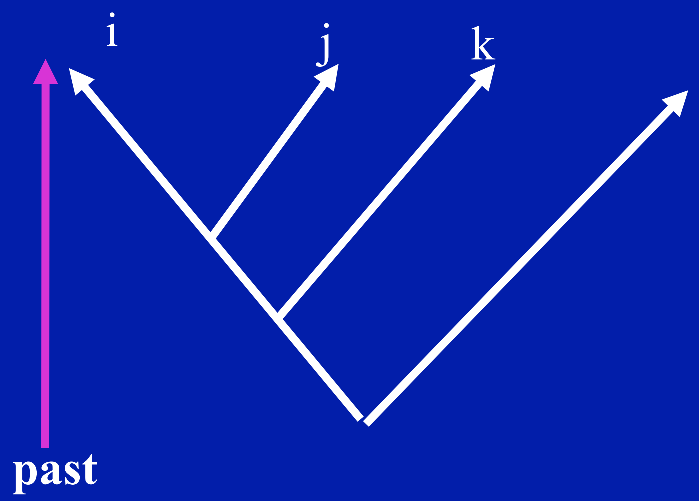
Phylogenetic Tree Construction

Clock-like Distances

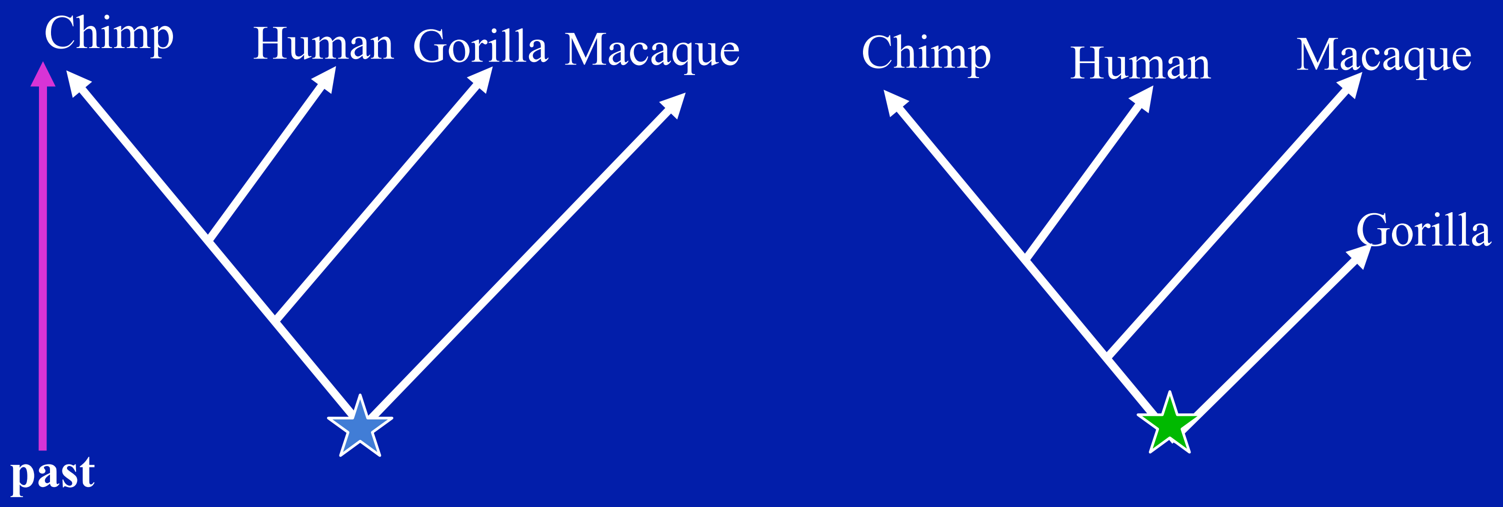


$D_{ik} = D_{jk}$ and $D_{ij} < D_{ik}$ guaranteed for any leaf nodes i, j, k. I.e. two distances must be equal, and the third must be smaller.

- Assume a rooted tree, and assert that the edge length is just the length of time of divergence from root. “Ultrametric”

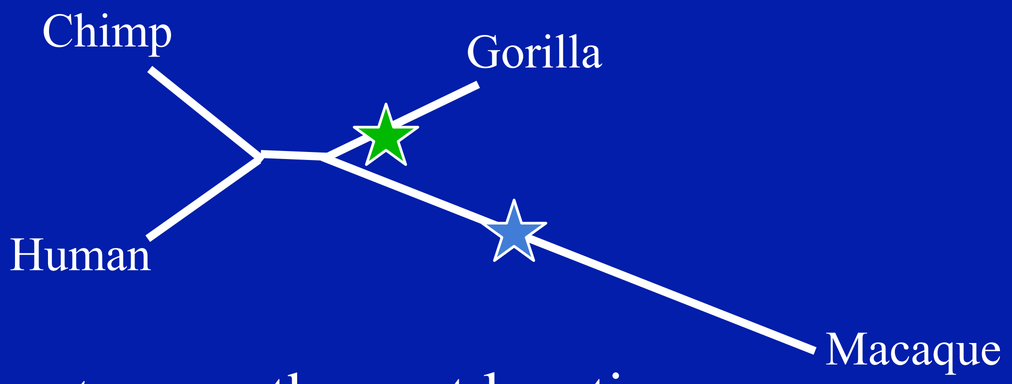
Choosing a Root Location?

Rooted Trees



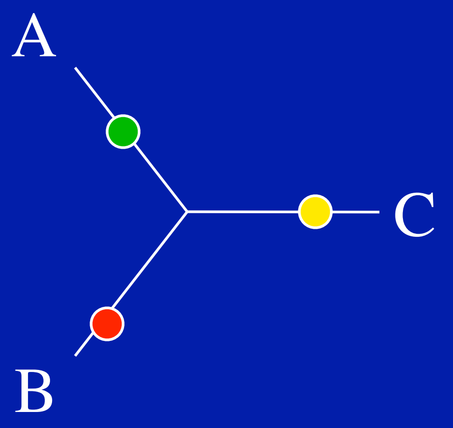
A rooted tree is just an unrooted tree with a *root location* chosen on one branch.

Unrooted Tree

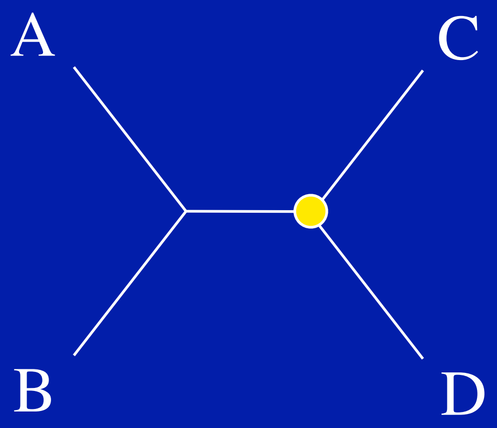


In the absence of an *outgroup*, the root location is often unclear and therefore arbitrary.

How Many Possible Trees?

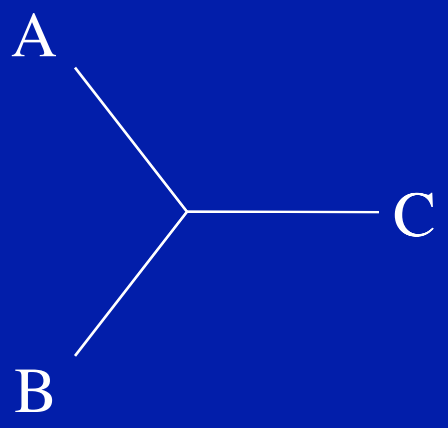


For *n*=3, only one possible tree (shuffling A,B,C has no effect on tree topology).

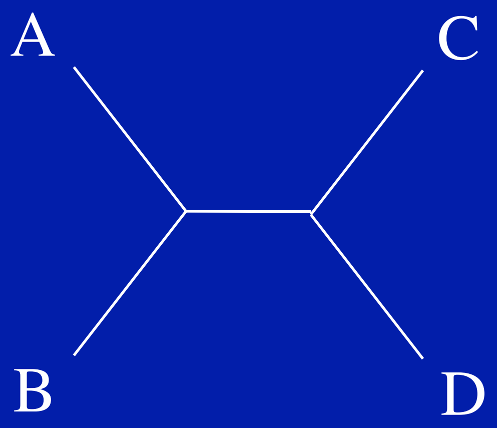


In general, a new leaf can be added to any of the existing internal or branches.

How Many Possible Trees?

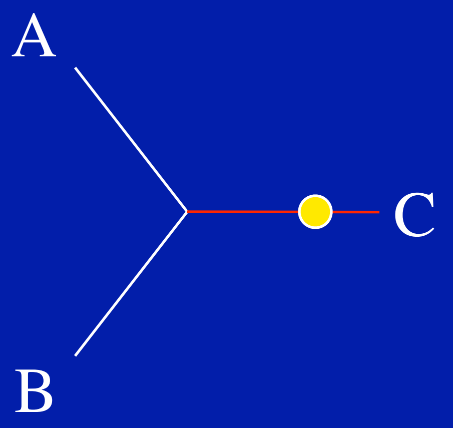


For *n*=3, only one possible tree (shuffling A,B,C has no effect on tree topology). Tree has 3 “leaves” and 3 “branches”

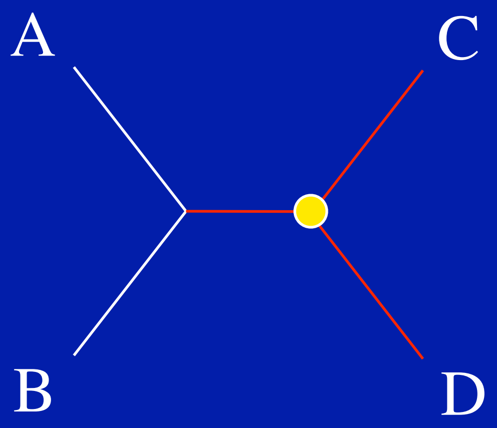


For *n*=4, three possible ways of partnering D with A, B, or C. Tree has 4 leaves and 4+1 branches.

How Many Branches?



For n>3, each new leaf adds *two* new branches.



So for n leaves, the number of branches is:
$m = 2(n\text{-}3)+3 = 2n\text{-}3$

Number of Unrooted Trees b_n

- So $b_n = (2(n\text{-}1)\text{-}3)b_{n\text{-}1} = (2n\text{-}5)b_{n\text{-}1}$
- In general, for $n \geq 3$

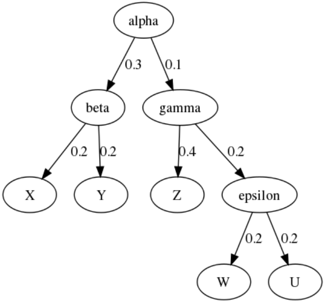
$$\boxed{b_n = \frac{(2n-5)!}{(n-3)!2^{n-3}}}$$

which grows astronomically, becoming impractical even for small n

Viterbi on Trees

Differences vs. a linear HMM:

- unrooted trees are undirected graphs, whereas HMM is directed. Let's restrict our attention to rooted trees (which have directed edges).
- instead of each variable transitioning to a single "next" variable, it has two "descendant" variables in a binary tree.
- instead of each variable emiting an observation, only the "terminal" nodes (i.e. modern sequences) are observed.



Which Direction to Optimize?

- following directed edges in "forward" order, starting from the root: maximizing $p(\alpha \to \beta \to \gamma \to ...)$ (where α is the root) is straightforward but not meaningful (no observations on this path, so no basis for inference).
- following directed edges in "reverse" order, starting from the leaves: define the *descendants* of α as D_α (i.e. all nodes below α in the tree). Consider the probability $p(D_\alpha|\alpha)$. Since D_α always includes leaf nodes (observations), this gives us a basis for inferences about α.

Viterbi Maximization

- Define $p^*(D_\alpha|\alpha)$ as the maximum probability of the descendants D_α out of all possible values of the descendants D_α, given a specific value of the ancestor α.
- Thanks to the Markov property it can be maximized recursively.
- Say α has two child nodes β, γ. Given a specific value of α, we then find the values of β, γ that maximize the total probability of the descendants of α:

$$p^*(D_\alpha|\alpha) = \mathrm{Max}\left[p^*(D_\beta|\beta)p(\beta|\alpha)p^*(D_\gamma|\gamma)p(\gamma|\alpha)\right]$$

Viterbi at the Leaf Nodes

Let's compute $p^*(D_\alpha|\alpha)$ for the simplest possible case, where both child nodes of α are observed (i.e. leaf nodes; call them X, Y). Furthermore, assume that all of the variables (sequences) consist of just a single nucleotide (e.g. $X = \text{A}, Y = \text{G}$). Write an equation for how you would compute the Viterbi maximum probability $p^*(D_\alpha|\alpha)$ in this case.

Viterbi at the Leaf Nodes Answer

- X, Y have no descendants, so in effect $p^*(D_X|X) = 1$.
- Furthermore, since X,Y are directly observed, only one state (the observed state) is allowed for them, so there is nothing to maximize over.

$$p^*(D_\alpha|\alpha) = p(X|\alpha)p(Y|\alpha)$$

Viterbi at the Root

We wish to find the values of all our variables (i.e. all internal nodes) $\alpha, \beta, \gamma, ...$ that maximize the total probability of the tree $p(\alpha, \beta, \gamma, ..., X, Y, Z, ...)$, where $X, Y, Z, ...$ are the observed sequences (leaf nodes).

- Say you are given $p^*(D_\alpha|\alpha)$ for the root node α.
- How would you find the value of α that maximizes $p(\alpha, \beta, \gamma, ..., X, Y, Z, ...)$, and the associated maximum probability $p^*(\alpha, \beta, \gamma, ..., X, Y, Z, ...)$?

Viterbi at the Root Answer

By the chain rule and Markov property,

$$p(\alpha,\beta,\gamma,...,X,Y,Z,...) = p(\alpha)p(\beta,\gamma,...,X,Y,Z,...|\alpha) = p(\alpha)p(D_\alpha|\alpha)$$

We find its maximum by finding the value of α that maximizes

$$p^*(\alpha,\beta,\gamma,...,X,Y,Z,...) = p(\alpha)p^*(D_\alpha|\alpha)$$

E.g. for a single nucleotide site, under the Jukes-Cantor model, $p(\alpha) = \frac{1}{4}$.

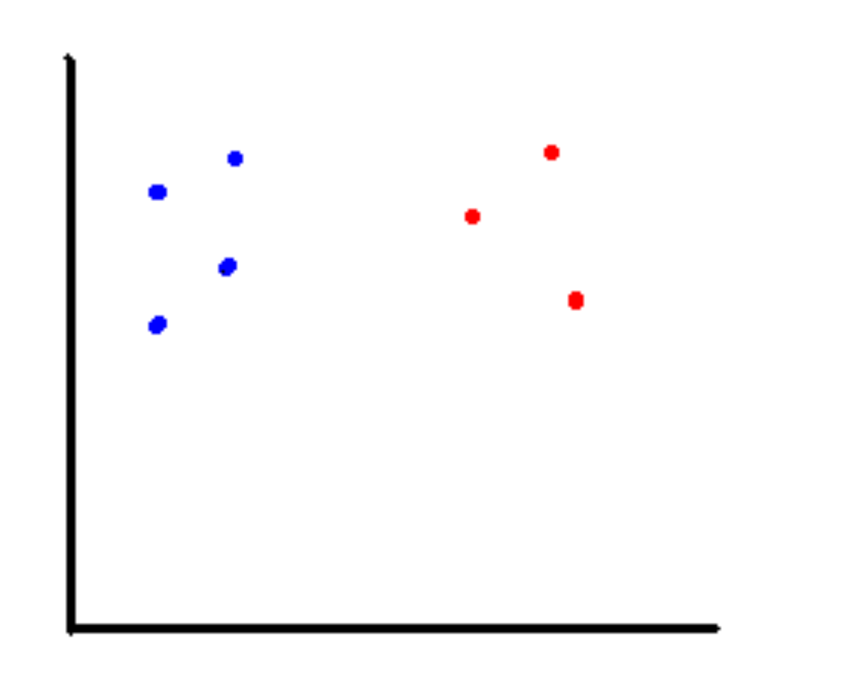
Finding the Most Parsimonious Tree

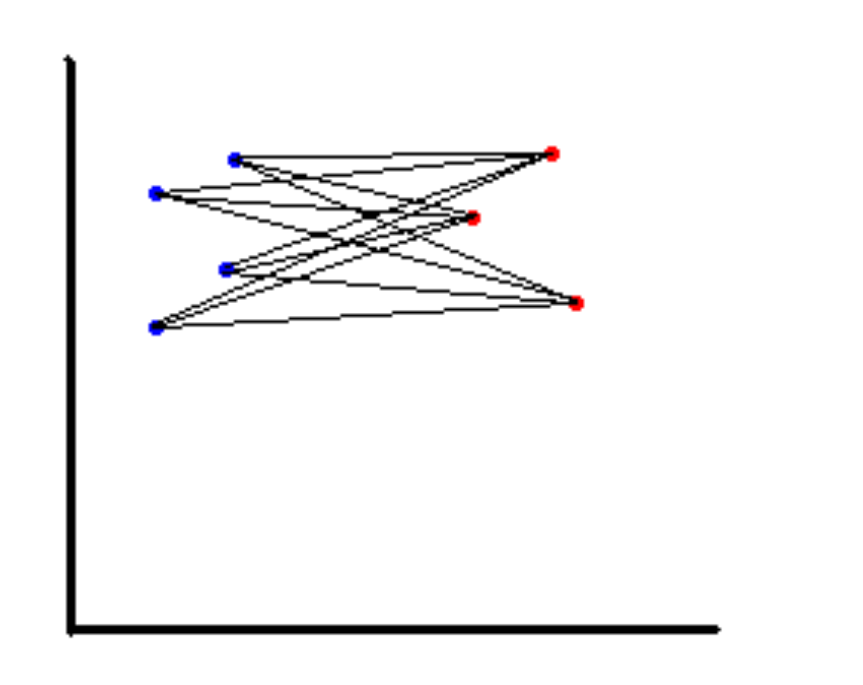
- Theoretically, enumerating all possible trees and calculating the number of mutations required on each, would yield the most parsimonious tree.
- Too many trees to enumerate exhaustively.
- Must exclude whole families of trees that are provably not optimal, or use heuristics.

Heuristic: Hierarchical Clustering

- Same basic method as used for clustering gene expression data.
- Requires *distance metric* for each pair.
- *Agglomerative*: join sequences that are closest together first
- *Mean distance*: the distance between a pair of clusters is the average of the distance between their members.

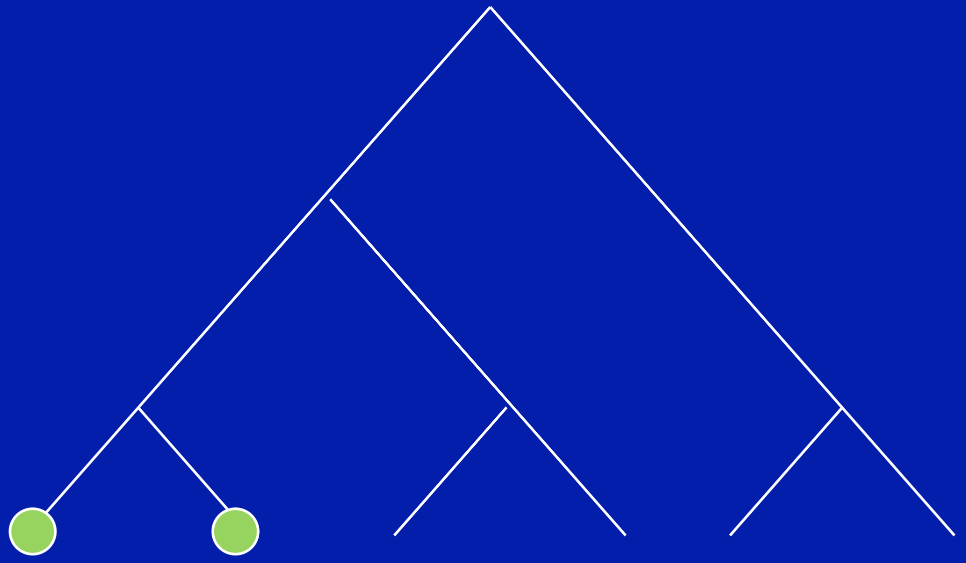
Unweighted Pair Group Method with Arithmetic mean (UPGMA)





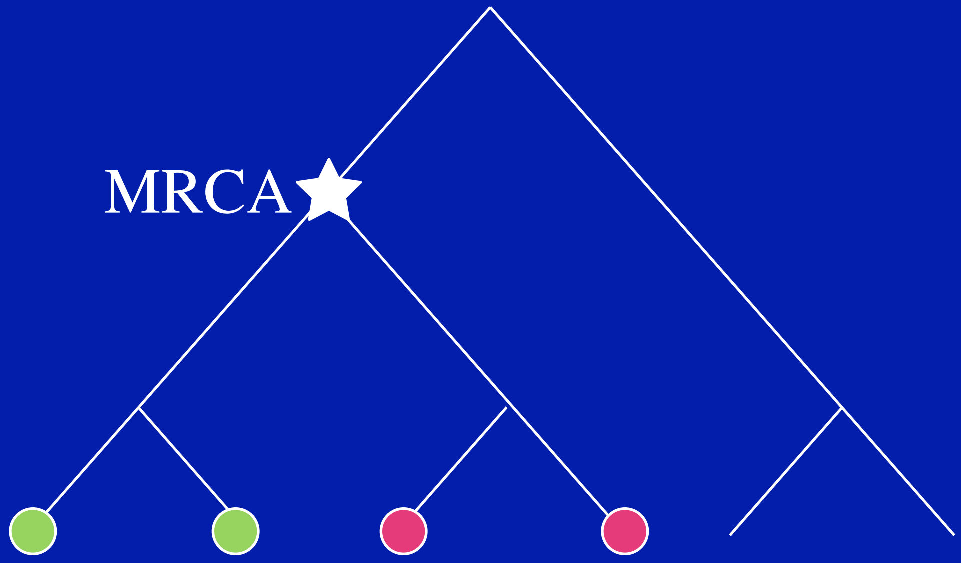
$$d_{GA}(G,H) = \sum_{i \in G} \sum_{i' \in H} d_{ii'} / (N_G N_H)$$

Ultrametric Distances Make it Easy to Find Neighbors



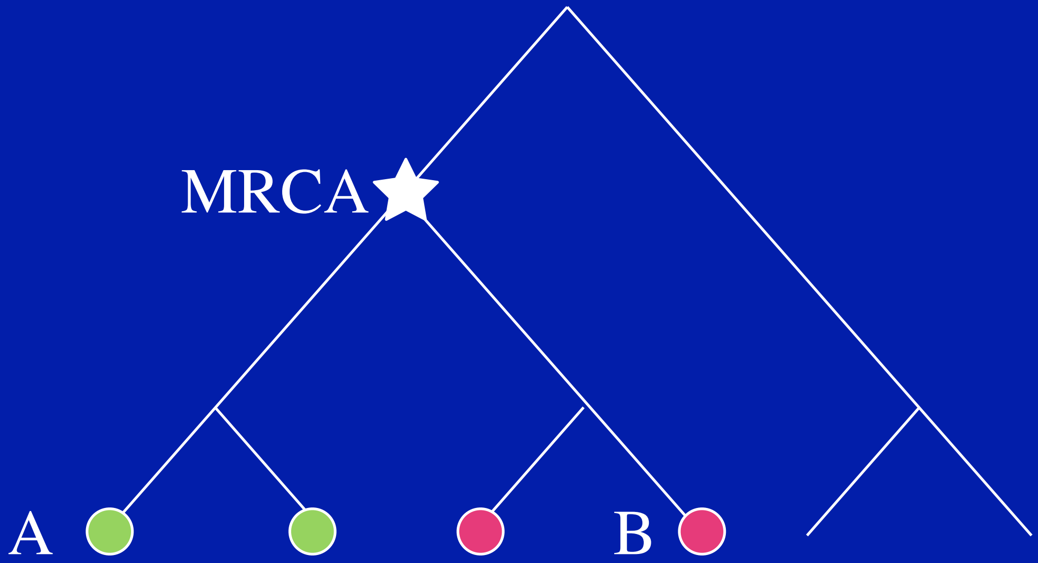
By the ultrametric assumption, the pair of nodes A B with minimum distance D(A,B) must be neighbors. (If another node C was A’s neighbor instead, then D(A,C)<D(A,B)).

UPGMA Fits the Ultrametric Distance Assumption



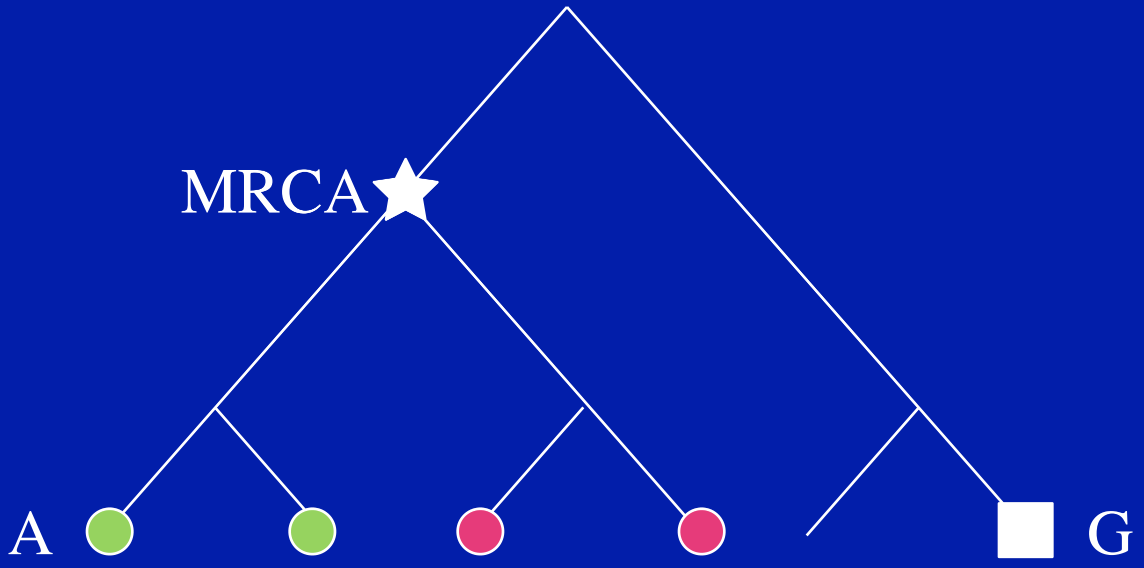
By ultrametric assumption, children descended from same most recent common ancestor (MRCA) must be same distance from each other. So the UPGMA average should equal that distance.

UPGMA Fits the Ultrametric Distance Assumption



Furthermore, the distance of all children from the MRCA should be equal. Therefore D(A, MRCA) = D(A,B)/2

UPGMA Fits the Ultrametric Distance Assumption



Finally, the distances from the MRCA to any node G that is *not* a child of the MRCA should be equal. Easy to calculate: D (MRCA,G) = D(A,G) - D(A,MRCA)

UPGMA Algorithm: iteratively join neighbors

Create initial list: each sequence is a separate cluster

While not all clusters joined

- Compute distances for all cluster pairs (UPGMA)
- Choose closest pair of clusters
- Join that pair into a single cluster: add a new node M midway between nodes representing the pair of clusters.

Example: Manhattan Distance

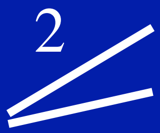
Characters

```
Chimp      ..AGCTAAAGGGTCAGGGGAAGGGCA..
Gorilla    ..AGCATAGGGGTCAGGGGAAAGGCT..
Human      ..AGCAAAAGGGTCAGGGGAAGGGGA..
Macaque    ..AGCTCATCGGTAAGGAGAAAGGAT..
Orangutan  ..AGCCCATCGGTCAGGAGAAAGGAT..
Squirrel   ..AGCGGACCGGTAAGGAGAAAGGAC..
```

Distances

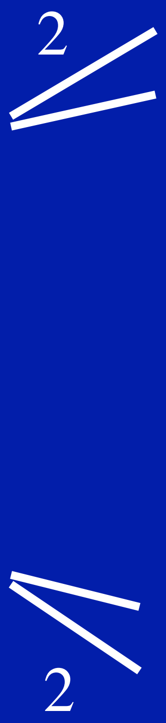
	Chimp	Gor	Hum	Mac	Orang	Sq
Chimp		5	2	8	8	9
Gor			5	7	6	8
Hum				9	8	9
Mac					2	4
Orang						5

Hierarchical Clustering



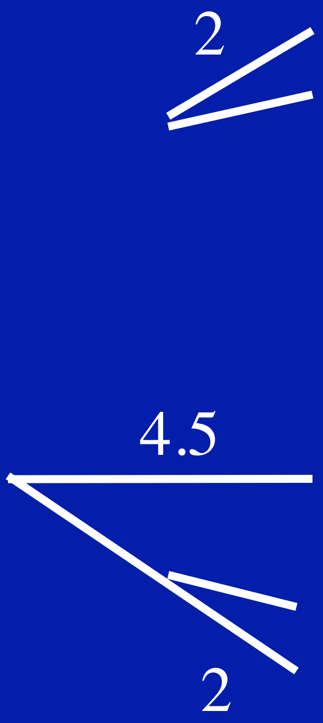
Chimp	..AGCTAAAGGGTCAGGGGAAGGGCA..
Human	..AGCAAAAGGGTCAGGGGAAGGGGA..
Gorilla	..AGCATAGGGGTCAGGGGAAAGGCT..
Squirrel	..AGCGGACCGGTAAGGAGAAAGGAC..
Macaque	..AGCTCATCGGTAAGGAGAAAGGAT..
Orangutan	..AGCCCATCGGTCAGGAGAAAGGAT..

Hierarchical Clustering



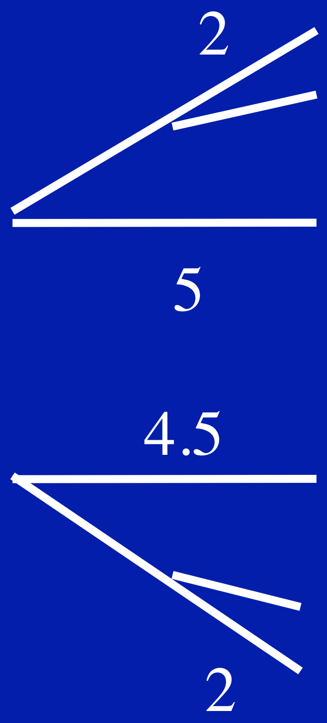
```
Chimp      ..AGCTAAAGGGTCAGGGGAAGGGCA..
Human      ..AGCAAAAGGGTCAGGGGAAGGGGA..

Gorilla    ..AGCATAGGGGTCAGGGGAAAGGCT..



Squirrel   ..AGCGGACCGGTAAGGAGAAAGGAC..

Macaque    ..AGCTCATCGGTAAGGAGAAAGGAT..
Orangutan  ..AGCCCATCGGTCAGGAGAAAGGAT..
```

Hierarchical Clustering



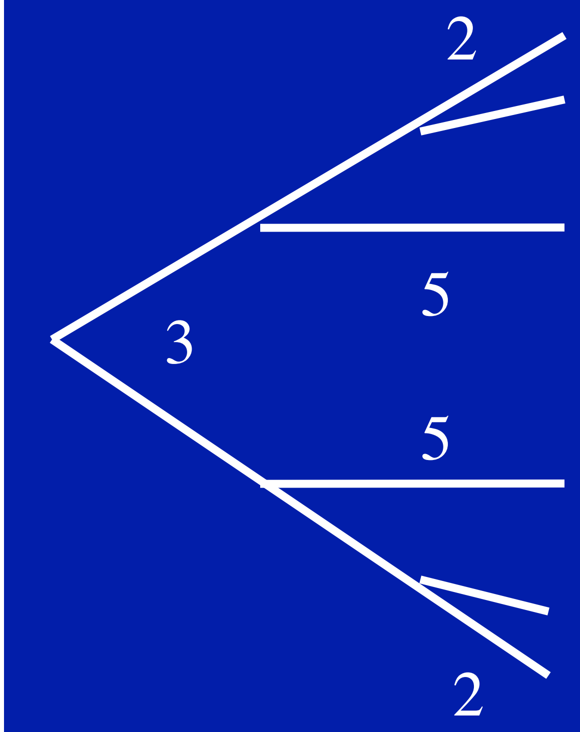
```
Chimp      ..AGCTAAAGGGTCAGGGGAAGGGCA..
Human      ..AGCAAAAGGGTCAGGGGAAGGGGA..

Gorilla    ..AGCATAGGGGTCAGGGGAAAGGCT..



Squirrel   ..AGCGGACCGGTAAGGAGAAAGGAC..

Macaque    ..AGCTCATCGGTAAGGAGAAAGGAT..
Orangutan  ..AGCCCATCGGTCAGGAGAAAGGAT..
```

Hierarchical Clustering



```
Chimp      ..AGCTAAAGGGTCAGGGGAAGGGCA..
Human      ..AGCAAAAGGGTCAGGGGAAGGGGA..

Gorilla    ..AGCATAGGGGTCAGGGGAAAGGCT..



Squirrel   ..AGCGGACCGGTAAGGAGAAAGGAC..

Macaque    ..AGCTCATCGGTAAGGAGAAAGGAT..
Orangutan  ..AGCCCATCGGTCAGGAGAAAGGAT..
```

Hierarchical Clustering



```
Chimp      ..AGCTAAAGGGTCAGGGGAAGGGCA..
Human      ..AGCAAAAGGGTCAGGGGAAGGGGA..

Gorilla    ..AGCATAGGGGTCAGGGGAAAGGCT..



Squirrel   ..AGCGGACCGGTAAGGAGAAAGGAC..

Macaque    ..AGCTCATCGGTAAGGAGAAAGGAT..
Orangutan  ..AGCCCATCGGTCAGGAGAAAGGAT..
```

Tree Reduction

UPGMA is a recursive algorithm that depends on a *reduction* step that replaces the selected pair of clusters with a new cluster, prior to applying the tree construction algorithm recursively on the "reduced" set of clusters.

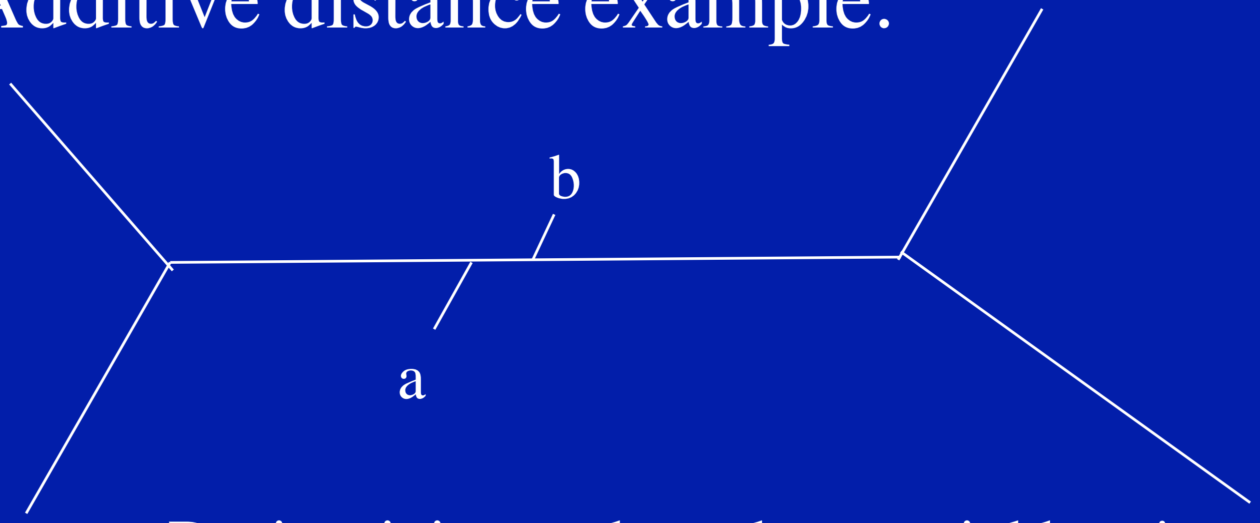
Let's consider the details of the reduction step for the algorithm for a simple tree of four sequences X,Y,Z,W. Say sequences X,Y are chosen as nearest neighbors. Indicate precisely how they would be replaced with a new cluster vertex C, by drawing a representative tree *before* and *after* this replacement.

Tree Reduction Answer

- In UPGMA, X,Y are replaced by a new node C whose distances to Z,W are just the averages of the corresponding distances from X and Y.
- Preserving distances in this way is vital for the correctness of subsequent clustering steps (which again look for the minimal distance pair).
- Note that C will itself (eventually) be clustered with some other cluster, to connect it to the rest of the tree.

Neighbor Joining Algorithm for Additive Distances

- If we drop the ultrametric distance assumption, we can no longer assume that the closest pair will be neighbors in the tree.
- Additive distance example:

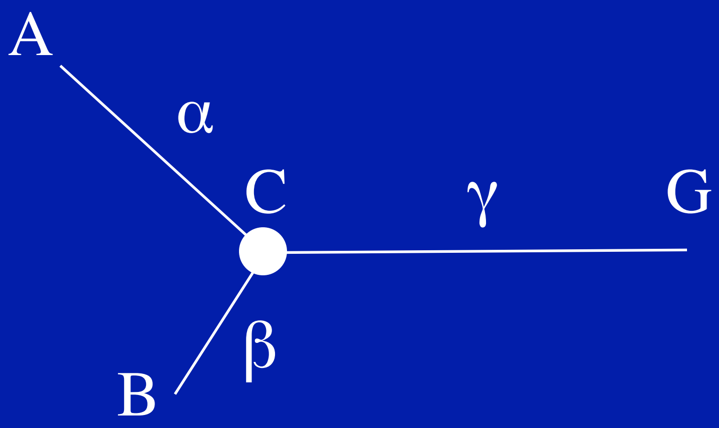


D_{ab} is minimum, but a,b not neighbors!

Additive Neighbor Metric

- Assume $D(C_1,C_2)$ is additive distance.
- $u(C) = \Sigma_{C'}D(C,C')/(N_c-2)$
- Measures distance from all other clusters
- $\delta(C_1,C_2) = D(C_1,C_2)-u(C_1)-u(C_2)$
- Can prove that if $\delta(C_1,C_2)$ is minimum, C_1 and C_2 are neighbors. Use this to iteratively build tree, similar to UPGMA.

Neighbor Distance Calculation



$D_{AB} + D_{BG} = \alpha + \beta + \beta + \gamma$
$D_{AG} = \alpha + \gamma$
$\beta = (D_{AB} + D_{BG} - D_{AG})/2$
$\alpha = (D_{AB} + D_{AG} - D_{BG})/2$

This is true in general for *any* third node G, so equivalent to
$\alpha = [D_{AB} + u(A) - u(B)]/2$
Since u(A) measures average distance of A to all other G.

Tree Reduction

Neighbor Joining is a recursive algorithm that depends on a *reduction* step that replaces the selected pair of clusters with a new cluster, prior to applying the tree construction algorithm recursively on the “reduced” set of clusters.

Let’s consider the details of the reduction step for the algorithm for a simple tree of four sequences X,Y,Z,W. Say sequences X,Y are chosen as nearest neighbors. Indicate precisely how they would be replaced with a new cluster vertex C, by drawing a representative tree *before* and *after* this replacement.

Tree Reduction Answer

- In NJ, X,Y are replaced by the vertex representing their junction (C). We must compute its distance versus all other sequences Z,W. We do this using additivity.
- This leaves us with a smaller tree to infer, using the exact same procedure (i.e. neighbor joining based on additivity).
- Note that C will itself (eventually) be clustered with some other cluster, to connect it to the rest of the tree.

Neighbor Joining Algorithm

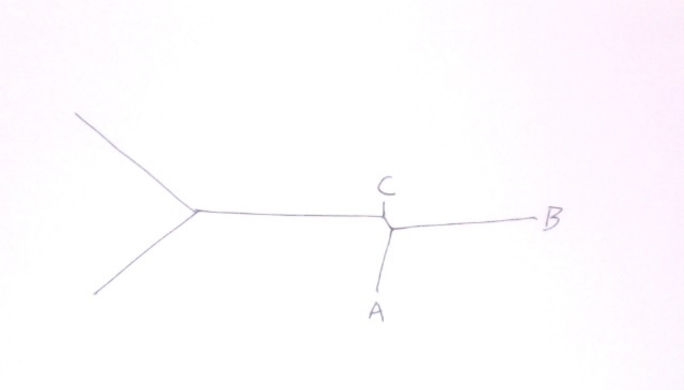
Create initial list: each sequence is a separate cluster

While not all clusters joined

- Compute $\delta(C_1,C_2)$ for all cluster pairs
- Choose pair of clusters with minimum $\delta(C_1,C_2)$
- Join that pair into a single cluster: add a new node C, new edge C_1-C with length $[D(C_1,C_2)+u(C_1)-u(C_2)]/2$, and new edge C_2-C.

NJ Tree Connectivity Significance?

Consider the following unrooted tree produced by Neighbor Joining. Note that sequences A,B are connected directly to each other (no intervening edges) even though $\delta_{AC} < \delta_{AB}$ and also $\delta_{BC} < \delta_{AB}$. Does this indicate that Neighbor Joining has made an evolutionarily incorrect tree? Does the fact that A,B are directly connected have any valid evolutionary meaning?



NJ Tree Connectivity Significance? Answer

- This is a valid unrooted tree. Neighbor Joining is completely insensitive to the lengths of the terminal edges. After all, the mutation rate could be different on different edges, stretching or shrinking them; we do not want that to fool the algorithm into inferring a different tree.
- The connectivity of the tree should reflect the actual evolutionary history of how they branched from each other. In this case, that means there is “extra period of history” connecting A,B to C (and the rest of the tree) than to each other.