

# Gene Sequence Analysis

## Lecture 4: **Phylogenetic Trees**

02/06/2024

Phuc Loi Luu, PhD

Updated from Dr. Morgan Langille

# Comparing MSAMs

Tool	Options	Algorithm	Alphabet
ClustalW	Defaults	Progressive	Amino Acid
Muscle	Defaults	Progressive (iterative)	Amino Acid
MAFFT	Defaults	Progressive (iterative)	Amino Acid
ProbCons	Defaults	Consistency	Amino Acid
ProbAlign	Defaults	Consistency	Amino Acid
Mummals	Defaults	Consistency/Structure	Amino Acid
Dialign-TX	Defaults	Greedy/Progressive	Amino Acid
Prank (AA)	+F (AA)	"Phylogenetically-aware"	Amino Acid
Prank	+F -codon	"Phylogenetically-aware"	Codon
BAlI-Phy	Model M0	Statistical Alignment	Codon
BAlI-Phy samples	Model M0	Statistical Alignment	Codon
BAlI-Phy integrated	Model M0	Statistical Alignment	Codon

**BAlI-Phy**

**MUMMALS**

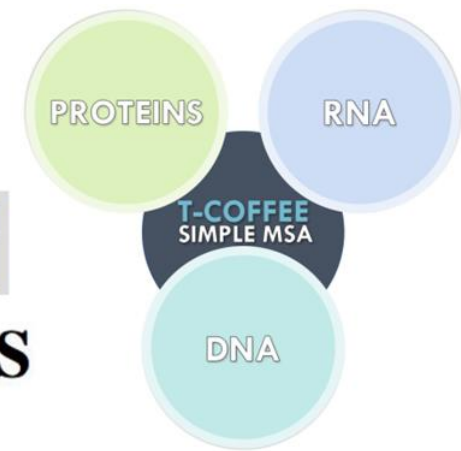
**MASS**

Multiple Alignment by Secondary Structures



**PROBCONS**

Probabilistic Consistency-based Multiple Alignment of Amino Acid Sequences



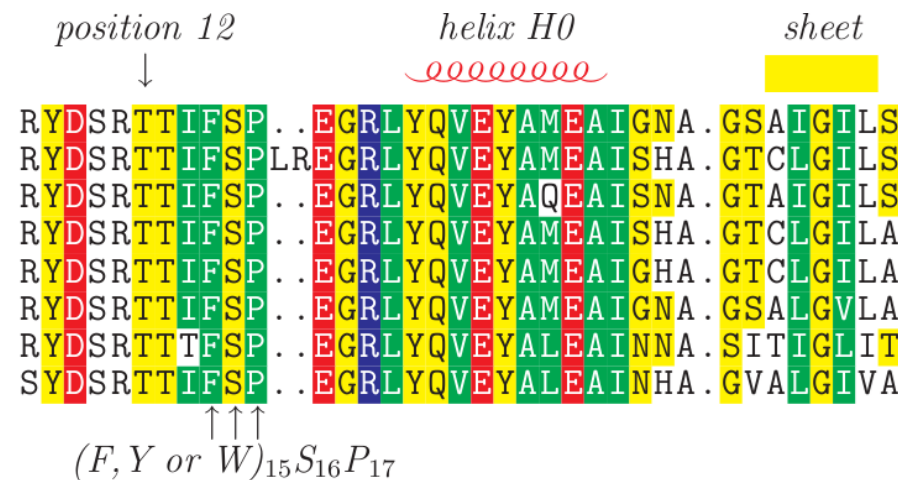
**ESPrIpt 3.0**

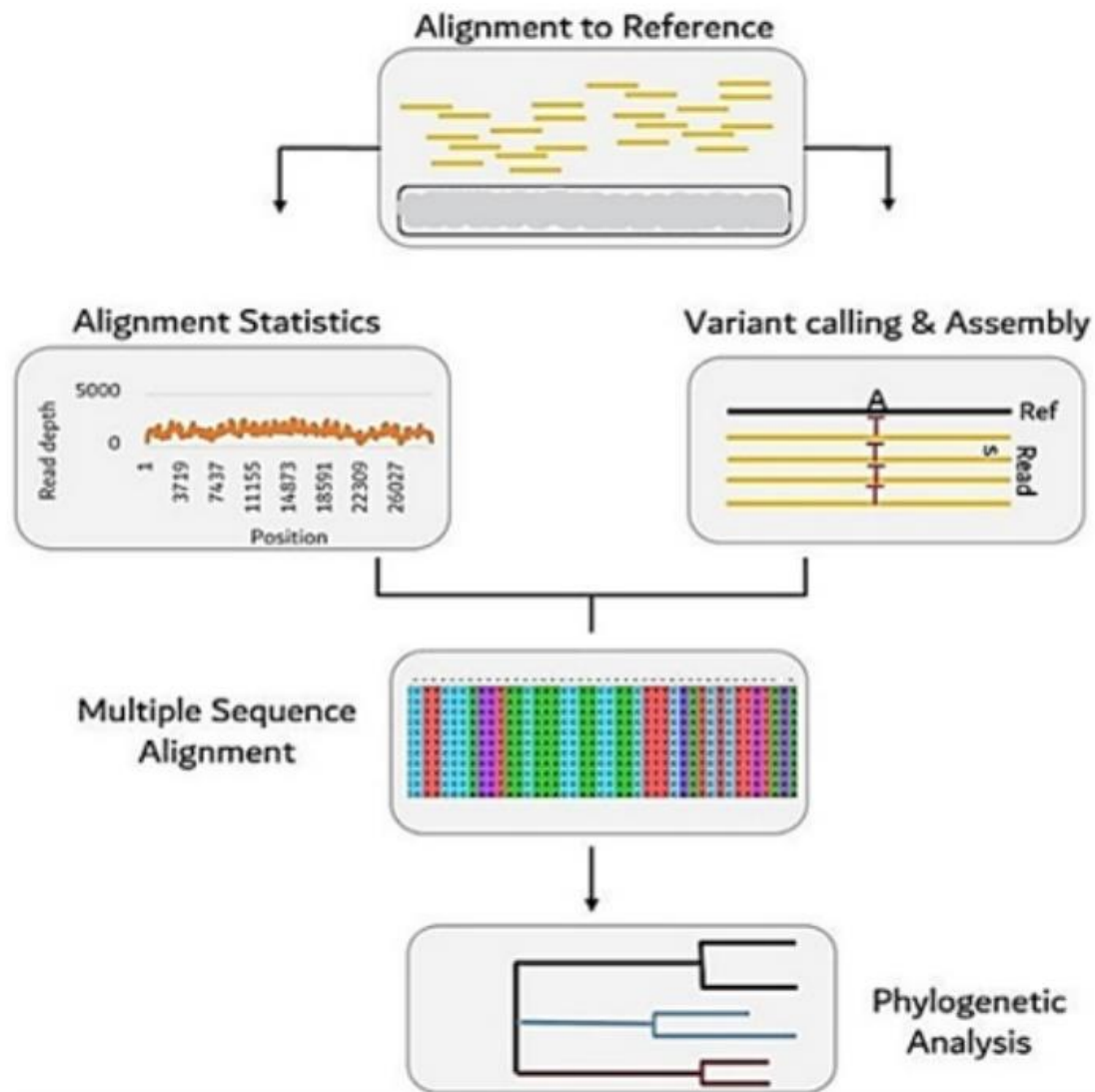
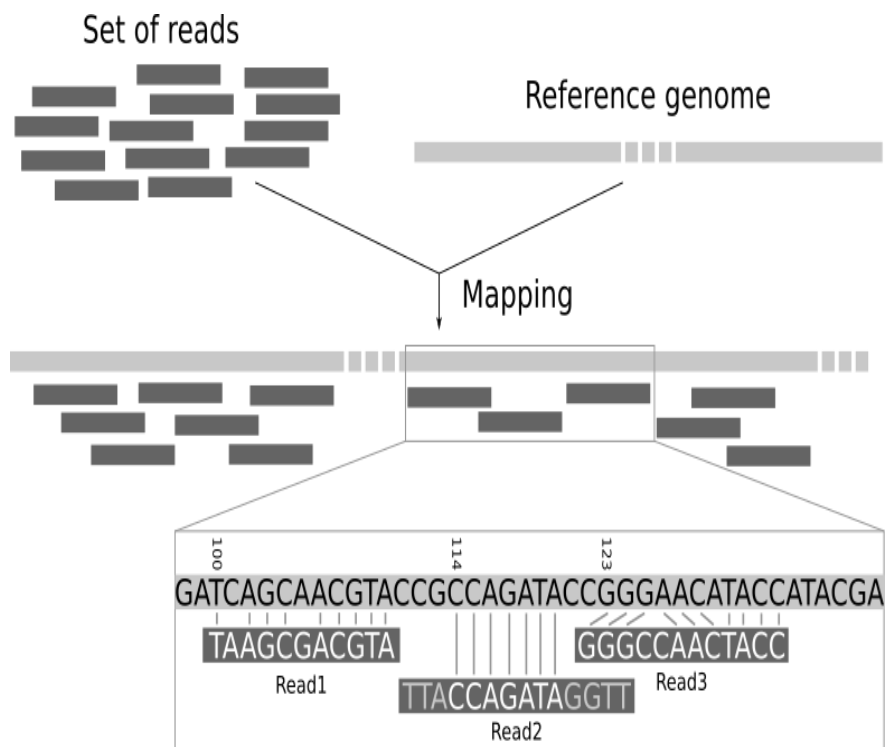


MAFFT version 7

Multiple alignment program for amino acid or nucleotide sequences

Blackburne & Whelan. 2013. *Mol. Biol. Evol.* 30(3):642–653.







# Light-weight AVS – AliView & SeaView

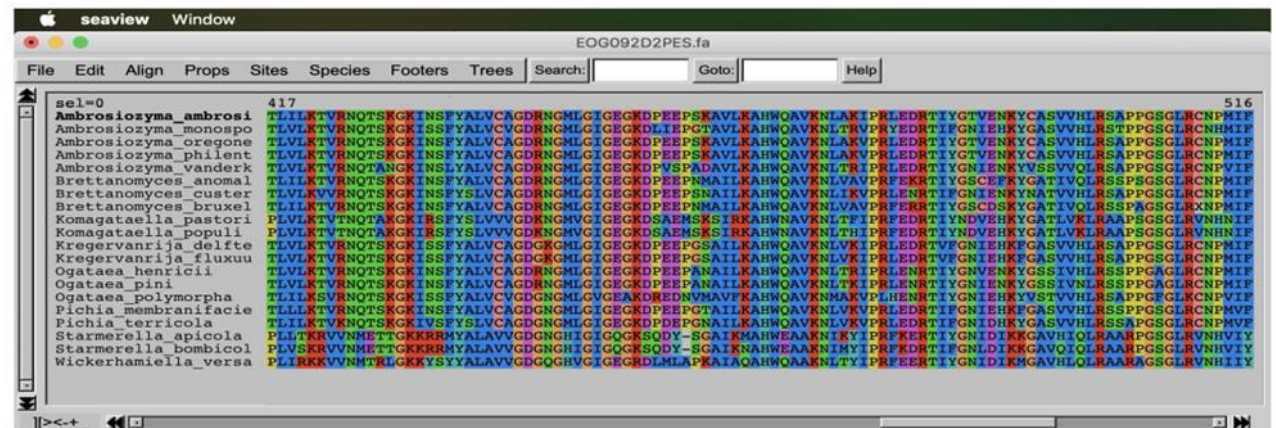
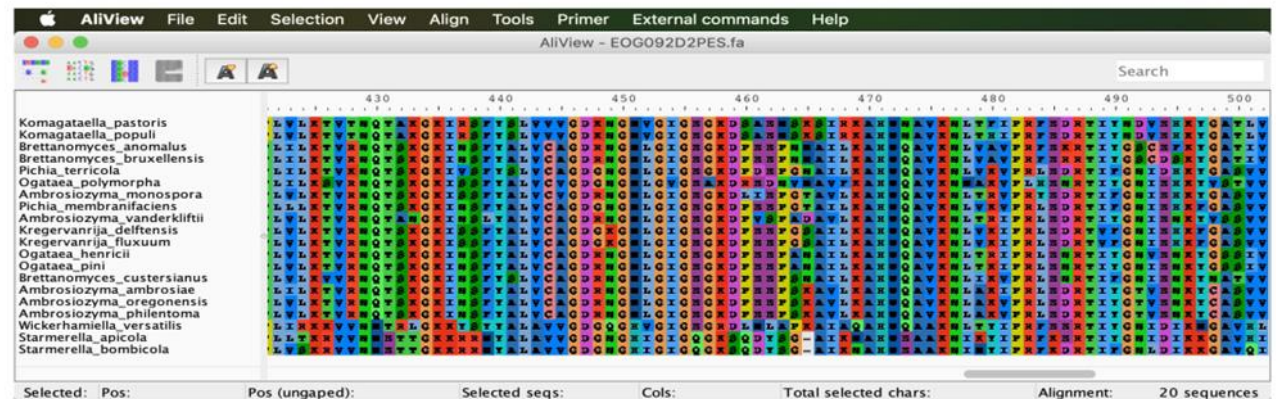
**AliView** → Larsson, A. (2014). AliView: a fast and lightweight alignment viewer and editor for large data sets.

*Bioinformatics* 30(22):3276–3278.

<http://dx.doi.org/10.1093/bioinformatics/btu531>

**SeaView** → Gouy M., Guindon S. & Gascuel O. (2010) SeaView version 4 : a multiplatform graphical user interface for sequence alignment and phylogenetic tree building. *Molecular Biology and Evolution* 27(2):221–224.

<https://academic.oup.com/mbe/article/27/2/221/970247>

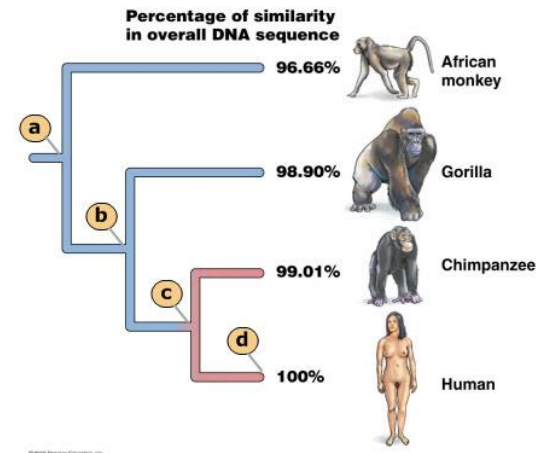


# Outline

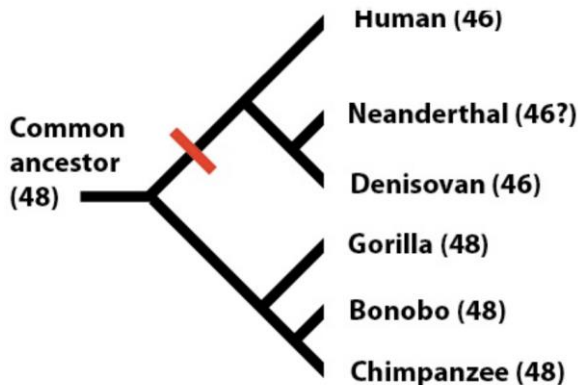
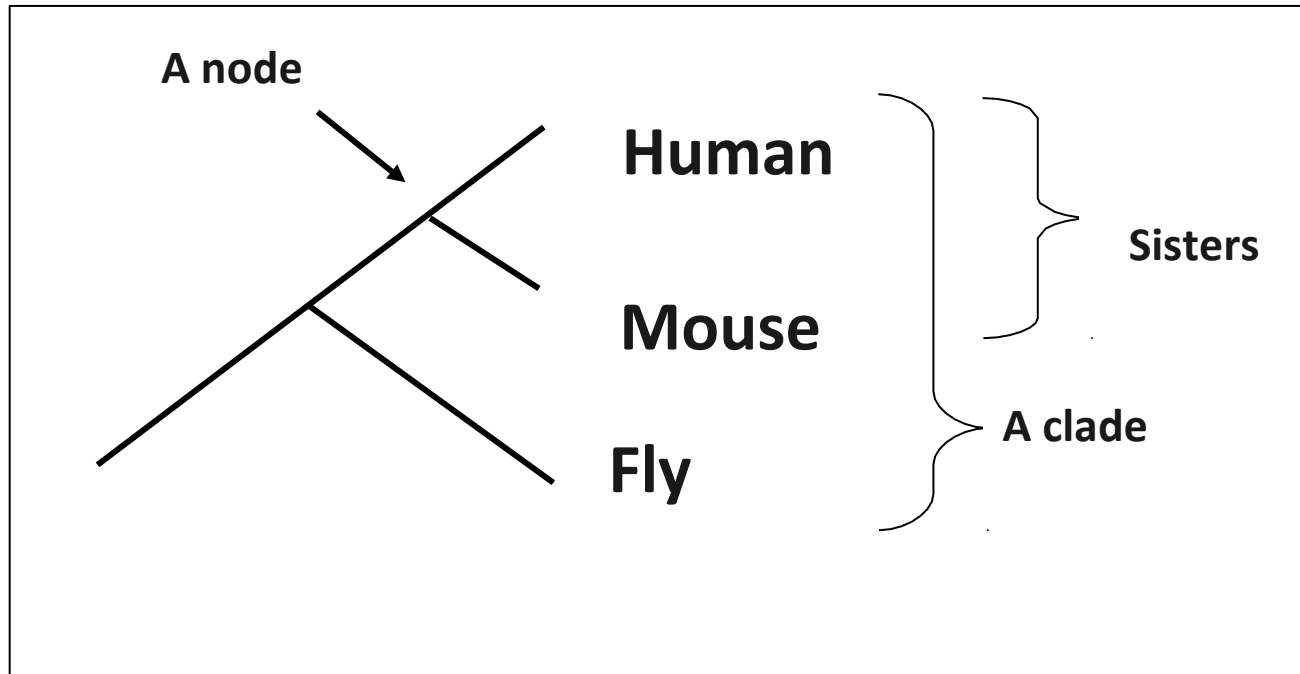
- Why we build trees?
- What is a tree?
  - Parsimony
  - Neighbour Joining (distance based)
  - Maximum Likelihood & Bayesian
- Bootstrapping
- Softwares
- Tree file formats


# Why do we build trees?

- Show relationship of related organisms
- Gene trees often used to infer species trees
- Gene family history
  - Duplication
  - Lateral Gene Transfer
- Determining ancestral states of certain traits
- Testing or removing phylogenetic signal

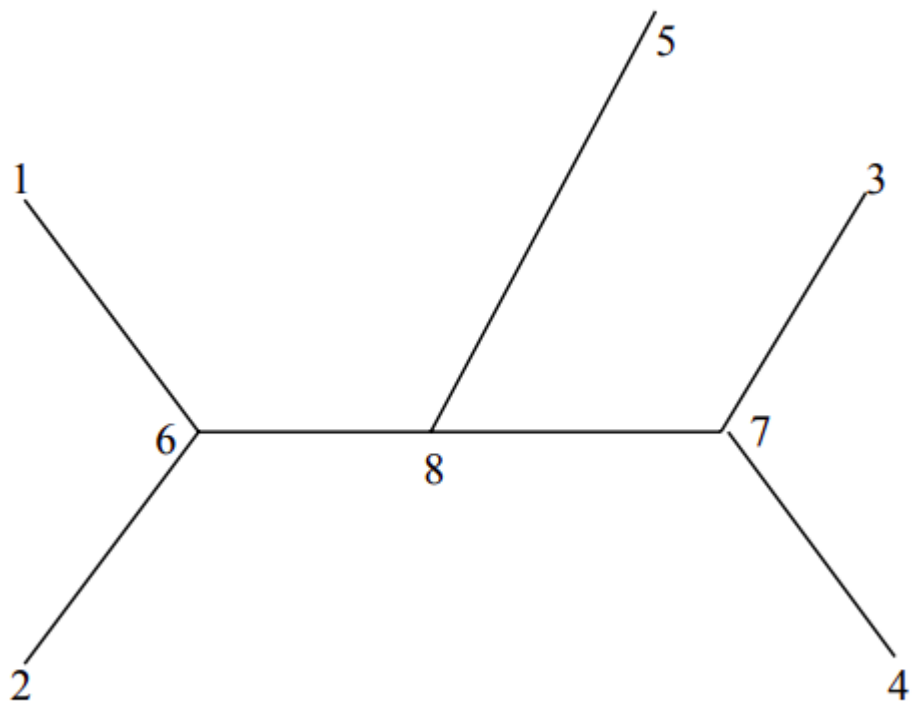


# A phylogenetic tree

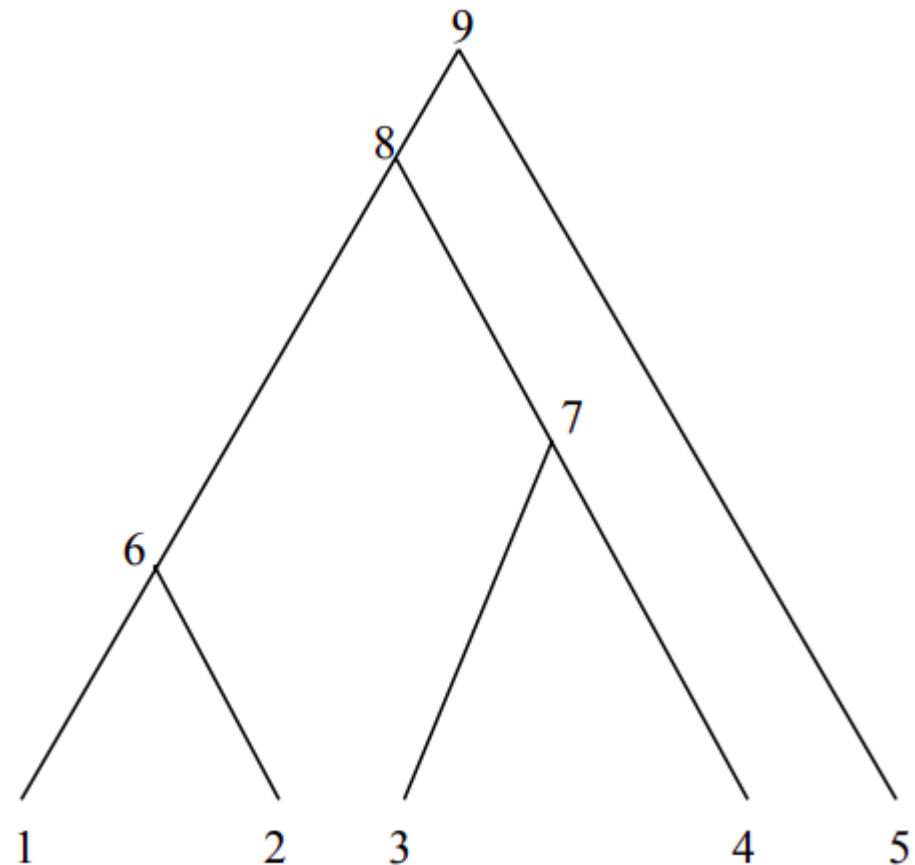


 Chromosome fusion leading to human chromosome 2

# A phylogenetic tree



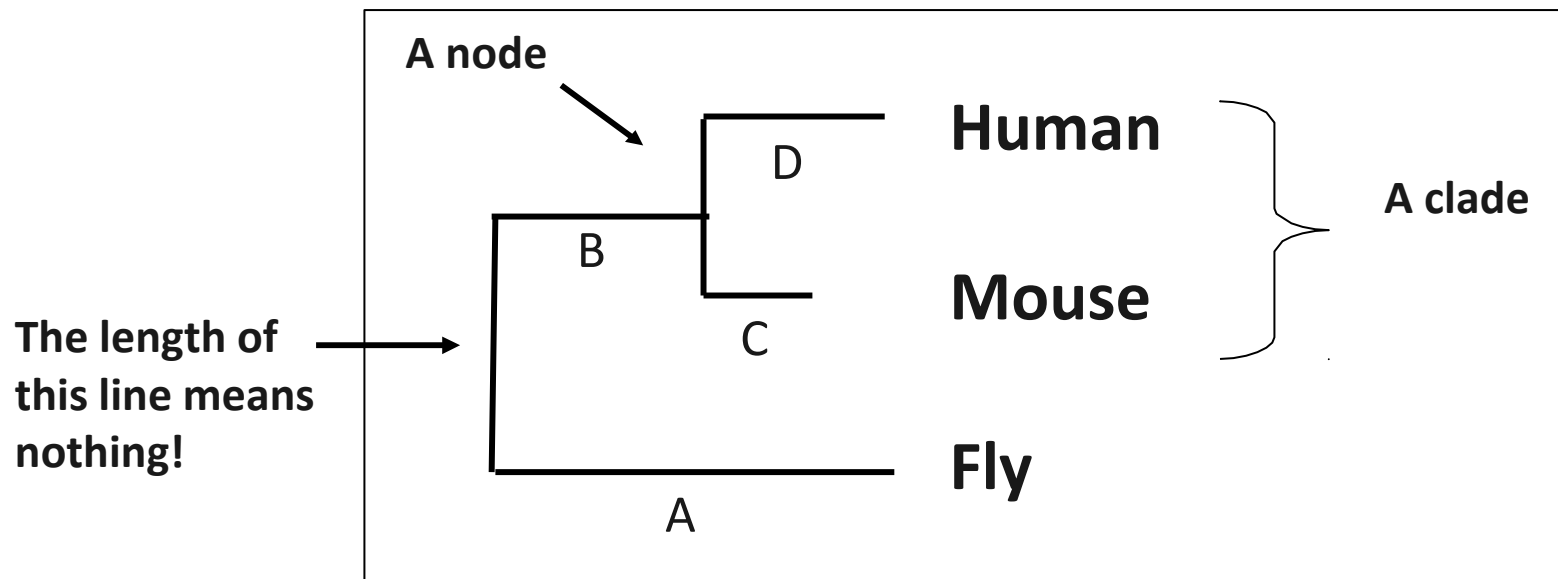
(a) Unrooted tree



(b) Rooted tree



# A phylogenetic tree with branch lengths



**Branch length can be significant...**

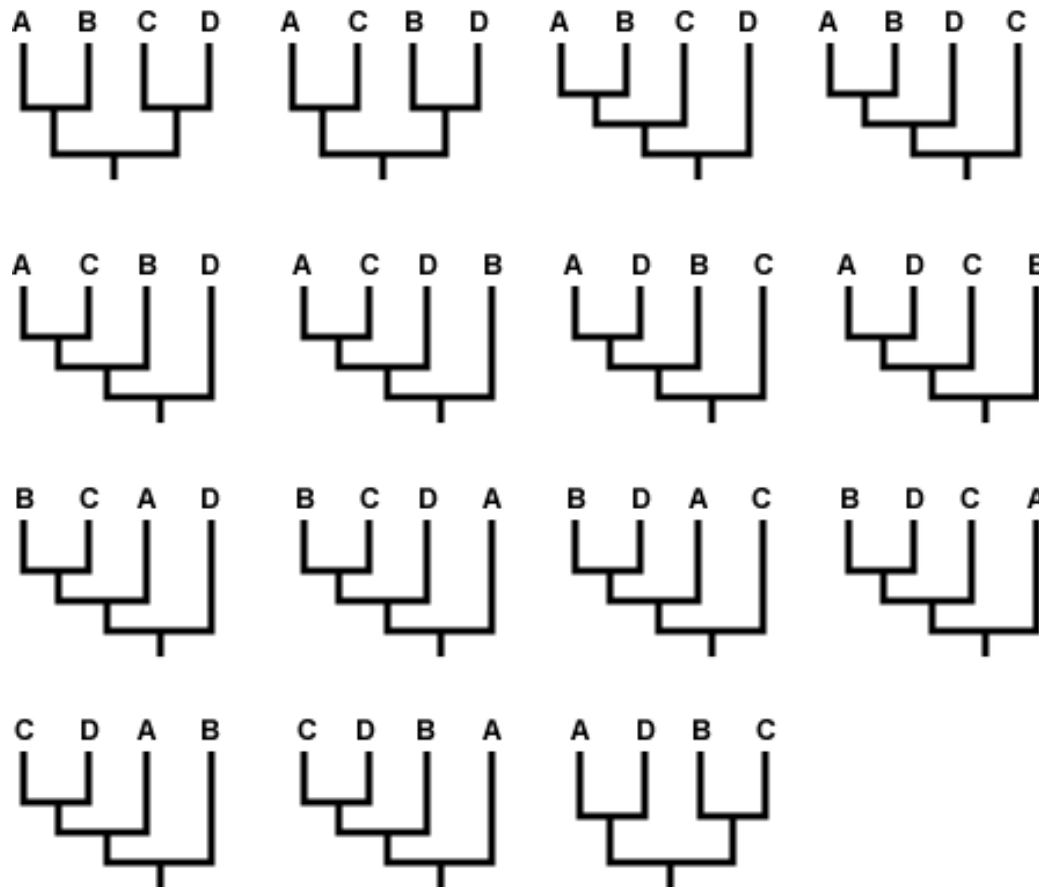
**In this case the analysis suggests that the mouse sequence/taxon is slightly more similar to fly than human is to fly**

**(i.e. sum of branches  $A+B+C$  is less than sum of  $A+B+D$ )**

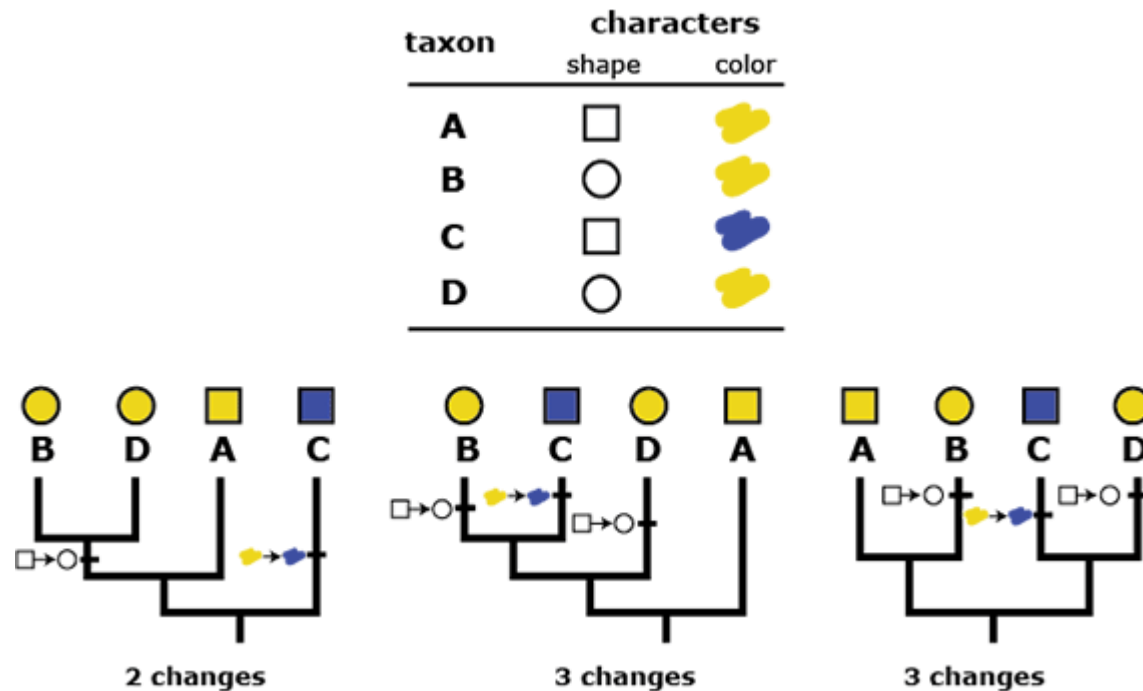
# Parsimony

- The tree implying the least number of changes in character states (most parsimonious) is the best.
- Note: Does not determine branch lengths

There are 15 different ways that those taxa could be related with Four taxa: A, B, C, and D



There are 15 different ways that those taxa could be related with Four taxa: A, B, C, and D



# Example Alignment

		Alignment column											
		1	2	3	4	5	6	7	8	9	10	11	12
OTUs	1	G	C	A	A	A	A	A	A	A	C	T	T
	2	G	C	A	A	A	A	A	A	A	C	C	T
	3	G	C	A	A	A	A	A	A	A	A	A	C
	4	A	C	A	G	G	A	G	G	A	A	A	A
	5	A	A	C	A	A	G	A	A	C	A	A	A

# Parsimony Example

Alignment column

1 2 3 4 5 6 7 8 9 10 11 12

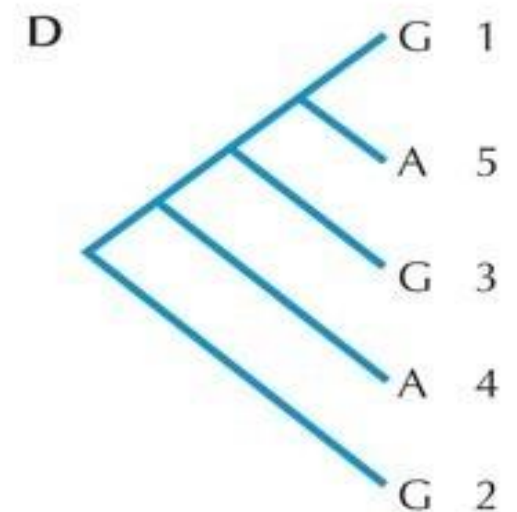
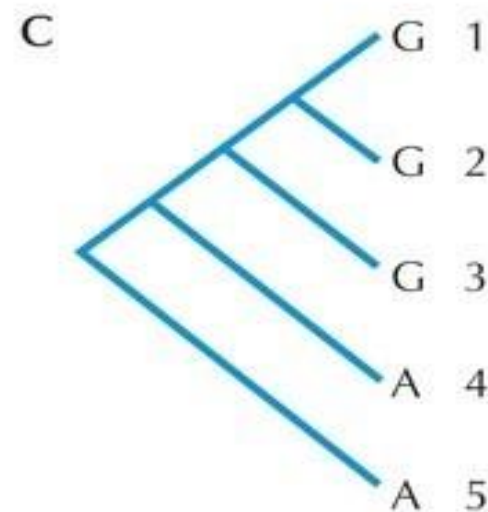
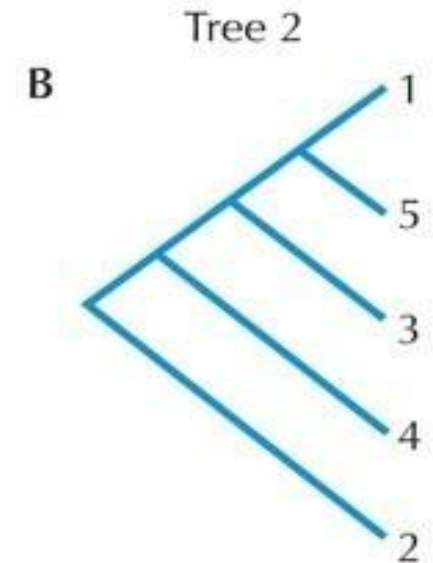
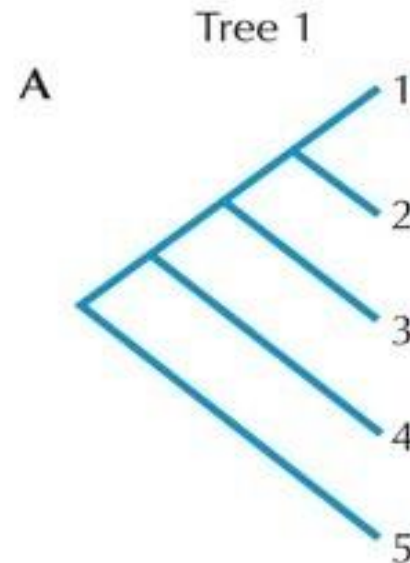
1 G C A A A A A A C T T

2 G C A A A A A A C C T

3 G C A A A A A A A A C

4 A C A G G A G G A A A A

5 A A C A A G A A C A A A

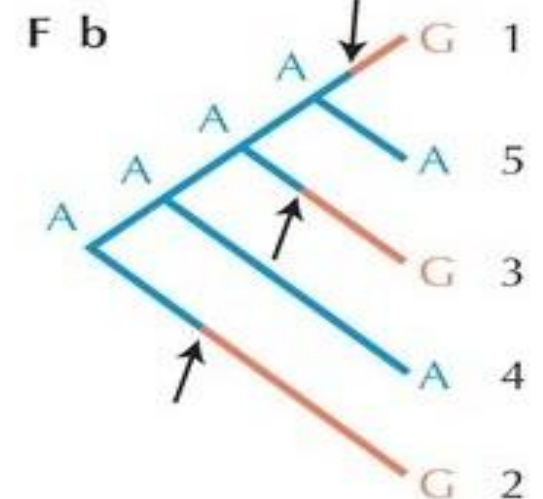
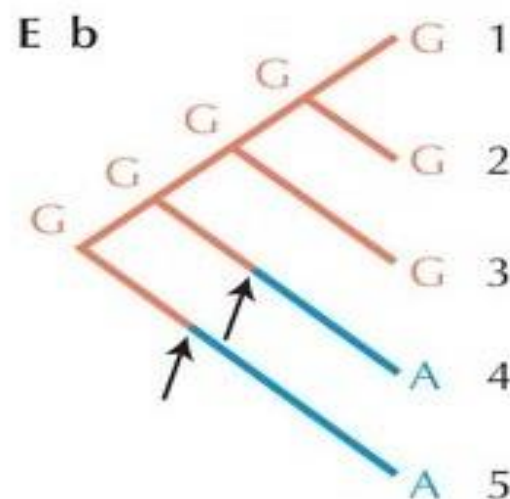
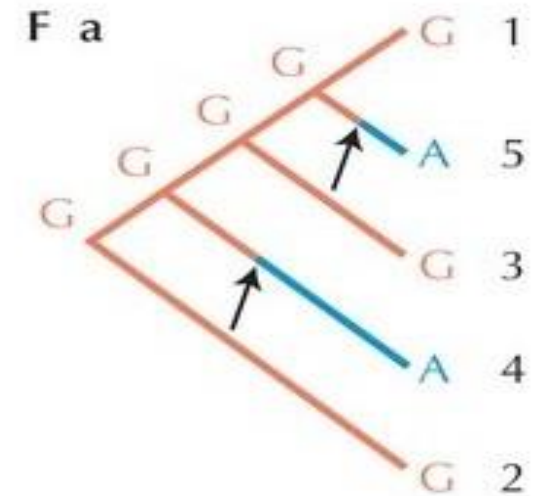
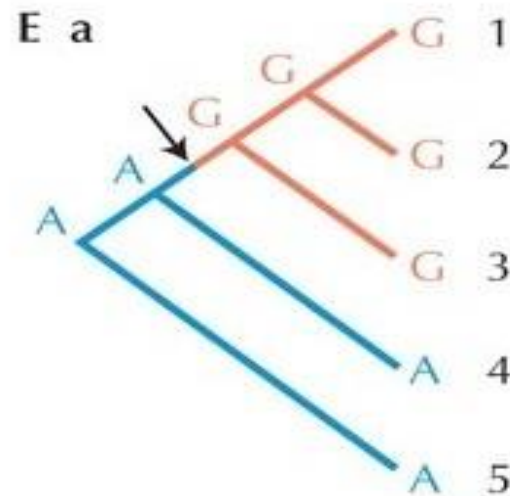




# Parsimony Example

Alignment column

1	2	3	4	5	6	7	8	9	10	11	12
G	C	A	A	A	A	A	A	A	C	T	T
G	C	A	A	A	A	A	A	A	C	C	T
G	C	A	A	A	A	A	A	A	A	A	C
A	C	A	G	G	A	G	G	A	A	A	A
A	A	C	A	A	G	A	A	C	A	A	A



# Number of Trees

TABLE 27.5. Number of possible branching patterns versus number of OTUs

Taxa	Rooted Trees <sup>a</sup>	Unrooted Trees <sup>b</sup>
3	3	1
4	15	3
5	105	15
6	945	105
7	10,395	945
8	135,135	10,395
9	2,027,025	135,135
10	34,459,425	2,027,025

$$^a N_r = (2n - 3) \times (2n - 5) \times (2n - 7) \times \dots \times 3 \times 1 = (2n - 3)! / [2^{n-2} \times (n - 2)!].$$

$$^b N_u = (2n - 5) \times (2n - 7) \times \dots \times 3 \times 1 = (2n - 5)! / [2^{n-3} \times (n - 3)!].$$

Scoring every single tree not possible! “Tree Searching” algorithms used.

# Distance Based Method

- Start with a distance matrix between every pair of sequences

TABLE 27.6. Distance matrix

OTUs	A	B	C	D	E	F
A	0	2	4	6	6	8
B	2	0	4	6	6	8
C	4	4	0	6	6	8
D	6	6	6	0	4	8
E	6	6	6	4	0	8
F	8	8	8	8	8	0

# Distance Based Methods

- Unweighted Pair Group method with Arithmetic Mean (UPGMA)
- Neighbour Joining

## Advantage:

- Computationally Very Fast
- Often included in MSA programs

# Building an UPGMA Tree - An example

	A	B	C	D	E	F	G
A							
B	19.00						
C	27.00	31.00					
D	8.00	18.00	26.00				
E	33.00	36.00	41.00	31.00			
F	18.00	1.00	32.00	17.00	35.00		
G	13.00	13.00	29.00	14.00	28.00	12.00	

# Building an UPGMA Tree - An example

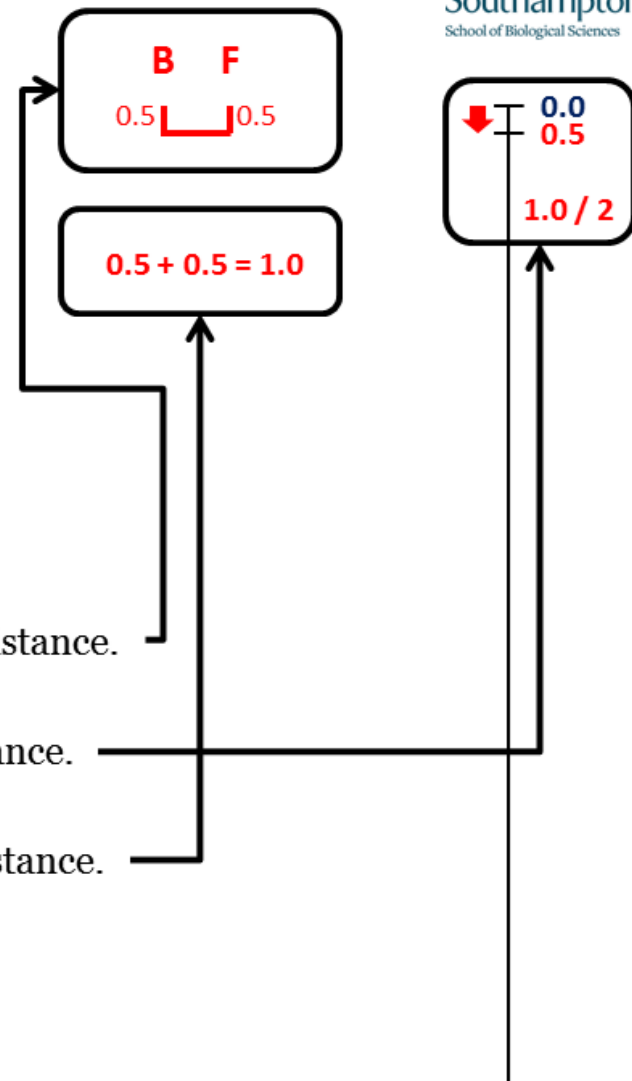
	A	B	C	D	E	F	G
A							
B	19.00						
C	27.00	31.00					
D	8.00	18.00	26.00				
E	33.00	36.00	41.00	31.00			
F	18.00	1.00	32.00	17.00	35.00		
G	13.00	12.00	29.00	14.00	28.00	12.00	

1. Find the shortest pairwise distance.

2. Join two sequences/groups with shortest distance.

3. Depth of new branch =  $\frac{1}{2}$  shortest distance.

4. Tip-to-tip path length = shortest distance.





# Building an UPGMA Tree - An example

	A	B	C	D	E	F	G
A							
B	19.00						
C	27.00	31.00					
D	8.00	18.00	26.00				
E	33.00	36.00	41.00	31.00			
F	18.00	1.00	32.00	17.00	35.00		
G	13.00	13.00	29.00	14.00	28.00	12.00	

**B F**  
0.5  0.5

5. Calculate mean pairwise distances with other sequences in new matrix.

	A	BF	C	D	E	G
A						
BF	18.50					
C	27.00	31.50				
D	8.00	17.50	26.00			
E	33.00	35.50	41.00	31.00		
G	13.00	12.50	29.00	14.00	28.00	

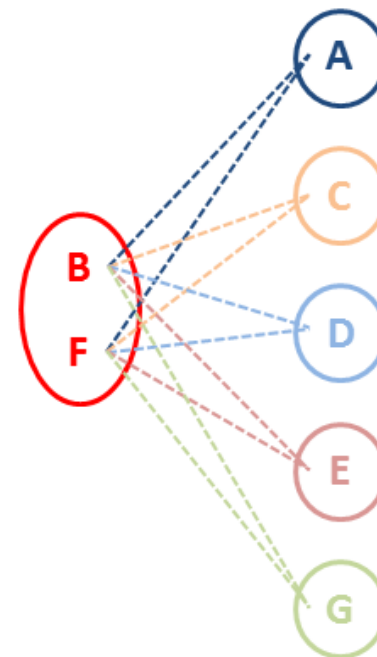
$$(19 + 18) / 2 = 18.5$$

$$(31 + 32) / 2 = 31.5$$

$$(18 + 17) / 2 = 17.5$$

$$(36 + 35) / 2 = 35.5$$

$$(13 + 12) / 2 = 12.5$$



0.0  
0.5

# Building an UPGMA Tree - An example

	A	B	C	D	E	F	G
A							
B	19.00						
C	27.00	31.00					
D	8.00	18.00	26.00				
E	33.00	36.00	41.00	31.00			
F	18.00	1.00	32.00	17.00	35.00		
G	13.00	13.00	29.00	14.00	28.00	12.00	

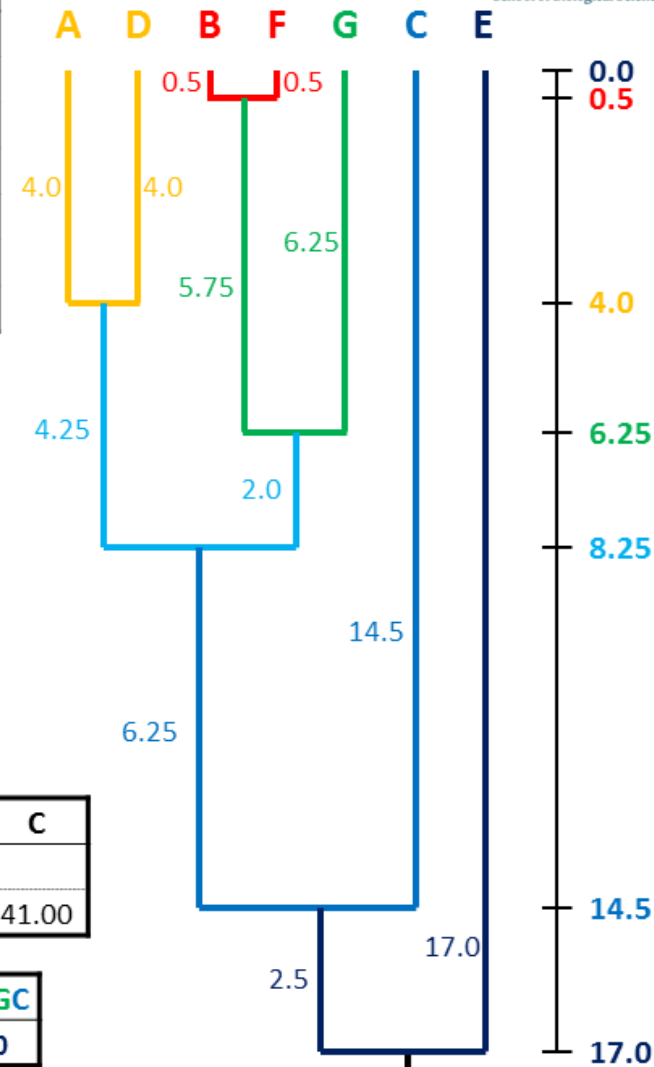
	A	BF	C	D	E
BF	18.50				
C	27.00	31.50			
D	8.00	17.50	26.00		
E	33.00	35.50	41.00	31.00	
G	13.00	12.50	29.00	14.00	28.00

	AD	BF	C	E
BF	18.00			
C	26.50	31.50		
E	32.00	35.50	41.00	
G	13.50	12.50	29.00	28.00

	AD	BFG	C
BFG	16.50		
C	26.50	30.67	
E	32.00	33.00	41.00

	ADBFG	C
C	29.00	
E	32.60	41.00

	ADBFGC
E	34.00



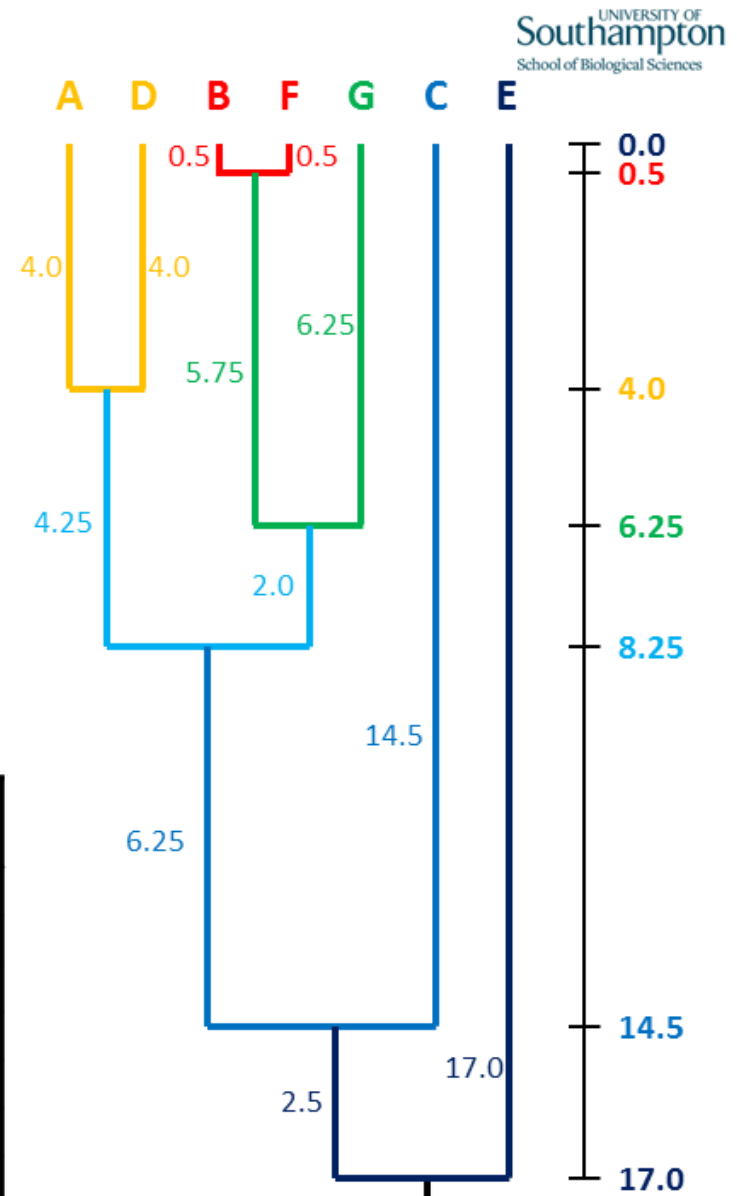
# Building an UPGMA Tree - An example

	A	B	C	D	E	F	G
A							
B	19.00						
C	27.00	31.00					
D	8.00	18.00	26.00				
E	33.00	36.00	41.00	31.00			
F	18.00	1.00	32.00	17.00	35.00		
G	13.00	13.00	29.00	14.00	28.00	12.00	

The source data for this worked example is a selection of Cytochrome C distances from Table 3 of one of the seminal phylogenetic papers: Fitch WM & Margoliash E (1967). Construction of phylogenetic trees. *Science* **155**:279-84.

<http://www.ncbi.nlm.nih.gov/pubmed/5334057>

	Turtle	Man	Tuna	Chicken	Moth	Monkey	Dog
	A	B	C	D	E	F	G
Turtle							
Man	19						
Tuna	27	31					
Chicken	8	18	26				
Moth	33	36	41	31			
Monkey	18	1	32	17	35		
Dog	13	13	29	14	28	12	

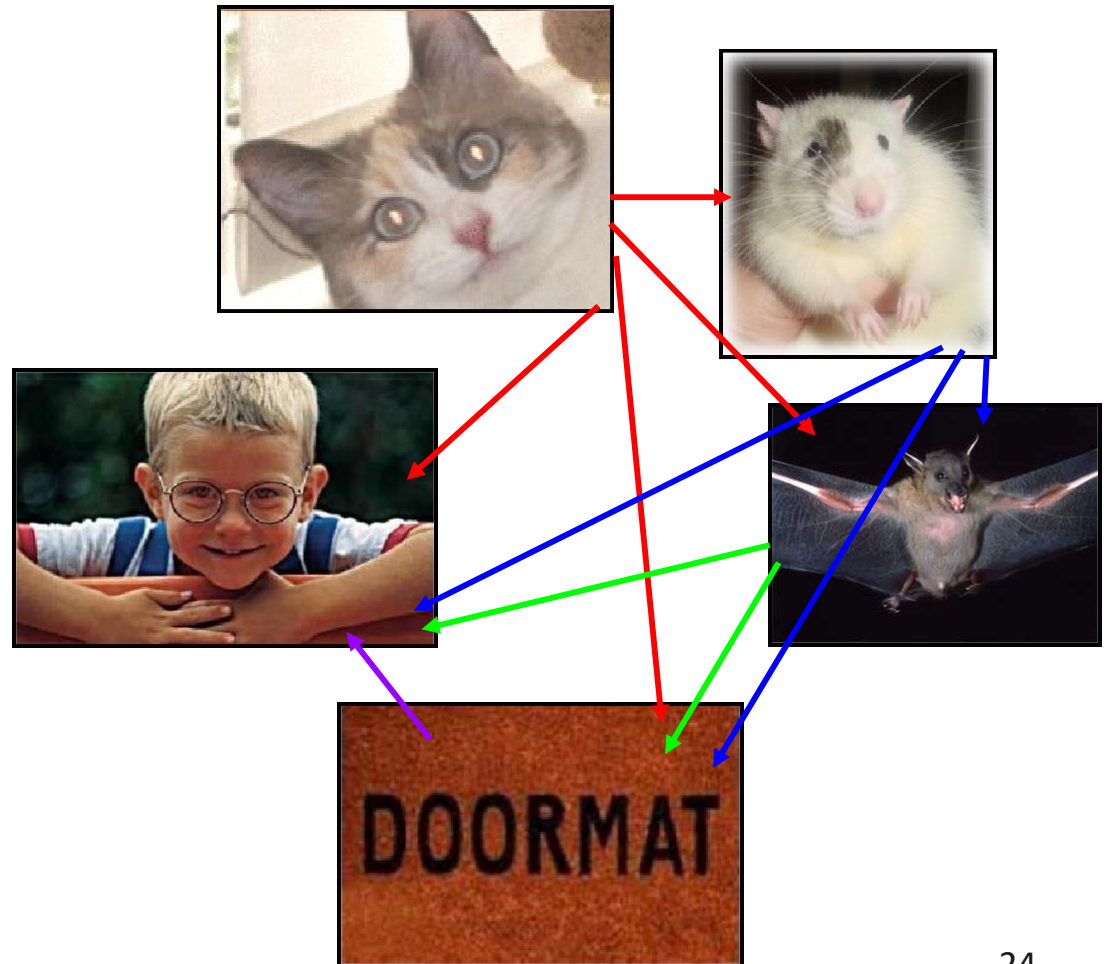


# Building an NJ Tree - An Example

*Cbw protein from cat, rat, bat, mat and Matt*

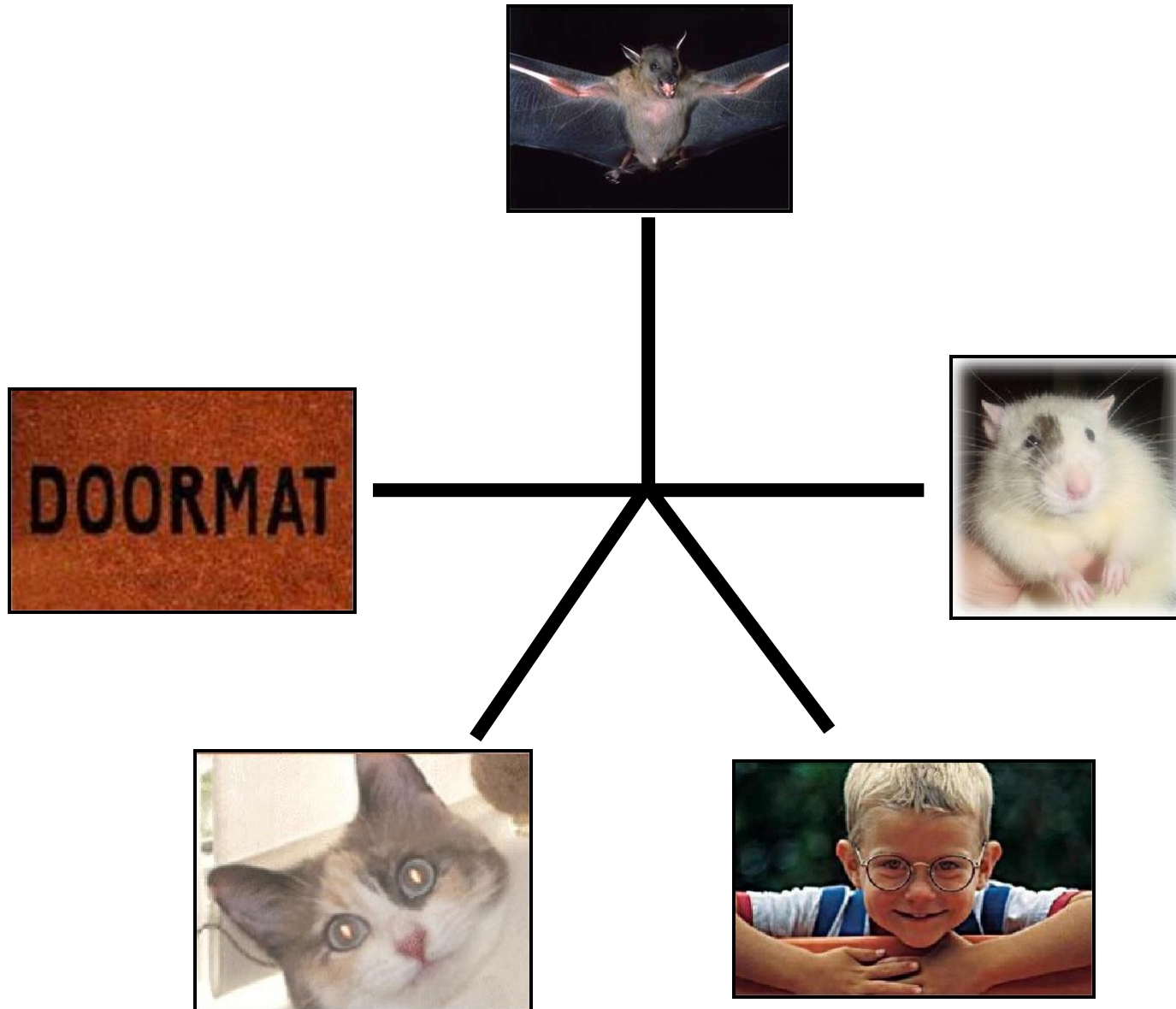
1. Compare all sequences to each other
2. Assign divergence values to each pair
3. Assemble the values in a distance matrix

	Cat	Rat	Bat	Mat
Cat	-			
Rat	0.7	-		
Bat	0.8	0.2	-	
Mat	1.0	0.8	0.8	-
Matt	0.6	0.4	0.5	0.9



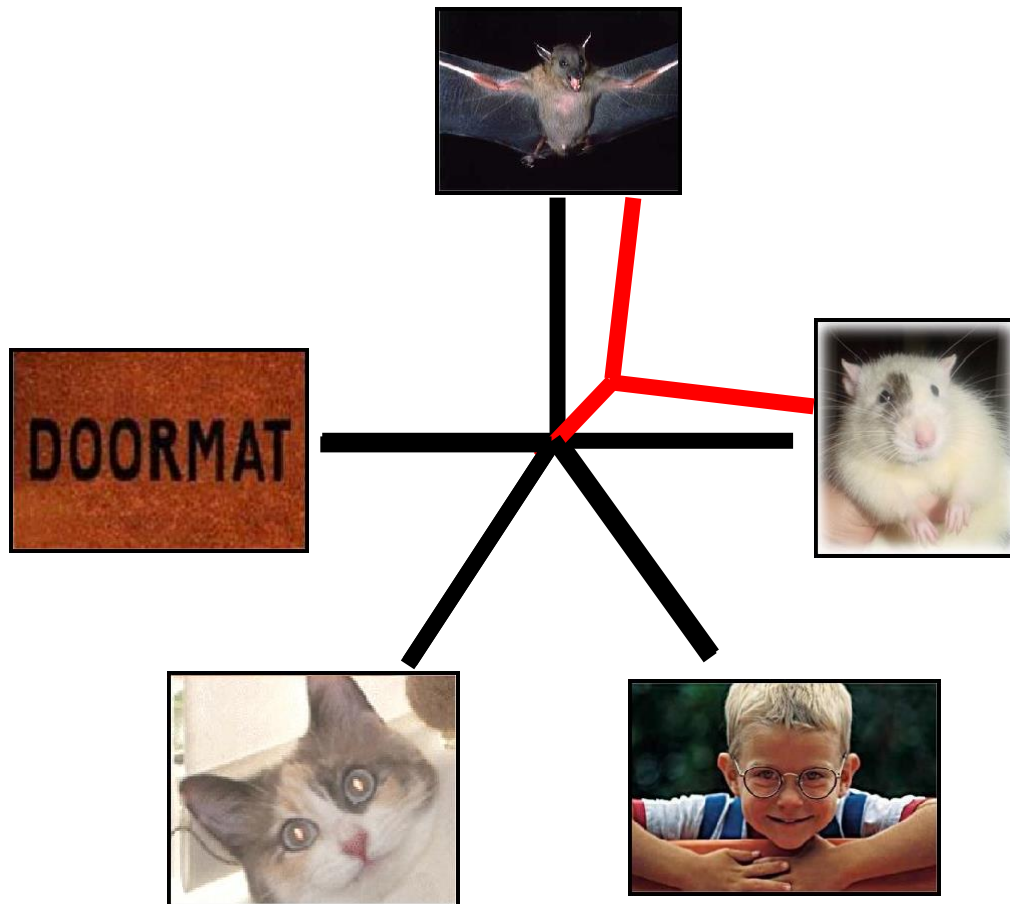
# Building a NJ Tree

4. Arrange the subjects in a “star” phylogeny



# Building a NJ Tree

5. Fuse the two branches with the least divergence



	Cat	Rat	Bat	Mat
Cat	-			
Rat	0.7	-		
Bat	0.8	0.2	-	
Mat	1.0	0.8	0.8	-
Matt	0.6	0.4	0.5	0.9



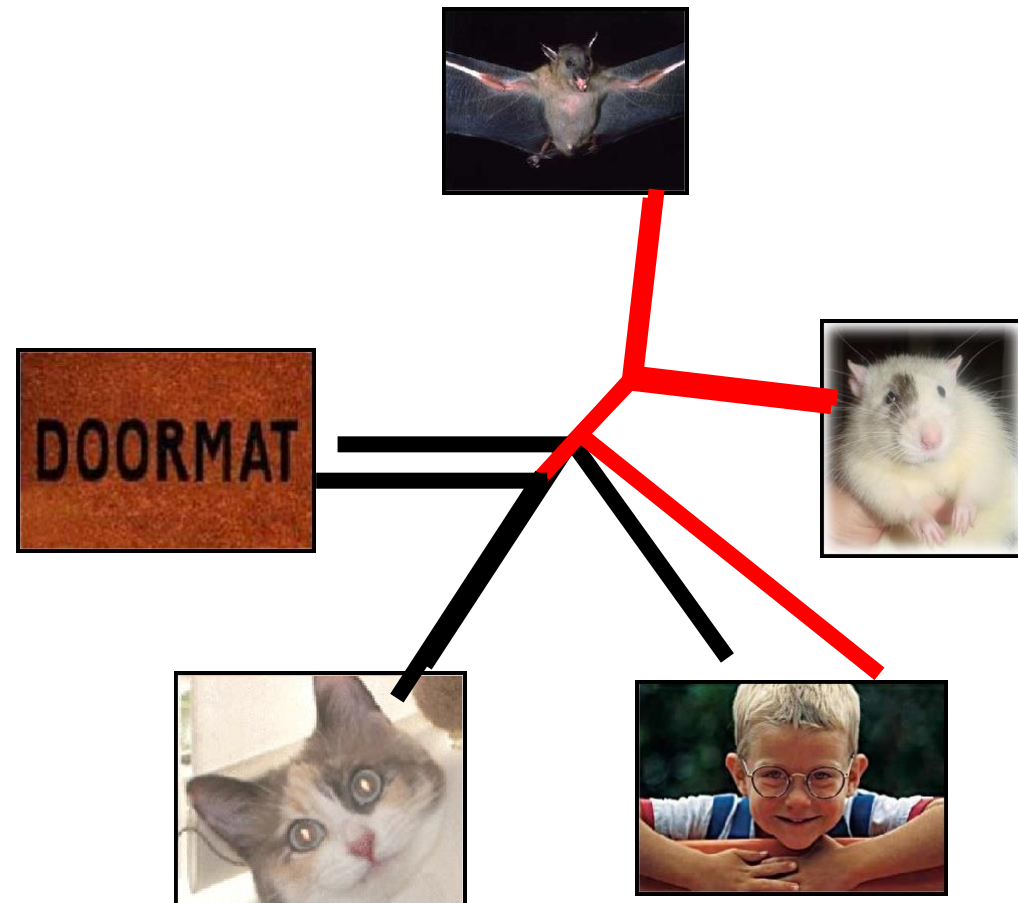
# Building a NJ Tree

6. Create a new distance matrix using the fusion consensus sequence

7. Fuse the next two closest sequences

8. Repeat until tree completed

	Cat	RatBat	Mat
Cat	-		
RatBat	0.75	-	
Mat	1.0	0.8	-
Matt	0.6	0.45	0.9



# UPGMA VERSUS NEIGHBOR JOINING TREE

## UPGMA

A straightforward approach for constructing a rooted phylogenetic tree from a distance matrix

Developed by Sokal and Michener in 1958

An agglomerative hierarchical clustering method based on the average linkage method

Builds a rooted phylogenetic tree

Requires the distances to be ultrametric

As it assumes equal rates of evolution, branch tips come out equal (same branch length from the root to the tips)

A simple and fast method

An unreliable method

## NEIGHBOR JOINING TREE

The new approach for constructing a phylogenetic tree, which is unrooted through a star tree

Developed by Naruya Saitou and Masatoshi Nei in 1987

An iterative clustering method based on the minimum-evolution criterion

Builds an unrooted phylogenetic tree

Requires the distances to be additive

Allows unequal rates of evolution, the branch lengths are proportional to the amount of change

A comparatively a rapid method

Produces better results

Visit [www.PEDIAA.com](http://www.PEDIAA.com)

<https://pediaa.com/difference-between-upgma-and-neighbor-joining-tree/>

# Maximum Likelihood (ML) & Bayesian

- Much more statistical based
- Provides probability of a particular tree (not just the answer)
- Similar to Parsimony in that different trees are “scored”, but scores are “likelihoods”
- Branch lengths are estimated.
- Various models of sequence evolution can be used and tested
- Much more computationally challenging

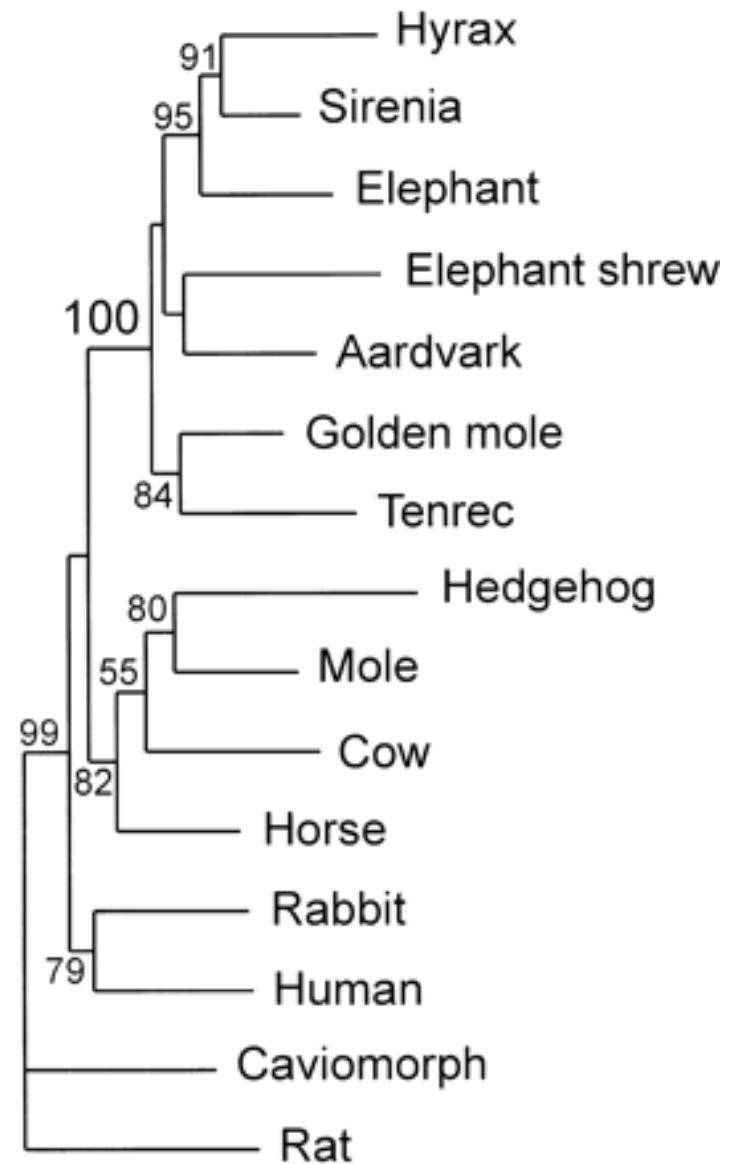
# Bootstrapping

-The number of times a particular branch is formed in the tree (out of the X times the analysis is done)

-High bootstrap values don't mean that your tree is the true tree!

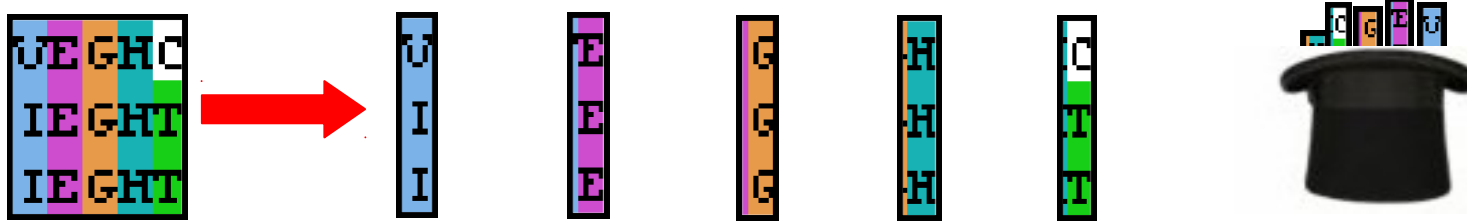
-Bootstrap is a measure of how well your data supports the tree

-Bad data, bad alignment, or bad model will still can give high bootstrap values

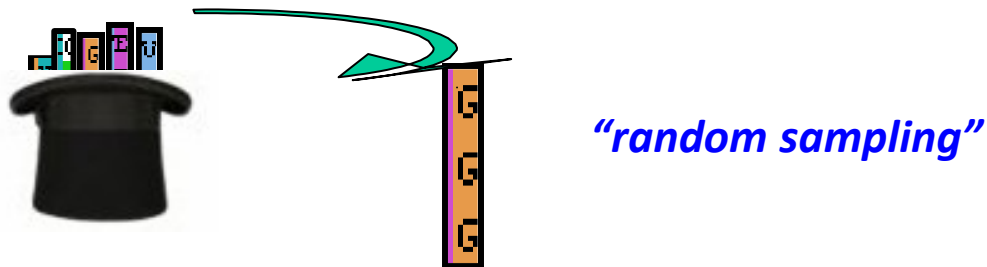


# Bootstrapping – The Picture Version

1. Slice original MSA of  $Y$  residues into  $Y$  columns, put the columns into a hat



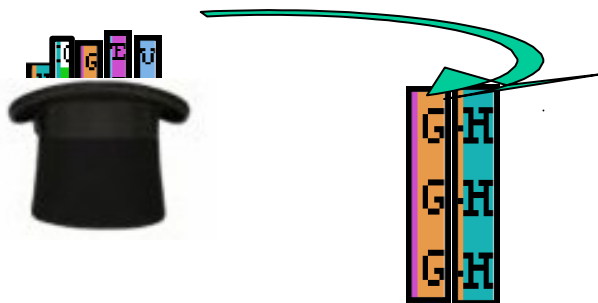
2. Pull out a random column, place it in column #1 of your new test set



3. Put the column back in the hat

*"with replacement"*

4. Pull another column from the hat, place it in column #2 in the test set, put it back



5. Repeat until a *pseudo-dataset* of  $Y$  columns has been made

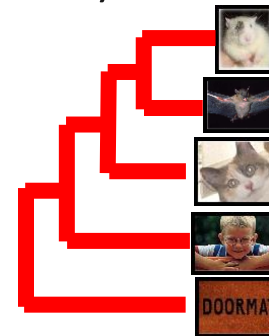
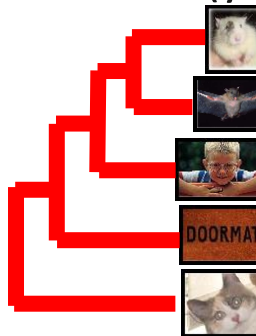
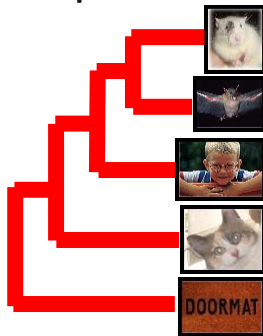
G	H	G	E	V
G	H	G	E	I
G	H	G	E	I

# Bootstrapping

- Repeat  $N$  number of times to generate  $N$  pseudo-datasets

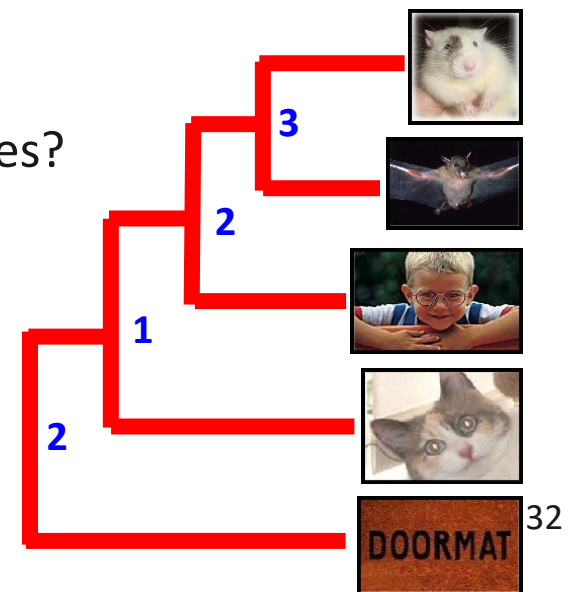


- For each pseudo-dataset, draw a tree (yields  $N$  trees)



- Compare your tree to all  $N$  trees. How often do the branching orders in your tree appear in the  $N$  pseudo-trees?

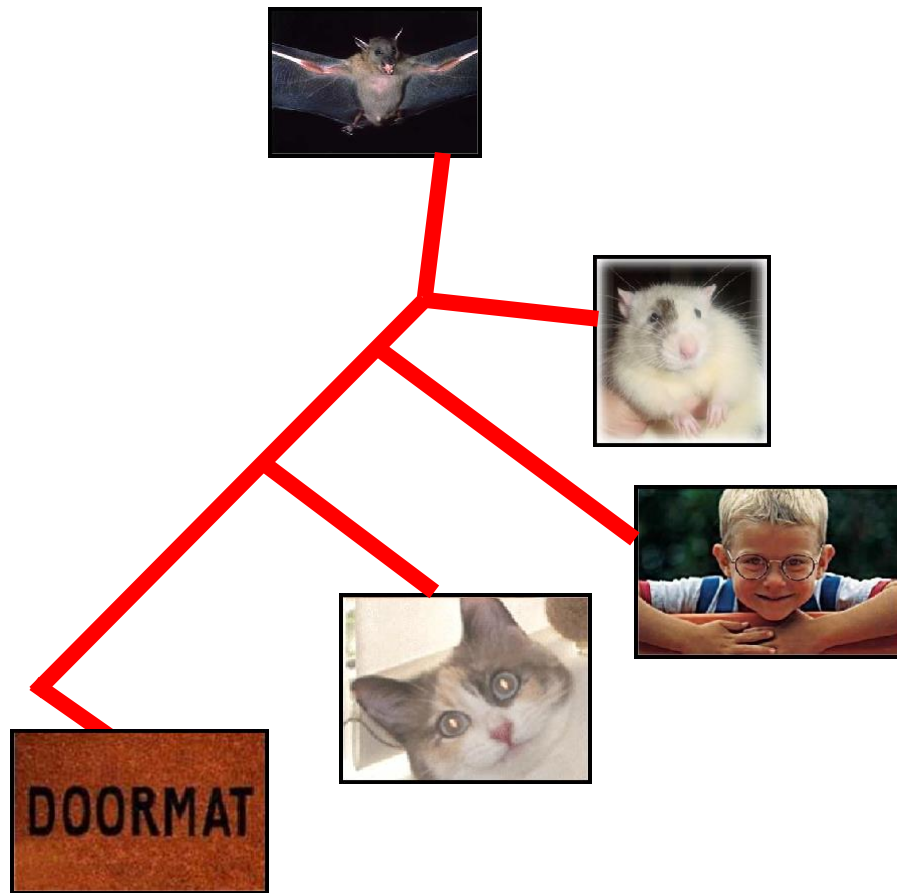
On branches of your tree, write # of times that branch appeared in your pseudo-dataset trees



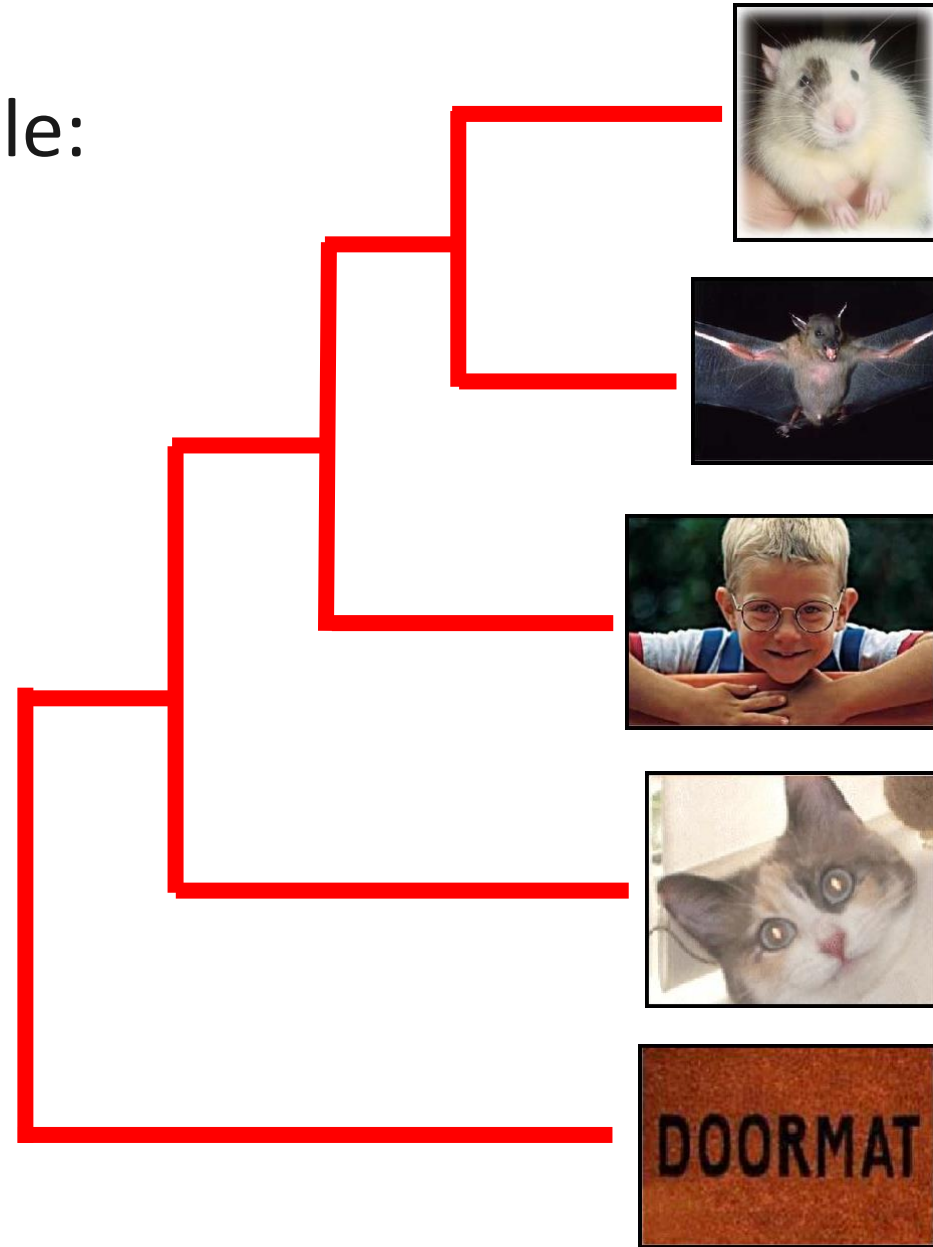


# A Completed Tree

- Alternative displays are possible:



Unrooted



Rooted

# General Software

- A list of everything (>370 programs)
  - <http://evolution.genetics.washington.edu/phylip/software.html>
- General Packages
  - MEGA
  - PHYLIP

<http://www.phylogeny.fr>

# Specific Software

- Parsimony OR Distance Based (e.g. NJ)
  - Clustal, MEGA, PHYLIP, etc
- ML
  - PhyML, Rax-ML (faster), FastTree (fastest)
- Bayesian
  - Mr. Bayes, BEAST

# Tree Viewing

- Archaeopteryx

<http://www.phylosoft.org/archaeopteryx/>

- FigTree

<http://tree.bio.ed.ac.uk/software/figtree/>

- Dendroscope

<http://ab.inf.uni-tuebingen.de/software/dendroscope/>

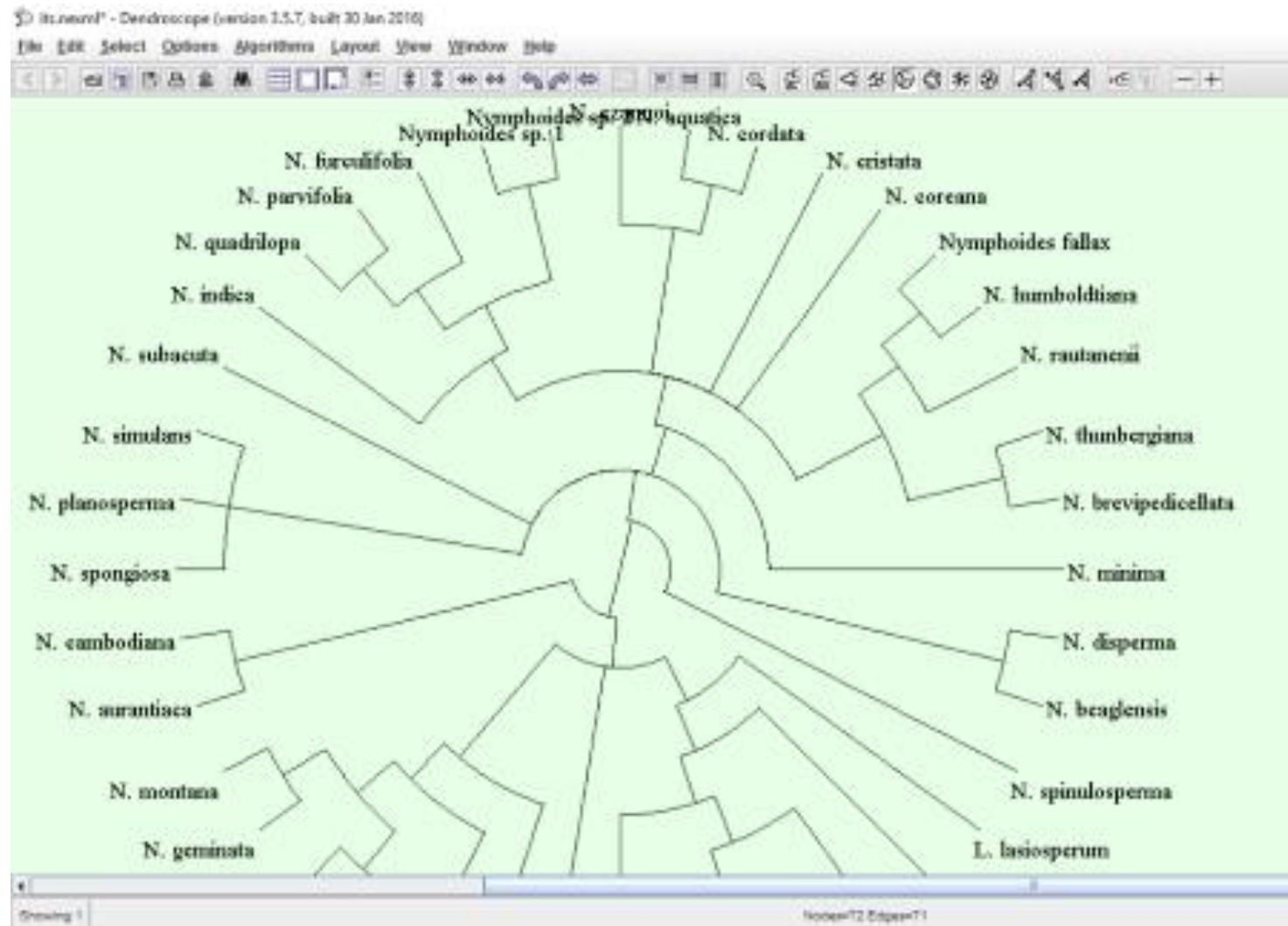
- iTOL

<http://itol.embl.de/>

[illegible]

<https://listoffreeware.com/free-phylogenetic-tree-viewer-software-windows/>

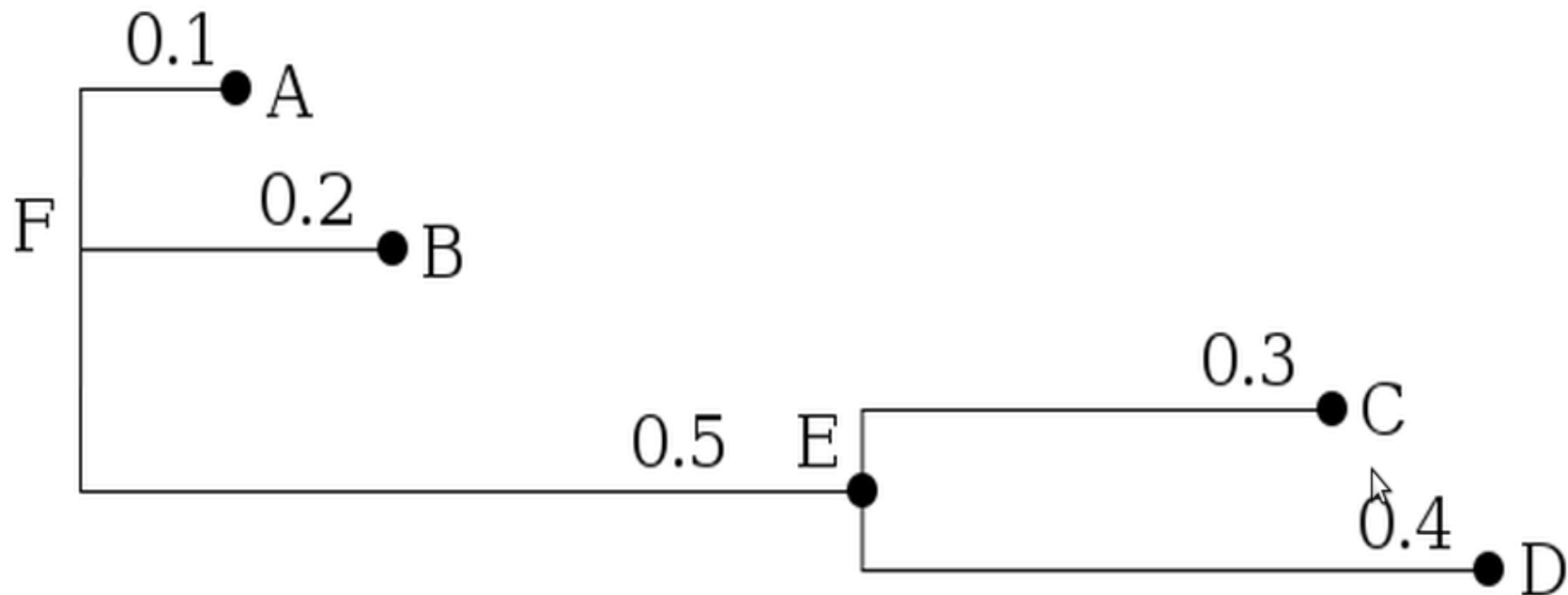
# Dendroscope



# Tree File Formats

- Newick
  - Simplest format
- NEXUS
  - More complex
  - Can handle multiple trees and MSAs all in one.

# Newick



could be represented in Newick format in several ways

```
(,,(,));  
(A,B,(C,D));  
(A,B,(C,D)E)F;  
(:0.1,:0.2,(0.3,0.4):0.5);  
(:0.1,:0.2,(0.3,0.4):0.5):0.0;  
(A:0.1,B:0.2,(C:0.3,D:0.4):0.5);  
(A:0.1,B:0.2,(C:0.3,D:0.4)E:0.5)F;  
((B:0.2,(C:0.3,D:0.4)E:0.5)F:0.1)A;
```

*no nodes are named*

*leaf nodes are named*

*all nodes are named*

*all but root node have a distance to parent*

*all have a distance to parent*

*distances and leaf names (popular)*

*distances and all names*

*a tree rooted on a leaf node (rare)*



# Nexus

```
#NEXUS
```

```
Begin trees; [Treefile saved Wed Jul 26 19:40:41 2000]
```

```
[output from your data run]
```

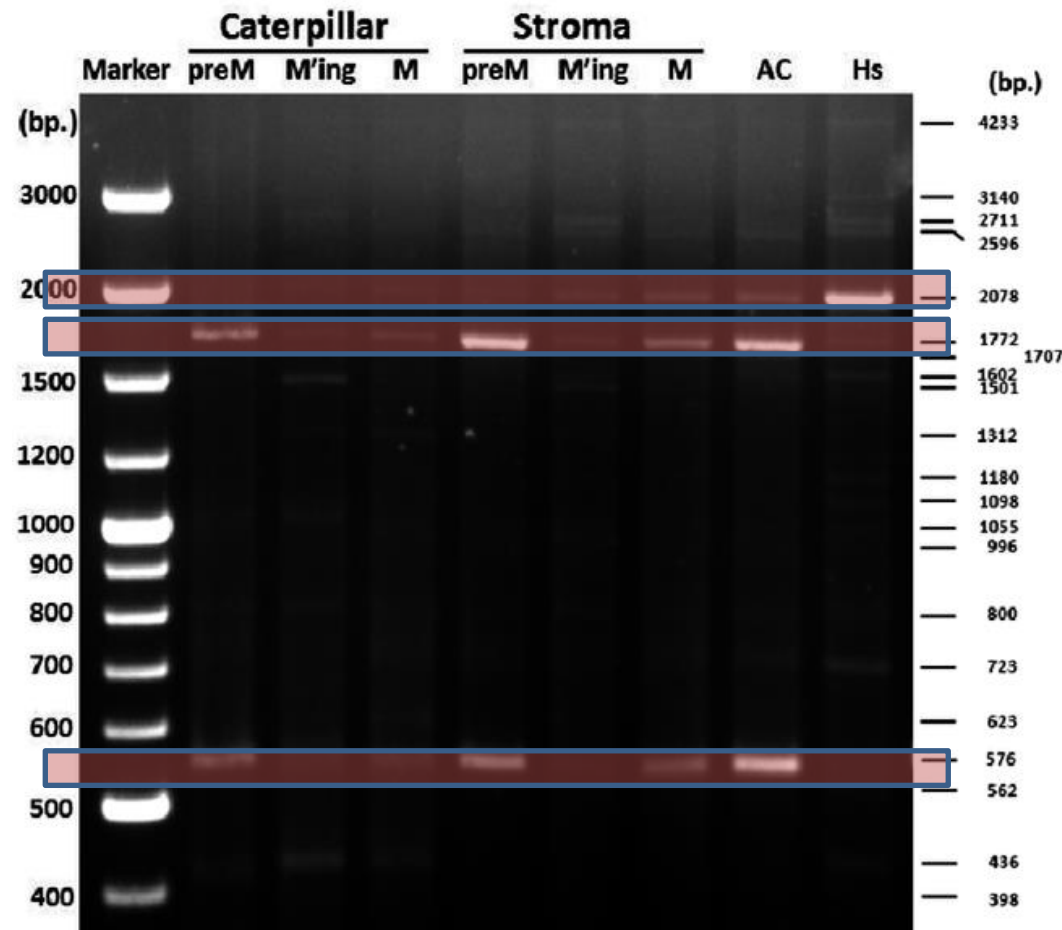
```
Translate
```

```
1 TRXecoli,  
2 TRXHomo,  
3 TRXSacch,  
4 erCaelA,  
5 erCaelB,  
6 erCaelC,  
7 erHomoA,  
8 erHomoB,  
9 erHomoC,  
10 erpCaelC  
;
```

```
tree PAUP_1 = [&U] (1,((2,3),(((4,10),(5,8)),(6,9)),7))) ;
```

```
End;
```

# Build phylogenetic tree using NJ method on RAPD data



	2078	1772	576
preMC	0	1	1
MingC	0	0	0
MC	0	1	0
preMS	0	1	1
MingS	0	0	0
MS	1	1	1
AC	1	1	1
Hs	1	0	0

