Biol5705
Module: Gene Sequence Analysis

Lecture 4
Trees

Dr. Morgan Langille

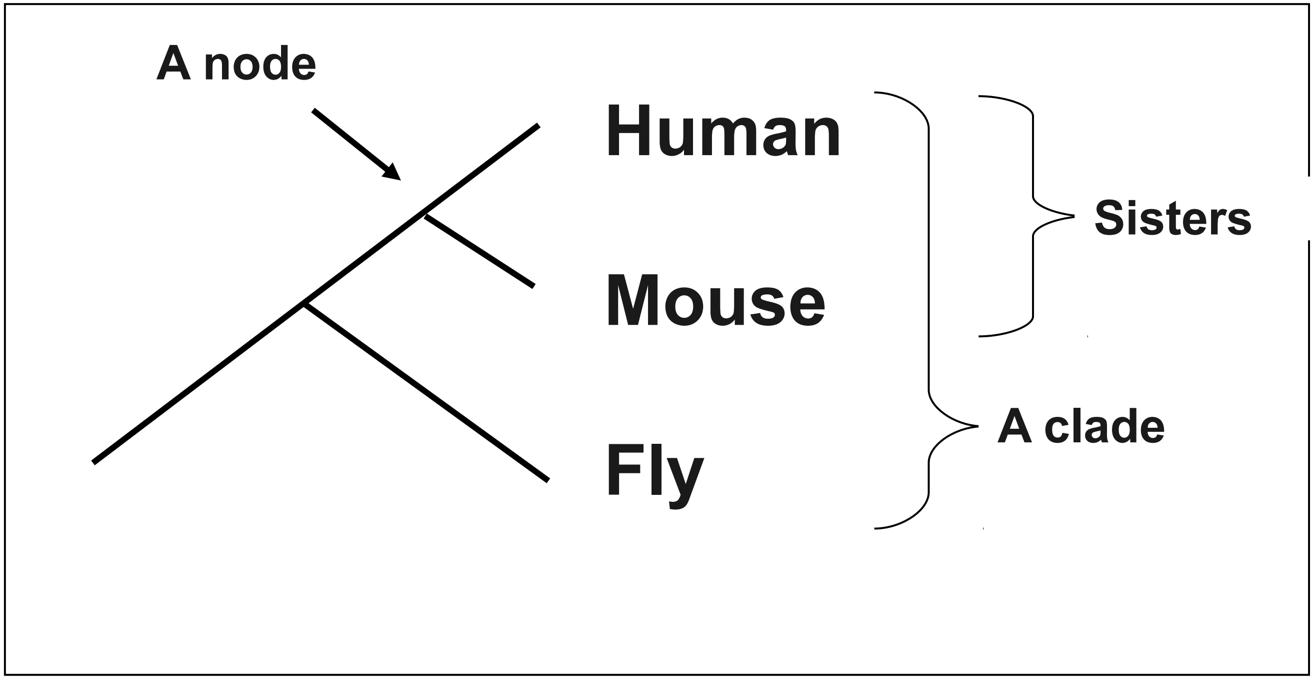
Outline

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

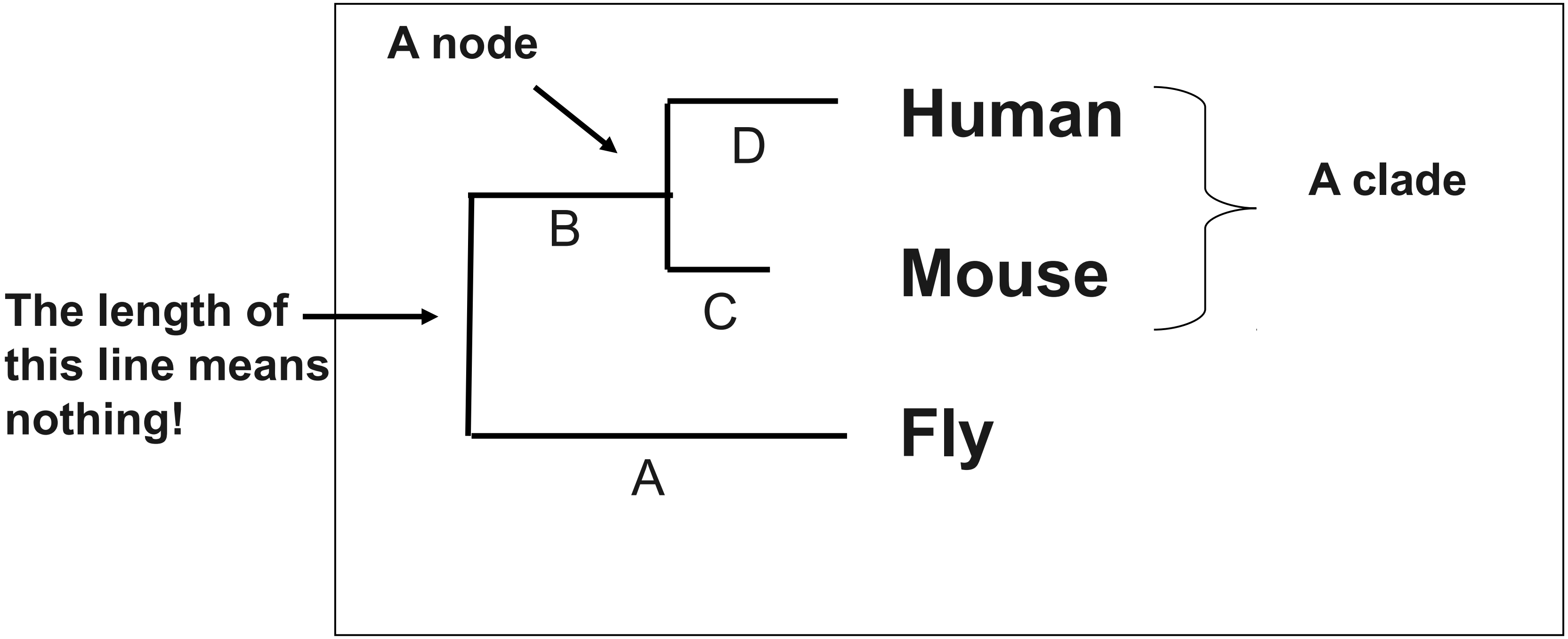
Why do we build trees?

- 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



A phylogenetic tree with branch lengths



A node

A clade

The length of this line means nothing!

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

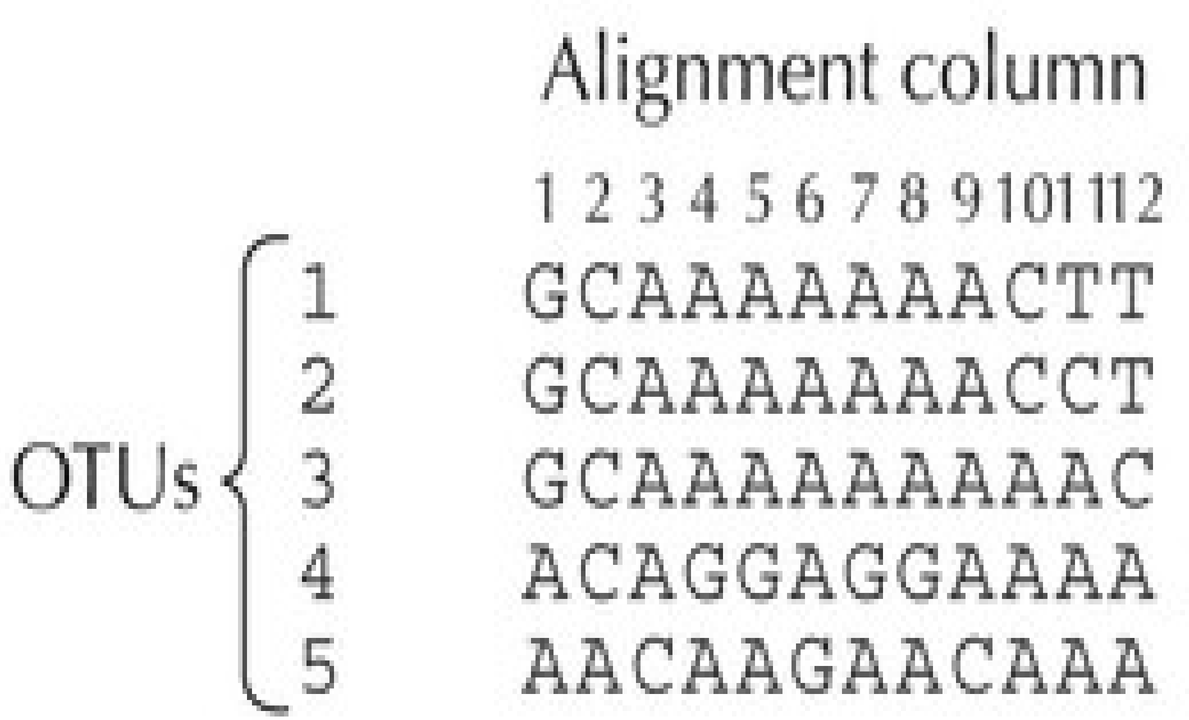
(i.e. sum of branches A+B+C is less than sum of A+B+D)

Parsimony

- The tree implying the least number of changes in character states (most parsimonious) is the best.

- Note:
  - Does not determine branch lengths

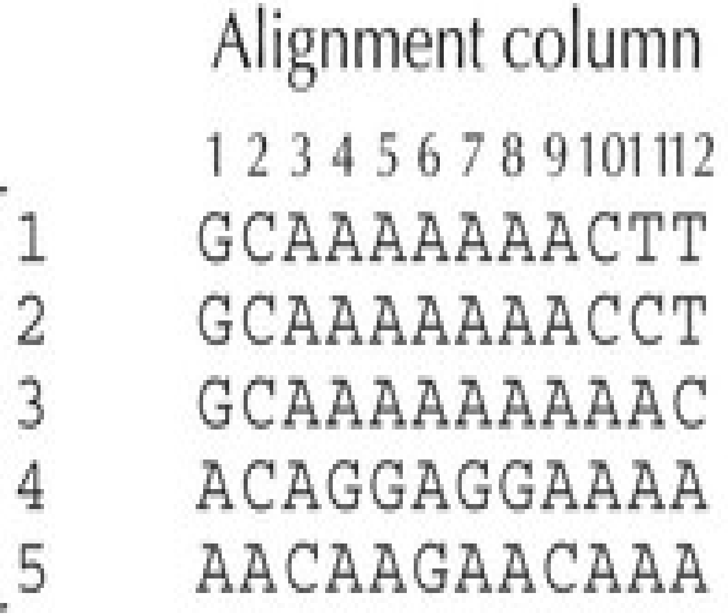
Example Alignment



Alignment column

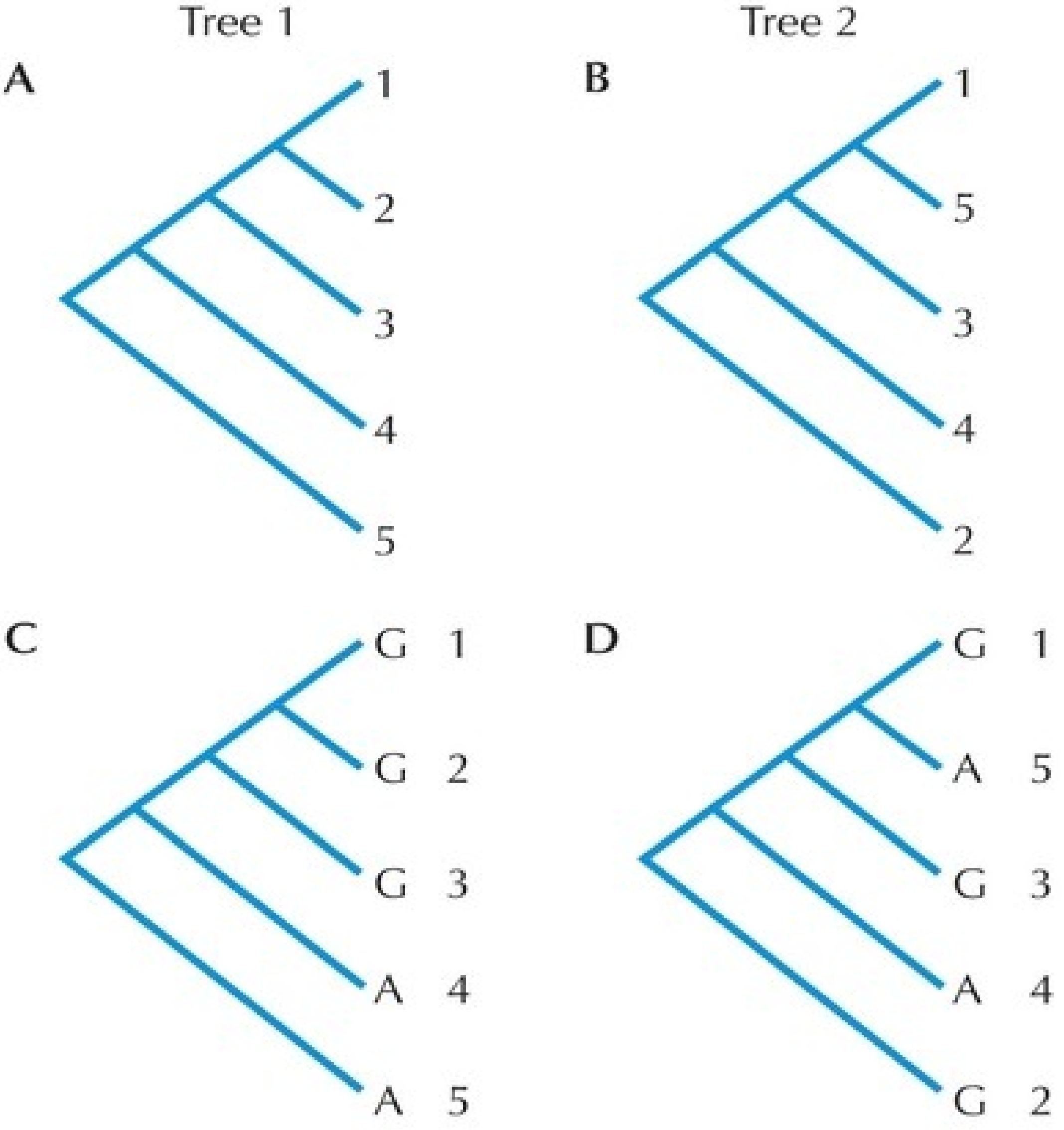
		1	2	3	4	5	6	7	8	9	10	11	12
OTUs	1	G	C	A	A	A	A	A	A	A	C	T	T
	2	G	C	A	A	A	A	A	A	A	C	C	T
	3	G	C	A	A	A	A	A	A	A	A	A	C
	4	A	C	A	G	G	A	G	G	A	A	A	A
	5	A	A	C	A	A	G	A	A	C	A	A	A

Parsimony Example

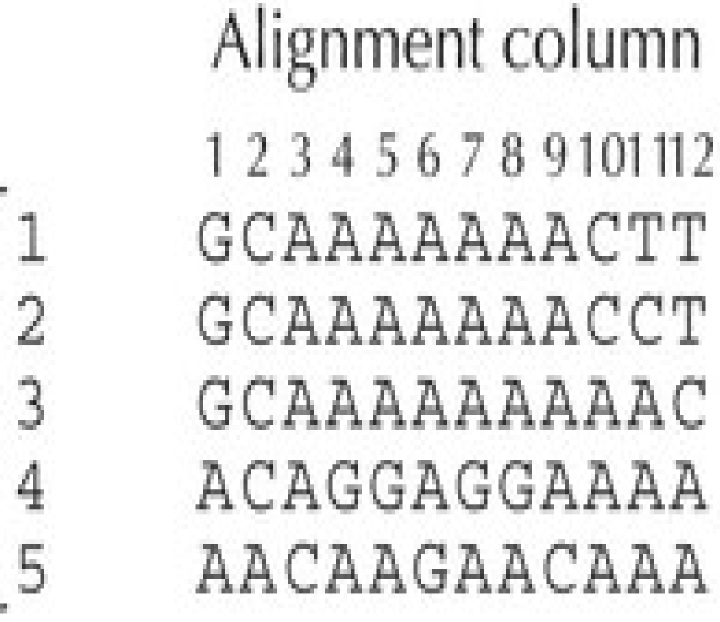


Alignment column

```
       1 2 3 4 5 6 7 8 9 10 11 12
1      GCAAAAAAACTT
2      GCAAAAAAACCT
3      GCAAAAAAAAAC
4      ACAGGAGGAAAA
5      AACAAGAACAAA
```

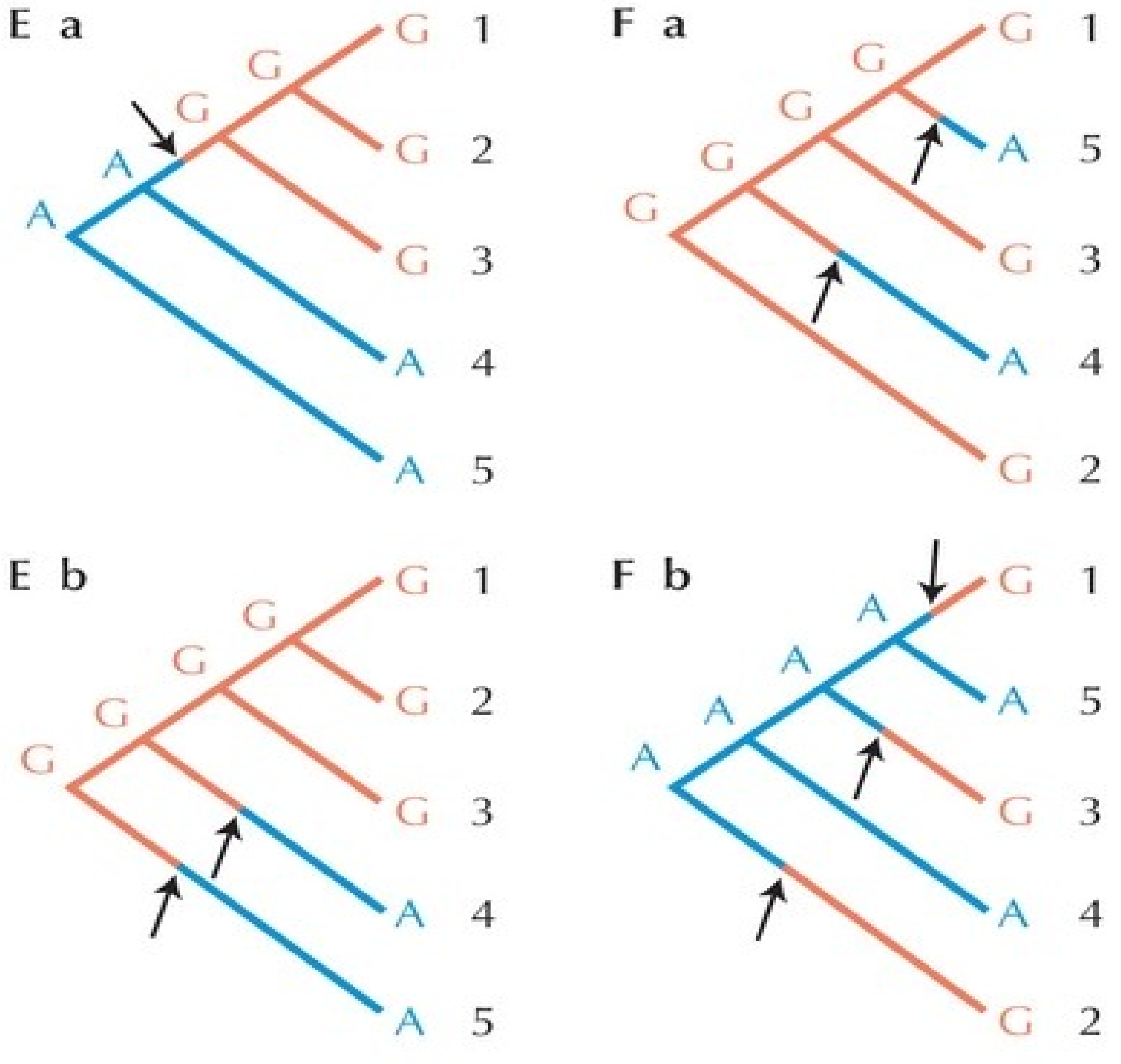


Parsimony Example



```
         Alignment column

         1 2 3 4 5 6 7 8 9 10 11 12
1        GCAAAAAAACTT
2        GCAAAAAAACCT
3        GCAAAAAAAAAC
4        ACAGGAGGAAAA
5        AACAAGAACAAA
```



Number of Trees

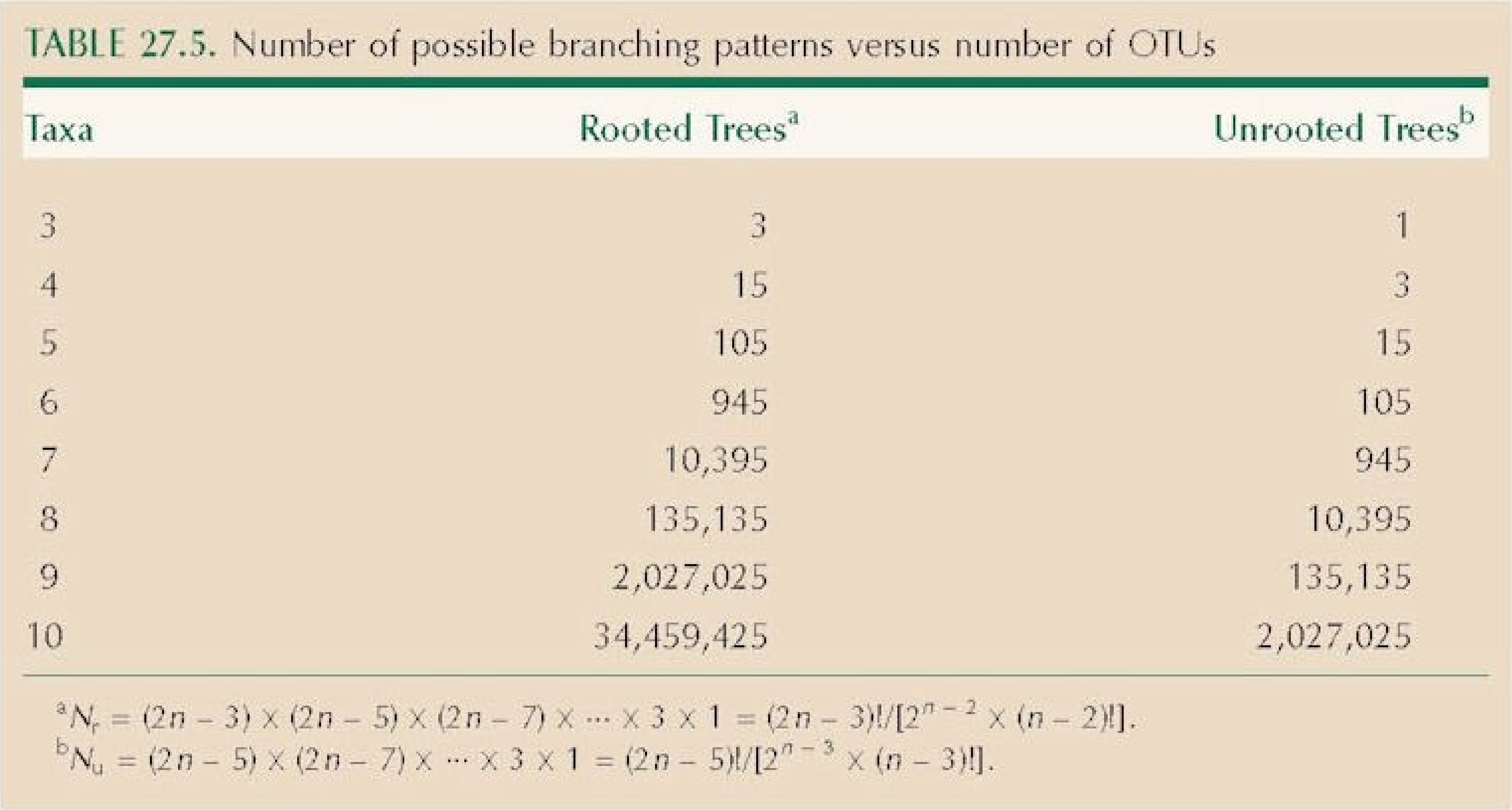


TABLE 27.5. Number of possible branching patterns versus number of OTUs

Taxa	Rooted Trees[a]	Unrooted Trees[b]
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] $N_r = (2n - 3) \times (2n - 5) \times (2n - 7) \times \cdots \times 3 \times 1 = (2n - 3)!/[2^{n-2} \times (n - 2)!]$.

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

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

Distance Based Method

- Start with a distance matrix between every pair of sequences

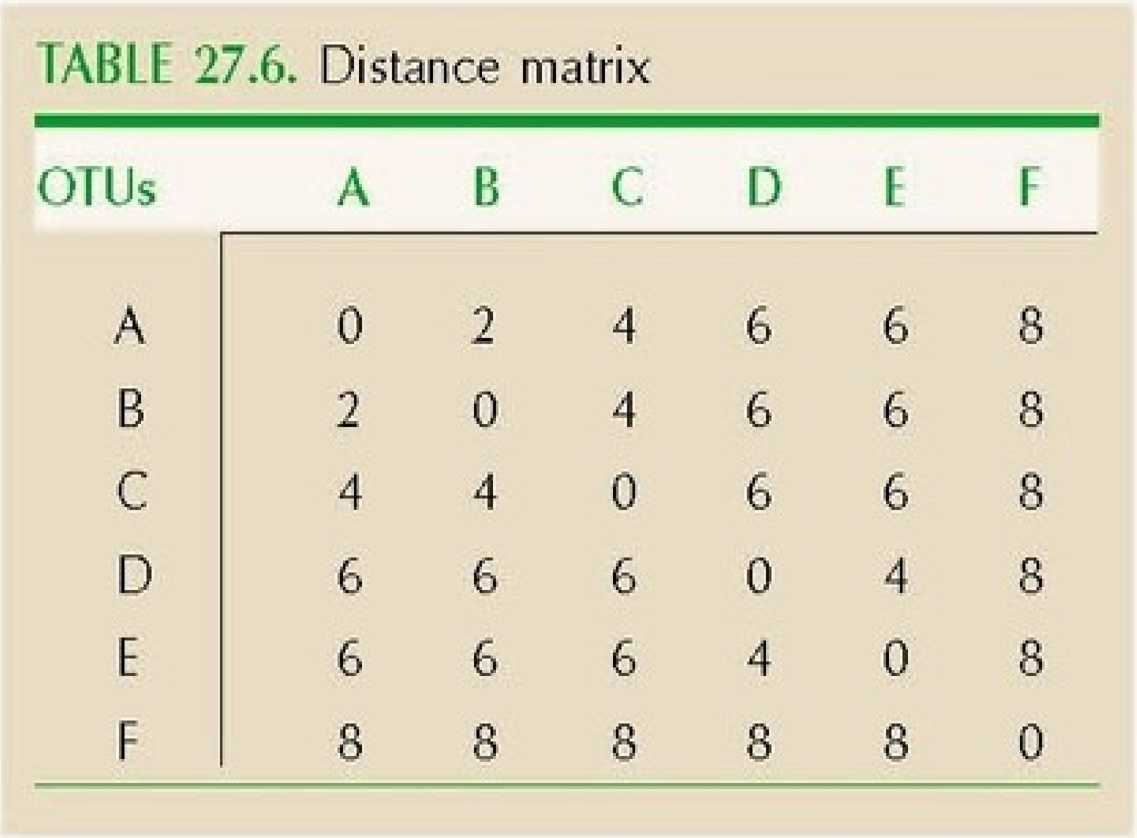


TABLE 27.6. Distance matrix

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

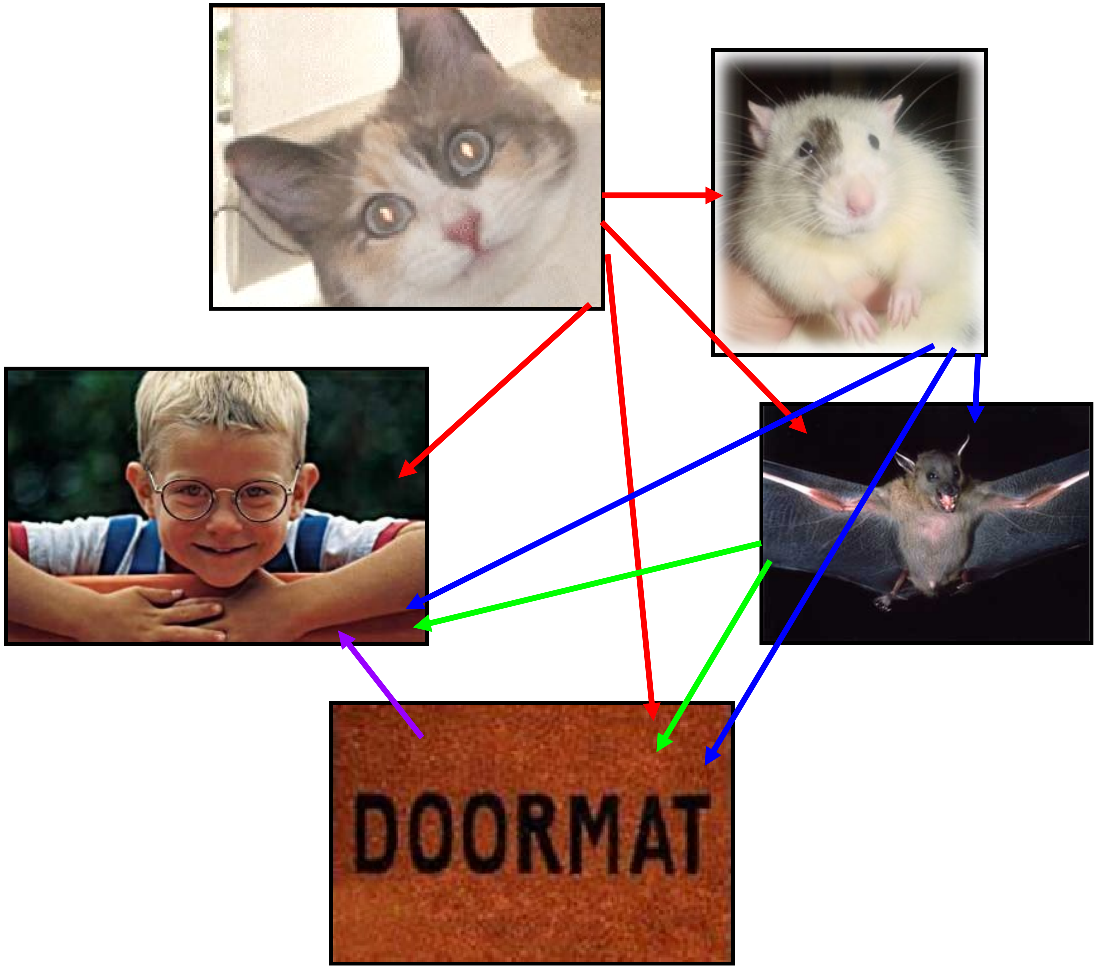
- unweighted pair group method with arithmetic mean (UPGMA)
- Neighbour Joining

- Computationally Very Fast
- Often included in MSA programs

Building an NJ Tree - An Example

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

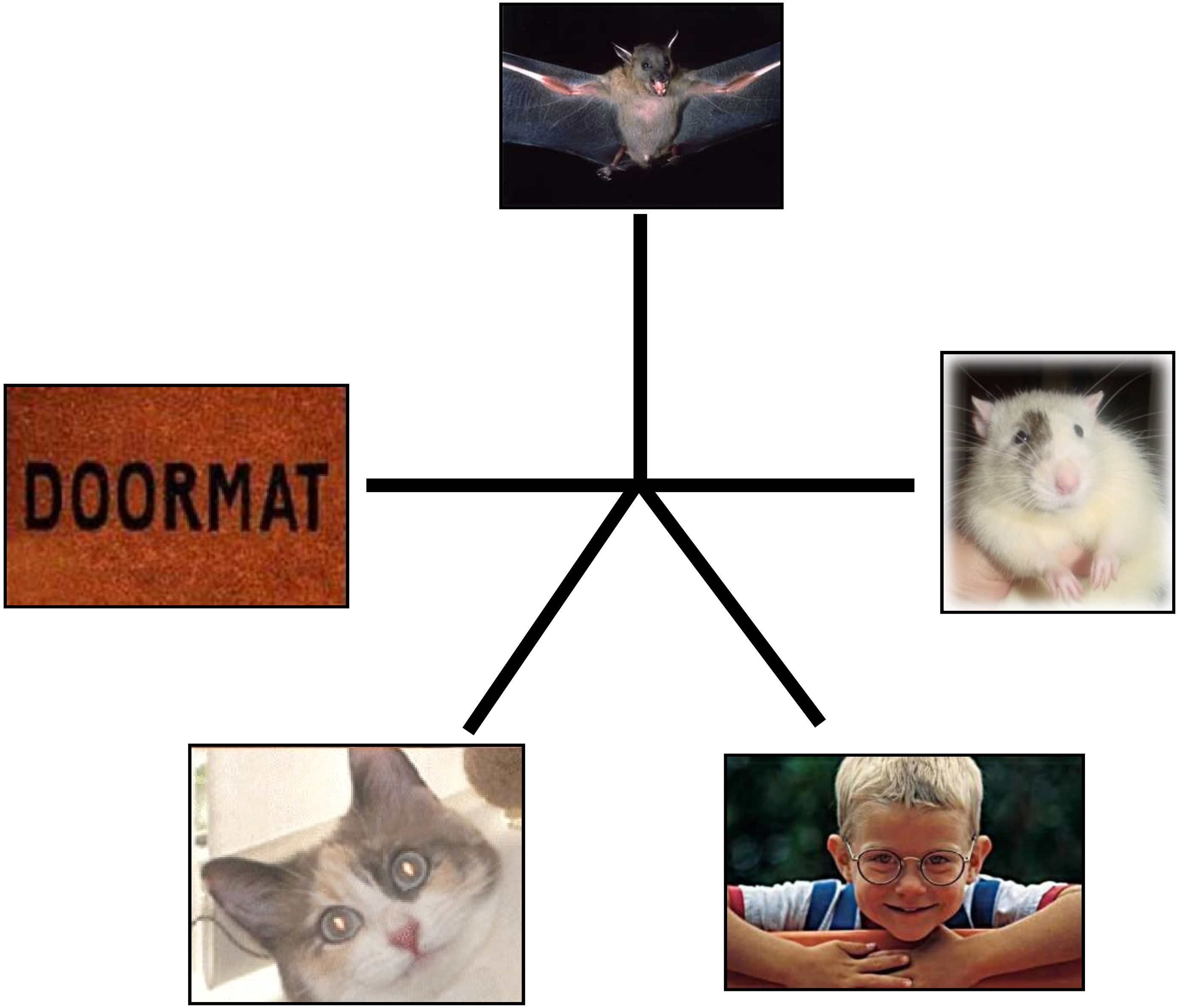
1. Compare all sequences to each other.
2. Assign divergence values to each pair
3. Assemble the values in a distance matrix



	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

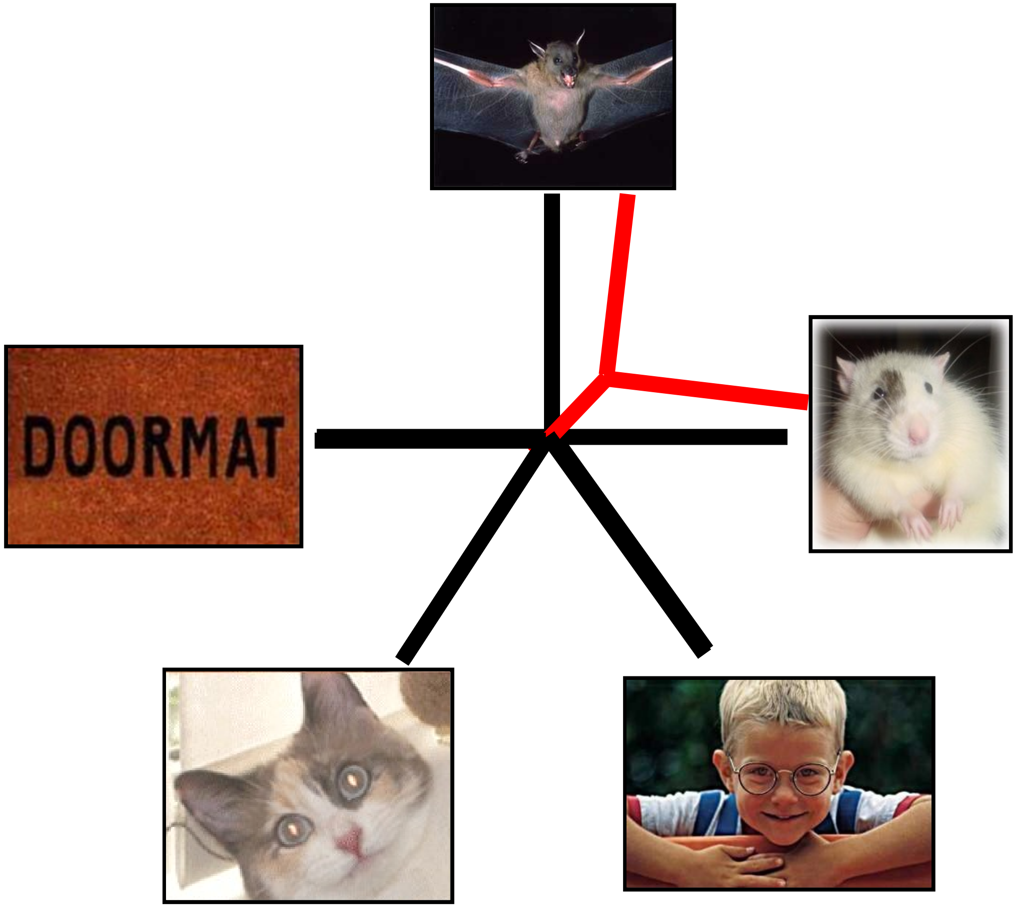
Building an NJ Tree

4. Arrange the subjects in a “star” phylogeny



Building an 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	0.3	0.8	-
Matt	0.6	0.4	0.5	0.9

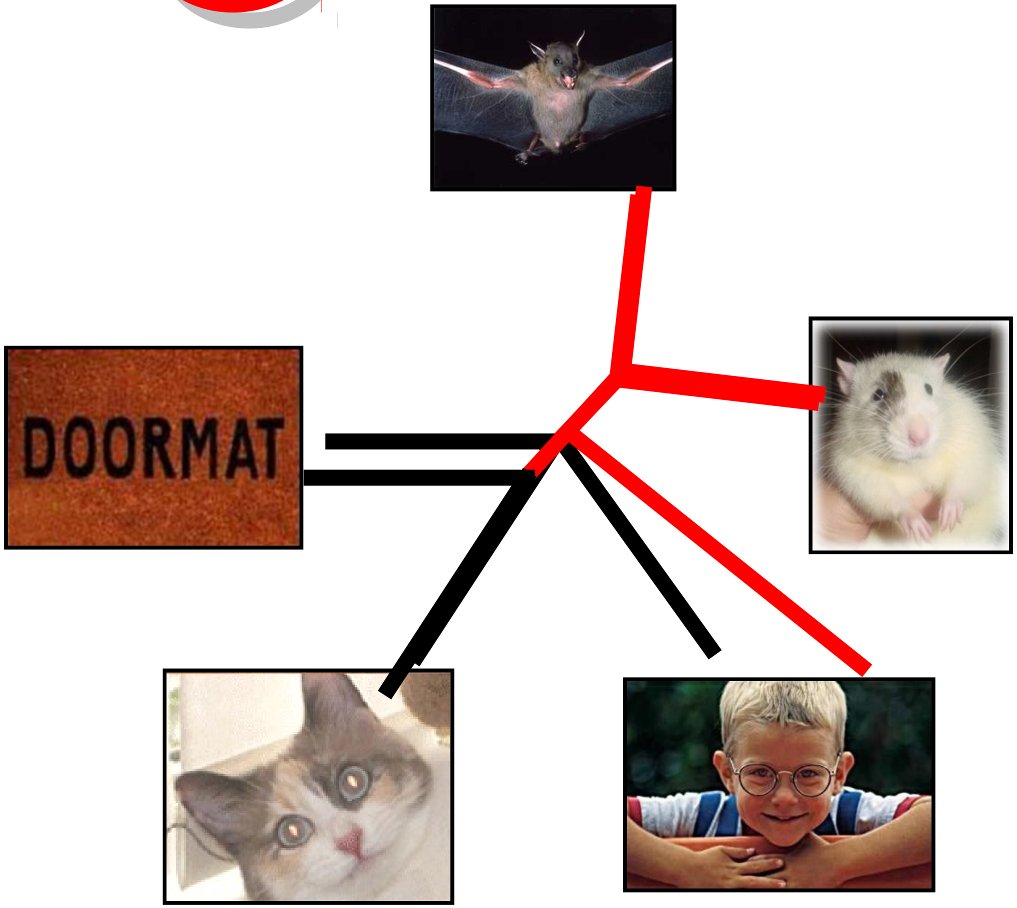
Building an NJ Tree

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

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

7. Fuse the next two closest sequences

8. Repeat until tree completed



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

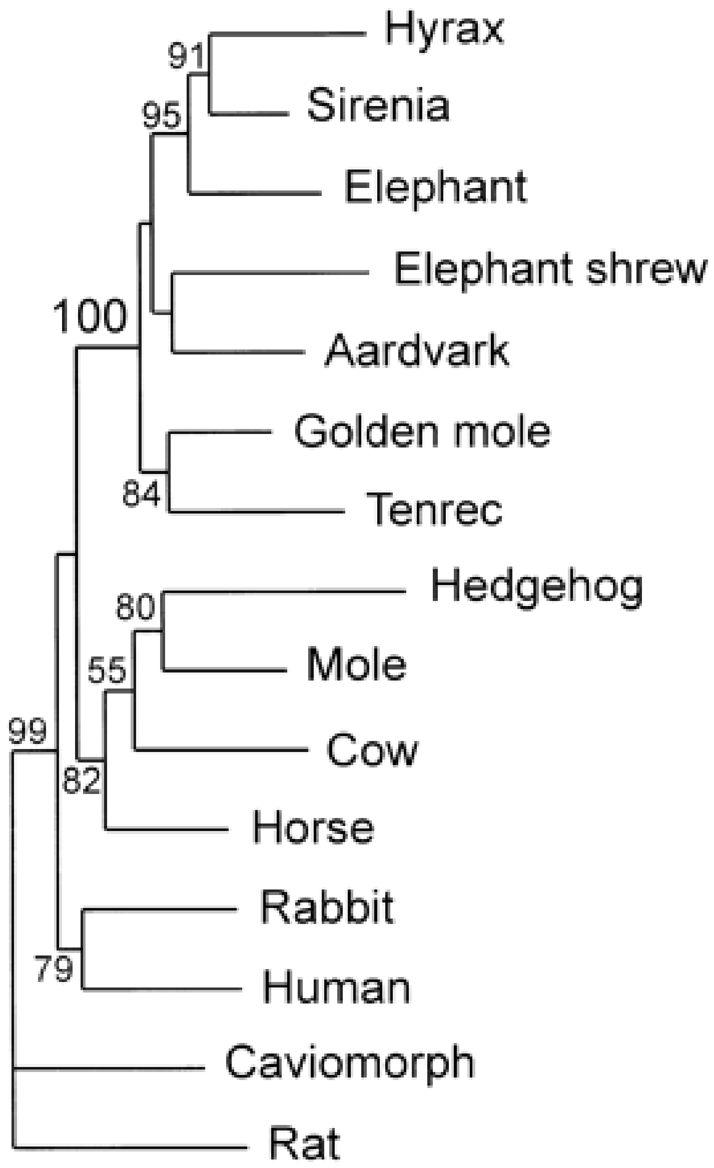
Bootstrapping

- 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!

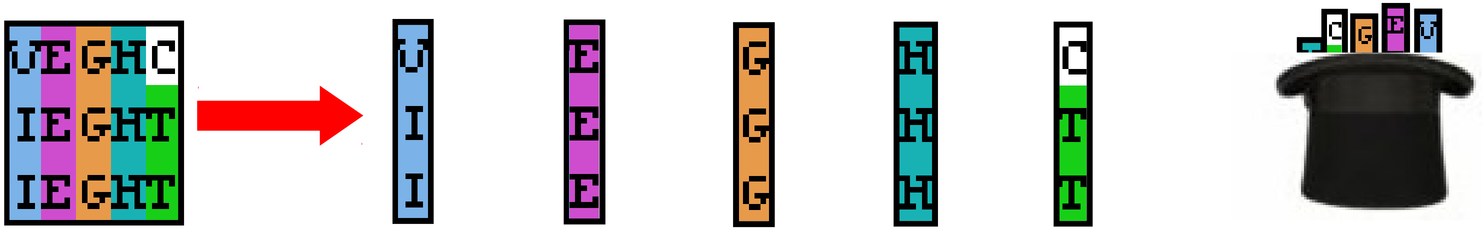
- Bootstrap is a measure of how well your data supports the tree

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

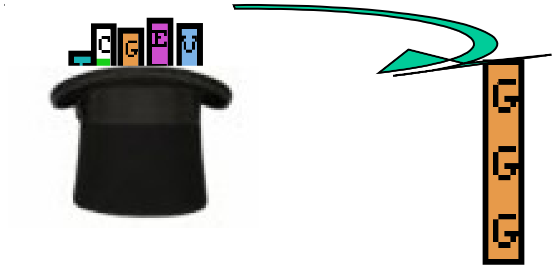


Bootstrapping – The Picture Version

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

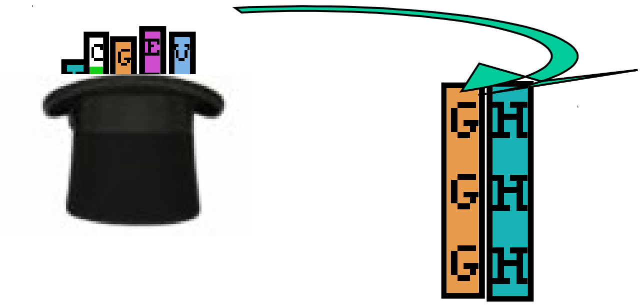


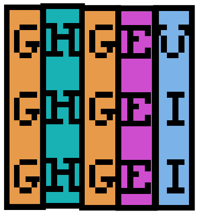
"random sampling"

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

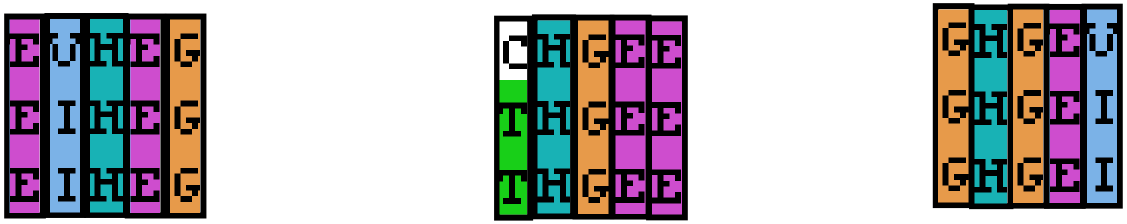




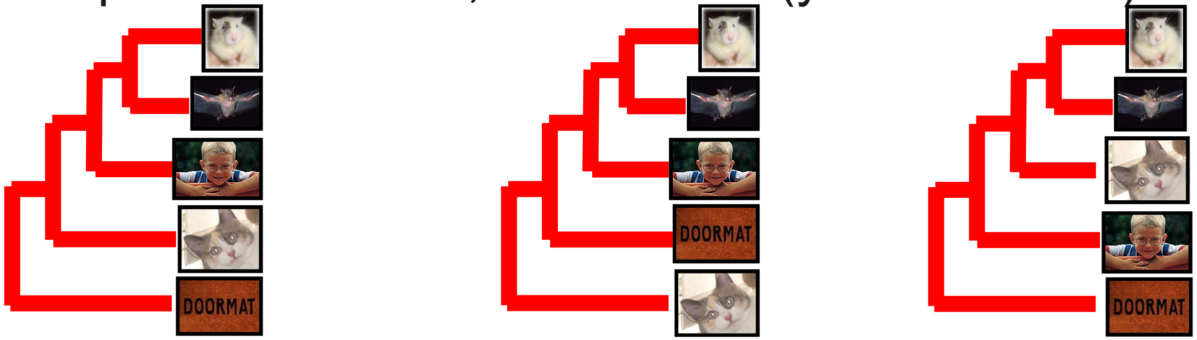
5. Repeat until a *pseudo-dataset* of *Y* columns has been made

Bootstrapping

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

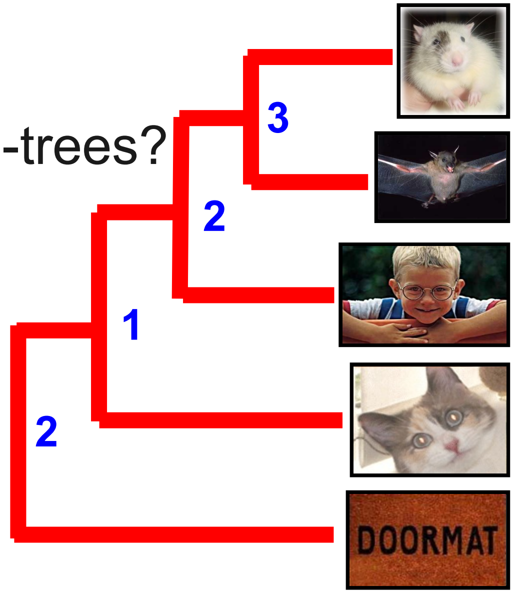


- For each pseudo-dataset, draw a tree (yields *N* trees)



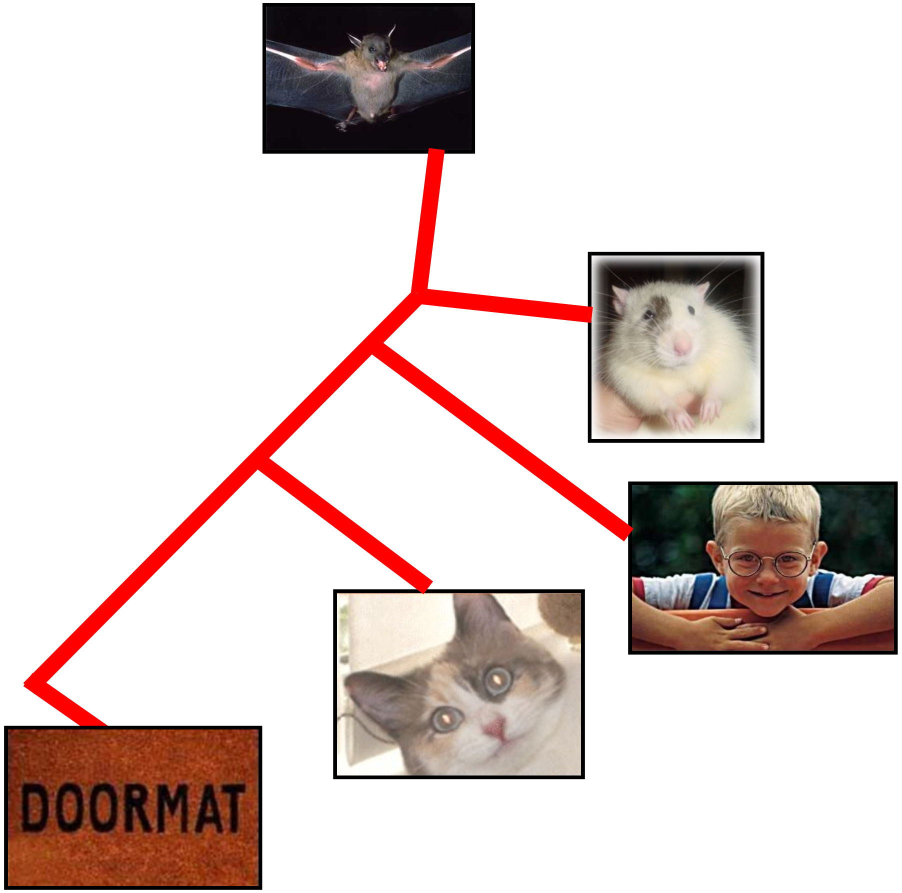
- Compare your tree to all *N* trees. How often do the branching orders in your tree appear in the *N* pseudo-trees?

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

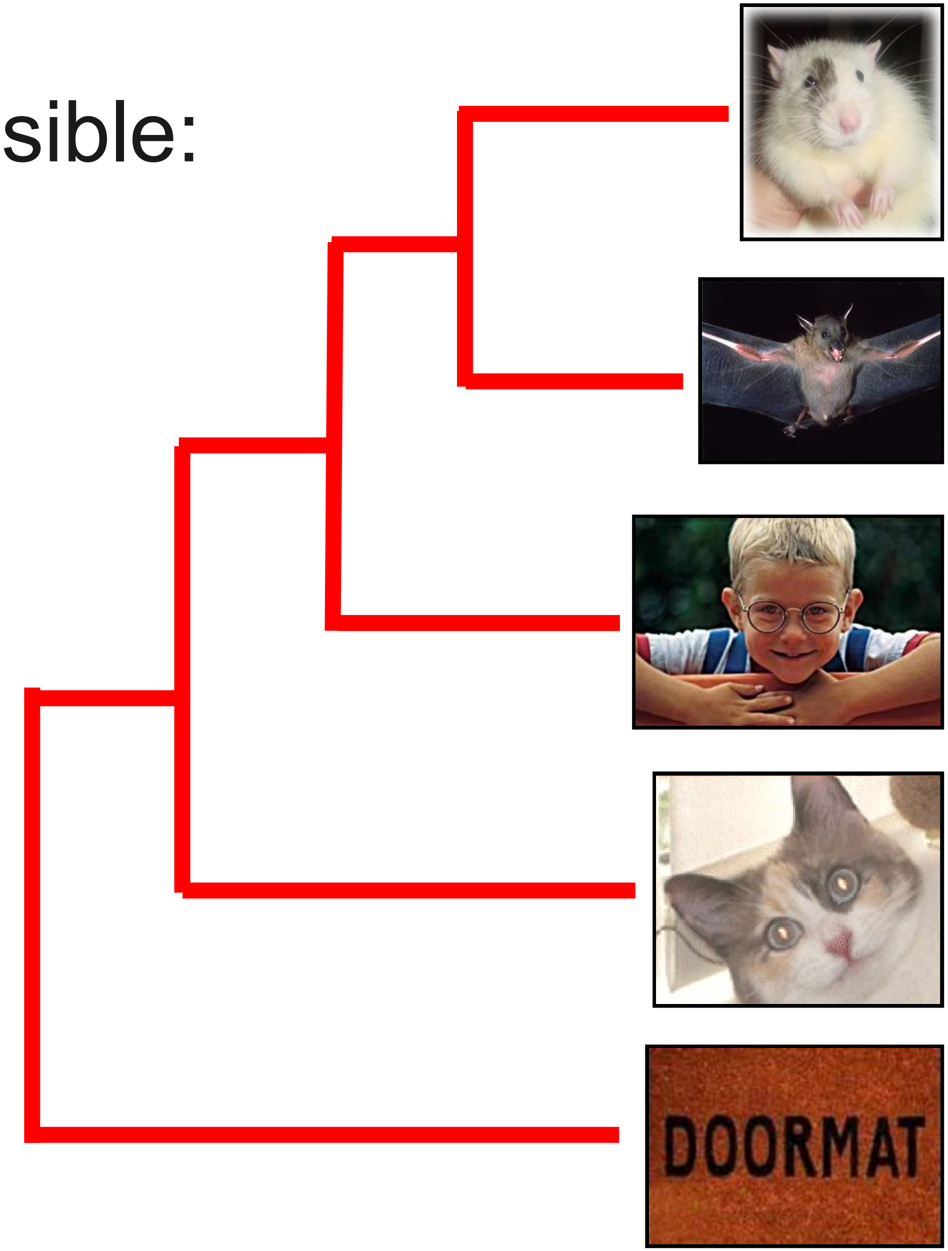


A Completed Tree

- 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

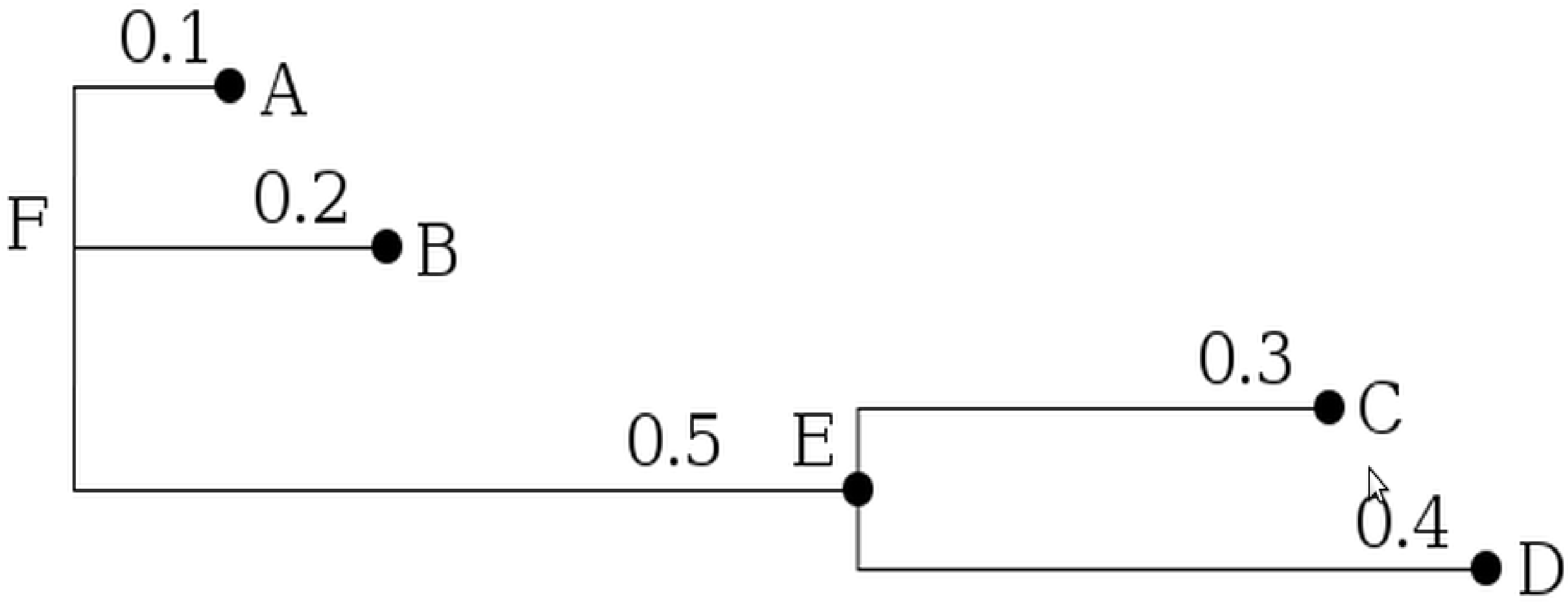
  http://ab.inf.uni-tuebingen.de/software/dendroscope/
- iTOL

  http://itol.embl.de/

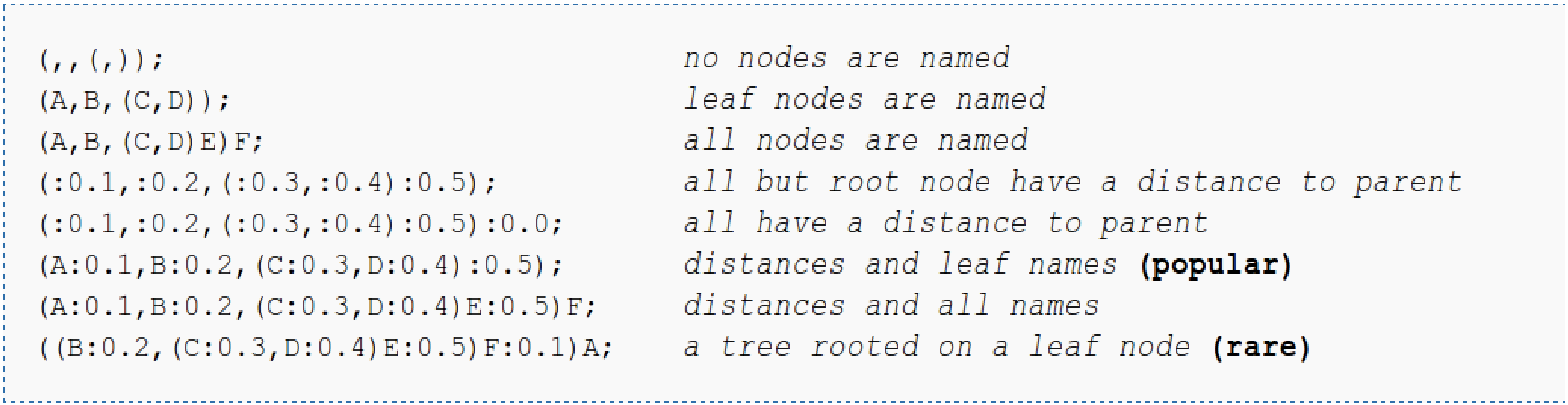
Tree File Formats

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

Newick

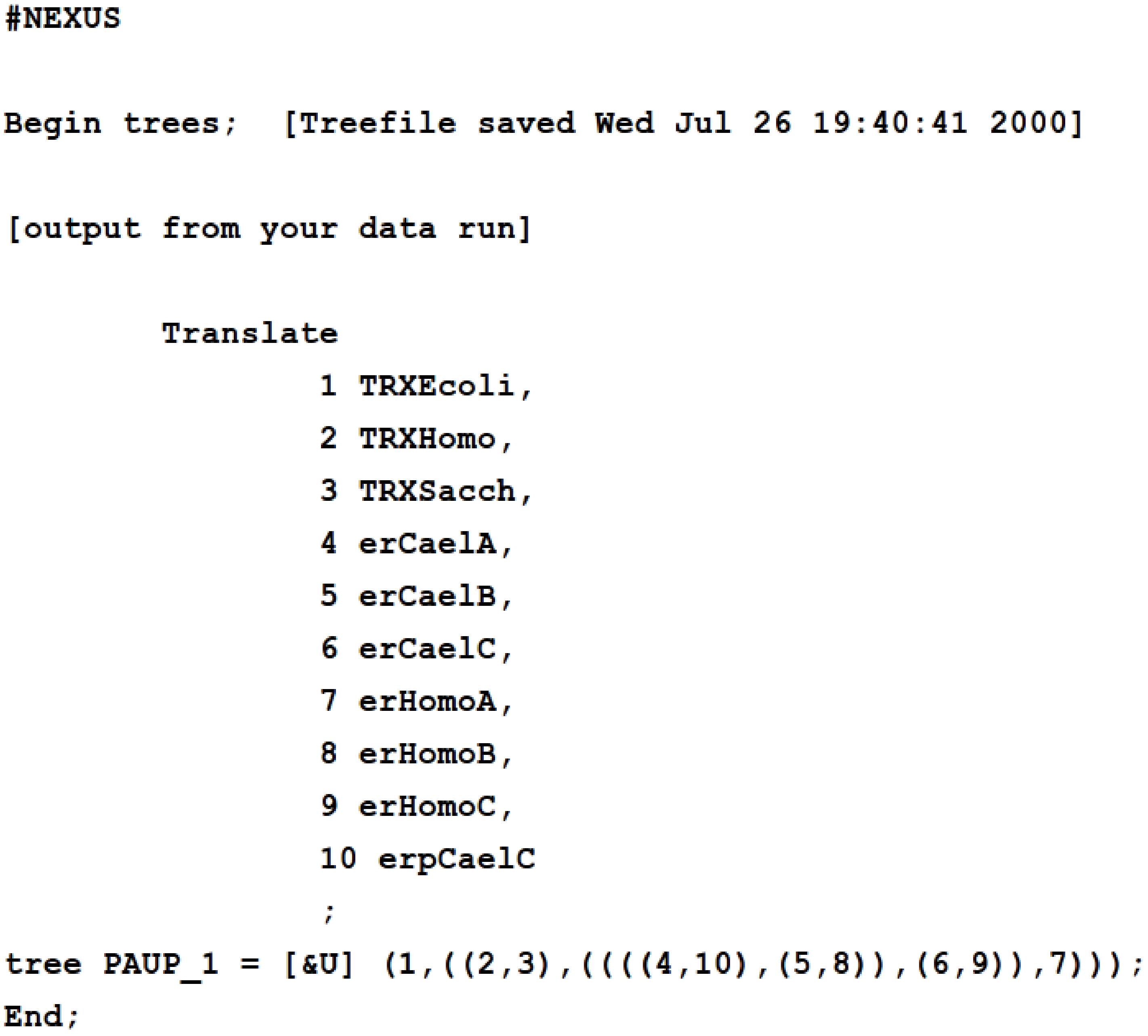


could be represented in Newick format in several ways



```
(,,(,));                               no nodes are named
(A,B,(C,D));                           leaf nodes are named
(A,B,(C,D)E)F;                         all nodes are named
(:0.1,:0.2,(:0.3,:0.4):0.5);           all but root node have a distance to parent
(:0.1,:0.2,(:0.3,:0.4):0.5):0.0;       all have a distance to parent
(A:0.1,B:0.2,(C:0.3,D:0.4):0.5);       distances and leaf names (popular)
(A:0.1,B:0.2,(C:0.3,D:0.4)E:0.5)F;     distances and all names
((B:0.2,(C:0.3,D:0.4)E:0.5)F:0.1)A;    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;
```