

# CSEP 590 A

## Spring 2013

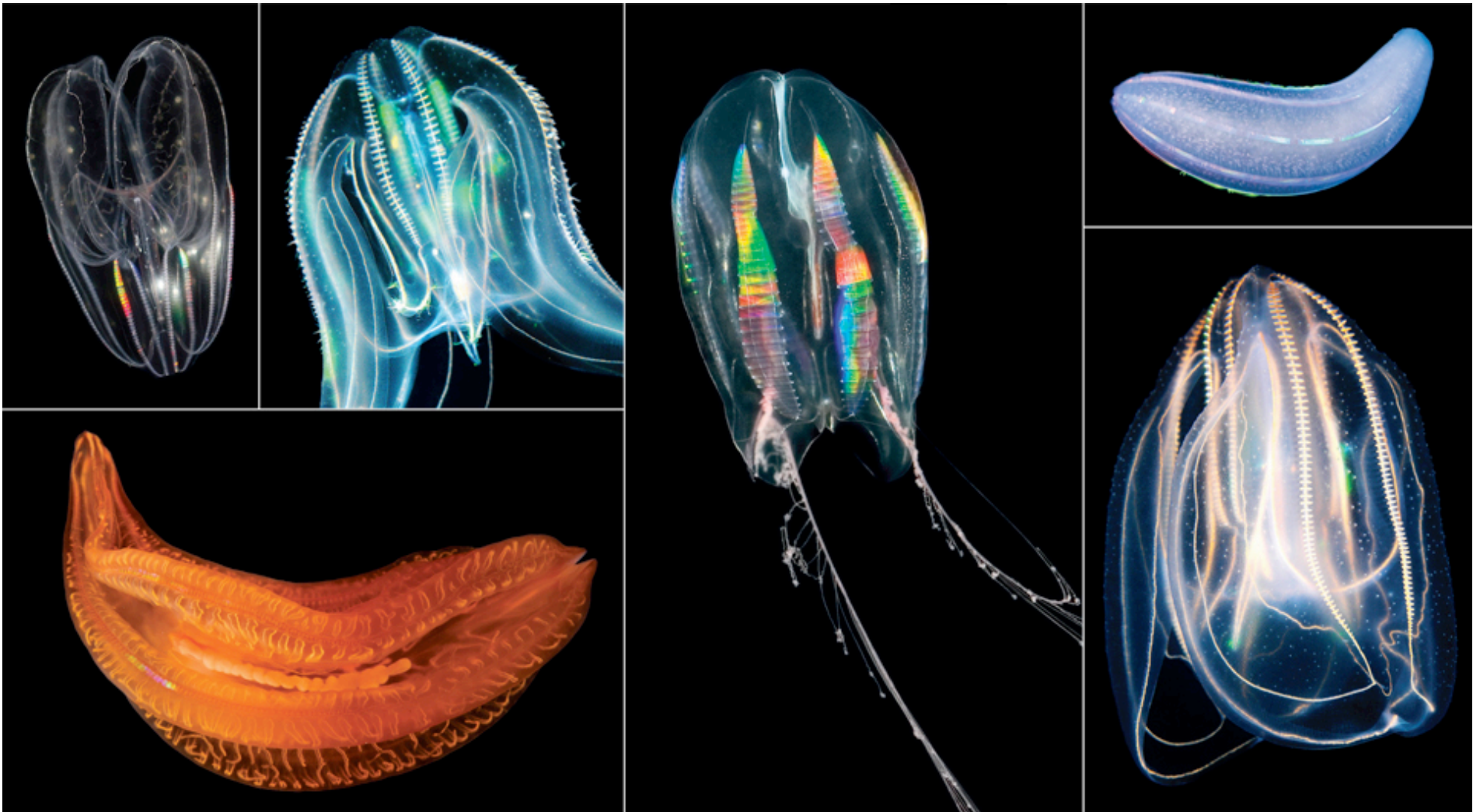
Phylogenies: Parsimony Plus a  
Tantalizing Taste of Likelihood

# Phylogenies (aka Evolutionary Trees)

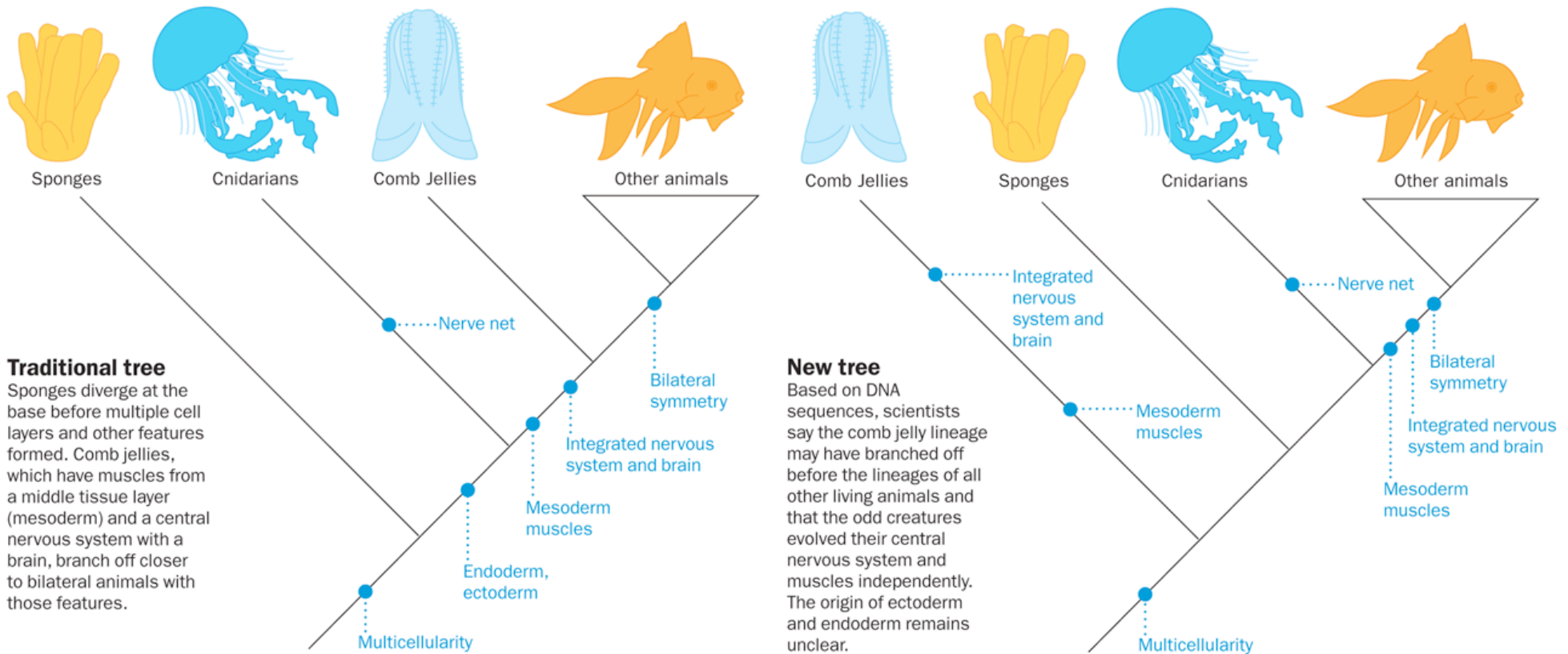
“Nothing in biology makes sense, except in the light of evolution”

-- Theodosius Dobzhansky, 1973

# Comb Jellies: Evolutionary enigma



[http://www.sciencenews.org/view/feature/id/350120/description/Evolutionary\\_enigmas](http://www.sciencenews.org/view/feature/id/350120/description/Evolutionary_enigmas)



## TREE OF LIFE

Diagrams depict the history of animal lineages as they evolved over time. Each branch represents a lineage that shares an ancestor with all of the animals that branch after the point where it splits from the tree. Biologists traditionally build trees by comparing species' anatomies; now they also compare DNA sequences.

	Comb jelly	Sponge	Cnidarian	Bilaterians
<b>DNA polymerase</b> important for cell replication	X	X	X	X
<b>Wnt hairpin 3</b> involved in embryonic development and cell division			X	X
<b>HOX</b> proteins pattern bodies during development and help form nerve cells			X	X
<b>microRNA</b> helps to regulate gene activity		X	X	X
<b>Drosha</b> cooperates with Pasha to make microRNA		X	X	X
<b>Pasha</b> cooperates with Drosha to make microRNA		X	X	X
<b>Voltage gated channels</b> (types L, N/P/Q and T) for nerve cell communication			X	X
<b>PAX Homeobox</b> proteins help embryos develop features such as eyes		X	X	X

A Complex Question:

Given data (sequences, anatomy, ...) infer the phylogeny

A Simpler Question:

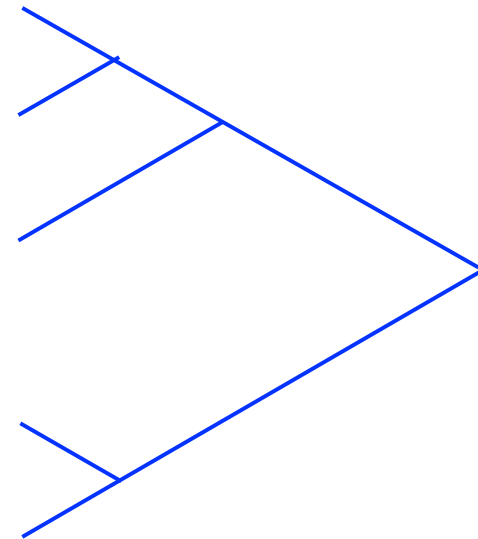
Given data *and a phylogeny*, evaluate “how much change” is needed to fit data to tree

(The former question is usually tackled by sampling tree topologies & comparing them by the later metric...)

# Parsimony

General idea ~ Occam's Razor:  
Given data where change is rare, prefer  
an explanation that requires few events

Human	A	T	G	A	T	...
Chimp	A	T	G	A	T	...
Gorilla	A	T	G	A	G	...
Rat	A	T	G	C	G	...
Mouse	A	T	G	C	T	...





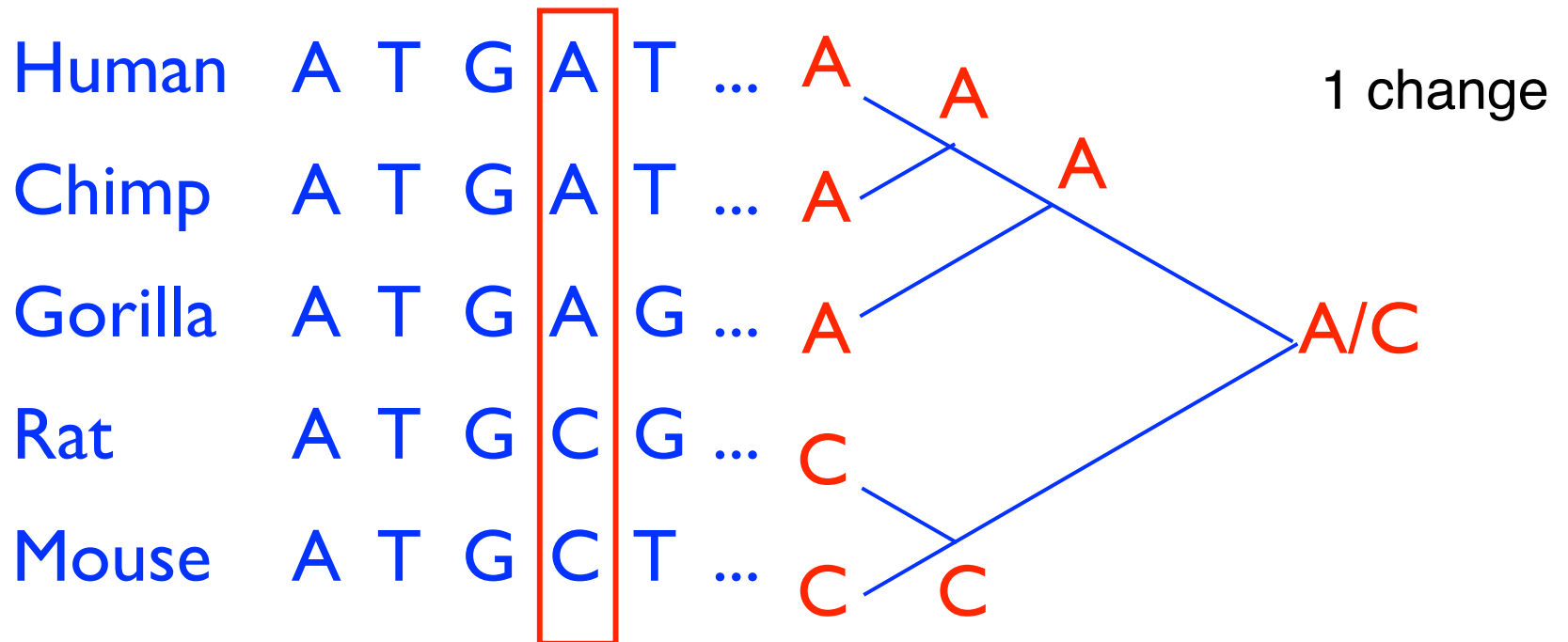






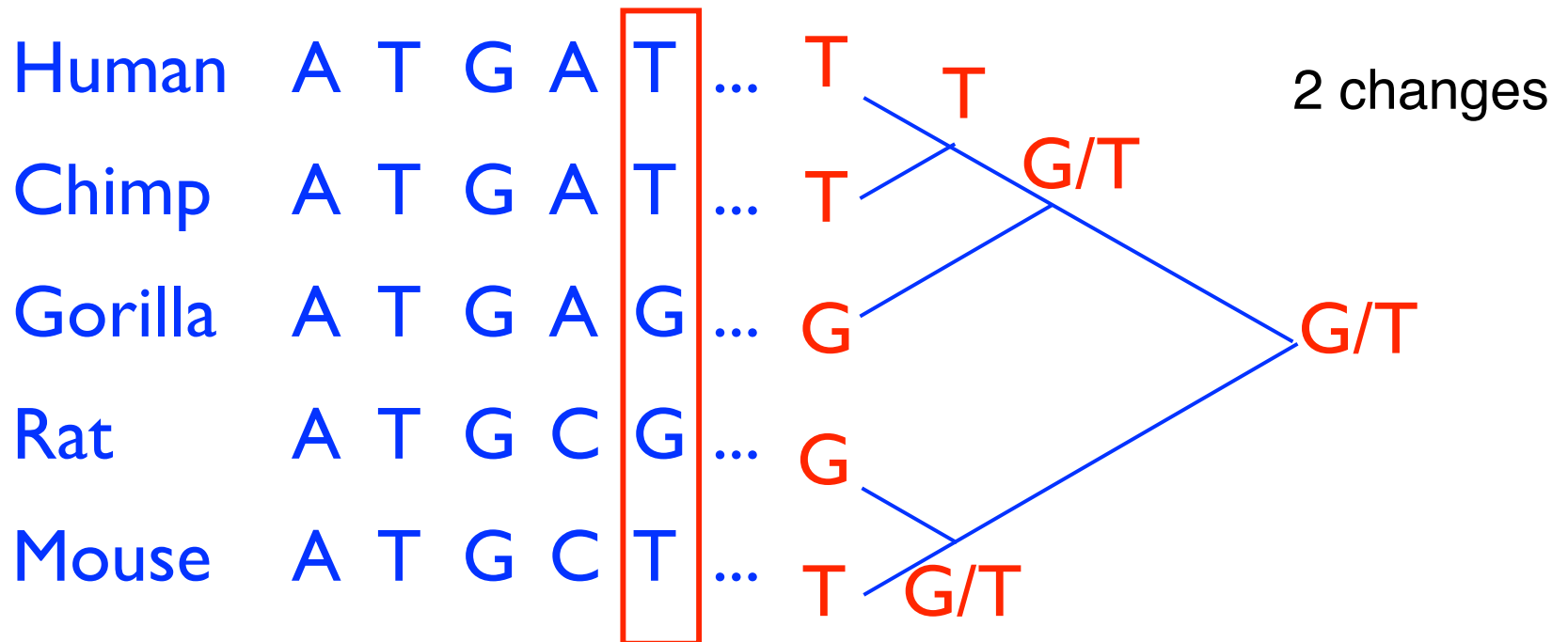
# Parsimony

General idea ~ Occam's Razor:  
Given data where change is rare, prefer  
an explanation that requires few events



# Parsimony

General idea ~ Occam's Razor:  
Given data where change is rare, prefer  
an explanation that requires few events



# Counting Events Parsimoniously

Lesson of example – no unique reconstruction

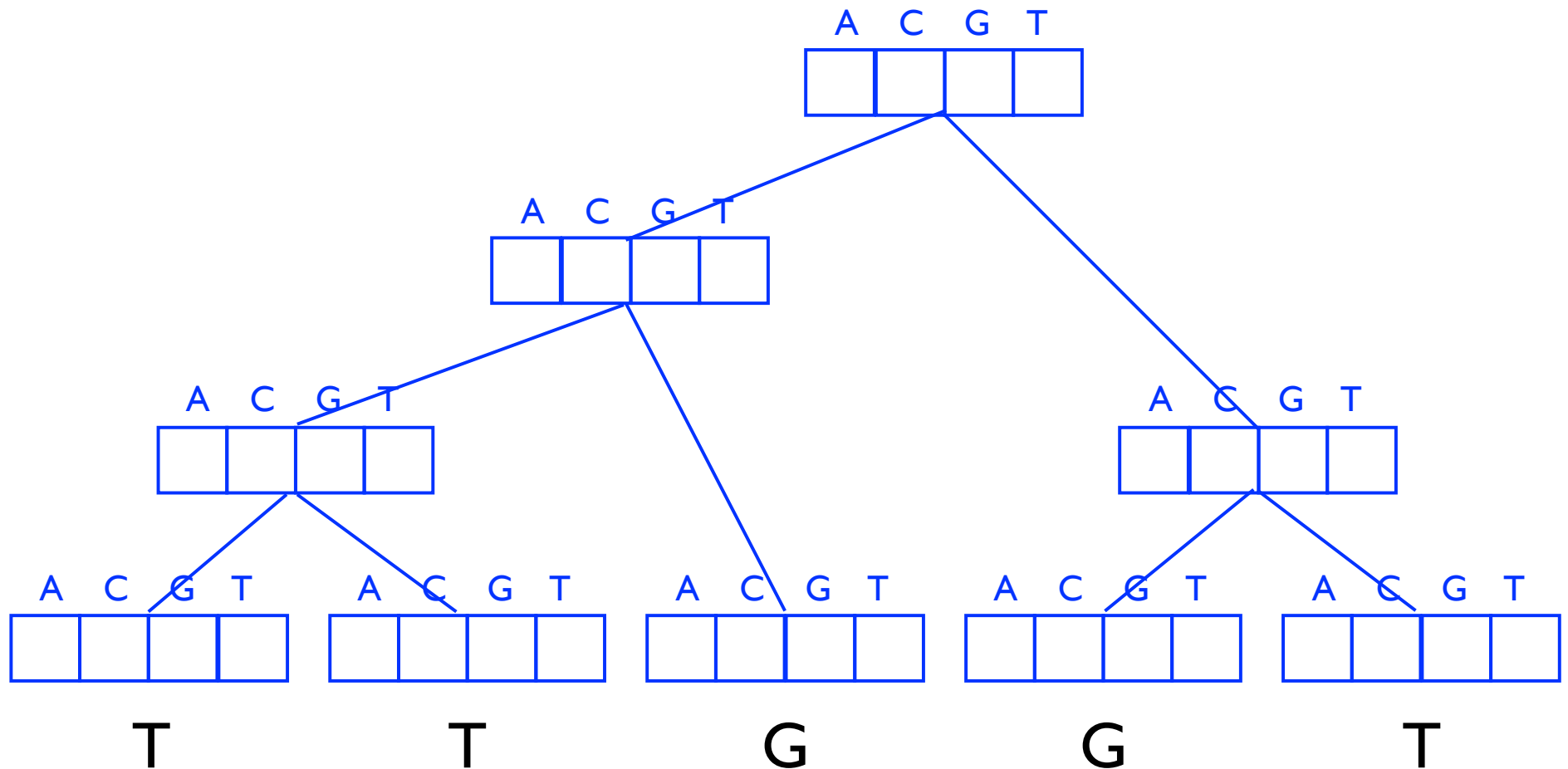
But there is a unique minimum number, of course

How to find it?

Early solutions 1965-75

# Sankoff & Rousseau, '75

$P_u(s)$  = best parsimony score of subtree rooted at node  $u$ , assuming  $u$  is labeled by character  $s$



# Sankoff-Rousseau Recurrence

$P_u(s)$  = best parsimony score of subtree rooted at node  $u$ , assuming  $u$  is labeled by character  $s$

For Leaf  $u$ :

$$P_u(s) = \begin{cases} 0 & \text{if } u \text{ is a leaf labeled } s \\ \infty & \text{if } u \text{ is a leaf not labeled } s \end{cases}$$

For Internal node  $u$ :

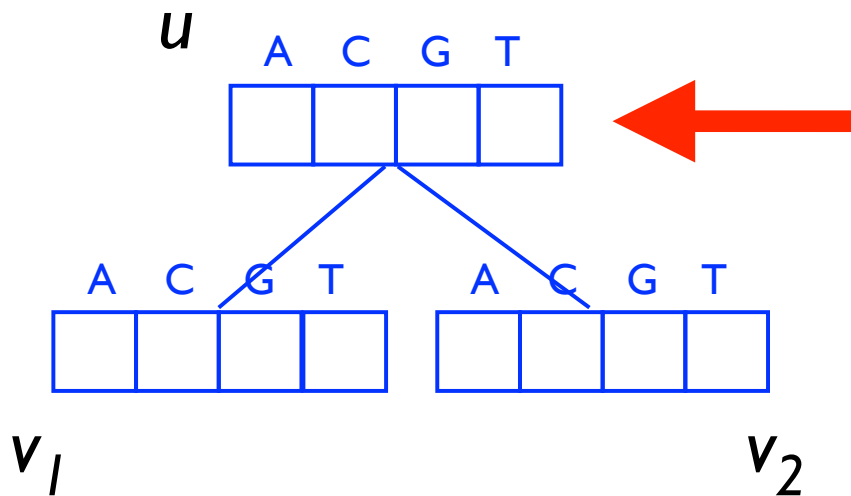
$$P_u(s) = \sum_{v \in \text{child}(u)} \min_{t \in \{A, C, G, T\}} \text{cost}(s, t) + P_v(t)$$

Time:  $O(\text{alphabet}^2 \times \text{tree size})$

# Sankoff & Rousseau, '75

$P_u(s)$  = best parsimony score of subtree rooted at node  $u$ , assuming  $u$  is labeled by character  $s$

$$P_u(s) = \sum_{v \in \text{child}(u)} \min_{t \in \{A, C, G, T\}} \text{cost}(s, t) + P_v(t)$$



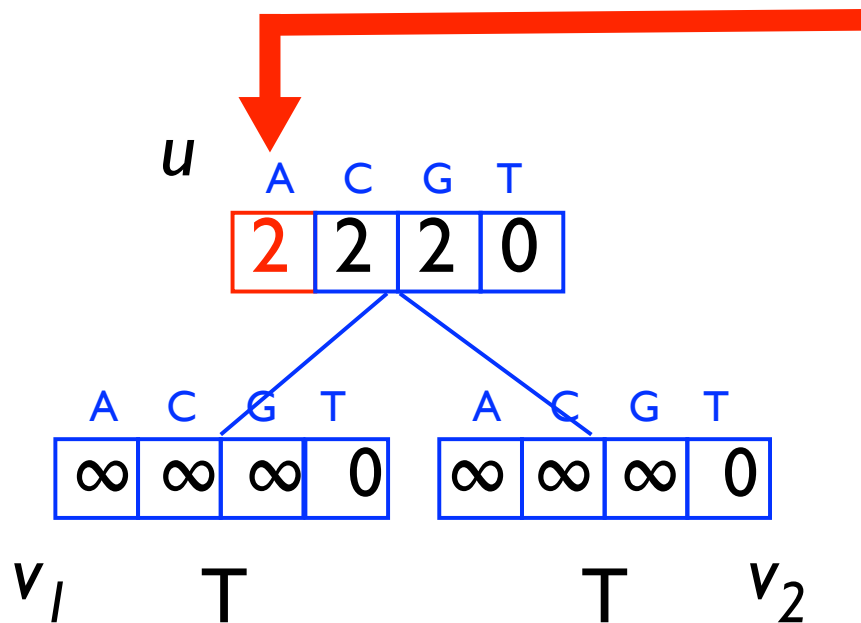
$s$	$v$	$t$	$\text{cost}(s, t) + P_v(t)$	min
	$v_1$	A		
		C		
		G		
		T		
	$v_2$	A		
		C		
		G		
		T		
sum: $P_u(s) =$				



# Sankoff & Rousseau, '75

$P_u(s)$  = best parsimony score of subtree rooted at node  $u$ , assuming  $u$  is labeled by character  $s$

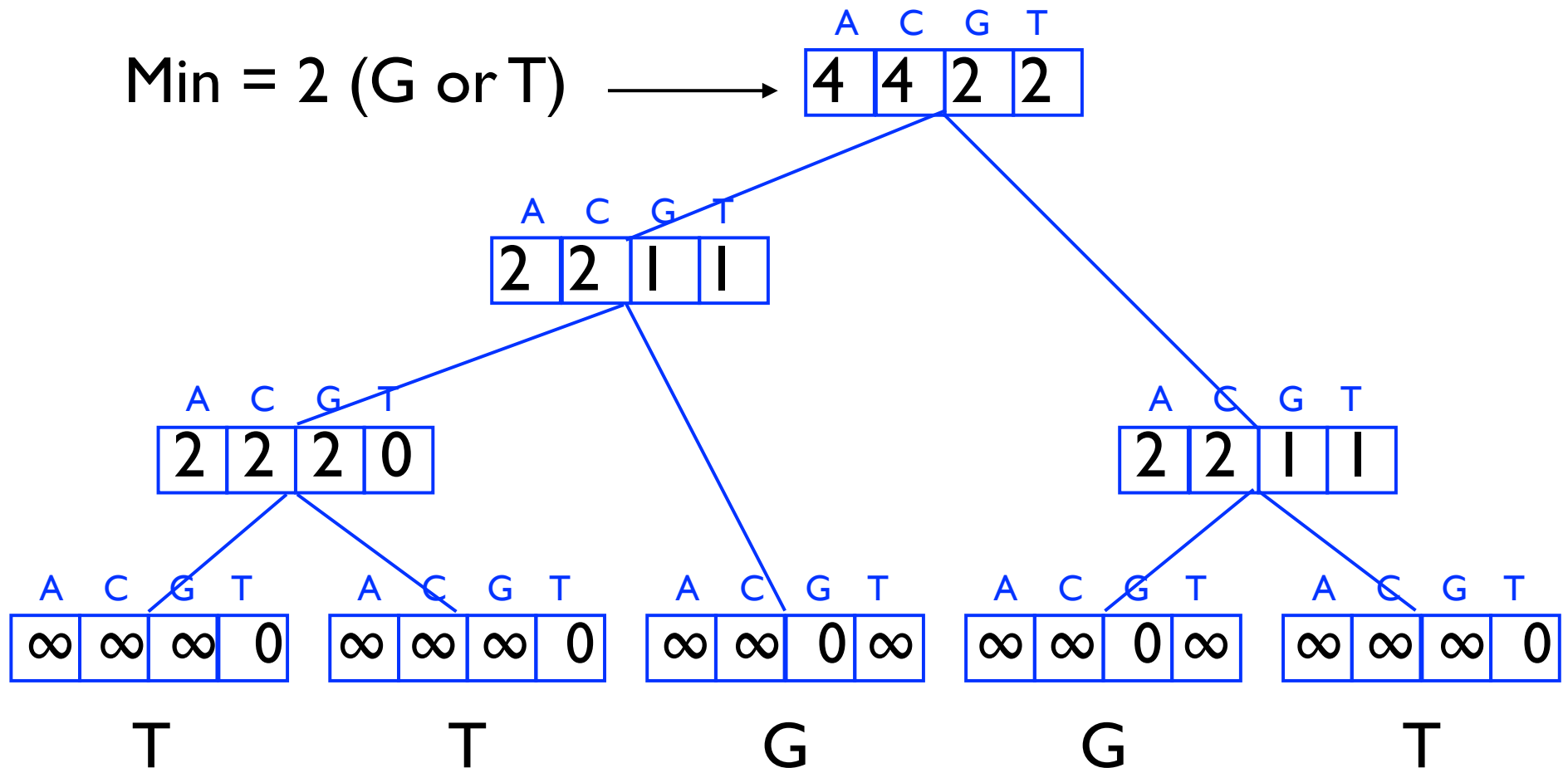
$$P_u(s) = \sum_{v \in \text{child}(u)} \min_{t \in \{A, C, G, T\}} \text{cost}(s, t) + P_v(t)$$



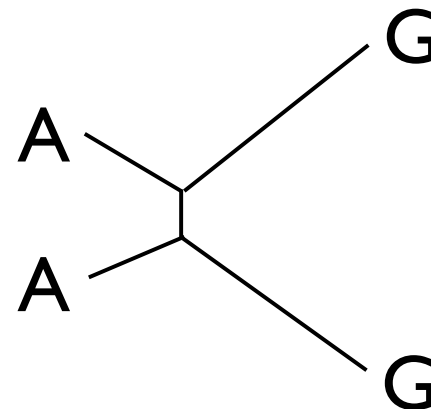
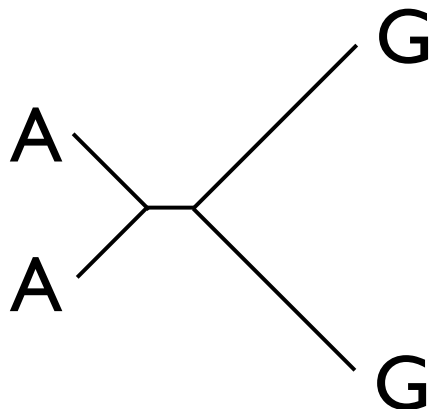
$s$	$v$	$t$	$\text{cost}(s, t) + P_v(t)$	min
A	$v_1$	A	$0 + \infty$	1
		C	$1 + \infty$	
		G	$1 + \infty$	
		T	$1 + 0$	
	$v_2$	A	$0 + \infty$	1
		C	$1 + \infty$	
		G	$1 + \infty$	
		T	$1 + 0$	
sum: $P_u(s) =$				2

# Sankoff & Rousseau, '75

$P_u(s)$  = best parsimony score of subtree rooted at node  $u$ , assuming  $u$  is labeled by character  $s$



# Which tree is better?



*Which has smaller parsimony score?*

*Which is more likely, assuming edge length proportional to evolutionary rate?*

# Parsimony – Generalities

Parsimony is not the best way to evaluate a phylogeny (maximum likelihood generally preferred - as previous slide suggests)

But it is a natural approach, works well in many cases, and is fast.

Finding the best tree: a much harder problem

Much is known about these problems; ***Inferring Phylogenies*** by Joe Felsenstein is a great resource.

# Phylogenetic Footprinting

A lovely extension of the above ideas. E.g., suppose promoters of orthologous genes in multiple species all contain (variants of) a common  $k$ -base transcription factor binding site. Roughly as above, but  $4^k$  table entries per node...

1. M Blanchette, B Schwikowski, M Tompa, [Algorithms for Phylogenetic Footprinting](#). *J Comp Biol*, vol. 9, no. 2, 2002, 211-223

2. M Blanchette and M Tompa, FootPrinter: a Program Designed for Phylogenetic Footprinting. *Nucleic Acids Research*, vol. 31, no. 13, July 2003, 3840-3842