

# *Tree Reconstruction*

## *Basic Principles of Phylogenetics*

***Distance***

***Parsimony***

***Compatibility***

***Inconsistency***

***Likelihood***

# Central Principles of Phylogeny Reconstruction

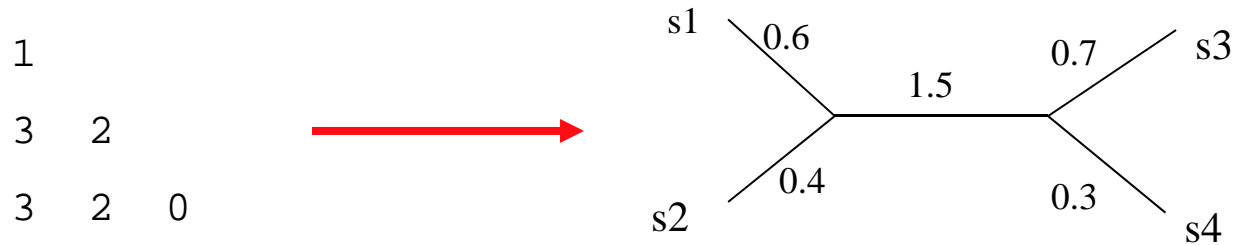
TTCAGT

TCCAGT

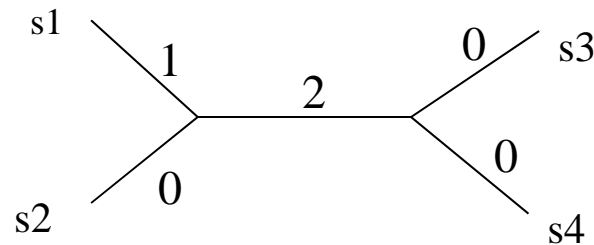
GCCAAT

GCCAAT

Distance

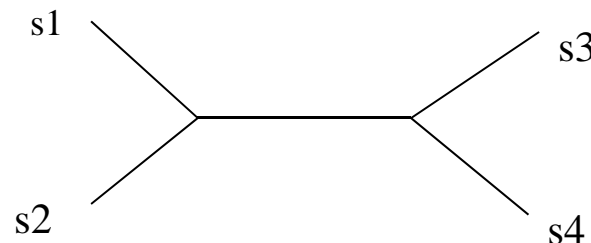


Parsimony



Total Weight: 3

Likelihood



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

Parameter estimates

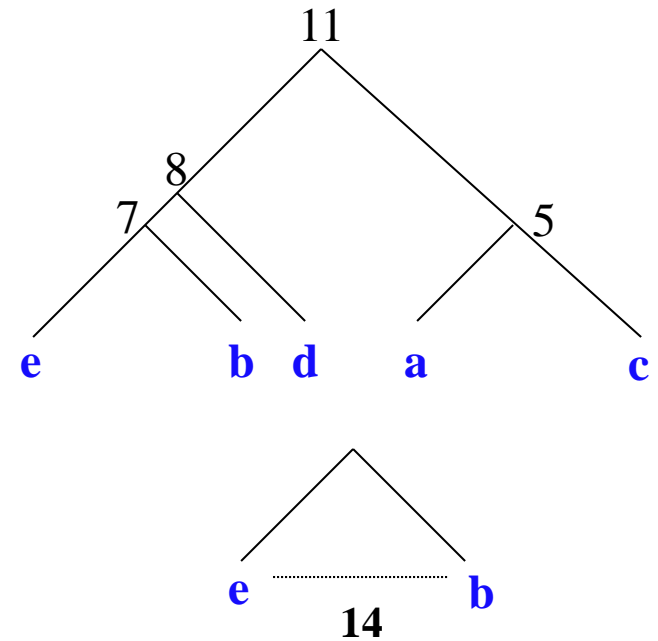
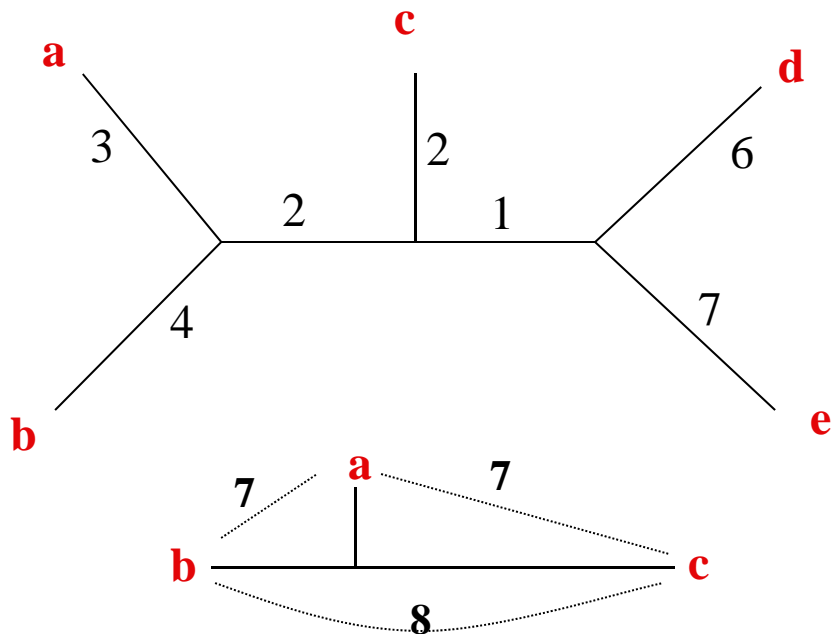
# From Distance to Phylogenies

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

**No Molecular clock**

Molecular clock

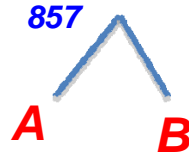
	a	b	c	d	e
a	-	22	10	22	22
b	7	-	22	16	14
c	7	8	-	22	22
d	12	13	9	-	16
e	13	14	10	13	-



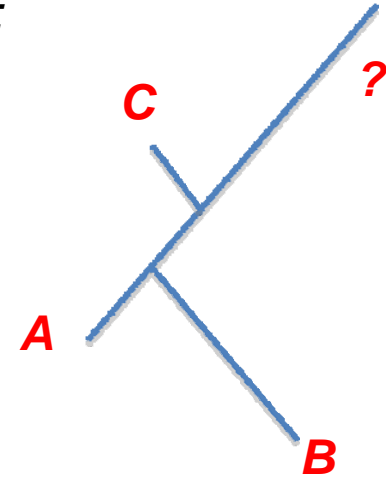
# UGPMA

*Unweighted Group Pairs Method using Arithmetic Averages*

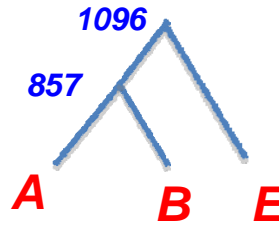
A	B	C	D	E
A	1715	2147	3091	2326
B		2991	3399	2058
C			2795	3943
D				4289
E				



*UGPMA can fail:*

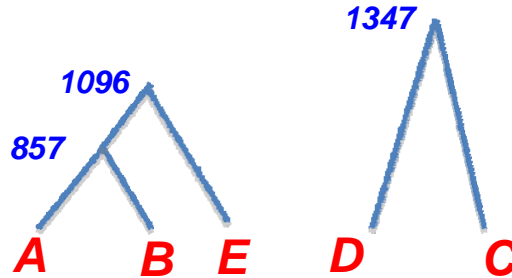


AB	C	D	E
AB	2529	3245	2192
C		2795	3943
D			4289
E			



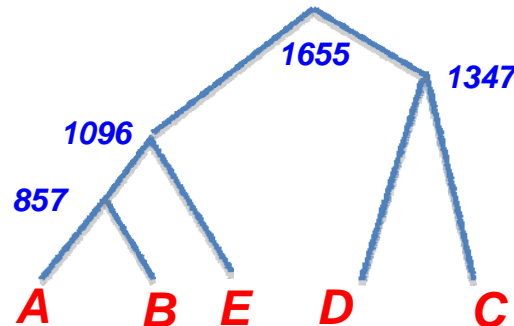
***A and B are siblings, but A and C are closest***

ABE	C	D
ABE	3027	3593
C		2795
D		



***Siblings will have***

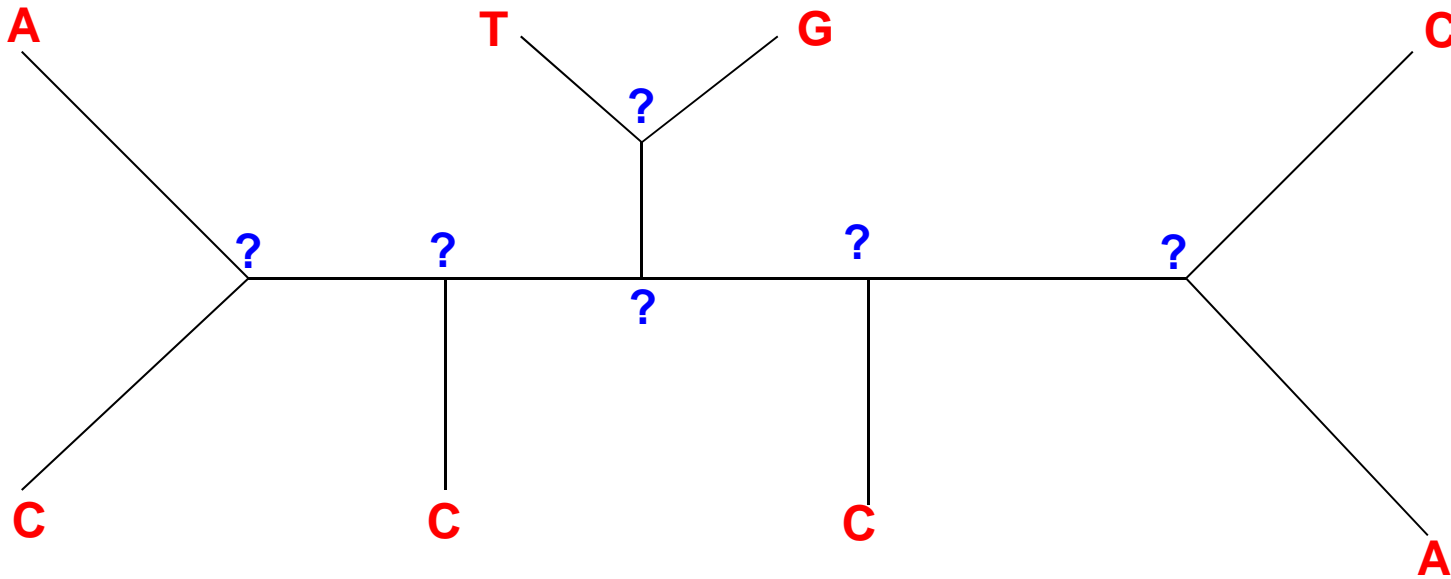
ABE	CD
ABE	3310
CD	



***[d(A, ?)+d(B, ?)-d(A, B)]/2***

***maximal.***

# Assignment to internal nodes: The simple way.

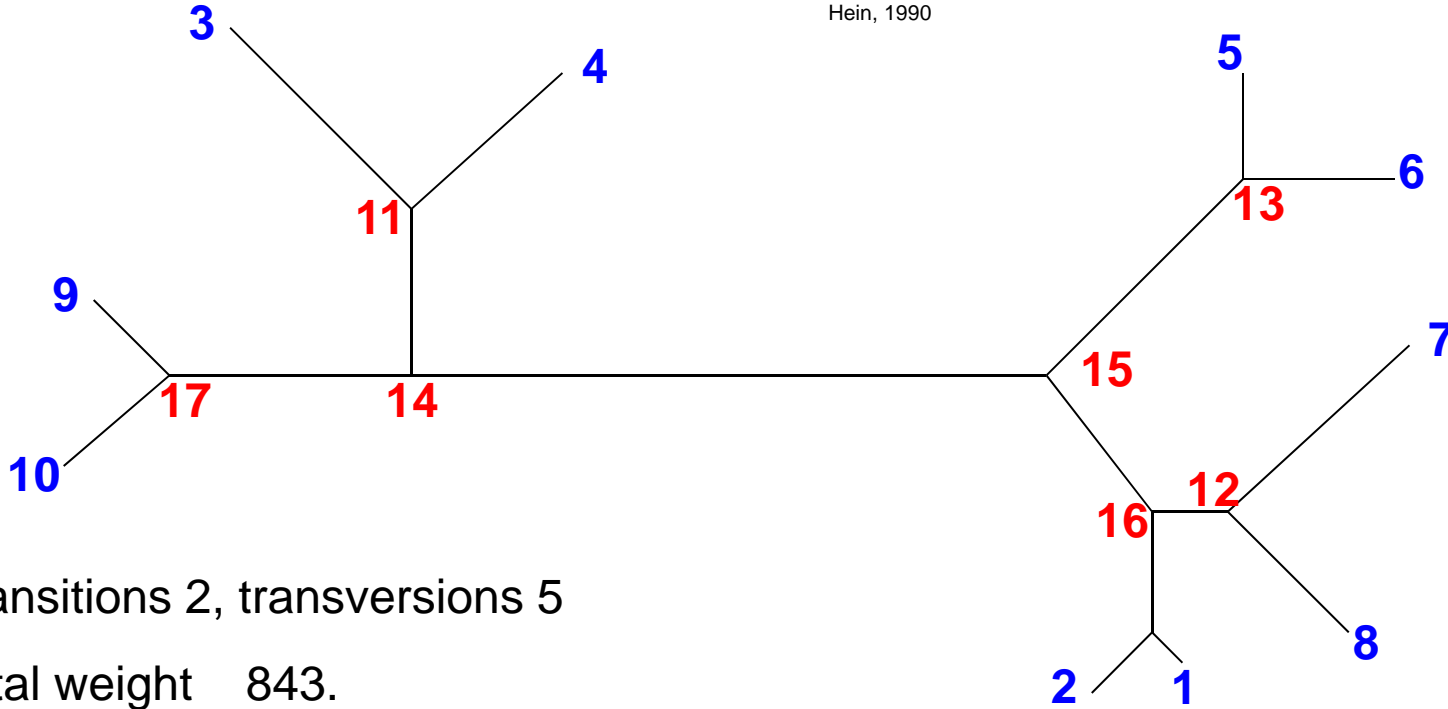


What is the cheapest assignment of nucleotides to internal nodes, given some (symmetric) distance function  $d(N_1, N_2)$ ??

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  $10^{12}$ .

# 5S RNA Alignment & Phylogeny

Hein, 1990



Transitions 2, transversions 5

Total weight 843.

```

10 tatt-ctggtgtcccaggcgtagaggaaccacaccgatccatctcgaacttgggtggtgaaactctgccgcggt--aaccaatact-cg-gg-gggggcct-gcggaaaaatagctcgatgccagga--ta
17 t--t-ctggtgtcccaggcgtagaggaaccacaccaatccatcccgaacttgggtggtgaaactctgctgcggt--ga-cgatact-tg-gg-gggagccc-atggaaaaatagctcgatgccagga--t-
9 t--t-ctggtgtctcaggcgtggaggaaccacaccaatccatcccgaacttgggtggtgaaactctattgcggt--ga-cgatactgta-gg-ggaagccc-atggaaaaatagctcgacgccagga--t-
14 t----ctggtggccatggcgtagaggaaacaccccaccccataccgaactcggcagttaaagctctgctgcgcc--ga-tggtact-tg-gg-gggagccc-ctgggaaaaataggacgctgccag-a--t-
3 t----ctggtgatgatggcggaggggacacaccggtcccataccgaacacggccttaagccctccagcgcc--aa-tggtact-tgctc-cgcagggag-cgggagagtaggacgtcgccag-g--c-
11 t----ctggtggcgtggcgaagaggacacaccggtcccataccgaacacggcagttaaagctctccagcgcc--ga-tggtact-tg-gg-ggcagtcg-ctgggagagtaggacgctgccag-g--c-
4 t----ctggtggcgtatagcagaaggtcacaccggtcccataccgaacacgggaagttaaagctctcagcgcc--ga-tggtagt-ta-gg-ggctgtccc-ctgtgagagtaggacgctgccag-g--c-
15 g----cctgcggccatagcacccgtgaaagcaccatccat--ccgaactcggcagttaaagcaggttgccgccaga--tagtact-tg-ggtgggagaccgctgggaaacctggatgctgcaag-c--t-
8 g----cctacggccatcccaccctggtaacgcccgatctcgt-ctgatctcggaagctaaagcaggtcgggcctggt--tagtact-tg-gatgggagacctcctgggaataccgggtgctgtagg-ct-t-
12 g----cctacggccataccaccctgaaagcaccatcccgt--ccgatctgggaagttaaagcaggttgagcccagt--tagtact-tg-gatgggagaccgctgggaaacctgggtgctgtagg-c--t-
7 g----cttacgaccatcacggtgaatgacgcccacccgt--ccgatctggcaagttaaagcaggttgagtcaggt--tagtact-tg-gatcggagacggcctgggaaacctggatggtgtaag-c--t-
16 g----cctacggccatagcacccctgaaagcaccatcccgt--ccgatctgggaagttaaagcaggttgccgccagt--tagtact-tg-ggtgggagaccgctgggaaacctgggtgctgtagg-c--t-
1 a----tccacggccataggactctgaaagcactgcatcccgt--ccgatctgcaagttaaaccagagtaccgcccagt--tagtacc-ac-ggtgggggaccacgcccgaatcctgggtgctgt-gg-t--t-
18 a----tccacggccataggactctgaaagcaccgcatcccgt--ccgatctgcaagttaaaccagagtaccgcccagt--tagtacc-ac-ggtgggggaccacatgggaaacctgggtgctgt-gg-t--t-
2 a----tccacggccataggactctgaaagcaccgcatcccgt--ctgatctgcgcagttaaacacagtgccgacctagt--tagtacc-at-ggtgggggaccacatgggaaacctgggtgctgt-gg-t--t-
5 g---tgggtgcggtcataccagcgctaatgacccggatccat--cagaactccgcagttaaagcgcgcttgggccagaa--cagtagt-gg-gatgggtgacctccgggaagtcctggtgcccacc-c--c-
13 g----ggtgcggtcataaccagcgttaatgacccggatccat--cagaactccgcagttaaagcgcgcttgggccagcc--tagtact-ag-gatgggtgacctcctgggaagtcctgatgctgcacc-c--t-
6 g---ggtgcgatcataccagcgttaatgacccggatccat--cagaactccgcagttaaagcgcgcttgggttgagg--tagtact-ag-gatgggtgacctcctgggaagtcctaatatgacacc-c--t-
    
```

# Cost of a history - minimizing over internal states

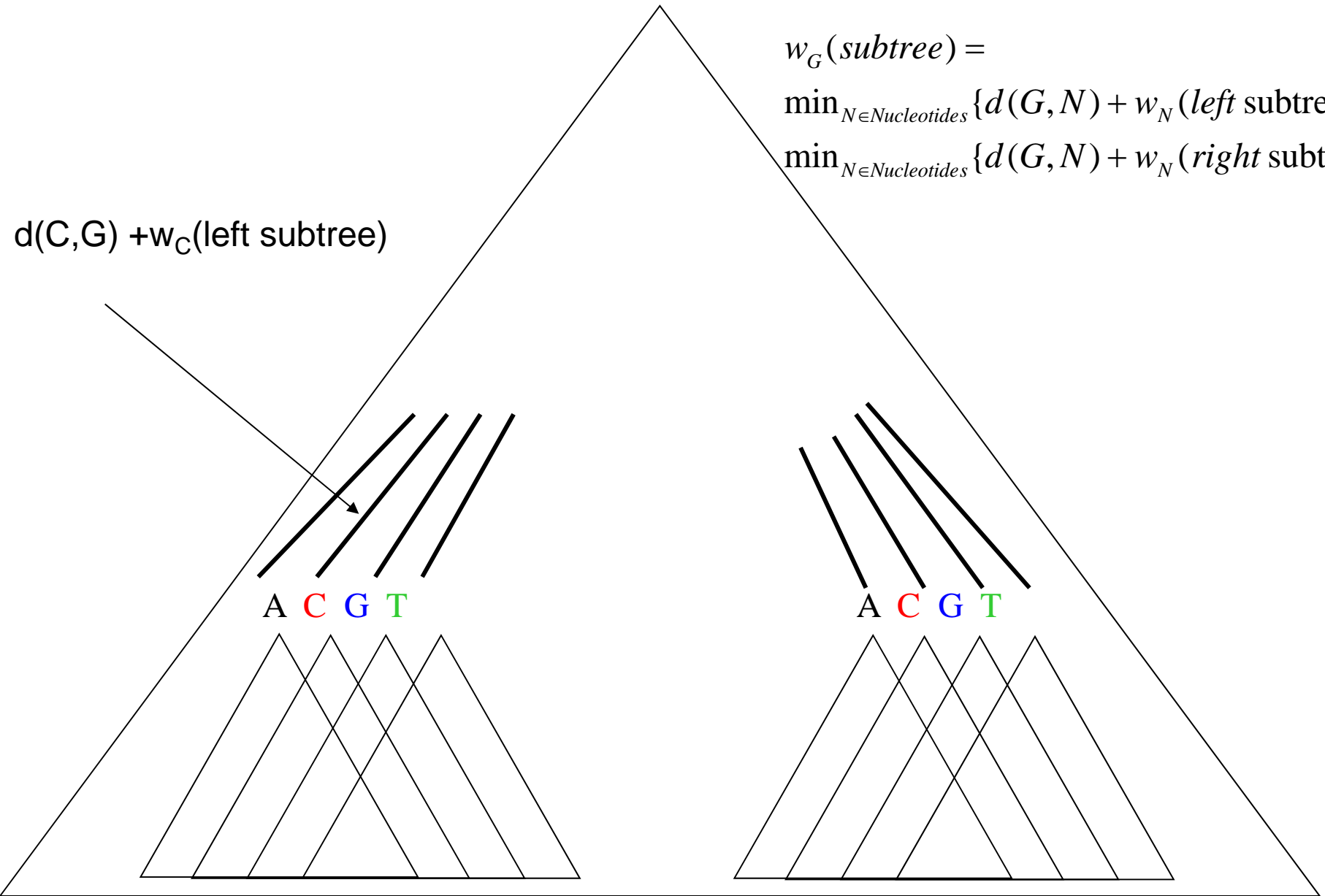
A C G T

$$w_G(\text{subtree}) =$$

$$\min_{N \in \text{Nucleotides}} \{d(G, N) + w_N(\text{left subtree})\} +$$

$$\min_{N \in \text{Nucleotides}} \{d(G, N) + w_N(\text{right subtree})\}$$

$d(C, G) + w_C(\text{left subtree})$



# Cost of a history – leaves (initialisation).

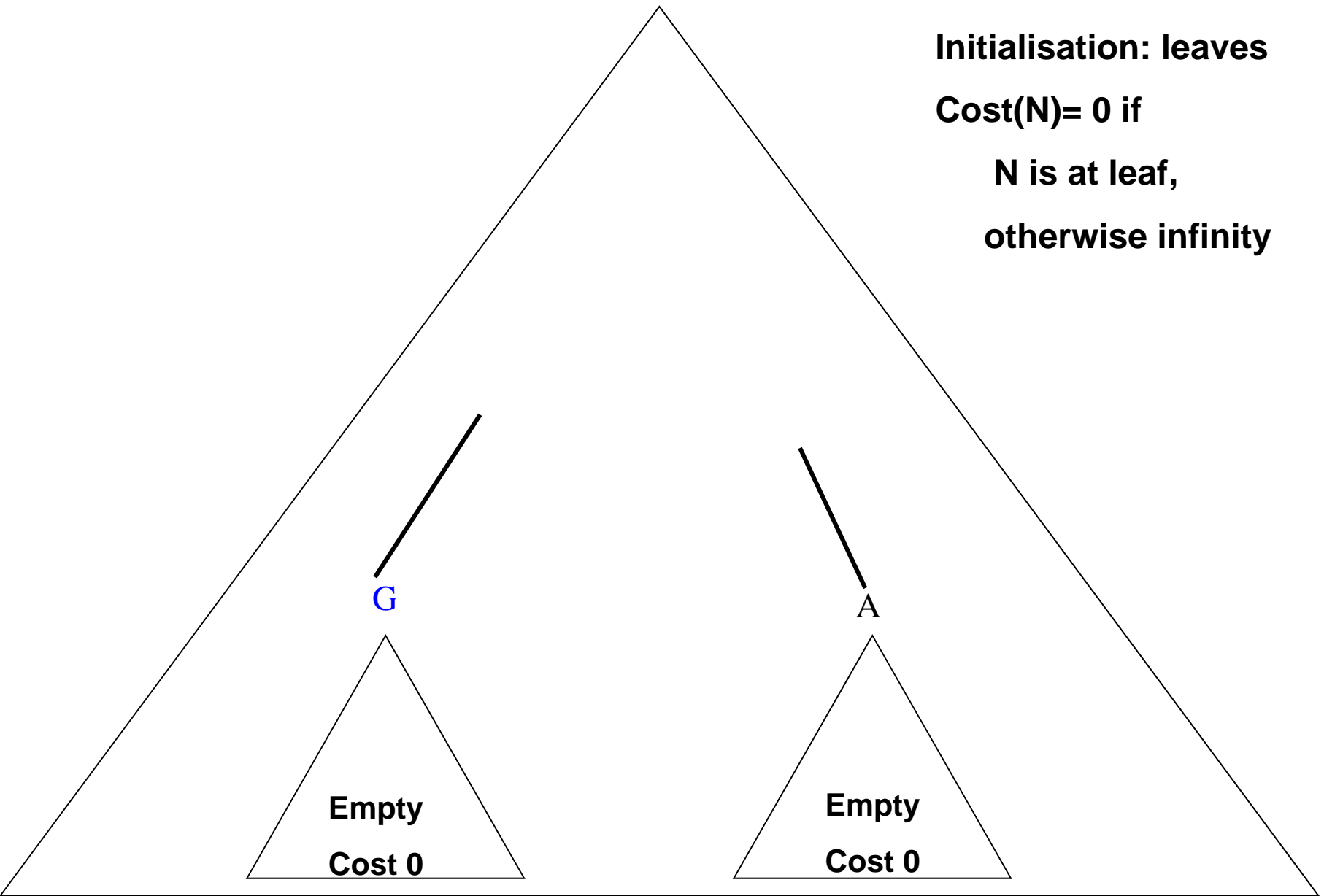
A C G T

Initialisation: leaves

$\text{Cost}(N) = 0$  if

N is at leaf,

otherwise infinity



# Compatibility and Branch Popping

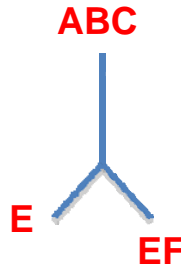
A GCACGTGCAGTTAGGA  
 B GCACGTGCAGTTAGGA  
 C TCTCGTGCAGTTAGGA  
 D TCTCATGCAATTAGGA  
 E TCTCATGCAATTATGA  
 F TCTCATGCAATTATGA



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

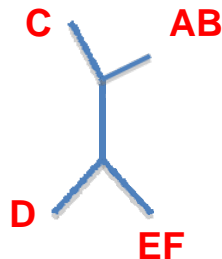
A GCACGTGCAGTTAGGA  
 B GCACGTGCAGTTAGGA  
 C TCTCGTGCAGTTAGGA  
 D TCTCATGCAATTAGGA  
 E TCTCATGCAATTATGA  
 F TCTCATGCAATTATGA



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

A GCACGTGCAGTTAGGA  
 B GCACGTGCAGTTAGGA  
 C TCTCGTGCAGTTAGGA  
 D TCTCATGCAATTAGGA  
 E TCTCATGCAATTATGA  
 F TCTCATGCAATTATGA

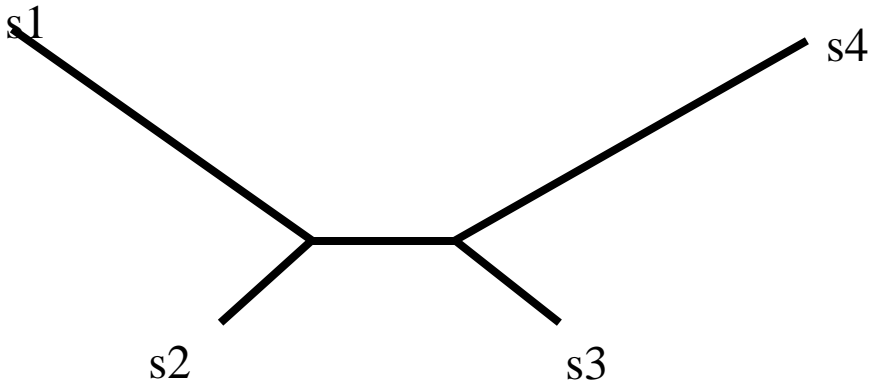


	1	2	3	4	5	6
1	+	?	?	?	?	?
2		+	?	?	?	?
3			+	?	?	?
4				+	?	?
5					+	?
6						+

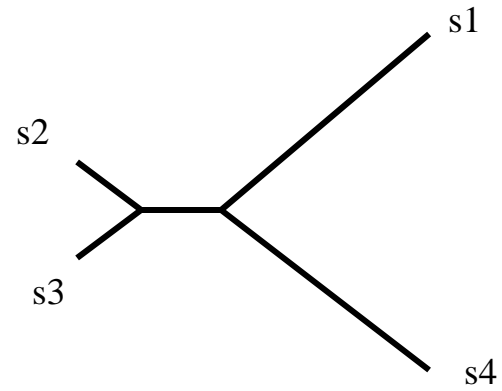
# The Felsenstein Zone

Felsenstein-Cavendar (1979)

True Tree



Reconstructed Tree



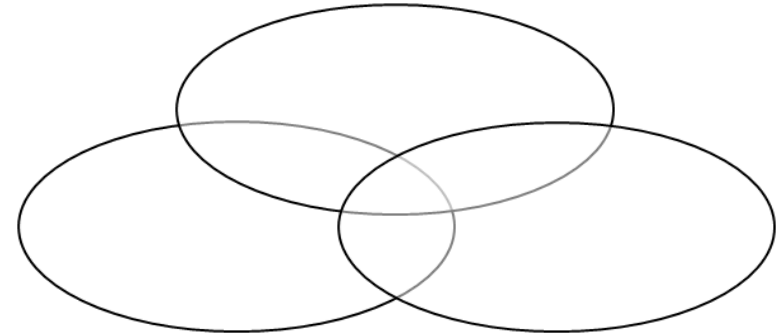
Patterns: (16 only 8 shown)

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*

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

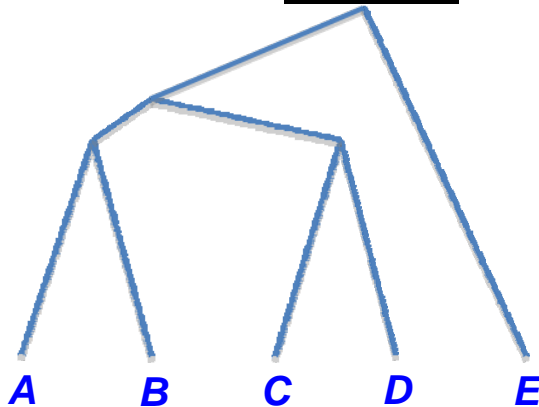


**Branch lengths –  $s$ , Bipartition lengths -  $q$**

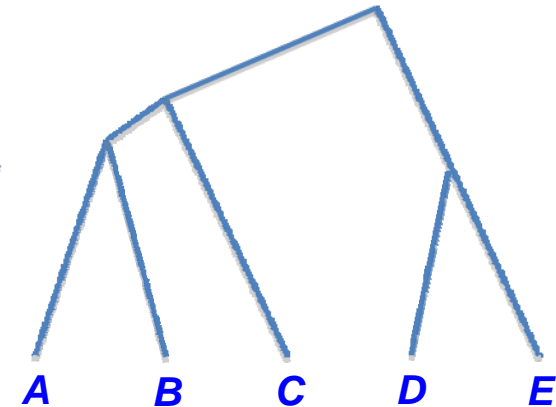
**From branch lengths to bipartitions  $q=Hs$**

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

**Inconsistency in presence of a Clock:**



*True Tree with Clock*

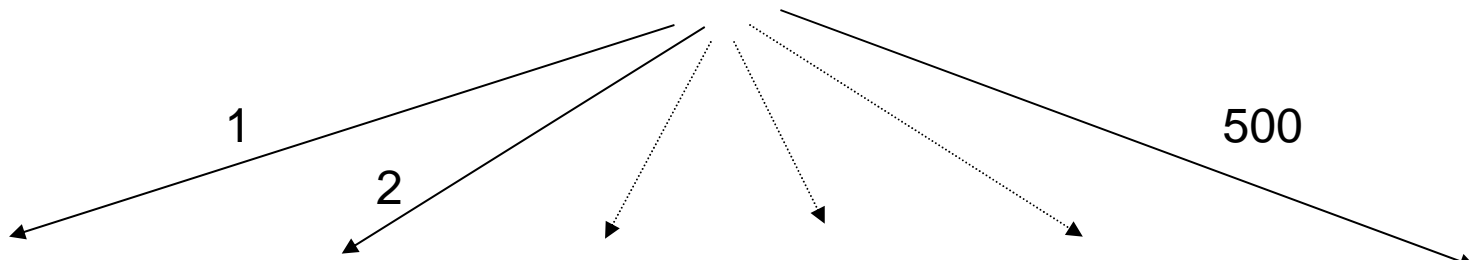


*More Likely Tree*

# Bootstrapping

Felsenstein (1985)

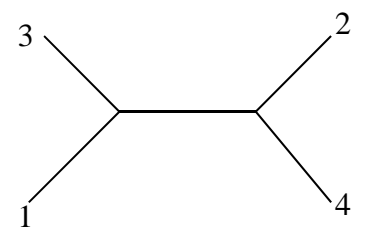
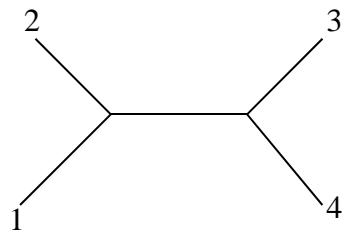
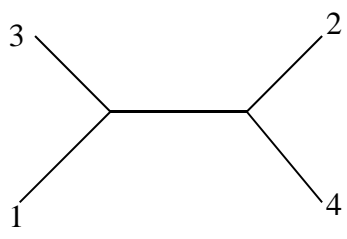
ATCTGTAGTCT  
ATCTGTAGTCT  
ATCTGTAGTCT  
ATCTGTAGTCT  
10230101201



ATCTGTAGTCT  
ATCTGTAGTCT  
ATCTGTAGTCT  
ATCTGTAGTCT

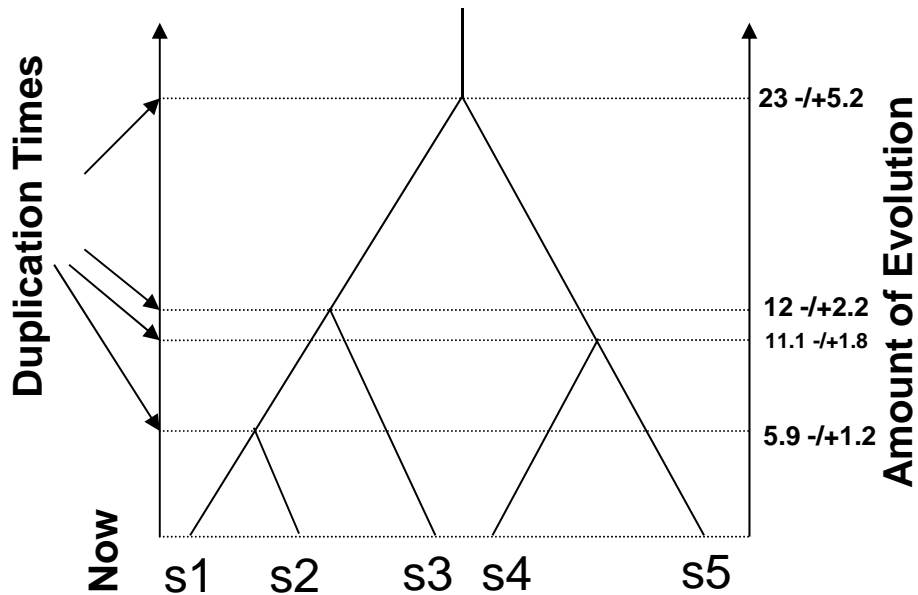
??????????  
??????????  
??????????  
??????????

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



# Output from Likelihood Method.

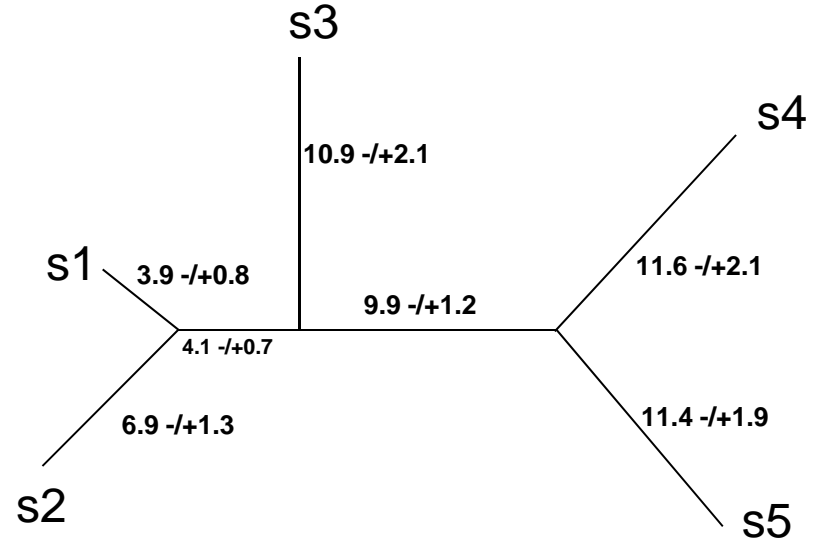
## Molecular Clock



$n-1$  heights estimated

Likelihood:  $7.9 \cdot 10^{-14}$   $\alpha, \beta = 0.31 \ 0.18$

## No Molecular Clock

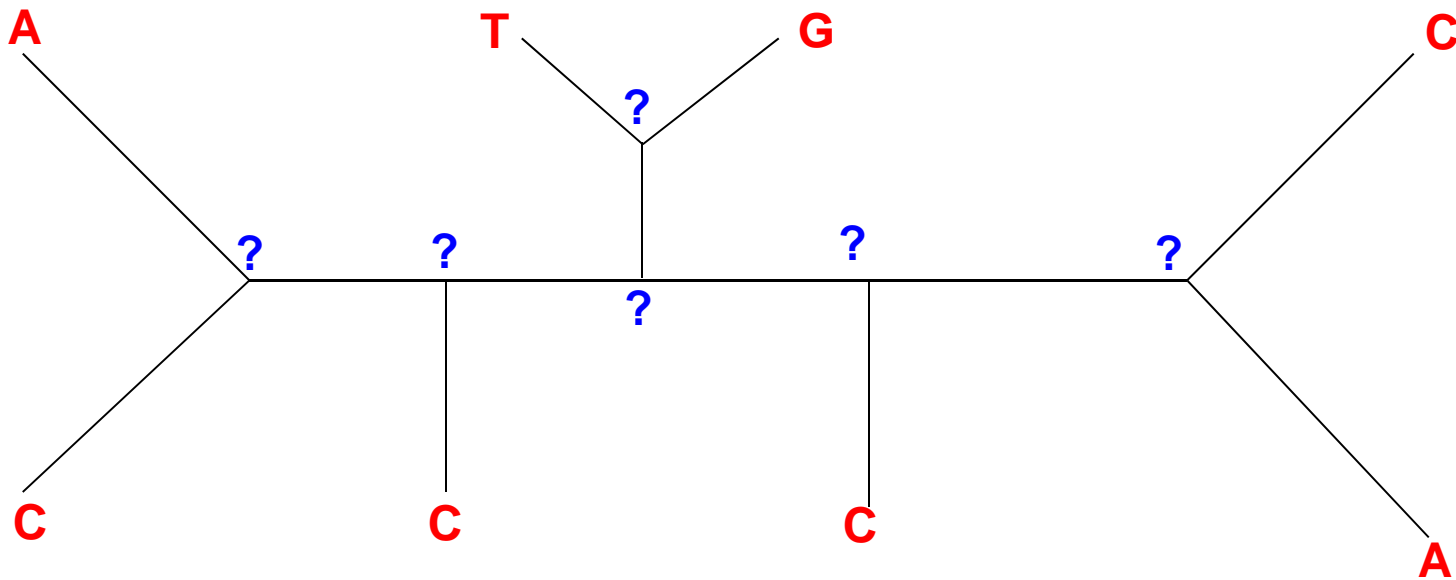


$2n-3$  lengths estimated

Likelihood:  $6.2 \cdot 10^{-12}$   $\alpha, \beta = 0.34 \ 0.16$

$\ln(7.9 \cdot 10^{-14}) - \ln(6.2 \cdot 10^{-12})$  is  $\chi^2$  - distributed with  $(n-2)$  degrees of freedom

# Assignment to internal nodes: The simple way.



If branch lengths and evolutionary process is known, what is the probability of nucleotides at the leaves?

Cctacggccatacca	<b>a</b>	ccctgaaagcaccatcccgt
Cttacgaccatatca	<b>c</b>	cgttgaatgcacgccatcccgt
Cctacggccatagca	<b>c</b>	ccctgaaagcaccatcccgt
Cccacggccatagga	<b>c</b>	ctctgaaagcactgcaccccgt
Tccacggccatagga	<b>a</b>	ctctgaaagcaccgcaccccgt
Ttccacggccatagg	<b>c</b>	actgtgaaagcaccgcaccccgt
Tggtgcggtcatacc	<b>g</b>	agcgctaatgcaccggatccca
Ggtgcggtcatacca	<b>t</b>	gcgttaatgcaccggatcccat

# Probability of leaf observations - summing over internal states

A C G T

$$P_G(\text{subtree}) =$$

$$\sum_{N \in \text{Nucleotides}} \{P(G \rightarrow N) \times P_N(\text{left subtree})\} \times$$

$$\sum_{N \in \text{Nucleotides}} \{P(G \rightarrow N) \times P_N(\text{right subtree})\}$$

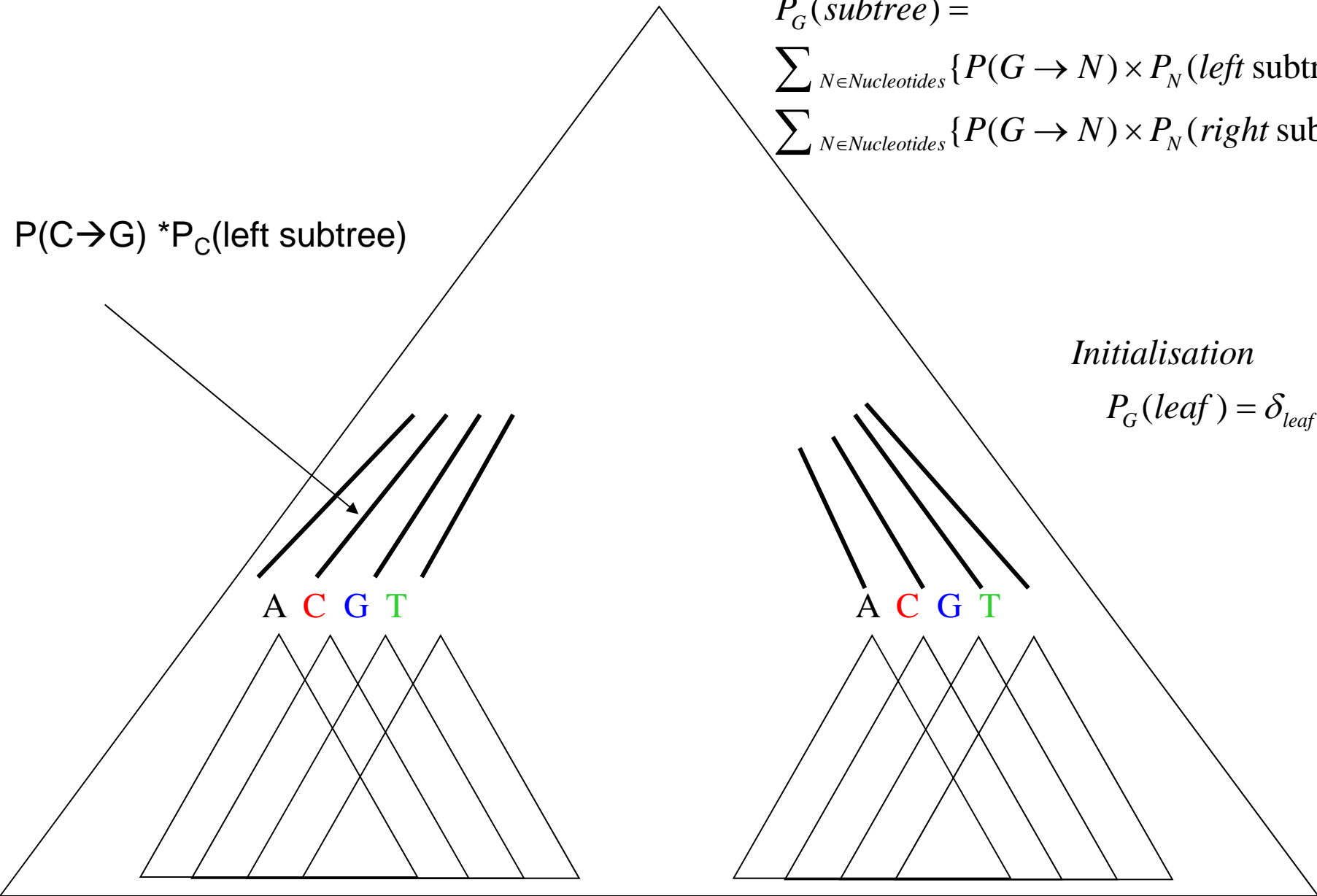
$P(C \rightarrow G) * P_C(\text{left subtree})$

*Initialisation*

$$P_G(\text{leaf}) = \delta_{\text{leaf}, G}$$

A C G T

A C G T



# Summary

## *Basic Principles of Phylogenetics*

***Distance***

***Parsimony***

***Compatibility***

***Inconsistency***

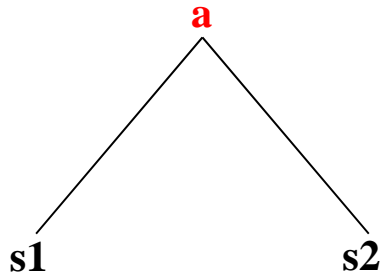
***Likelihood***

# The Molecular Clock

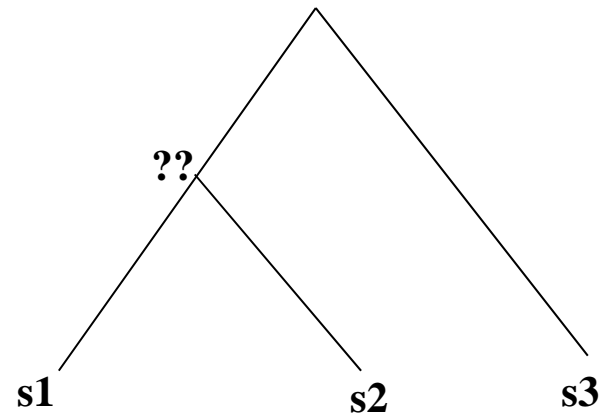
First noted by Zuckerkandl & Pauling (1964) as an empirical fact.

How can one detect it?

**Known Ancestor, a, at Time t**



**Unknown Ancestors**

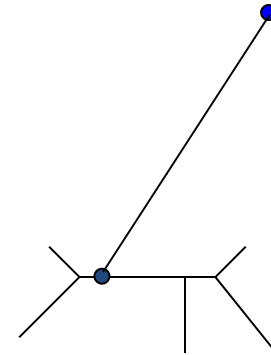


# Rootings

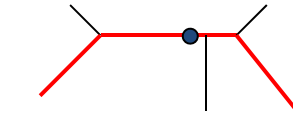
## Purpose

- 1) To give time direction in the phylogeny & most ancient point
- 2) To be able to define concepts such a monophyletic group.

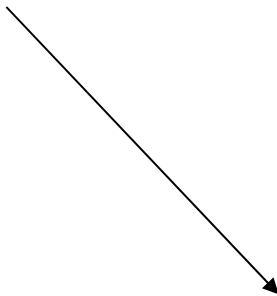
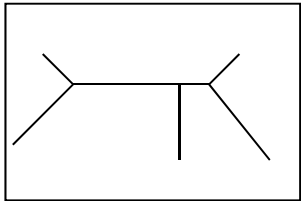
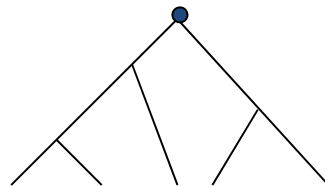
1) Outgroup: Enhance data set with sequence from a species definitely distant to all of them. It will be joined at the root of the original data



2) Midpoint: Find midpoint of longest path in tree.



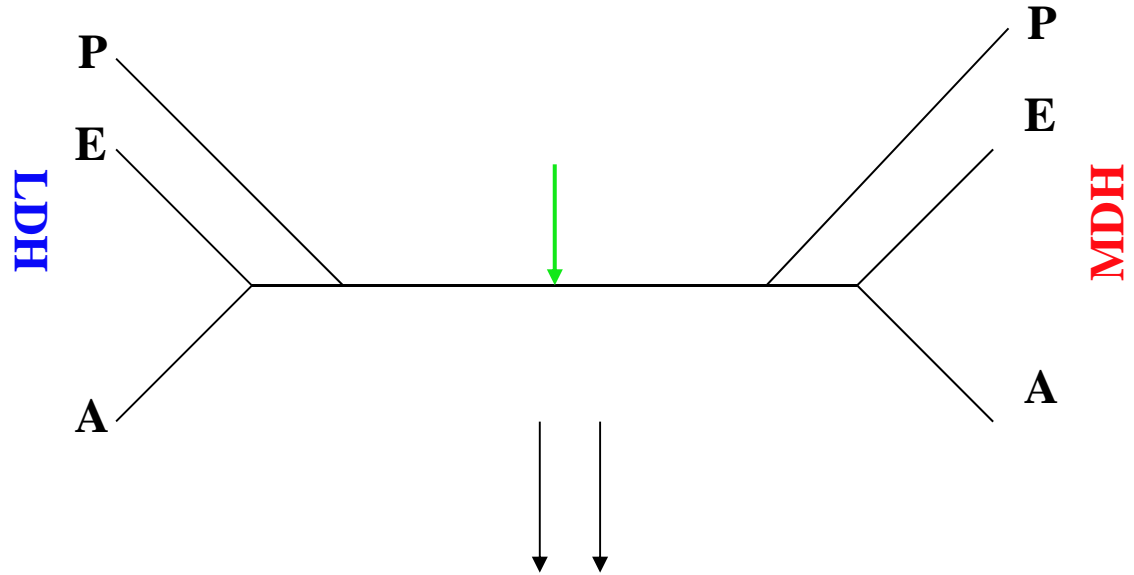
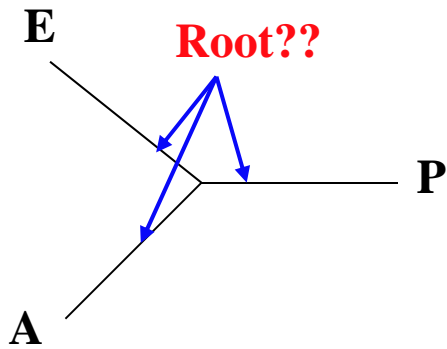
3) Assume Molecular Clock.



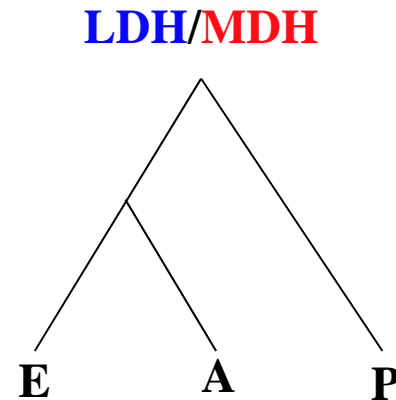
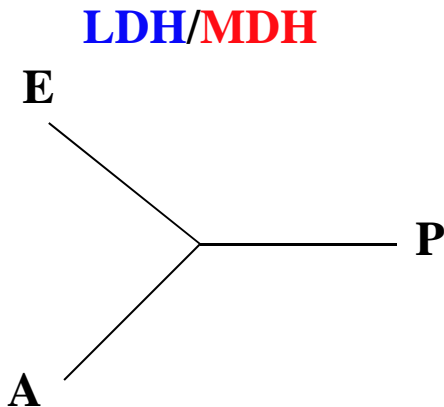
# Rooting the 3 kingdoms

3 billion years ago: no reliable clock - no outgroup

Given 2 set of homologous proteins, i.e. MDH & LDH can the archea, prokaria and eukaria be rooted?



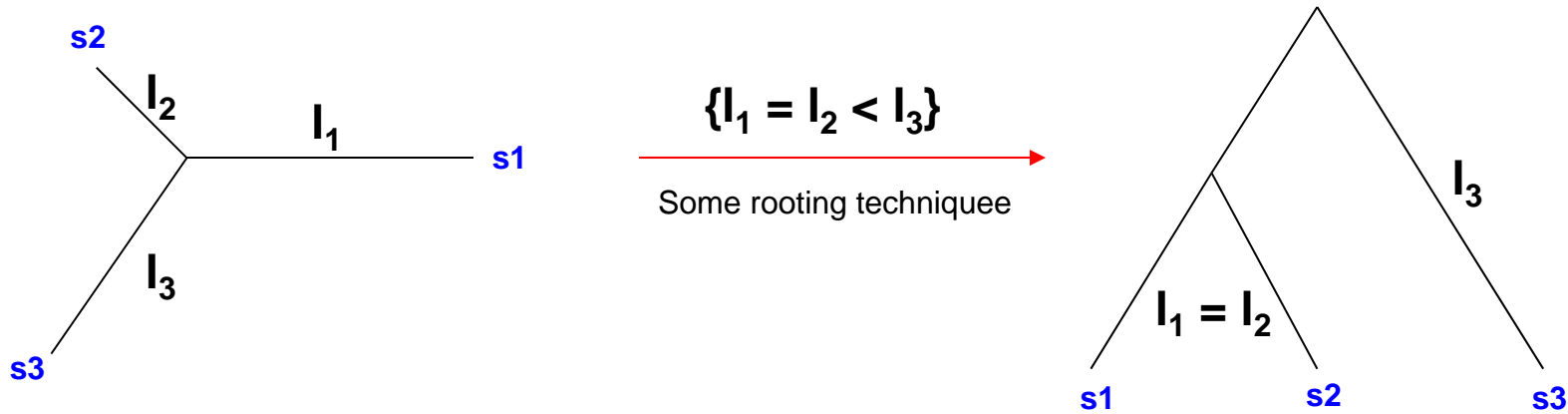
Given 2 set of homologous proteins, i.e. MDH & LDH can the archea, prokaria and eukaria be rooted?



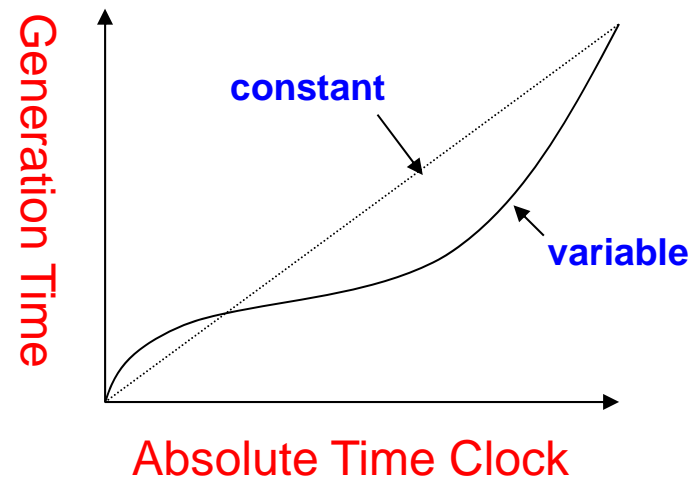
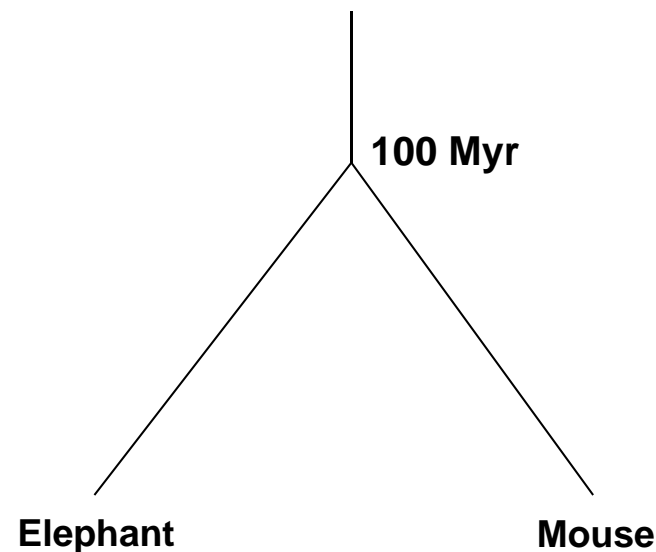
# The generation/year-time clock

Langley-Fitch, 1973

## Absolute Time Clock:

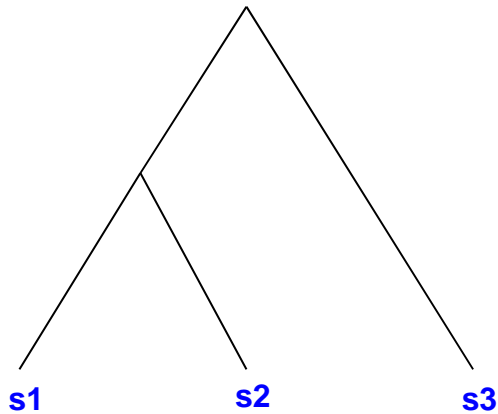


## Generation Time Clock:

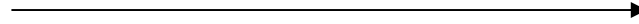


# The generation/year-time clock

Langley-Fitch, 1973



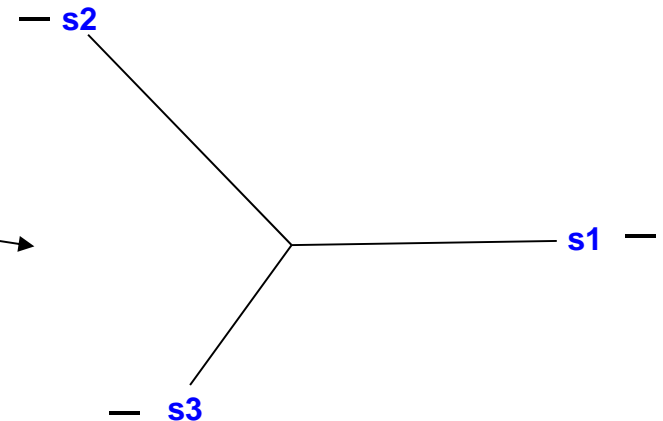
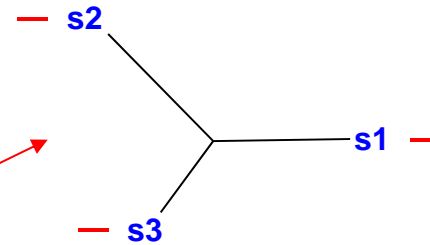
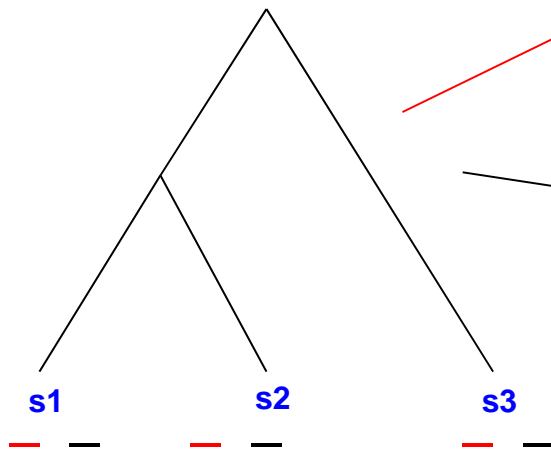
Generation Time Clock



Any Tree

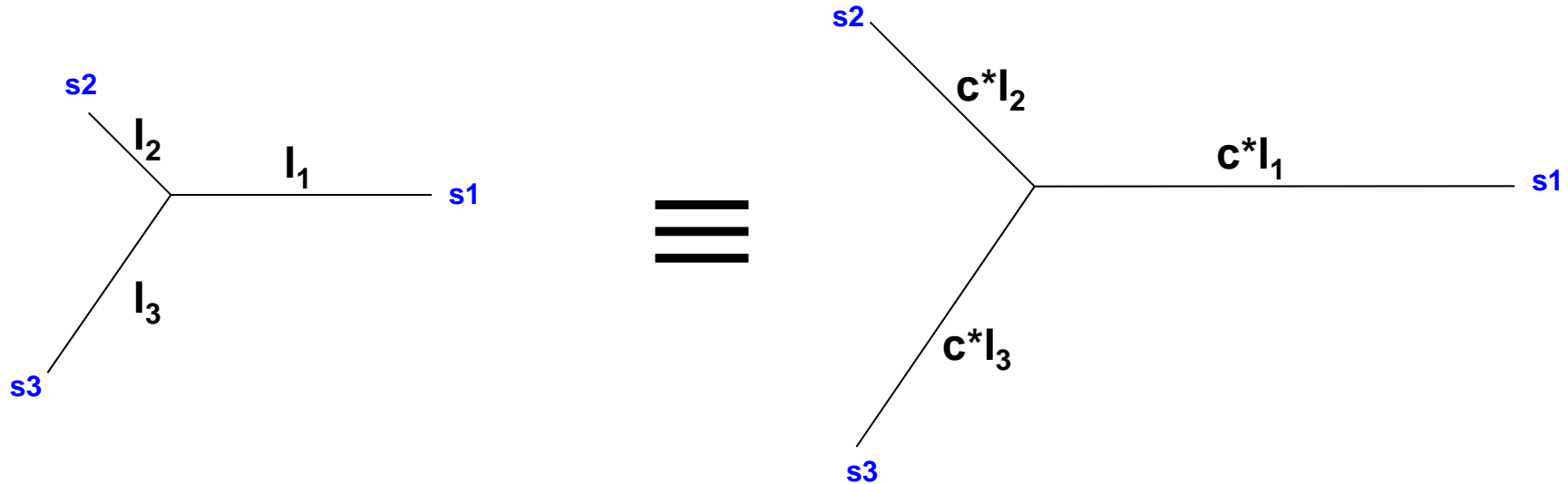
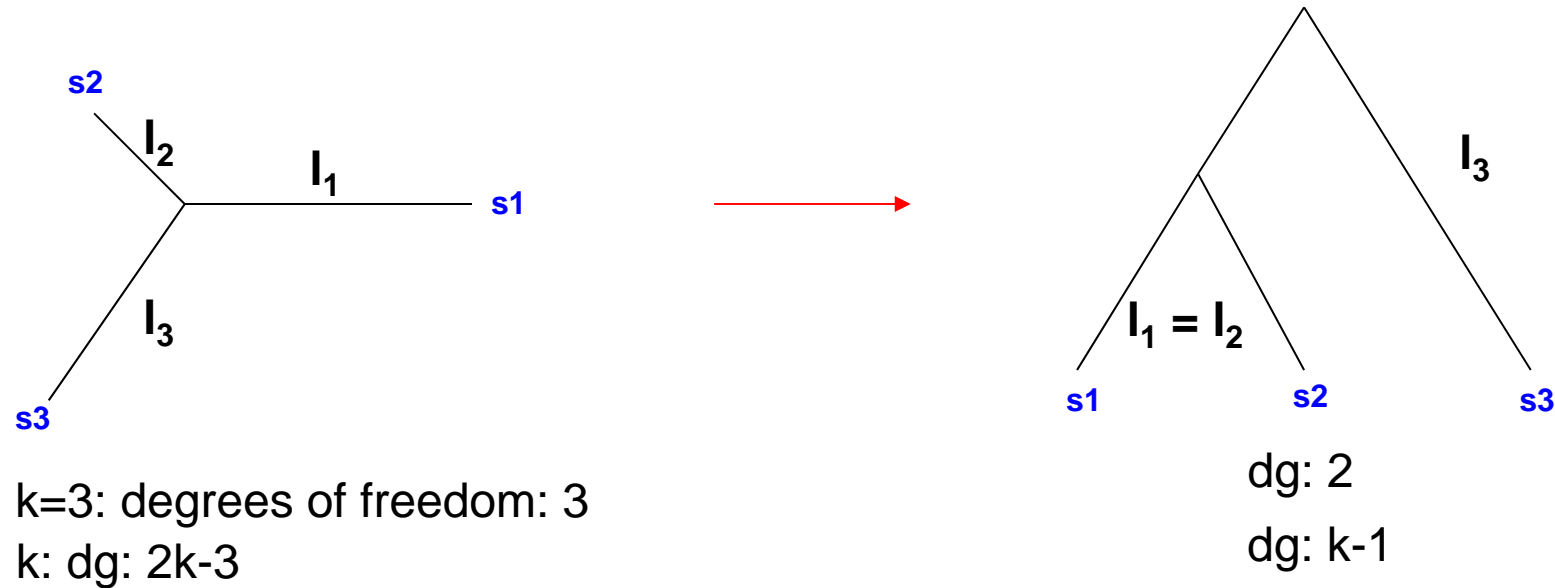
Can the generation time clock be tested?

Assume, a data set: 3 species, 2 sequences each



# The generation/year-time clock

Langley-Fitch, 1973



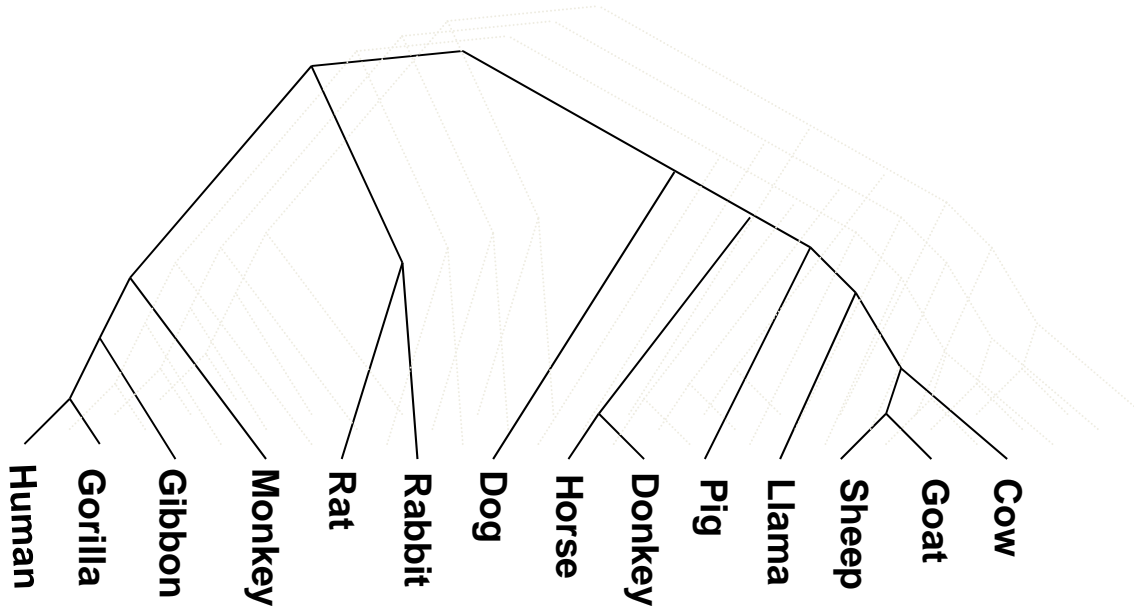
$k=3, t=2$ : dg=4

$k, t$ : dg =  $(2k-3)-(t-1)$

# $\alpha$ & $\beta$ – globin, cytochrome c, fibrinopeptide A & generation time clock

Langley-Fitch, 1973

## Fibrinopeptide A phylogeny:



### Relative rates

$\alpha$ -globin 0.342

$\beta$ -globin 0.452

cytochrome c 0.069

fibrinopeptide A 0.137

# Almost Clocks

(MJ Sanderson (1997) "A Nonparametric Approach to Estimating Divergence Times in the Absence of Rate Constancy" Mol.Biol.Evol.14.12.1218-31), J.L.Thorne et al. (1998): "Estimating the Rate of Evolution of the Rate of Evolution." Mol.Biol.Evol. 15(12).1647-57, JP Huelsenbeck et al. (2000) "A compound Poisson Process for Relaxing the Molecular Clock" Genetics 154.1879-92.)

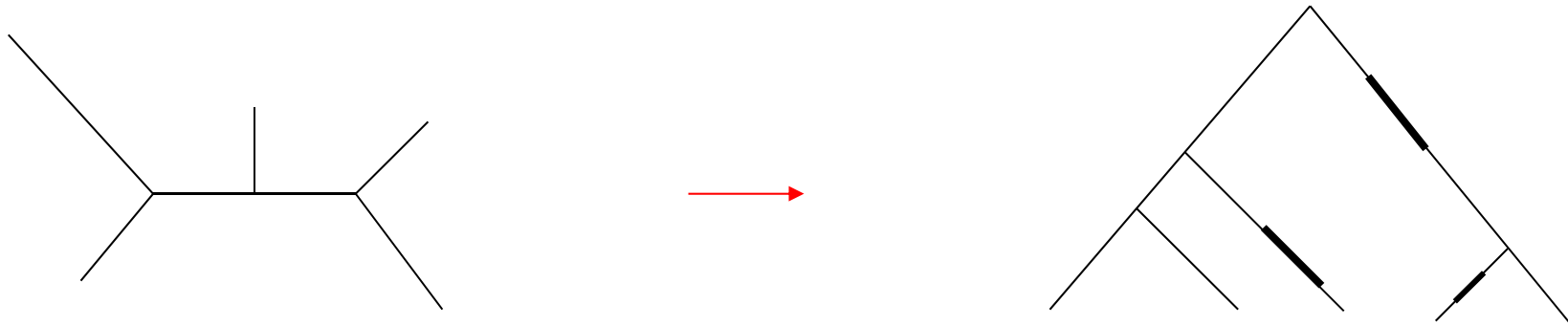
## I Smoothing a non-clock tree onto a clock tree (Sanderson)

## II Rate of Evolution of the rate of Evolution (Thorne et al.).

The rate of evolution can change at each bifurcation

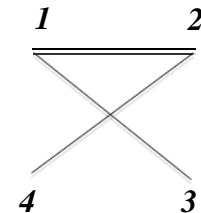
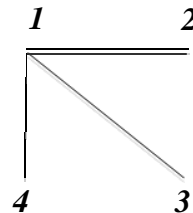
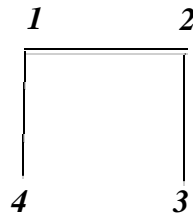
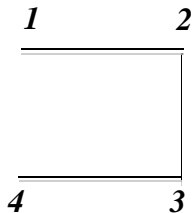
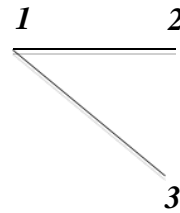
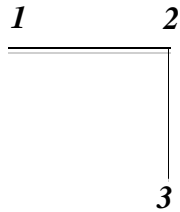
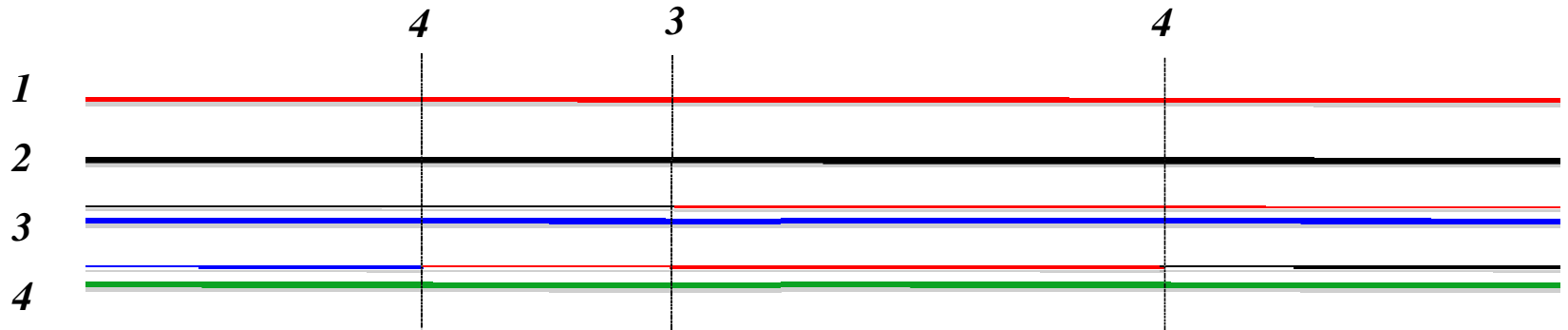
## III Relaxed Molecular Clock (Huelsenbeck et al.).

At random points in time, the rate changes by multiplying with random variable (gamma distributed)



Comment: Makes perfect sense. Testing no clock versus perfect is choosing between two unrealistic extremes.

# Li-Stephens



*Simplifications relative to the Ancestral Recombination Graph (ARG)*

*Local Trees are Spanning Trees – not phylogenies (Steiner Trees)*

*No non-ancestral bridges between ancestral material*

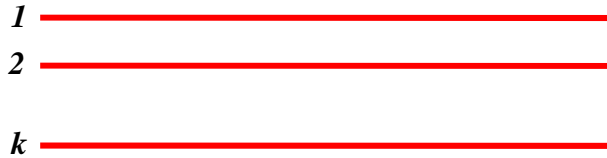


*Are there intermediates between Spanning Trees and Steiner Trees?*

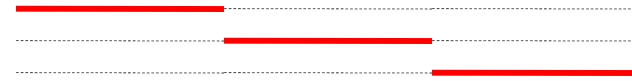
# FSA - Fast Statistical Alignment

Pachter, Holmes & Co

Data –  $k$  genomes/sequences:

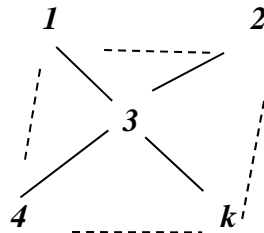


Iterative addition of homology statements to shrinking alignment:

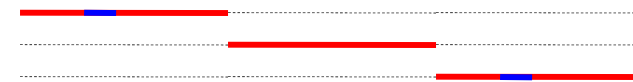


Spanning tree

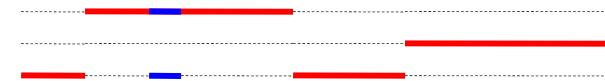
Additional edges



Add most certain homology statement from pairwise alignment compatible with present multiple alignment



An edge – a pairwise alignment

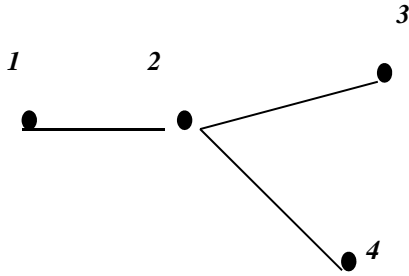


- 1,3 2,3 3,4 3,k
- 1,2 2,k 1,4 4,k

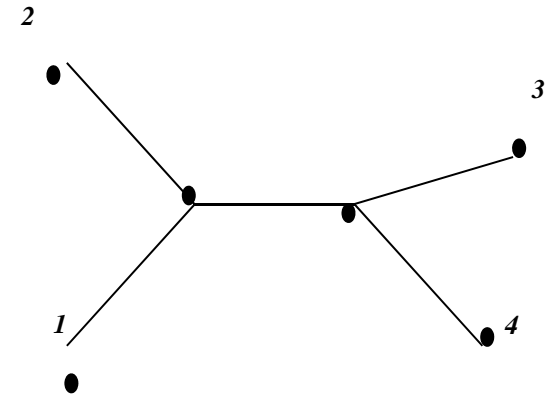
i. Conflicting homology statements cannot be added  
 ii. Some scoring on multiple sequence homology statements is used.

# Spannoids – $k$ -restricted Steiner Trees

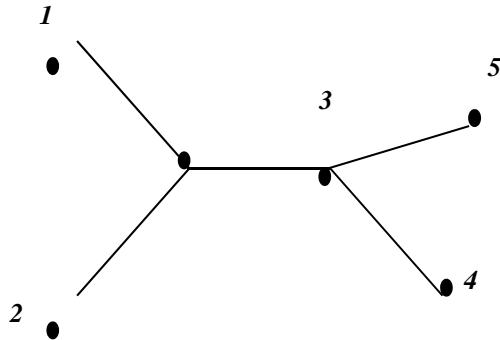
Baudis et al. (2000) Approximating Minimum Spanning Sets in Hypergraphs and Polymatroids



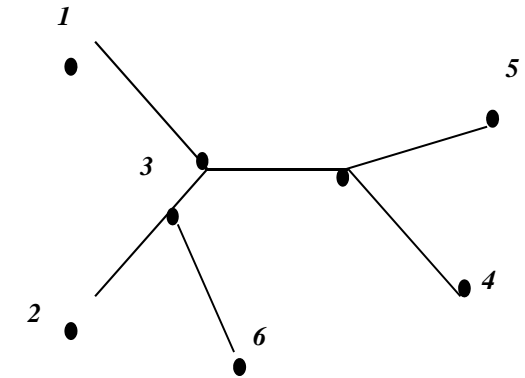
**Spanning tree**



**Steiner tree**



**1-Spannoid**



**2-Spannoid**

*Advantage: Decomposes large trees into small trees*

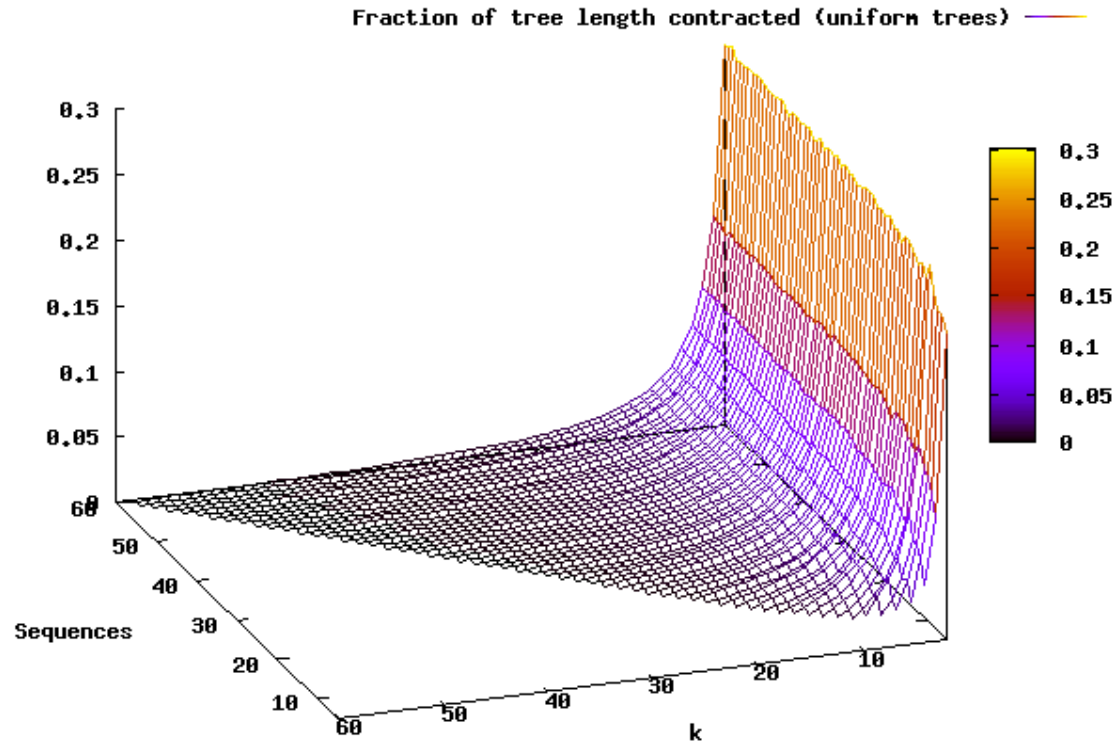
*Questions: How to find optimal spannoid?*

*How well do they approximate?*

# Example – Contraction of Simulated Coalescent Trees

## Simulation

- *Trees simulated from the coalescent*
- *Spannoid algorithm:*



## Conclusion

- *Approximation very good for  $k > 5$*
- *Not very dependent on sequence number*