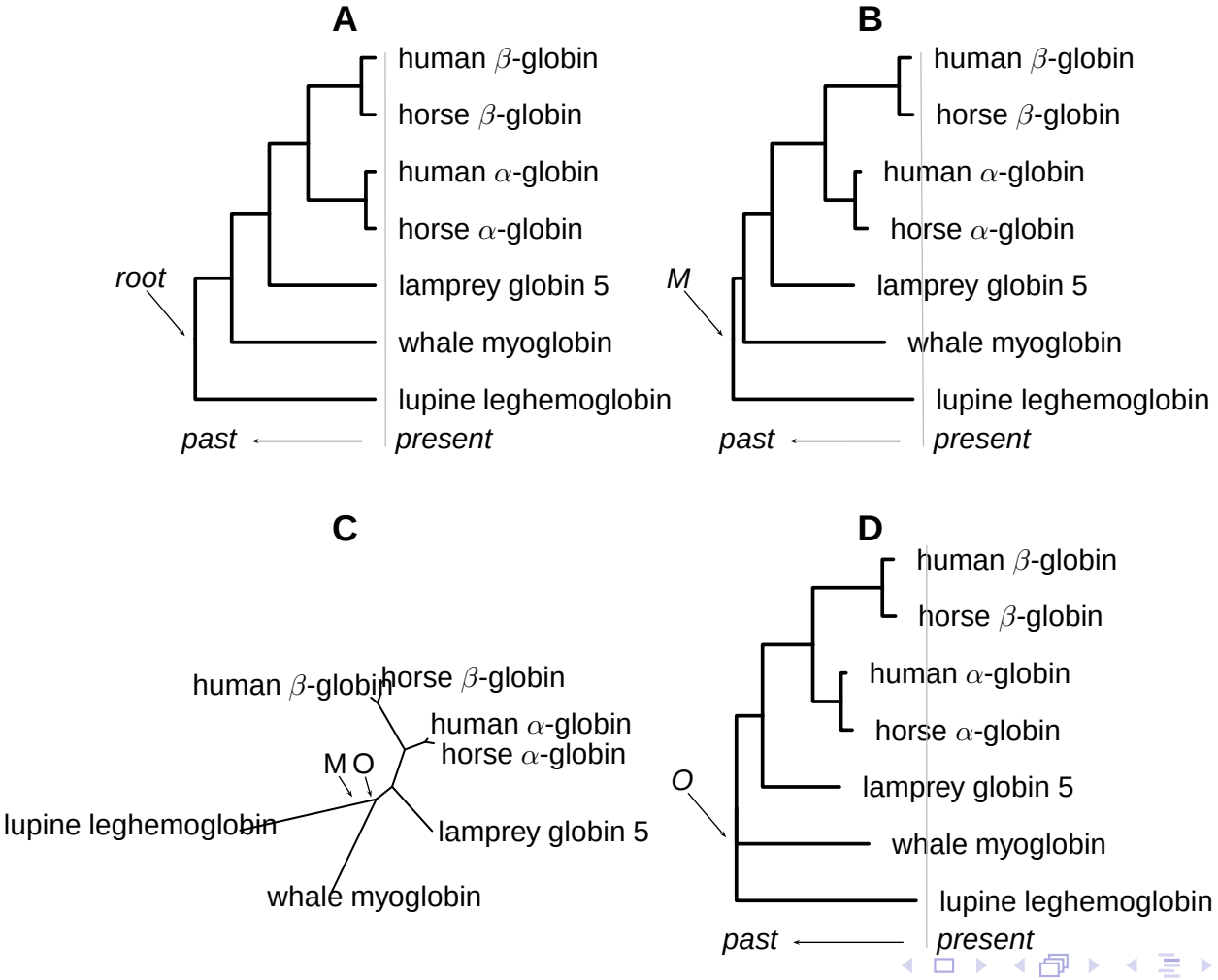
Introduction to Computational Biology; An Evolutionary Approach: Phylogeny

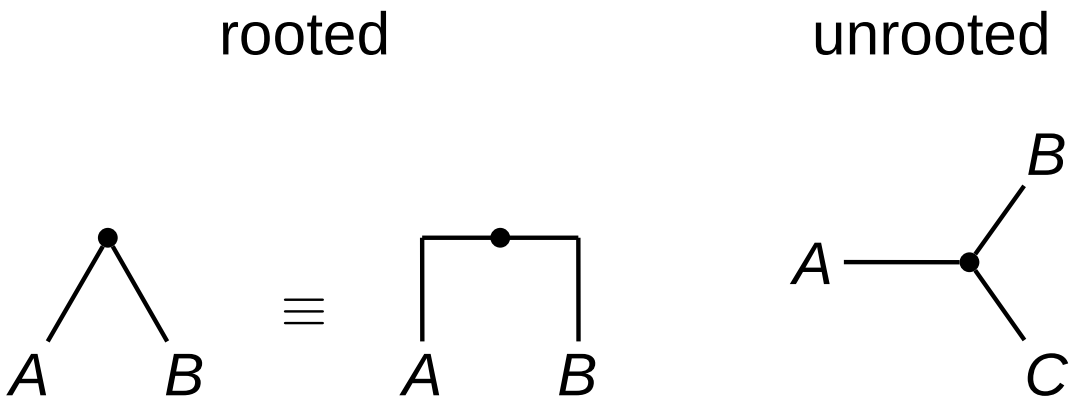
Bernhard Haubold & Thomas Wiehe

Outline

Clocks and Roots



Rooted and Unrooted Phylogenies

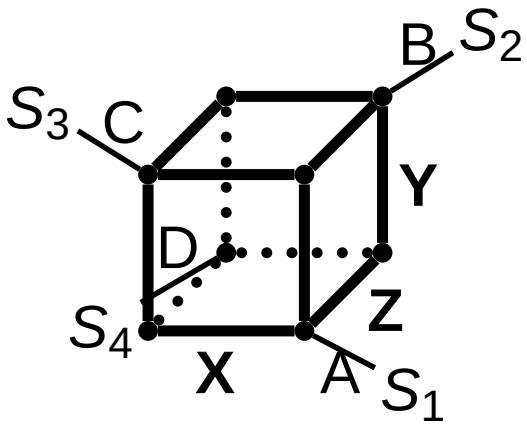


Statistical Geometry—1

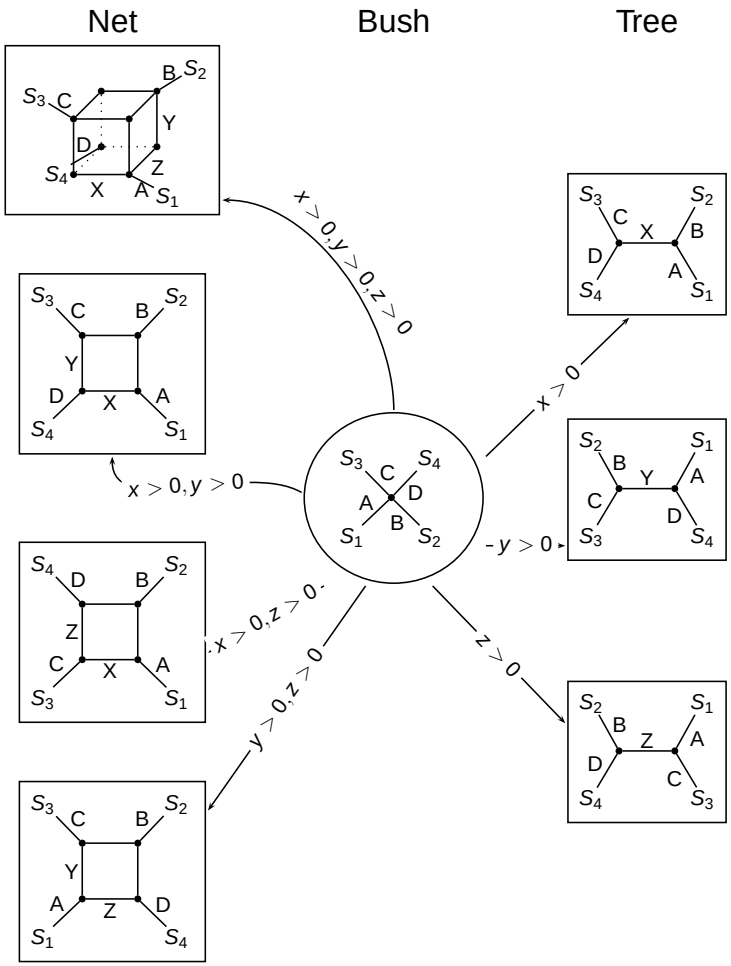
A

S_1	YRYYYYRR
S_2	YYRYYYYY
S_3	YYYRYRRY
S_4	YYYYRRYR
Class	0ABCDXYZ

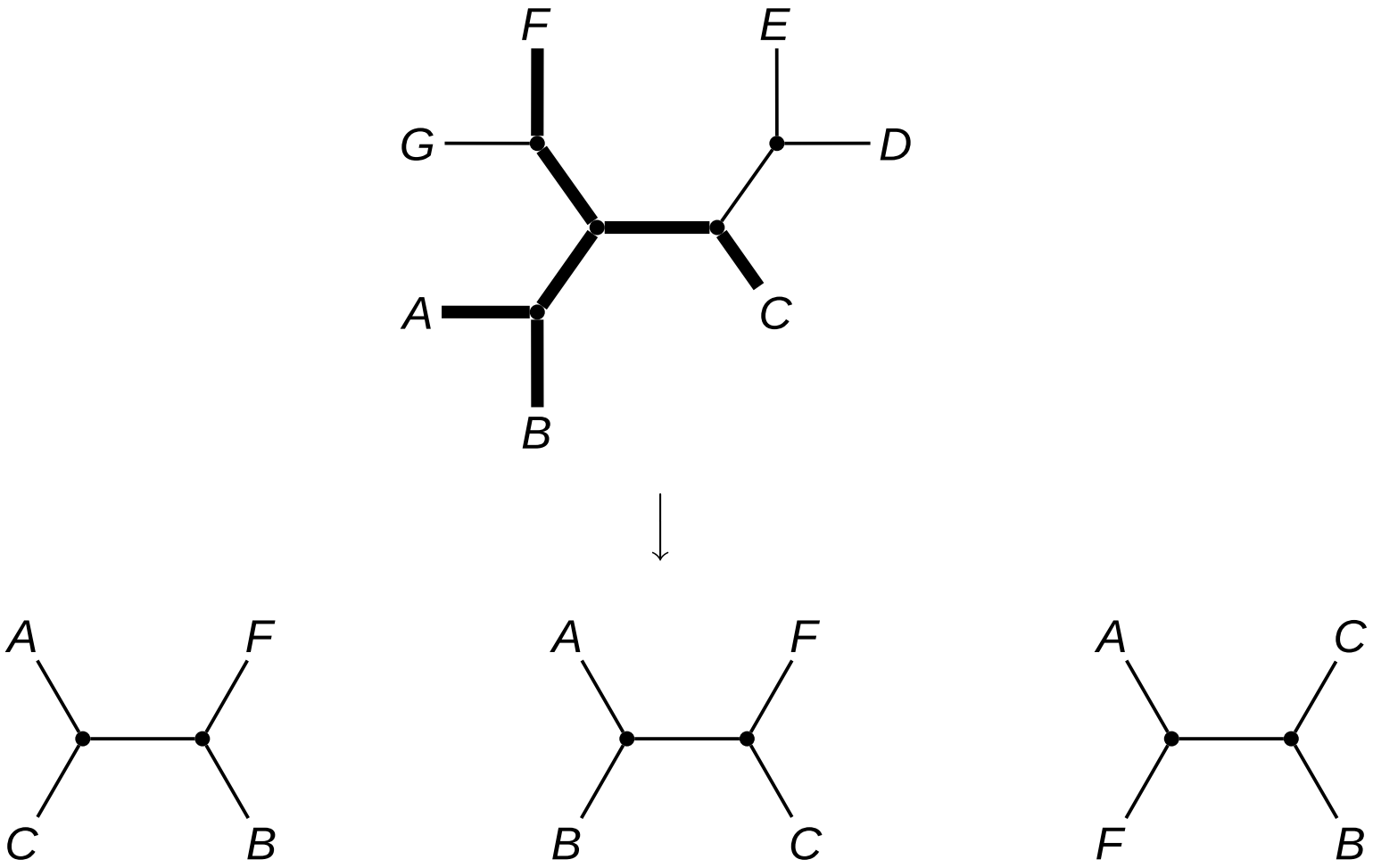
B



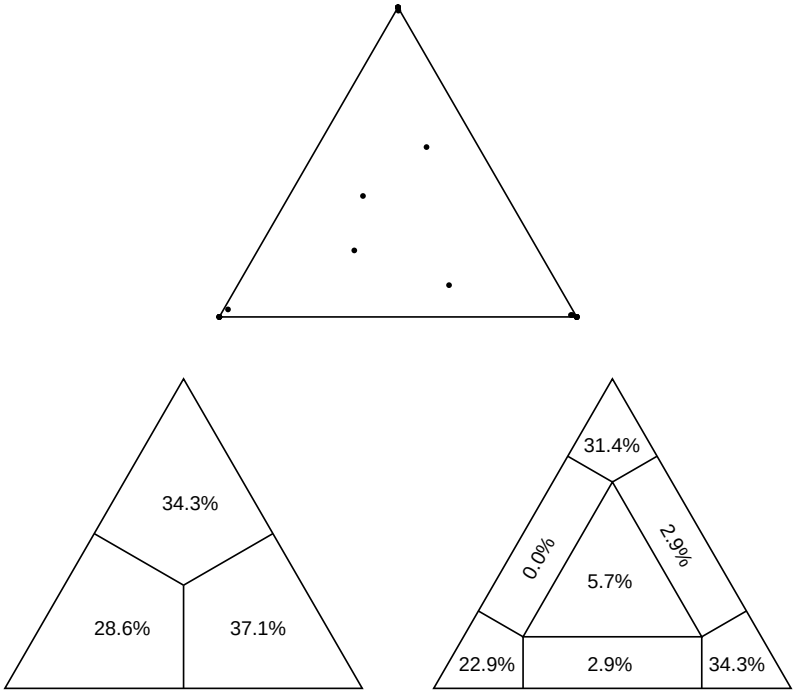
Statistical Geometry—2



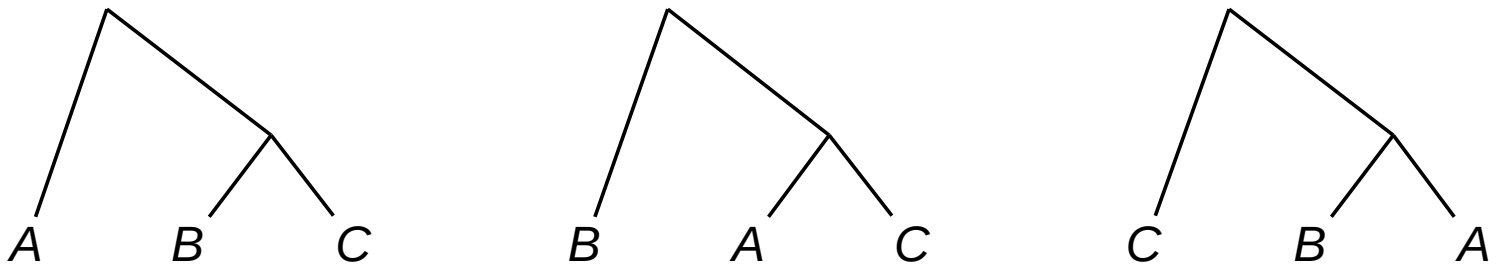
Quartet Analysis



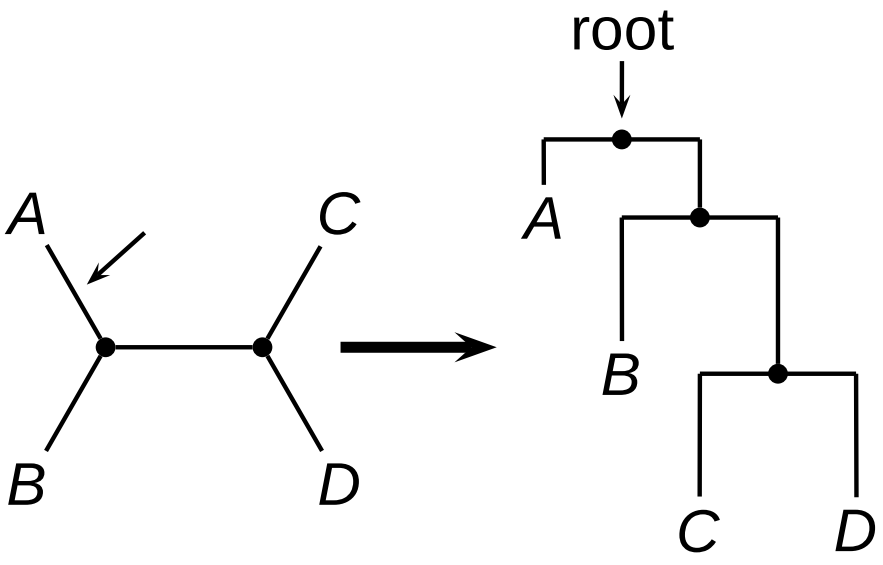
Likelihood Map



Rooted Phylogenies



Rooting



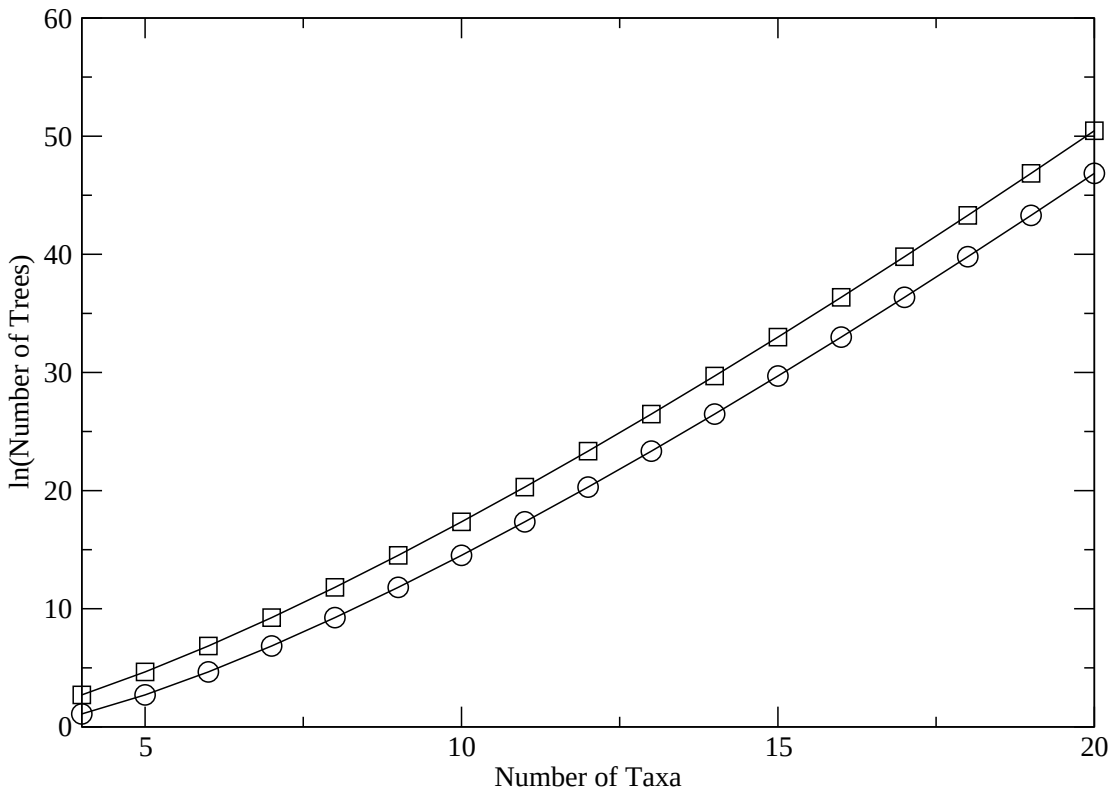
Number of Binary Phylogenies

$$N_u = 1 \times 3 \times 5 \times \ldots \times (2n-5) = \prod_{i=3}^{n}(2i-5)$$

$$N_r = 1 \times 3 \times 5 \times \ldots \times (2n-3) = \prod_{i=2}^{n}(2i-3).$$

$$N_r = (2n-3)N_u.$$

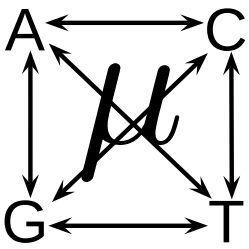
Number of Trees



Data

```
                                                  ↓
Human        GTAAATATAGTTTAACCAAAACATCAGATTGTGAATCTGACAACA   45
Chimpanzee   .....t........c........c.............g.c..c.   45
Gorilla      .....t........c........c.............g.t..c.   45
Orangutan    .....t........c........t.............a.t..t.   45
Gibbon       .....c........t........t.............a.c..t.   45

Human        GAGGCTTACGACCCCTTATTTACCG   70
Chimpanzee   .a.g.tcacg...c...at......   70
Gorilla      .a.g.tcaca...c...at......   70
Orangutan    .g.c.ccaca...c...at......   70
Gibbon       .a.g.tcgaa...t...gc......   70
```

Model

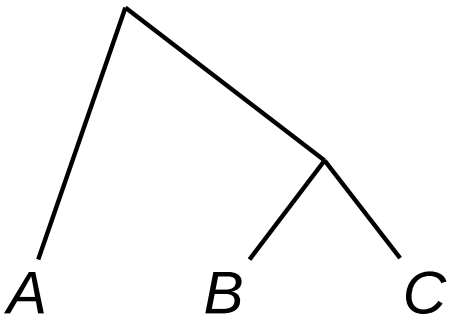


$$K(t) = -\frac{3}{4}\log(1 - \frac{4}{3}P_{\text{diff}}(t)),$$

Application Jukes-Cantor

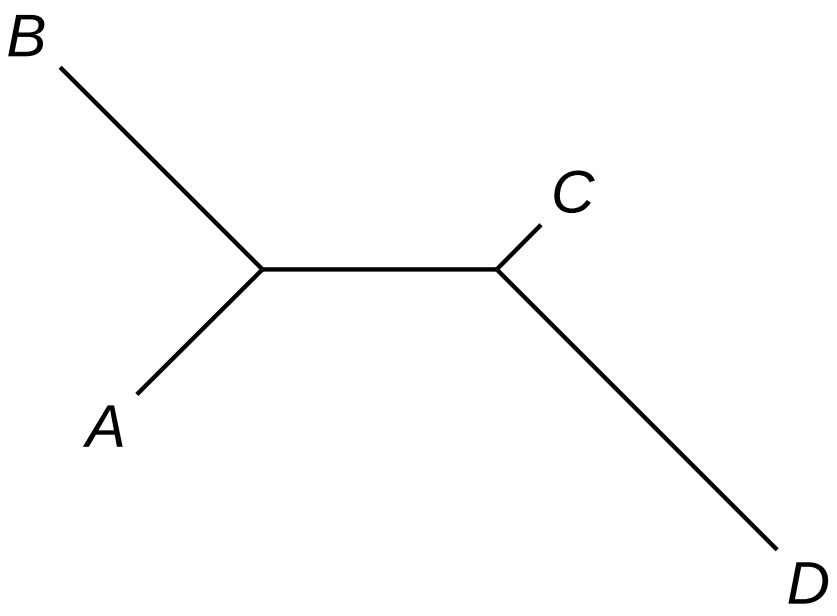
	Human	Chimpanzee	Gorilla	Orangutan	Gibbon
Human	-	0.014	0.044	0.141	0.195
Chimpanzee	0.014	-	0.029	0.124	0.176
Gorilla	0.043	0.029	-	0.091	0.176
Orangutan	0.129	0.114	0.086	-	0.176
Gibbon	0.171	0.157	0.157	0.157	-

Ultrametric Distances



$$d_{AB} = d_{AC} \geq d_{BC}$$

Additive Distances



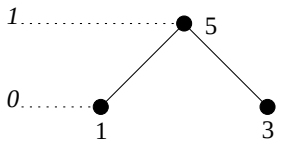
$$d_{AD} + d_{BC} = d_{BD} + d_{AC} \geq d_{AB} + d_{CD}$$

UPGMA — Algorithm

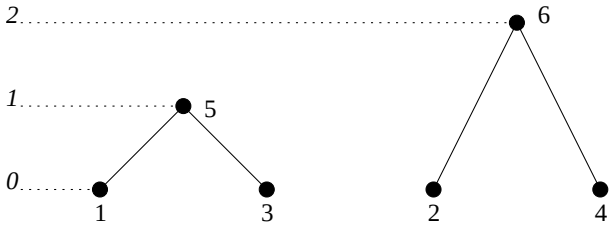
1. Input is a matrix of pairwise distances $D = (d_{ij})$ between a set of taxa $\mathcal{T} = \{1, 2, \ldots, n\}$
2. Find pair of taxa $\{i, j\}$ with smallest distance
3. Cluster taxa: $c \leftarrow \{i, j\}$
   Remove pair $\{i, j\}$ from set of taxa: $\mathcal{T} \leftarrow \mathcal{T} - \{i, j\}$
   Add new cluster to set of taxa: $\mathcal{T} \leftarrow \mathcal{T} \cup \{c\}$
   if $|\mathcal{T}| = 2$, stop
4. Compute the distances between c and all other taxa:
   $d_{ck} = \frac{d_{ik} + d_{jk}}{2}$, $k \in \mathcal{T}$
5. goto 2

UPGMA — Example

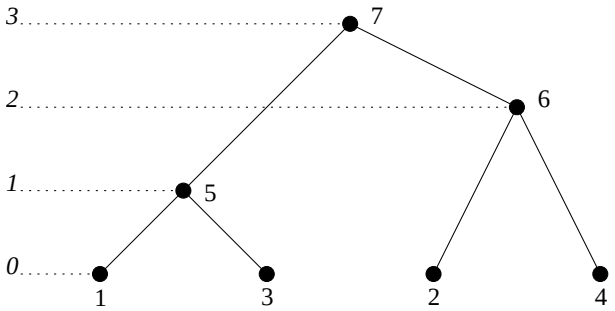
	1	2	3	4
1	-			
2	6	-		
3	2	6	-	
4	6	4	6	-



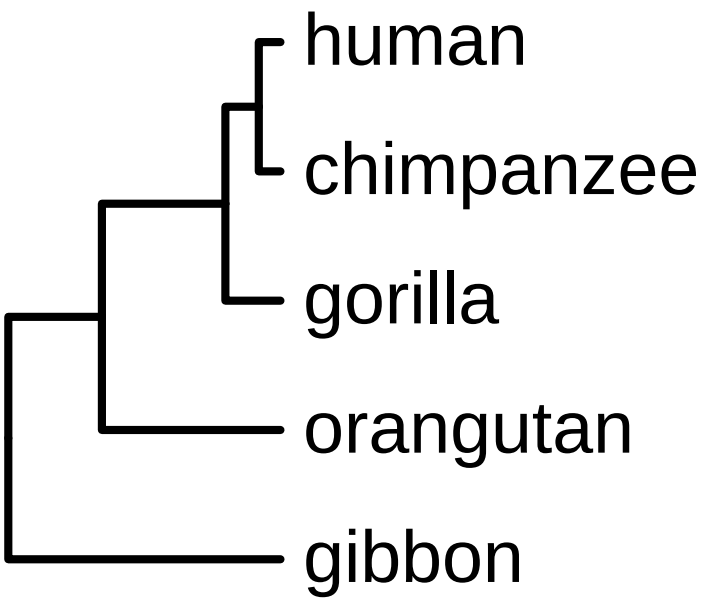
	5	2	4
5	-		
2	6	-	
4	6	4	-



	5	6
5	-	
6	6	-



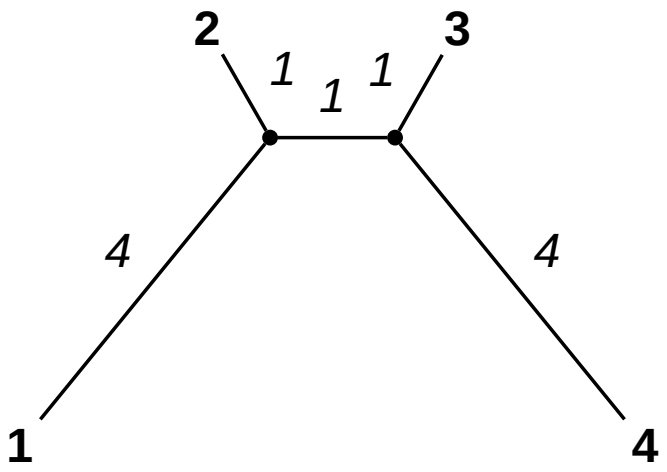
UPGMA — Application



Problem with UPGMA

Rate variation leads to erroneous phylogeny reconstruction.

A



B

	1	2	3	4
1	0	5	6	9
2		0	3	6
3			0	5
4				0

Neighbor Joining — Example

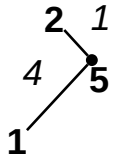
Step	Distance Matrix	Branch lengths	Tree

1

	1	2	3	4	r_i
1	0	5	6	9	20
2	-12	0	3	6	14
3	-11	-11	0	5	14
4	-11	-11	-12	0	20

$$d_{15} = \frac{1}{4}(2 \times 5 + 20 - 14) = 4$$

$$d_{25} = \frac{1}{4}(2 \times 5 - 20 + 14) = 1$$

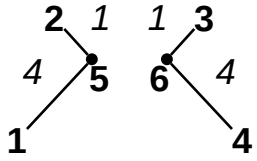


2

	3	4	5	r_i
3	0	5	2	7
4	-12	0	5	10
5	-12	-12	0	7

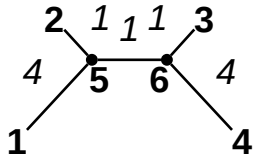
$$d_{36} = \frac{1}{2}(5 + 7 - 10) = 1$$

$$d_{46} = \frac{1}{2}(5 - 7 + 10) = 4$$

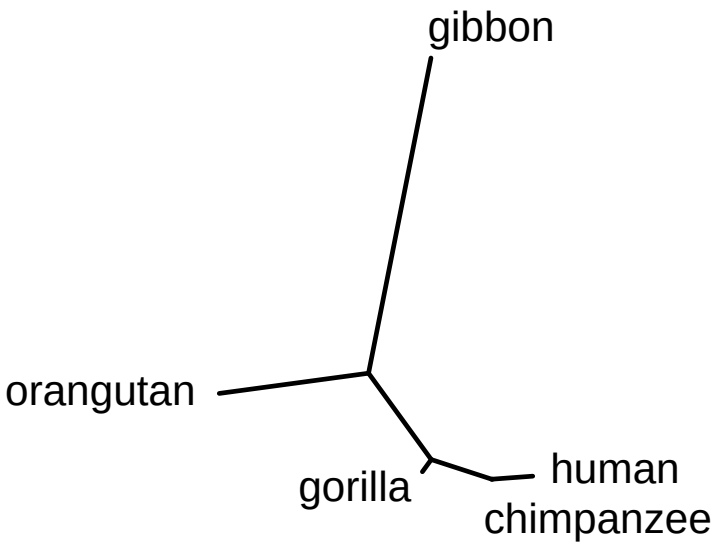


3

	5	6
5	0	1
6		0



Neighbor Joining — Application

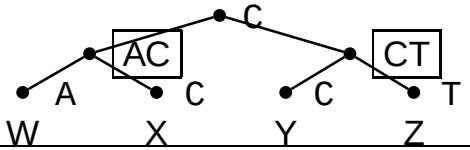
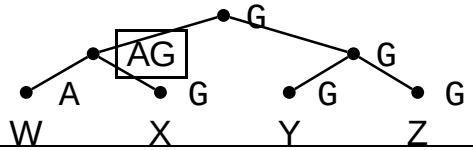
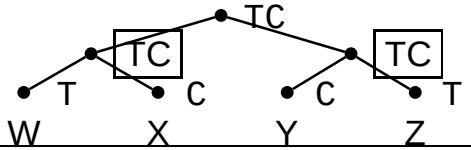
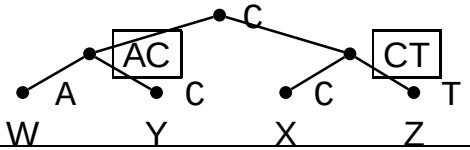
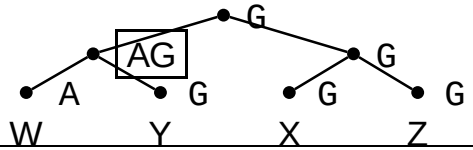
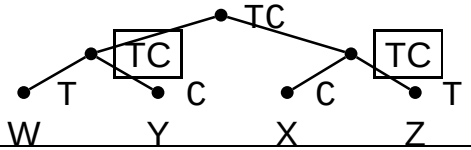
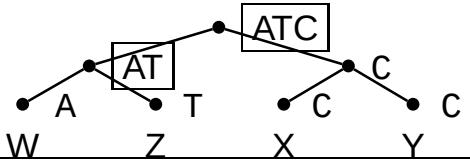
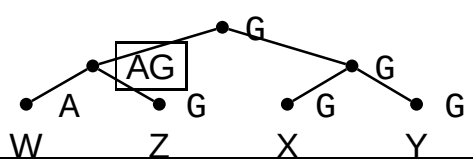
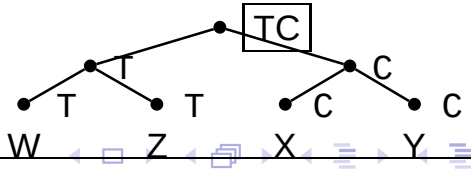


Maximum Parsimony — Algorithm

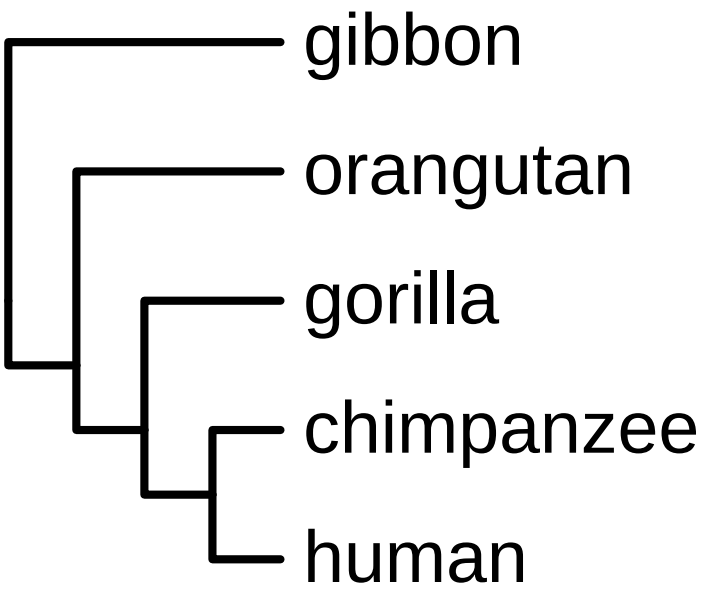
1. Label the leaves with the character of the corresponding taxon at P_j
2. Carry out a depth-first traversal of the tree and label each internal node by forming the intersection of the labels of its two child nodes. If this intersection is empty, the parent node is labelled by the union of the child labels, otherwise it is labelled by the intersection. In case of an empty intersection, $L(T)$ is increased by one.

Maximum Parsimony — Example

	P_1	P_2	P_3
W	A	A	T
X	C	G	C
Y	C	G	C
Z	T	G	T

	P_1	P_2	P_3	$L(T)$
T_1				5
T_2				5
T_3				4

Maximum Parsimony — Application



$$L(T) = 17$$

Parsimony Length, $L(T)$

Parsimony length can vary between

- Minimum: Number of variable positions in alignment, S
- Maximum: Half the number of branches times S

$$S \leq L(T) \leq S \times (n-1)$$

NB: $S \times (n-1)$ is usually much smaller than the number of possible trees $\Rightarrow$ cooptimal trees

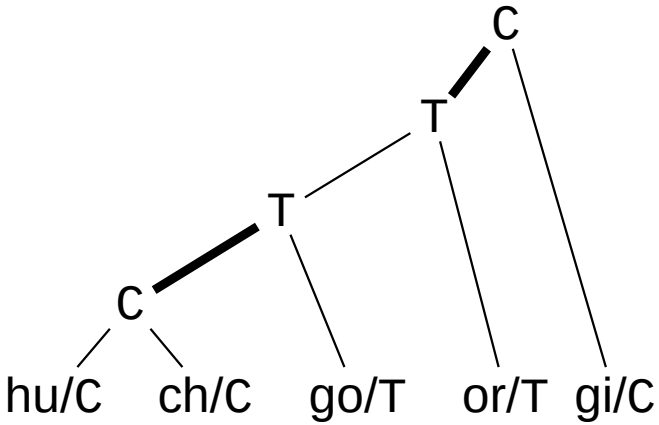
Data Revisited

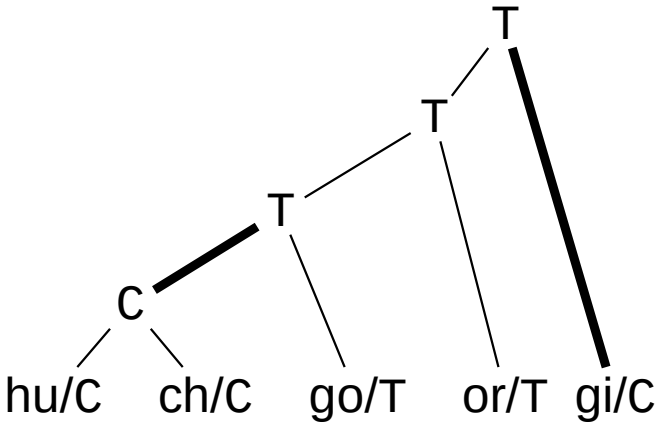
```
                                                  ↓
Human        GTAAATATAGTTTAACCAAAACATCAGATTGTGAATCTGACAACA   45
Chimpanzee   .....t........c........c.............g.c..c.   45
Gorilla      .....t........c........c.............g.t..c.   45
Orangutan    .....t........c........t.............a.t..t.   45
Gibbon       .....c........t........t.............a.c..t.   45

Human        GAGGCTTACGACCCCTTATTTACCG   70
Chimpanzee   .a.g.tcacg...c...at......   70
Gorilla      .a.g.tcaca...c...at......   70
Orangutan    .g.c.ccaca...c...at......   70
Gibbon       .a.g.tcgaa...t...gc......   70
```

Homoplasy — 1

- $S = 16$
- $L(T) = 17 \Rightarrow$ alignment contains position arisen by 2 mutations
- Homoplasy: $L(T) - S$

Homoplasy — 2





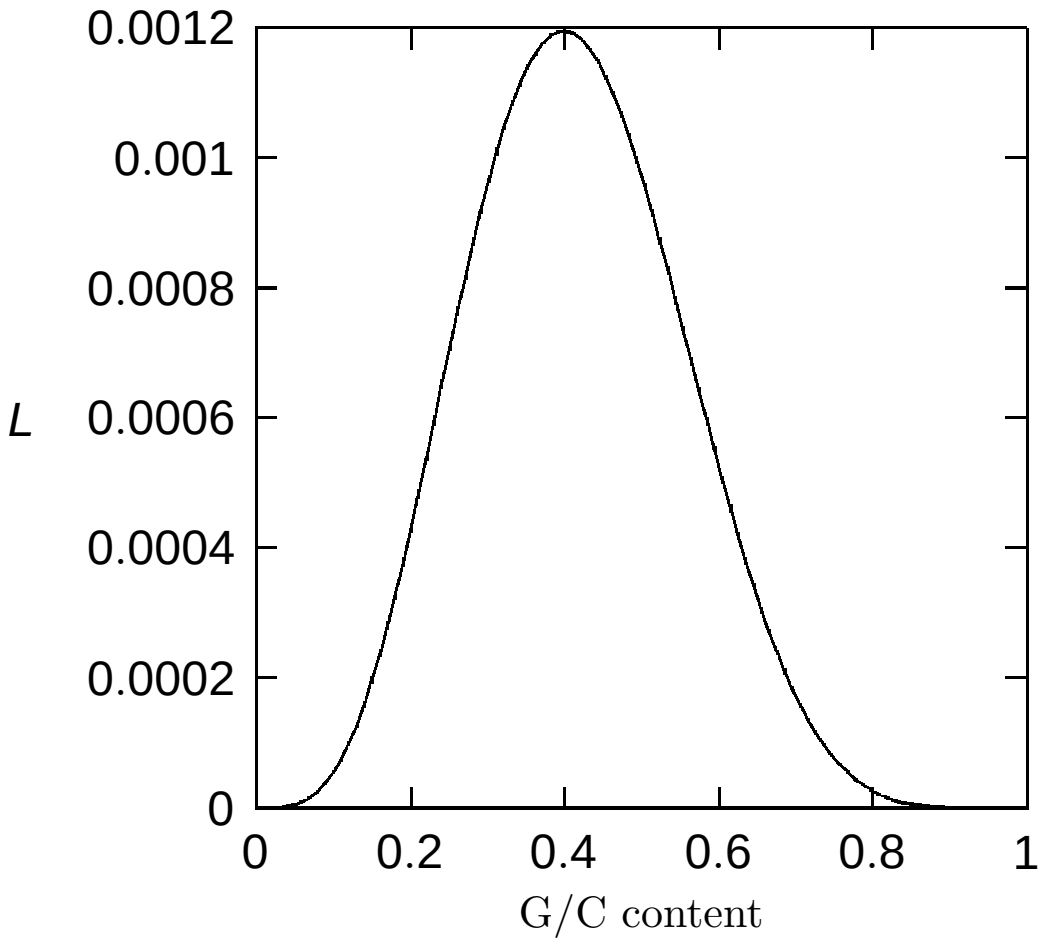
Maximum Likelihood — 1

1. Consider DNA sequence TACTTCCTGT
2. Infer G/C content p

Sequence contains 4 G/C & 6 A/T

$$L = P(D|p) = p^4(1-p)^6$$

Maximum Likelihood — 2



Maximum Likelihood — 3

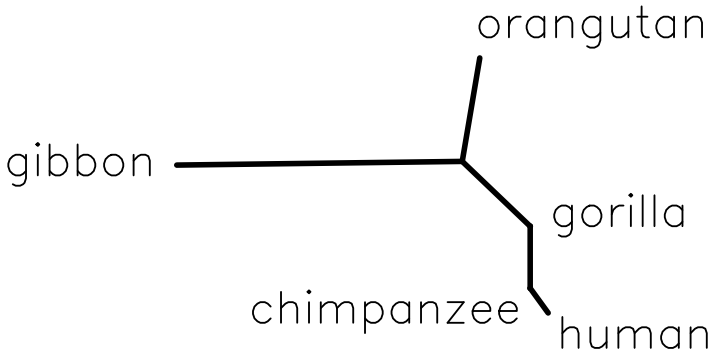
Find maximum of likelihood function:

$$\begin{aligned} L &= p^4(1-p)^6 \\ \ln(L) &= 4\ln p + 6\ln(1-p) \\ \frac{d(\ln L)}{dp} &= \frac{4}{p} - \frac{6}{1-p} = 0 \end{aligned}$$

$\Rightarrow \hat{p} = 2/5$.

Fraction of G/C residues is ML estimator of p.

Maximum Likelihood Phylogeny



- ((human, chimp),gorilla) $\ln(L) = -166.52$
- ((human,gorilla),chimp) $\ln(L) = -169.23$

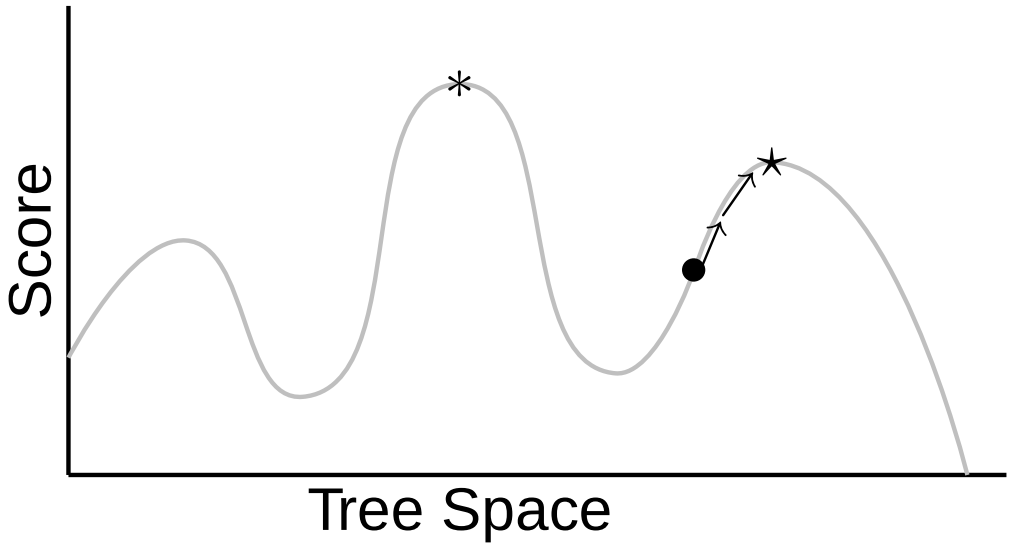
Likelihood-Ratio Test

$$R = 2 \ln \left(\frac{L_1}{L_2} \right) = 2(\ln L_1 - \ln L_2).$$

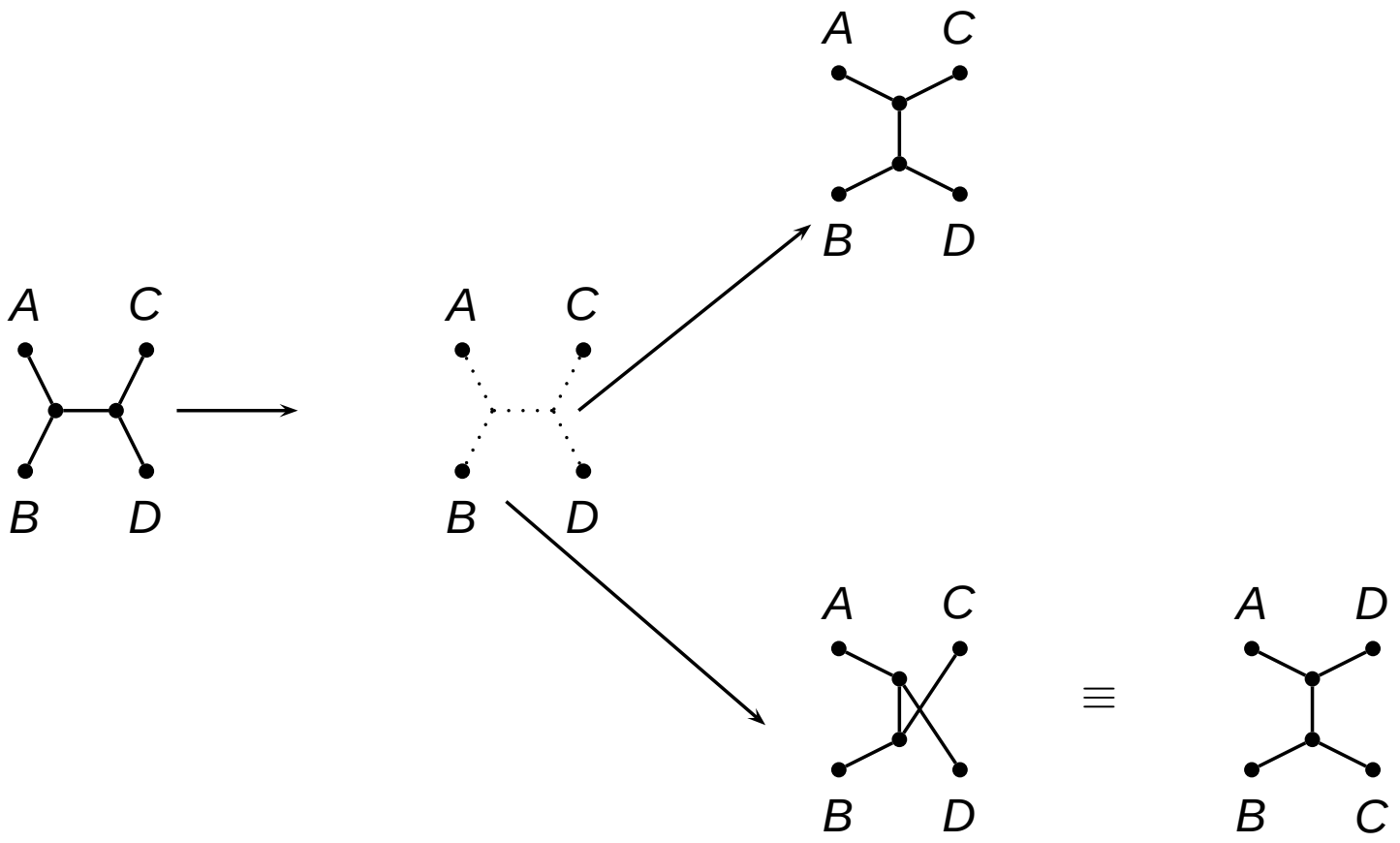
R approximately follows a χ^2-distribution. In example:

- $R = 2(169.23 - 166.52) = 5.42$
- phylogeny has $2n - 3 = 7$ branches whose lengths were estimated with a one-parameter substitution model $\Rightarrow$ 7 degrees of freedom
- $X_{0.05,7} = 14.07 \Rightarrow$ cannot distinguish between two trees.

Searching Through Tree Space

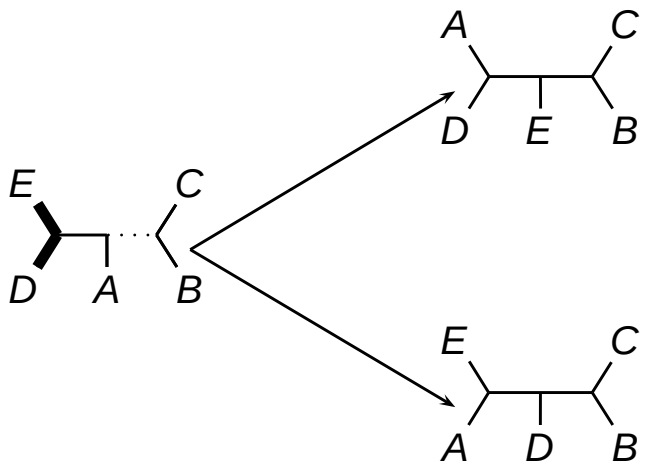


Nearest Neighbor Interchange

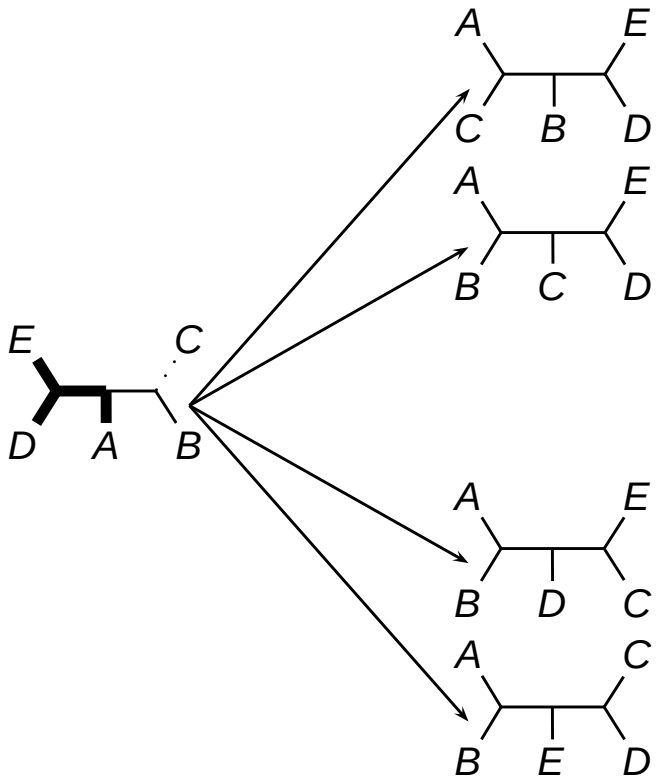


Subtree Pruning & Regrafting

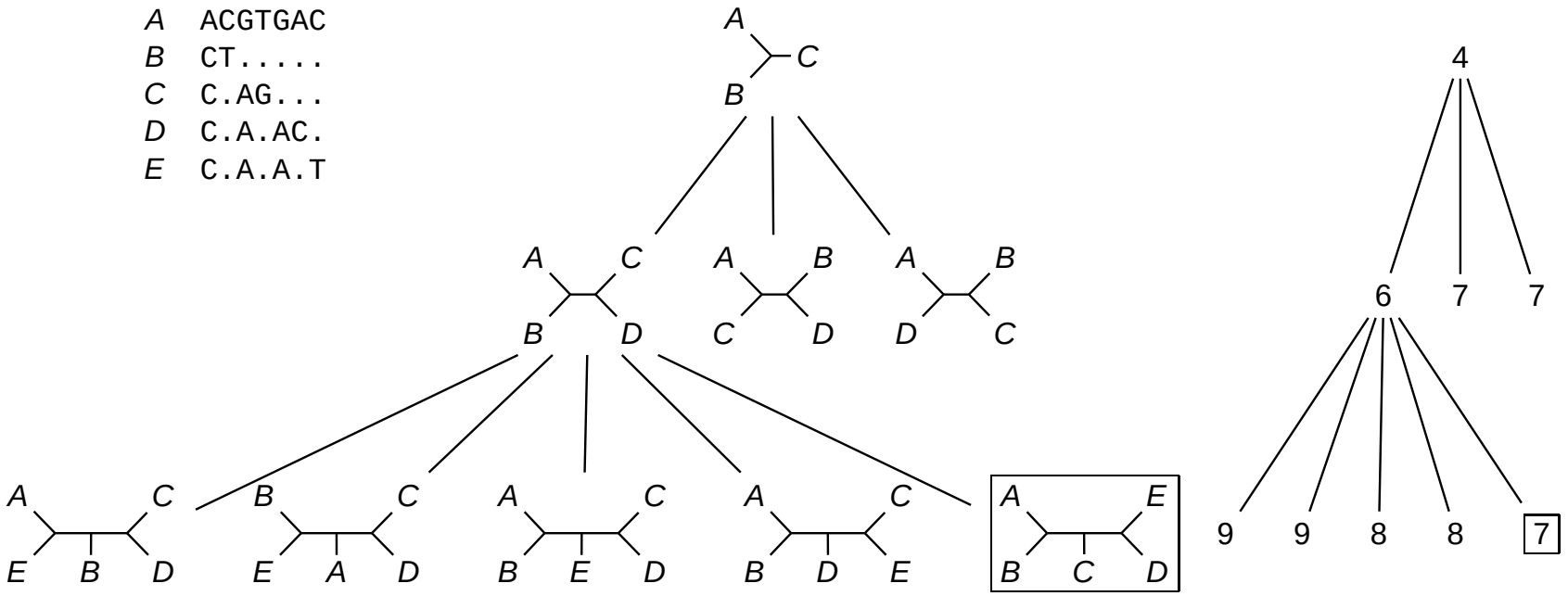
internal



external



Branch & Bound



Bootstrapping Phylogenies

Original sample

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

Bootstrap samples

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

Bootstrapped Phylogeny