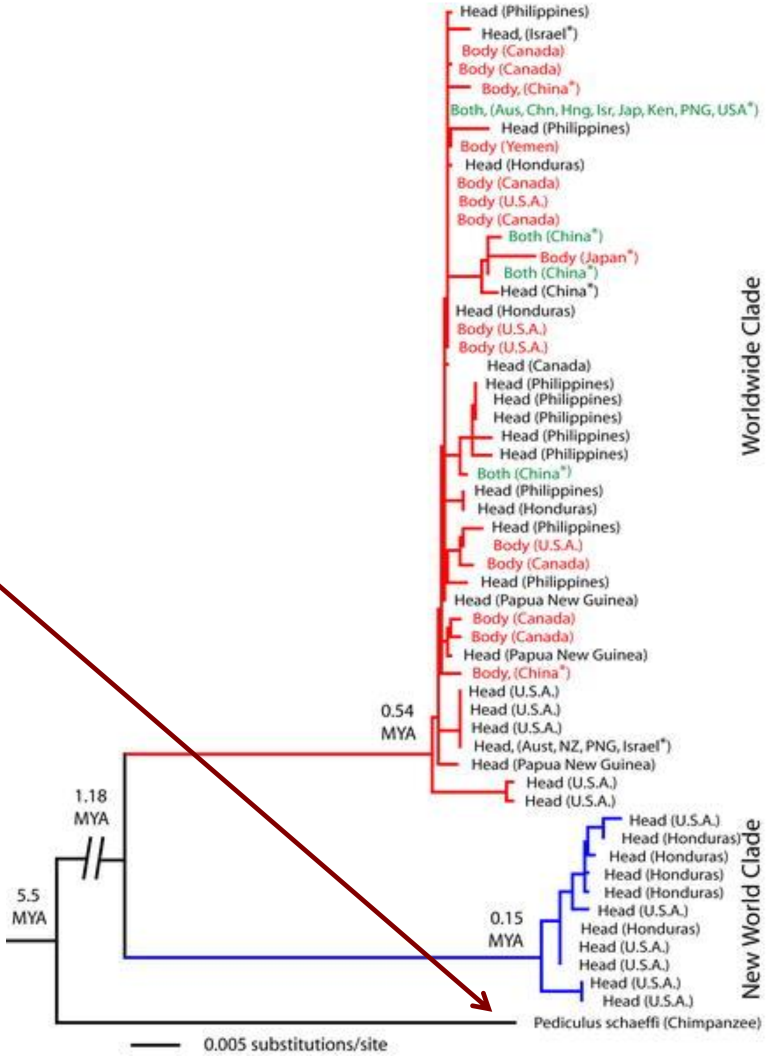
Rooting trees

- finding a root in an unrooted tree is sometimes accomplished by using an *outgroup*
- outgroup: a species known to be more distantly related to remaining species than they are to each other
- edge joining the outgroup to the rest of the tree is best candidate for root position



Rooting trees

chimpanzee lice used as outgroup in human lice study



Comments on distance-based methods

- if the given distance data is ultrametric (and these distances represent real distances), then UPGMA will identify the correct tree

- if the data is additive (and these distances represent real distances), then neighbor joining will identify the correct tree

- otherwise, the methods may not recover the correct tree, but they may still be reasonable heuristics

- neighbor joining is commonly used