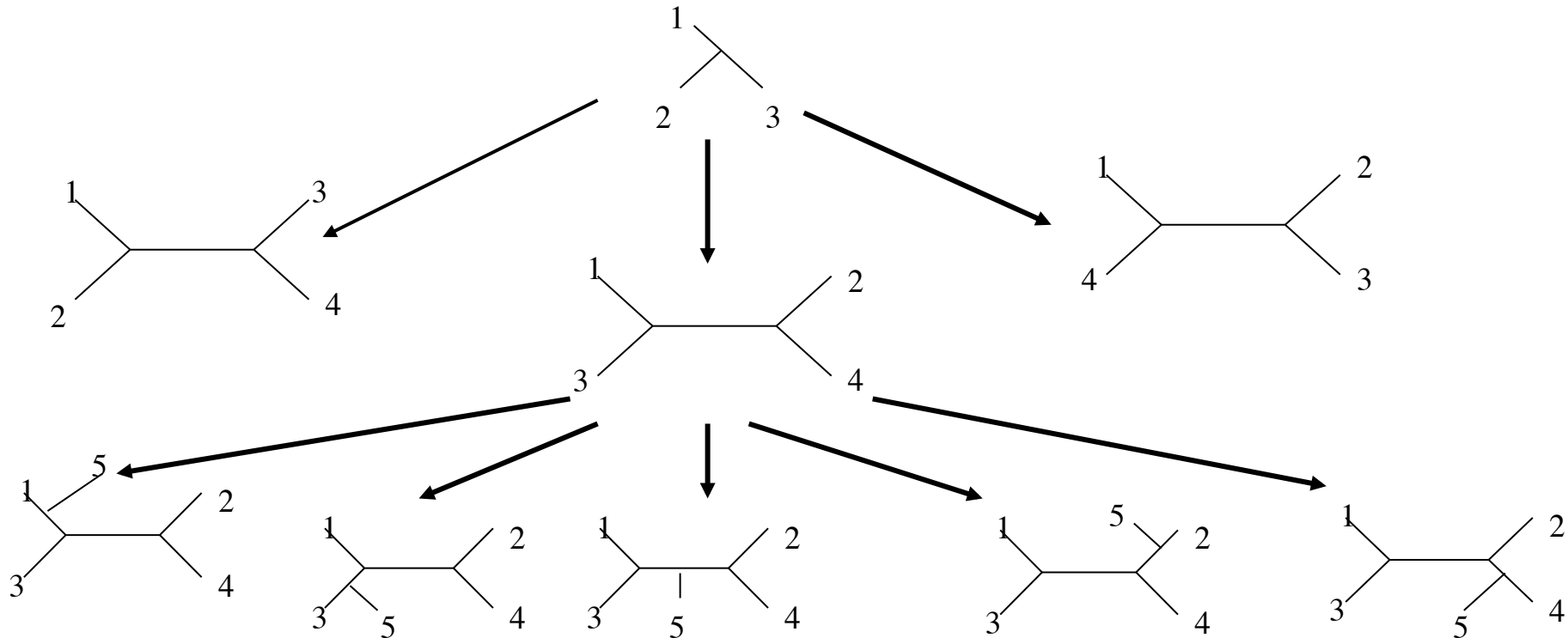


# Enumerating Trees: Unrooted & valency 3



**Recursion:**  $T_n = (2n-5) T_{n-1}$

**Initialisation:**  $T_1 = T_2 = T_3 = 1$

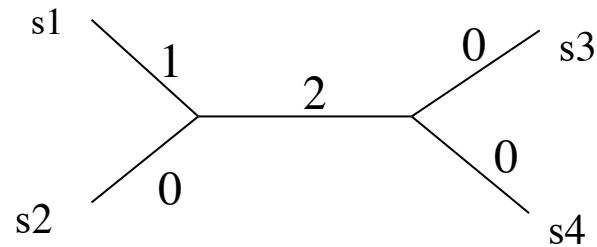
$$\prod_{j=3}^{n-1} (2j-3) = \frac{(2n-5)!}{(n-2)! 2^{n-2}}$$

4	5	6	7	8	9	10	15	20	
3	15	105	945	10345	1.4 10 <sup>5</sup>	2.0 10 <sup>6</sup>	7.9 10 <sup>12</sup>	2.2 10 <sup>20</sup>	

# Central Principles of Phylogeny Reconstruction

TTCAGT  
 TCCAGT  
 GCCAAT  
 GCCAAT

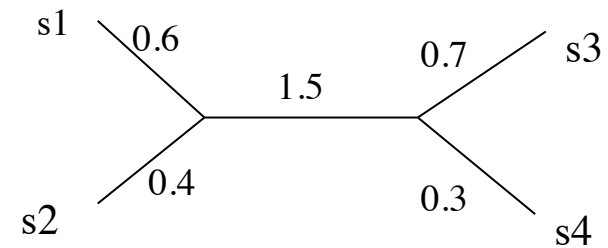
Parsimony



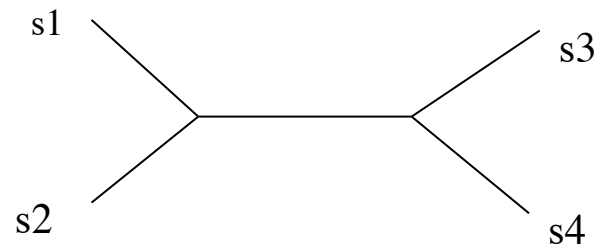
Total Weight: 3

Distance

1		
3	2	
3	2	0



Likelihood



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

Parameter estimates

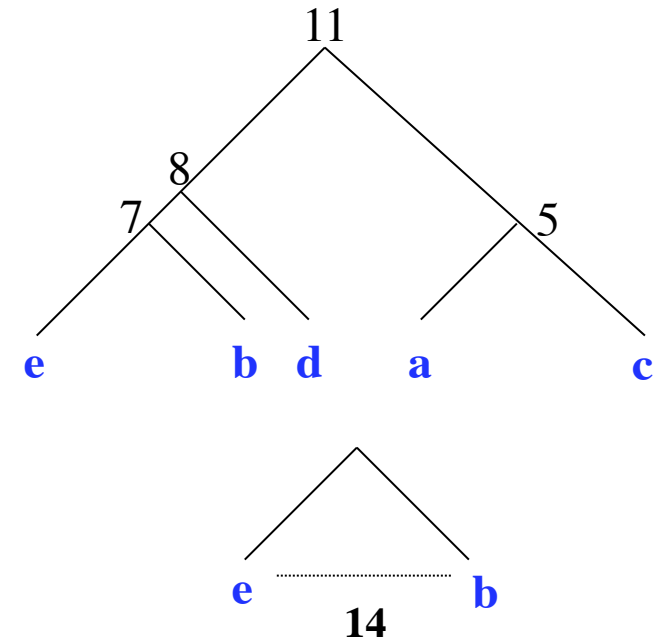
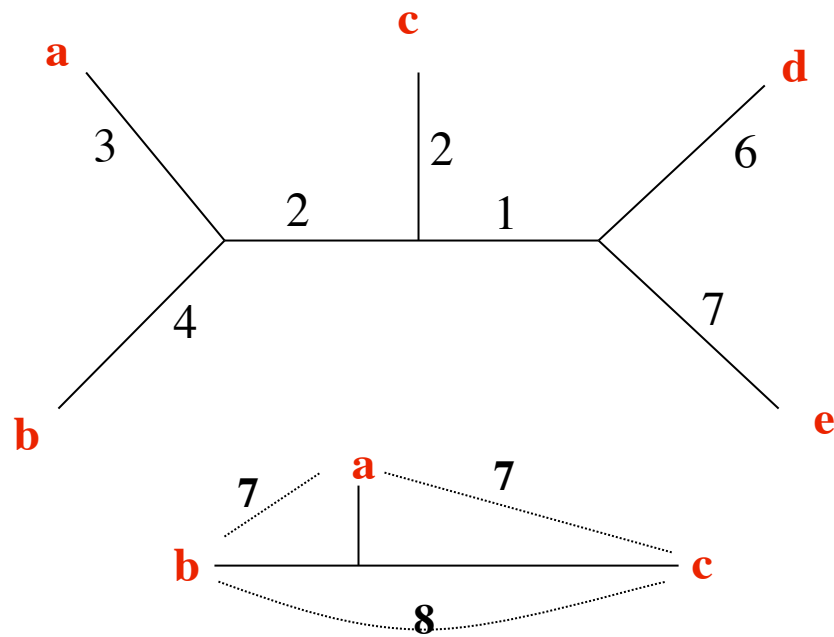
# From Distance to Phylogenies

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

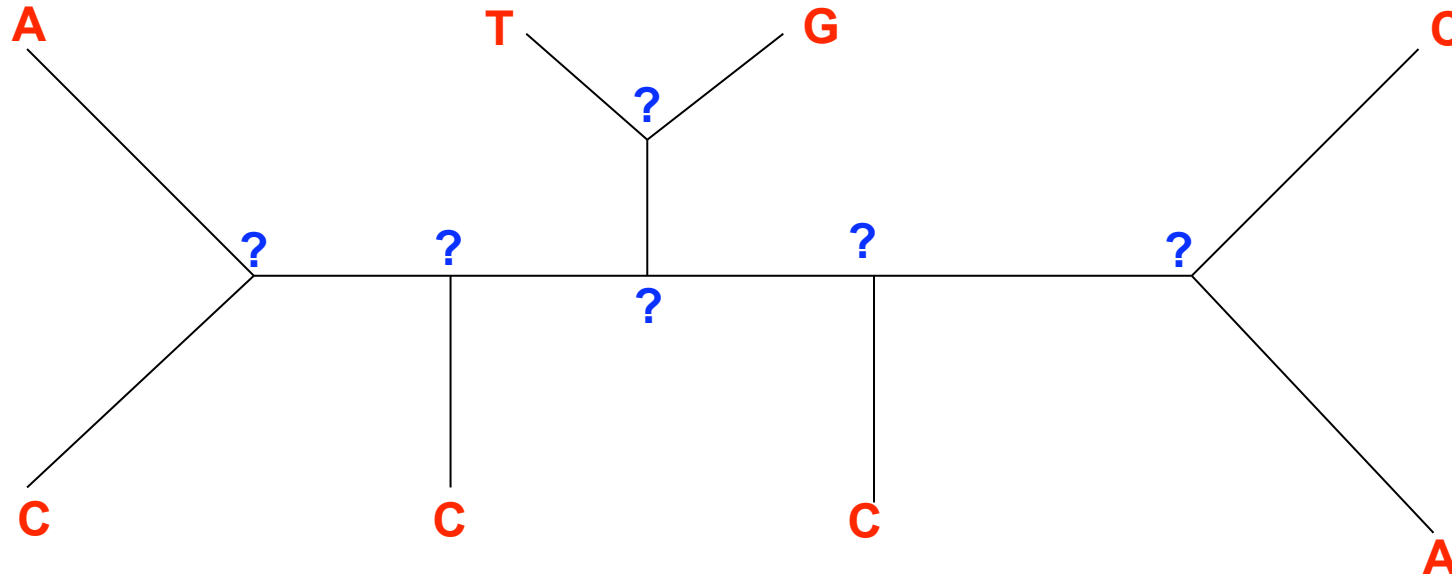
**No Molecular clock**

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	-



## Assignment to internal nodes: The simple way.

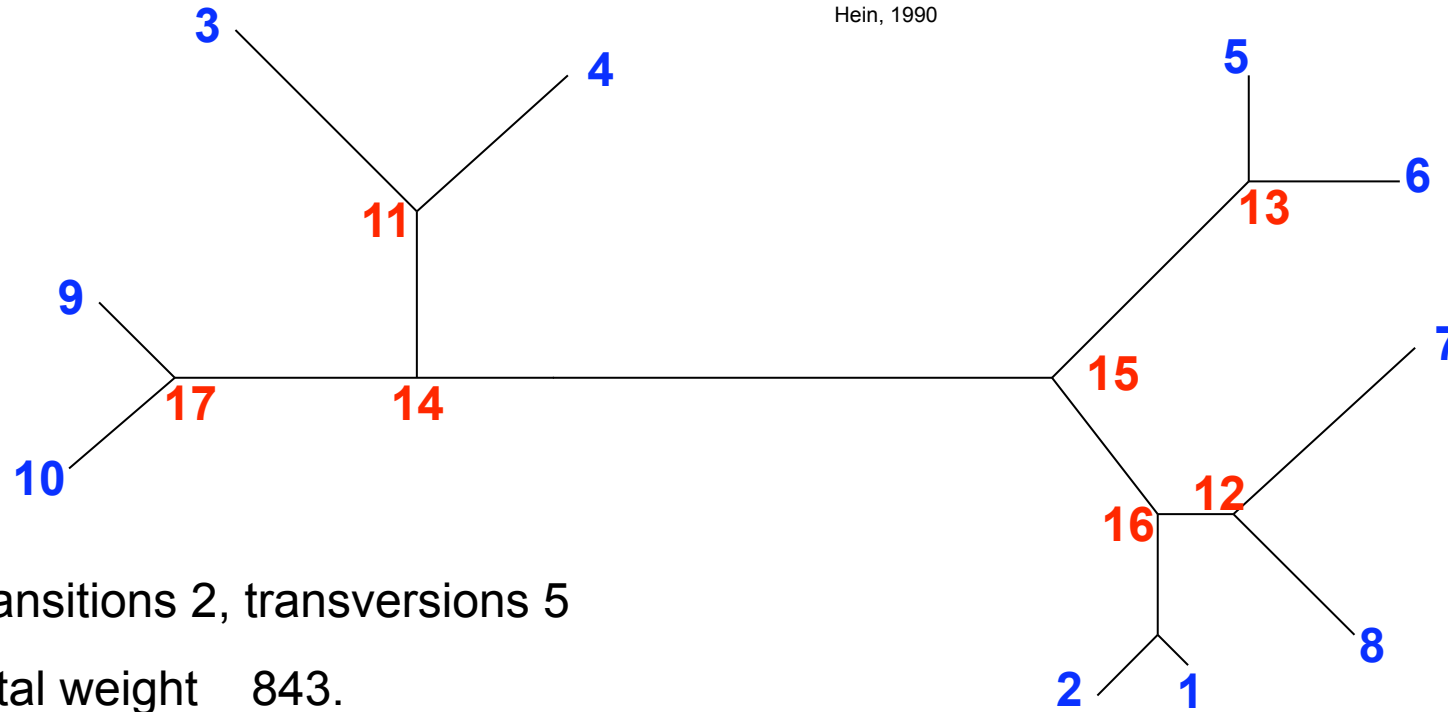


**What is the cheapest assignment of nucleotides to internal nodes, given some (symmetric) distance function  $d(N_1, N_2)$ ??**

**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.

```

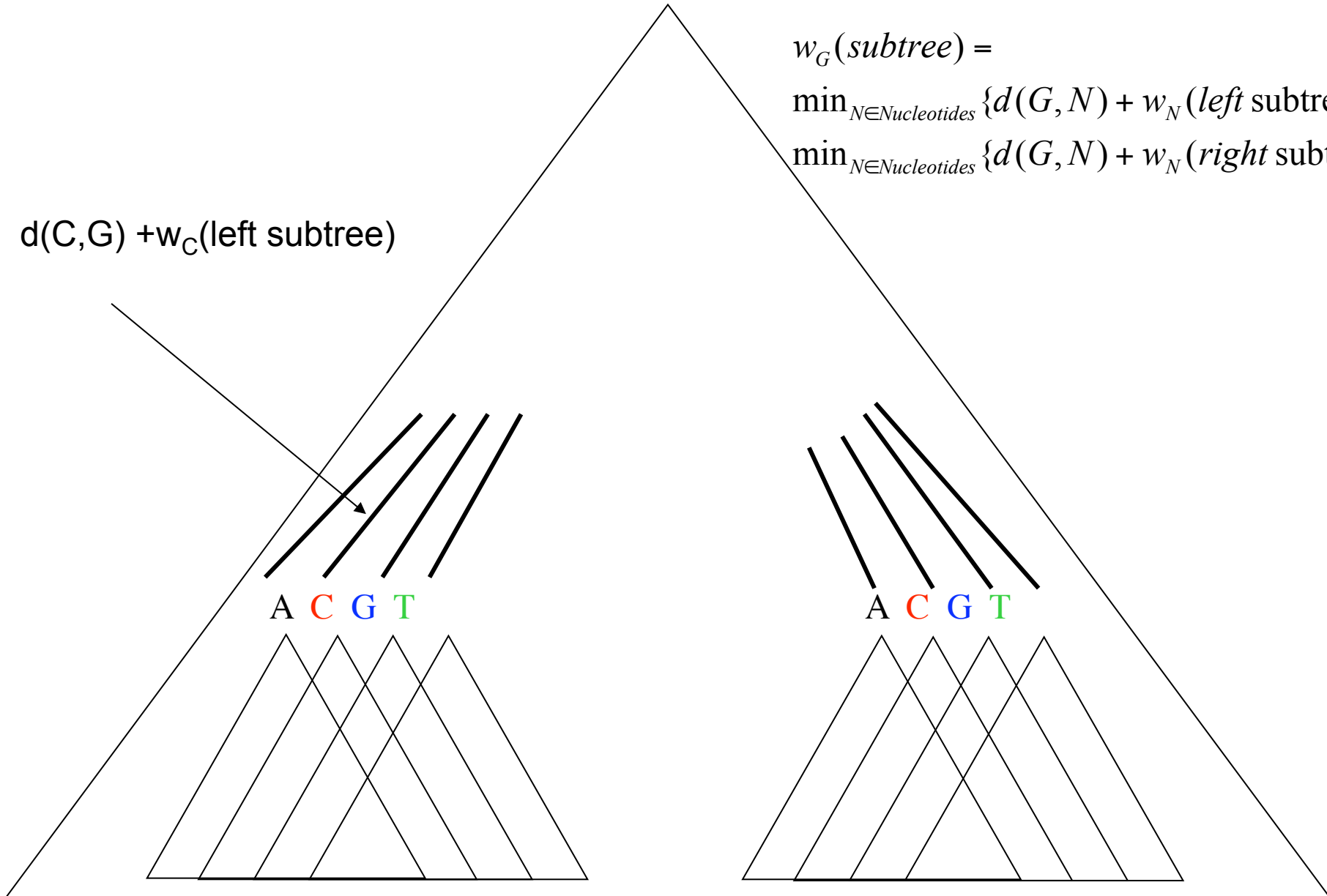
10 tatt-ctggtgtcccaggcgtagaggaaacacaccgatccatctcgaacttgggtggtgaaactctgccgcggt--aaccaatact-cg-gg-gggggccct-gcggaaaaatagctcgatgccagga--ta
17 t--t-ctggtgtcccaggcgtagaggaaacacaccaatccatcccgaacttgggtggtgaaactctgctgcggt--ga-cgatact-tg-gg-gggagcccg-atggaaaaatagctcgatgccagga--t-
9 t--t-ctggtgtctcaggcgtggaggaaacacaccaatccatcccgaacttgggtggtgaaactctattgcggt--ga-cgatactgta-gg-ggaagcccg-atggaaaaatagctcgacgccagga--t-
14 t----ctggtggccatggcgtagaggaaacaccccataccgaactcggcagttaagctctgctgcgcc--ga-tggtact-tg-gg-gggagcccg-ctgggaaaaataggacgctgccag-a--t-
3 t----ctggtgatgatggcgaggggacacaccggttcccataccgaacacggccgtaagccctccagcgcc--aa-tggtact-tgctc-cgcagggag-ccgggagagtaggacgctgccag-g--c-
11 t----ctggtggcgtggcgaagggacacaccggttcccataccgaacacggcagttaagctctccagcgcc--ga-tggtact-tg-gg-ggcagtcg-ctgggagagtaggacgctgccag-g--c-
4 t----ctggtggcgtagcgagaaggtcacaccggttcccataccgaacacggagttaagcttctcagcgcc--ga-tggtagt-ta-gg-ggctgtccc-ctgtgagagtaggacgctgccag-g--c-
15 g----cctgcgcccatagcaccgtgaaagcaccatcccgaactcggcagttaagcagcggttgcccaga-tagtact-tg-ggtgggagaccgctgggaaacctggatgctgcaag-c--t-
8 g----cctacggccatcccaccctggtaacgcccgatctcgt-ctgatctcgaagtaagcagggctgggcctggt-tagtact-tg-gatgggagacctcctgggaataccgggtgctgtagg-ct-t-
12 g----cctacggccataccaccctgaaagcaccatcccgaactcggcagttaagcagggttgagcccagt-tagtact-tg-gatgggagaccgctgggaatcctgggtgctgtagg-c--t-
7 g----cctacgaccatatacgttgatgacgccatcccgaactcggcagttaagcagcagttgagtcaggt-tagtact-tg-gatcggagaccgctgggaatcctggatggtgtaag-c--t-
16 g----cctacggccatagcaccctgaaagcaccatcccgaactcggcagttaagcagggttgcccaga-tagtact-tg-ggtgggagaccgctgggaatcctgggtgctgtagg-c--t-
1 a----tccacggccataggactctgaaagcactgcatcccgt-ccgatctgcaagtaaacagagtagcccccagt-tagtacc-ac-ggtggggaccacatgggaatcctgggtgctgt-gg-t--t-
18 a----tccacggccataggactctgaaagcaccgcatcccgt-ccgatctgcaagtaaacagagtagcccccagt-tagtacc-ac-ggtggggaccacatgggaatcctgggtgctgt-gg-t--t-
2 a----tccacggccataggactgtgaaagcaccgcatcccgt-ctgatctgcgcagttaaacacagtgccgcctagt-tagtacc-at-ggtggggaccacatgggaatcctgggtgctgt-gg-t--t-
5 g---tggtgcggtcataccagcgtaatgacccggatcccgaactcgcagttaagcgcgcttgggcccagaa-cagtact-gg-gatgggtgacctcccgggaagtccctggtgccacc-c--c-
13 g----ggtgcggtcataccagcgttaatgacccggatcccgaactcgcagttaagcgcgcttgggcccagaa-tagtact-ag-gatgggtgacctcctgggaagtccctgtagctgcacc-c--t-
6 g----ggtgcgatcataccagcgttaatgacccggatcccgaactcgcagttaagcgcgcttgggtggag-tagtact-ag-gatgggtgacctcctgggaagtccctggaatcctggaatattgcacc-c--tt-
    
```

# Cost of a history - minimizing over internal states

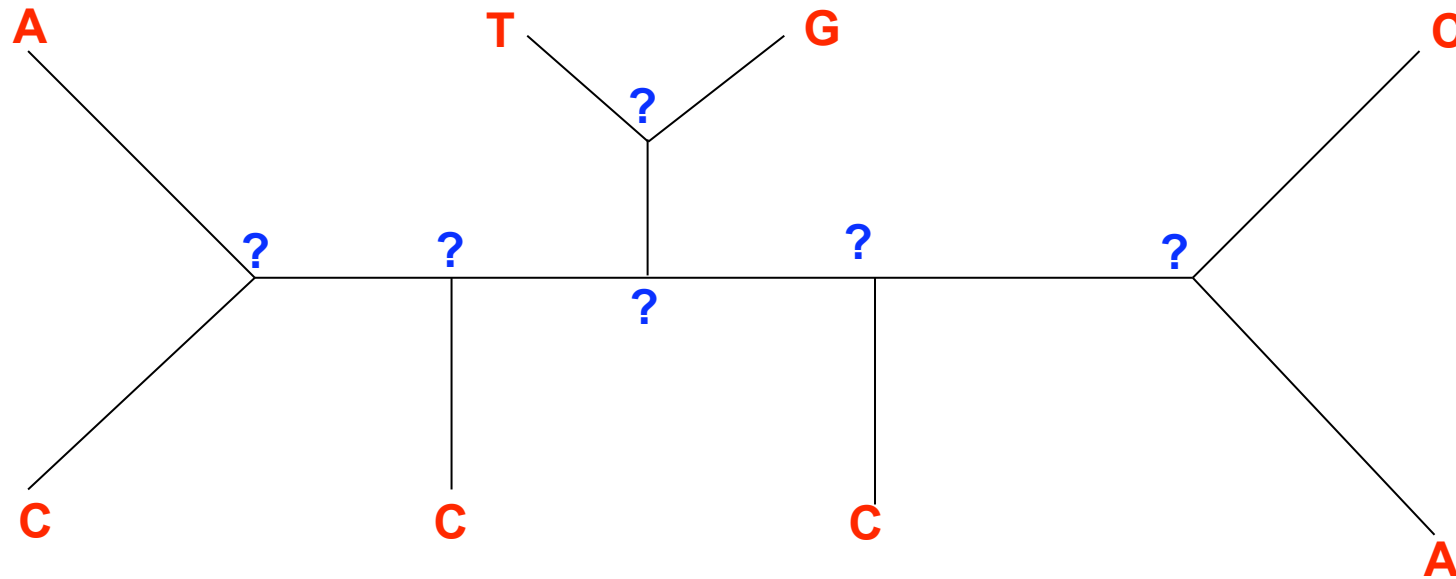
A C G T

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

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



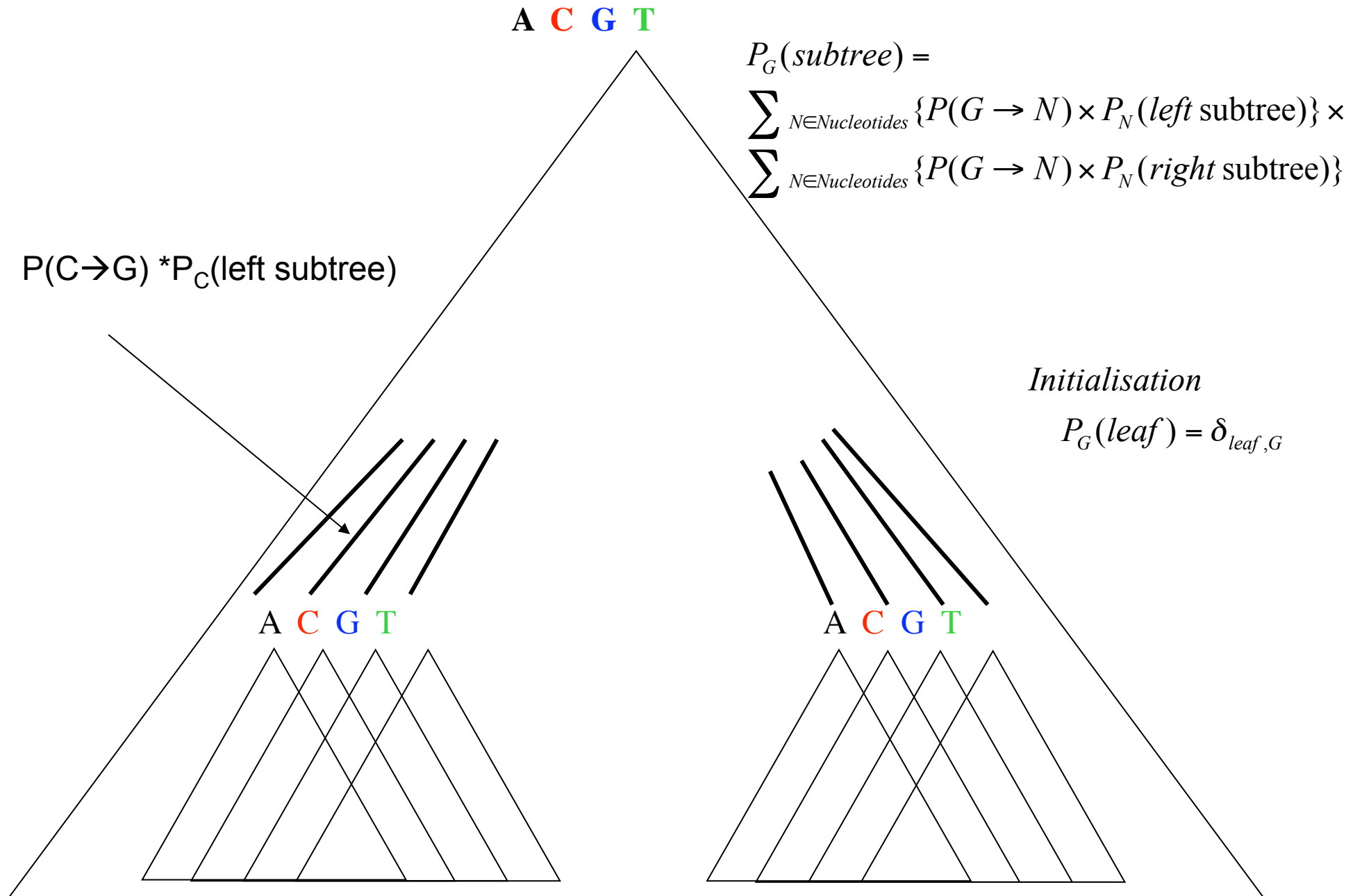
# Assignment to internal nodes: The simple way.



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

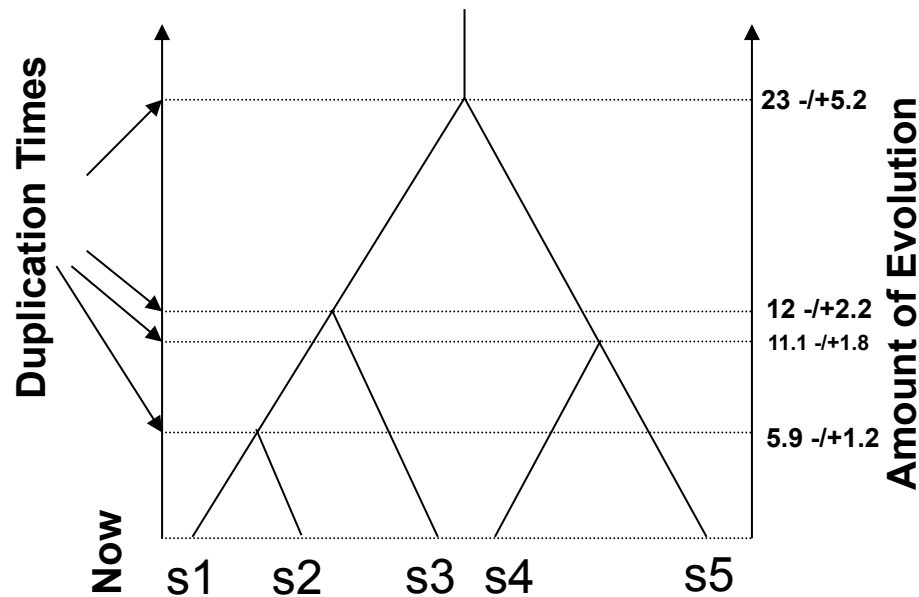
Cctacggccatacca	<b>a</b>	ccctgaaagcaccatcccgt
Cttacgaccatatca	<b>c</b>	cgttgaatgcacgccatcccgt
Cctacggccatagca	<b>c</b>	ccctgaaagcaccatcccgt
Cccacggccatagga	<b>c</b>	ctctgaaagcactgcatcccgt
Tccacggccatagga	<b>a</b>	ctctgaaagcaccgcatcccgt
Ttccacggccatagg	<b>c</b>	actgtgaaagcaccgcatcccgt
Tggtgcggtcatacc	<b>g</b>	agcgtaatgcaccggatccca
Ggtgcggtcatacca	<b>t</b>	gcgttaatgcaccggatcccat

# Probability of leaf observations - summing over internal states



# Output from Likelihood Method.

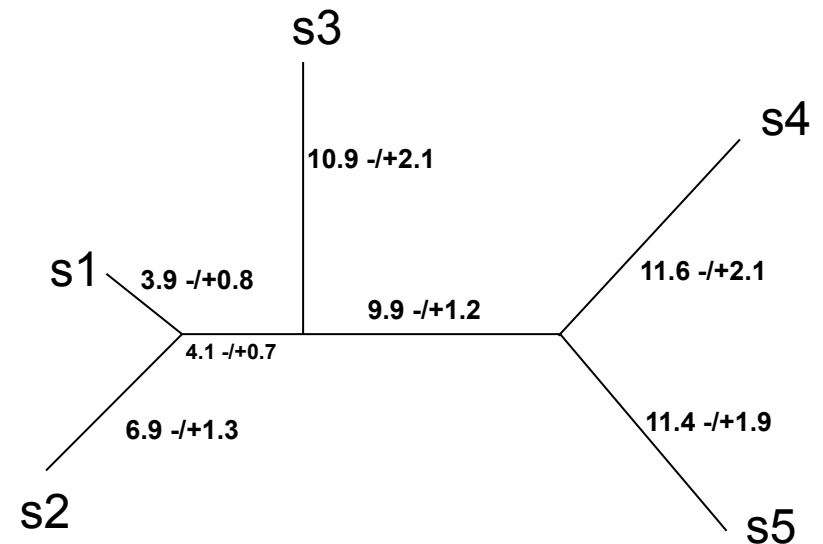
## Molecular Clock



$n-1$  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

# *Summary of Models and Phylogenies*

## **Models**

- The need for models for models of sequence evolution
- The assumptions of models
- The basic models
- selection and coding regions

## **Phylogenies**

- Enumeration of Phylogenies
- The Basic Principles
  - Distance
  - Parsimony
  - Likelihood