

Tree Reconstruction

Basic Principles of Phylogenetics

Distance

Parsimony

Compatibility

Inconsistency

Likelihood

Central Principles of Phylogeny Reconstruction

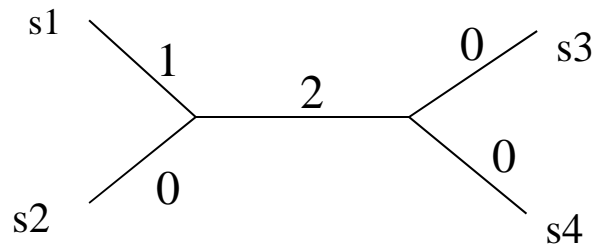
TTCAGT

TCCAGT

GCCAAT

GCCAAT

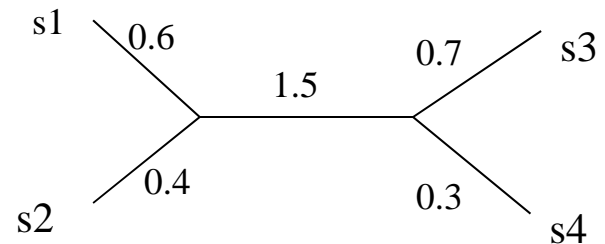
Parsimony



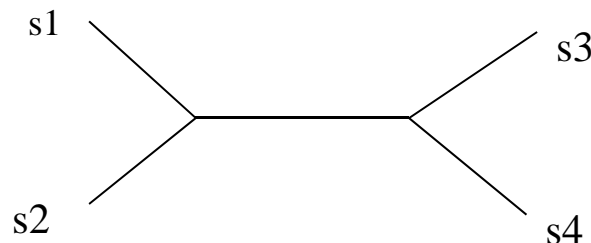
Total Weight: 3

Distance

1
3 2
3 2 0



Likelihood



$L=3.1 \times 10^{-7}$

Parameter estimates

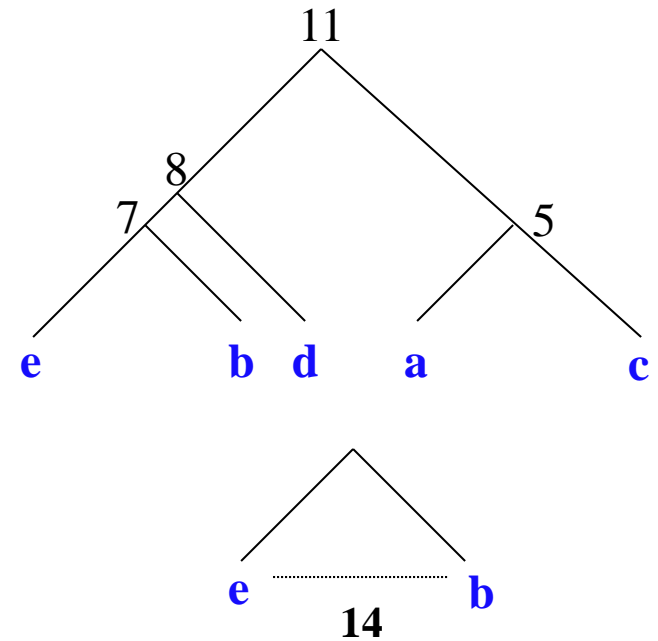
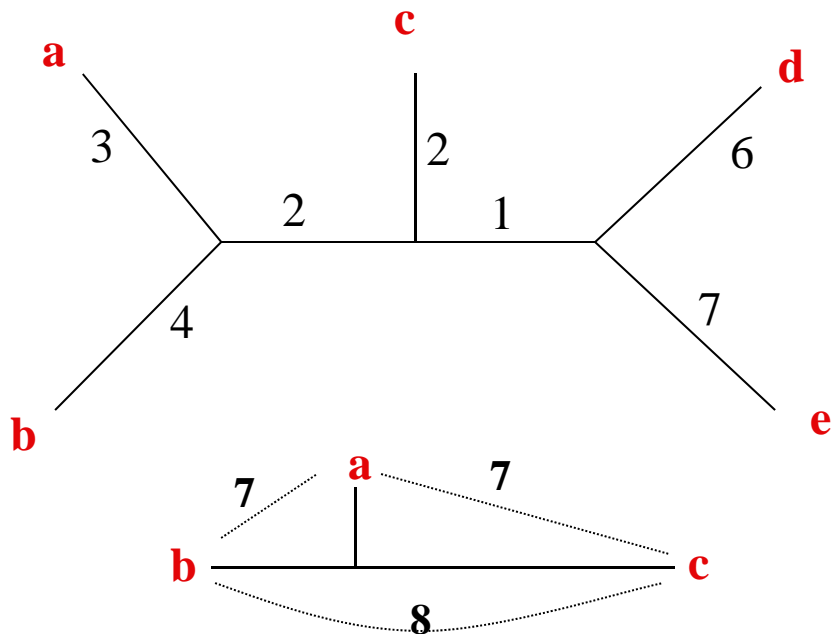
From Distance to Phylogenies

What is the relationship of a, b, c, d & e?

No Molecular clock

Molecular clock

	a	b	c	d	e
a	-	22	10	22	22
b	7	-	22	16	14
c	7	8	-	22	22
d	12	13	9	-	16
e	13	14	10	13	-



UGPMA

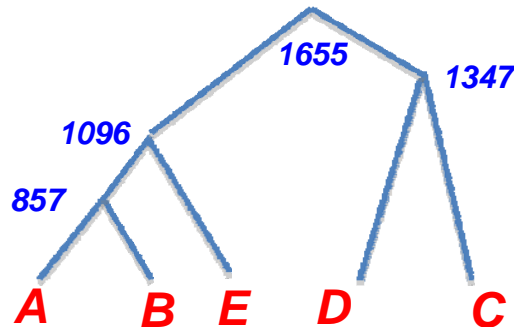
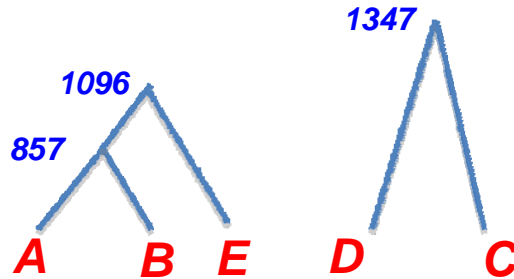
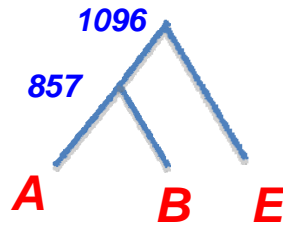
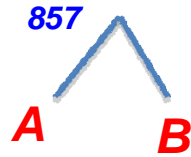
Unweighted Group Pairs Method using Arithmetic Averages

A	B	C	D	E
A	1715	2147	3091	2326
B		2991	3399	2058
C			2795	3943
D				4289
E				

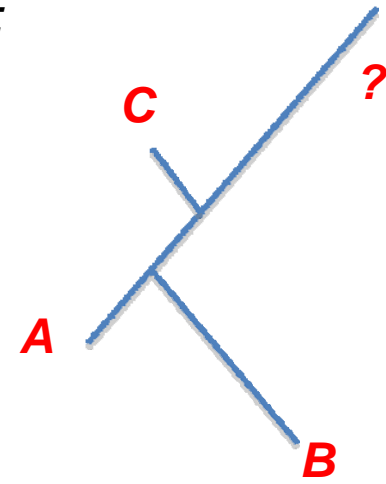
AB	C	D	E
AB	2529	3245	2192
C		2795	3943
D			4289
E			

ABE	C	D
ABE	3027	3593
C		2795
D		

ABE	CD
ABE	3310
CD	



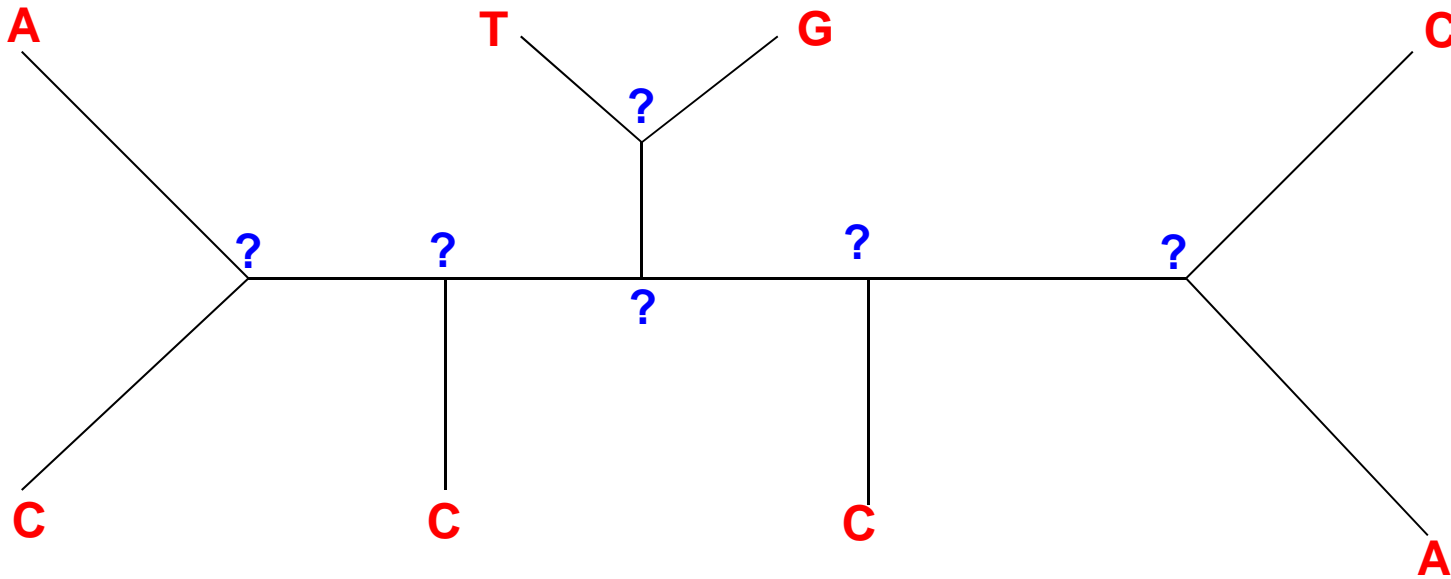
UGPMA can fail:



A and B are siblings, but A and C are closest

Siblings will have
 $[d(A, ?) + d(B, ?) - d(A, B)] / 2$
maximal.

Assignment to internal nodes: The simple way.



What is the cheapest assignment of nucleotides to internal nodes, given some (symmetric) distance function $d(N_1, N_2)$??

If there are k leaves, there are $k-2$ internal nodes and 4^{k-2} possible assignments of nucleotides. For $k=22$, this is more than 10^{12} .

Cost of a history – leaves (initialisation).

A C G T

Initialisation: leaves

Cost(N) = 0 if

N is at leaf,

otherwise infinity

G

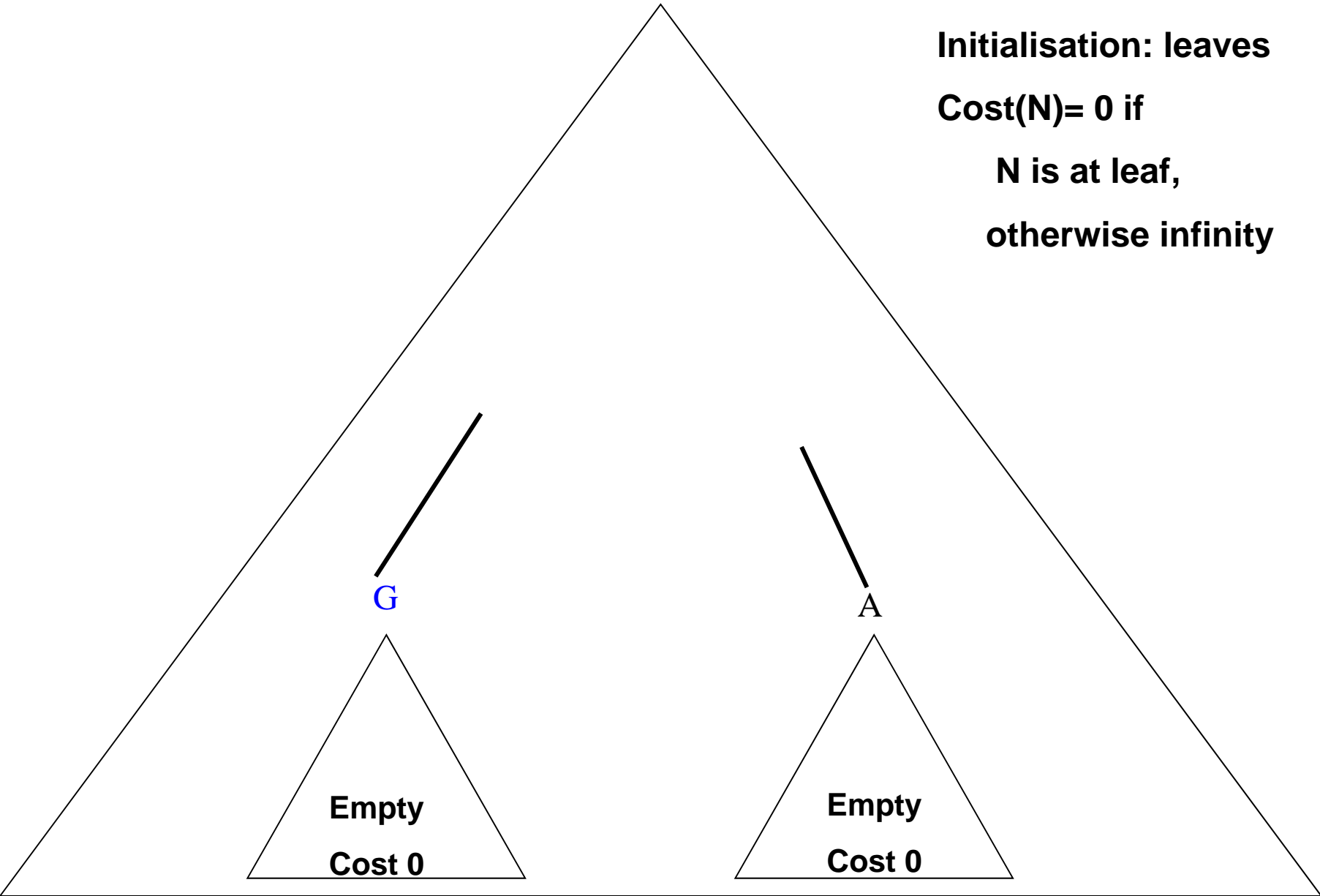
Empty

Cost 0

A

Empty

Cost 0



Compatibility and Branch Popping

A
 GCACGTGCAG**G**TTAGGA
B
 GCACGTGCAG**G**TTAGGA
C
 TCTCGTGCAG**G**TTAGGA
D
 TCTCATGCAATTAGGA
E TCTCATGCAATTATGA
A TCTCATGCAATTATGA
 GCACGTGCAG**G**TTAGGA
B
 GCACGTGCAG**G**TTAGGA
C
 TCTCGTGCAG**G**TTAGGA
D
 TCTCATGCAATTAGGA
E TCTCATGCAATTATGA
F TCTCATGCAATTATGA
A
 GCACGTGCAG**G**TTAGGA
B
 GCACGTGCAG**G**TTAGGA
C
 TCTCGTGCAG**G**TTAGGA
D
 TCTCATGCAATTAGGA
E TCTCATGCAATTATGA
F TCTCATGCAATTATGA

ABC



EFG

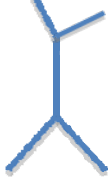
ABC



E

EF

C



AB

D

EF

Definition: Two columns can be placed on the same tree – each explained by 1 mutation.

This is equivalent to: In the two columns only 3 or the 4 possible character pairs are observed

Multistate Definition: The number of mutations needed to explain a pair of columns is the sum of the mutations needed to explain the individual columns

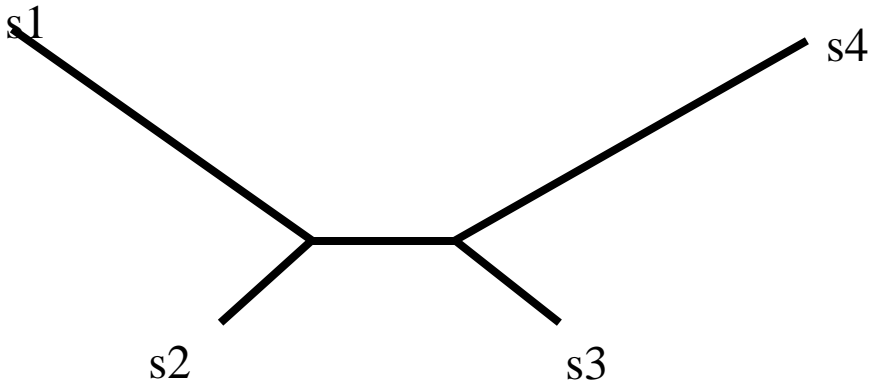
For imperfect data: Find the maximal compatible set of characters and then branch-pop

	1	2	3	4	5	6
1	+	?	?	?	?	?
2		+	?	?	?	?
3			+	?	?	?
4				+	?	?
5					+	?
6						+

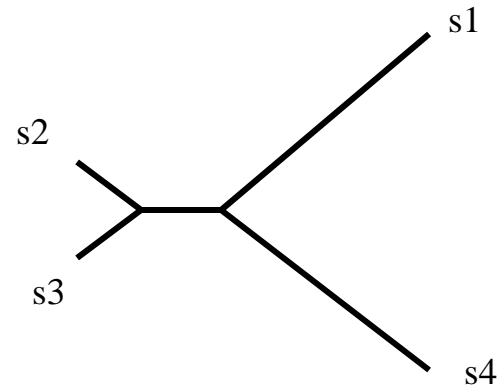
The Felsenstein Zone

Felsenstein-Cavendar (1979)

True Tree



Reconstructed Tree



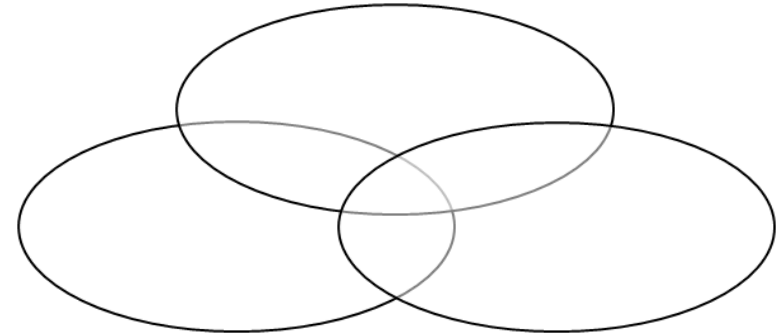
Patterns:(16 only 8 shown)

0	1	0	0	0	0	0	0
0	0	1	0	0	1	0	1
0	0	0	1	0	1	1	0
0	0	0	0	1	0	1	1

Hadamard Conjugation & binary characters on a tree

Closely related to inclusion-exclusion principle and Sieve Methods

$$H_1 = \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix} \quad H_k = \begin{pmatrix} H_{k-1} & H_{k-1} \\ H_{k-1} & -H_{k-1} \end{pmatrix}$$

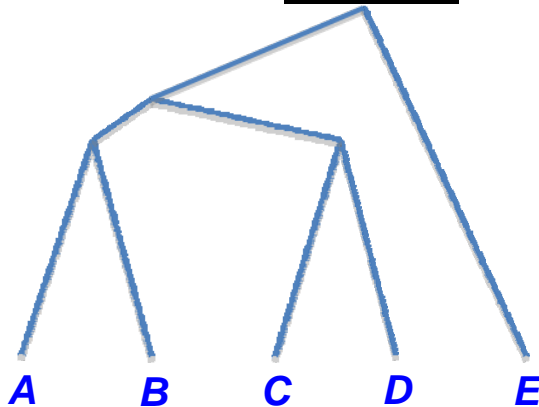


Branch lengths – s , Bipartition lengths - q

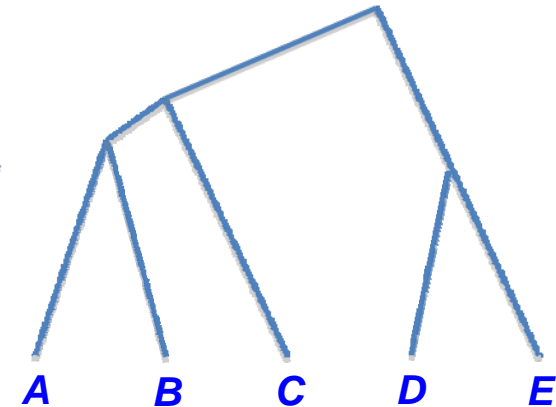
From branch lengths to bipartitions $q=Hs$

From bipartition to lengths $s=H^{-1} q$

Inconsistency in presence of a Clock:



True Tree with Clock

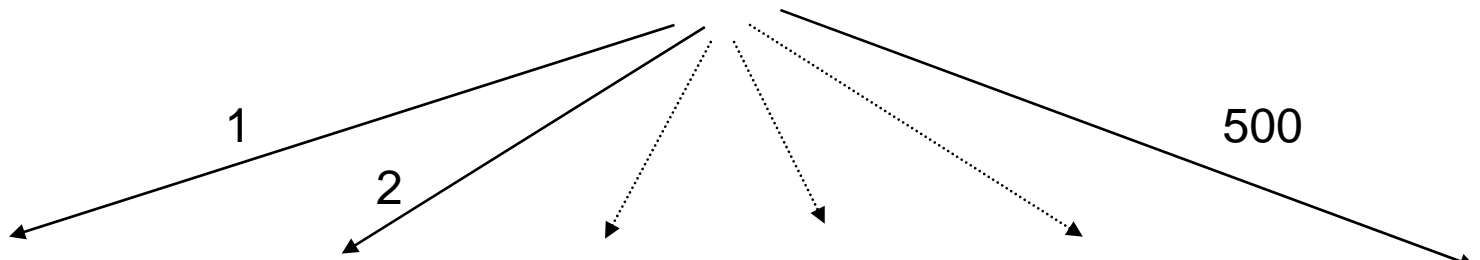


More Likely Tree

Bootstrapping

Felsenstein (1985)

ATCTGTAGTCT
ATCTGTAGTCT
ATCTGTAGTCT
ATCTGTAGTCT
10230101201



ATCTGTAGTCT
ATCTGTAGTCT
ATCTGTAGTCT
ATCTGTAGTCT

??????????
??????????
??????????
??????????

??????????
??????????
??????????
??????????

