

CSE 527

Lecture 11

Parsimony and
Phylogenetic Footprinting

Phylogenies (aka Evolutionary Trees)

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

-- Dobzhansky

- 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

Parsimony

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

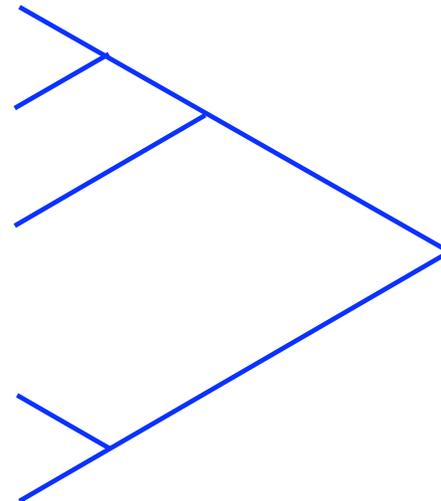
Human A T G A T ...

Chimp A T G A T ...

Gorilla A T G A G ...

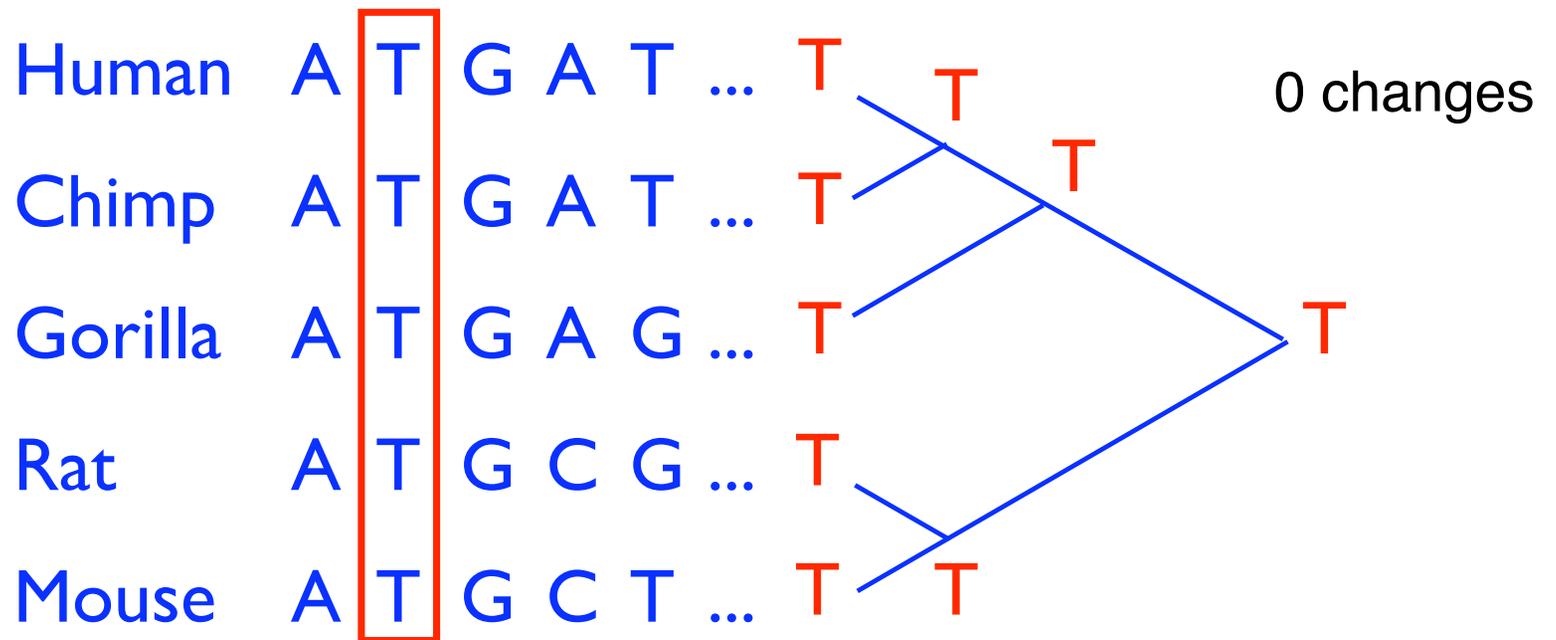
Rat A T G C G ...

Mouse A T G C T ...



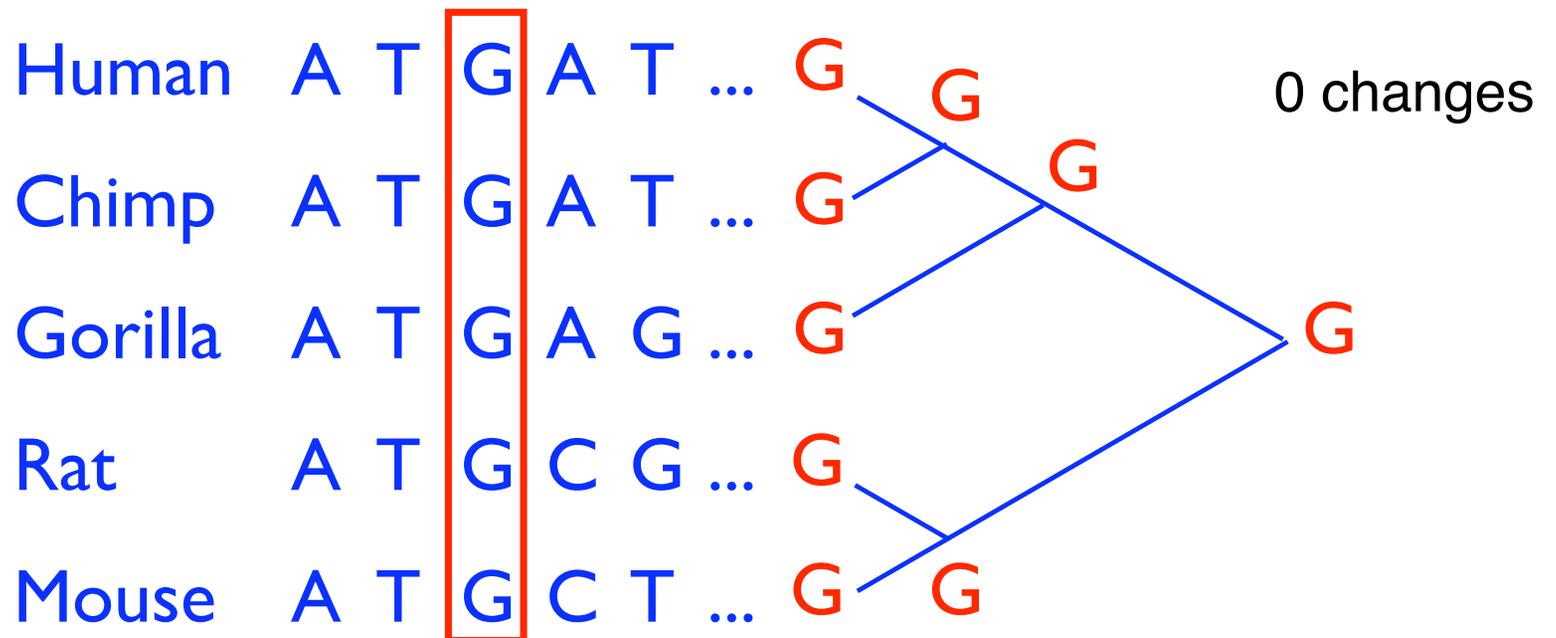
Parsimony

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



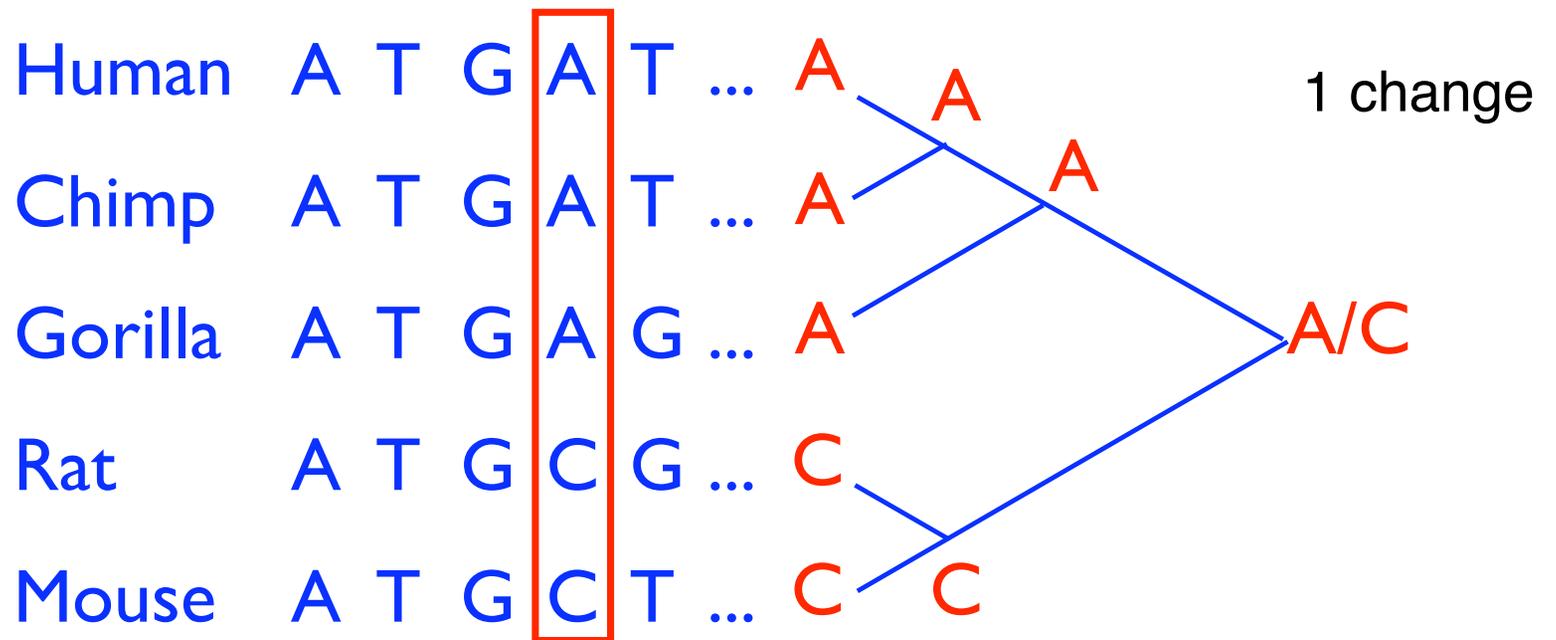
Parsimony

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



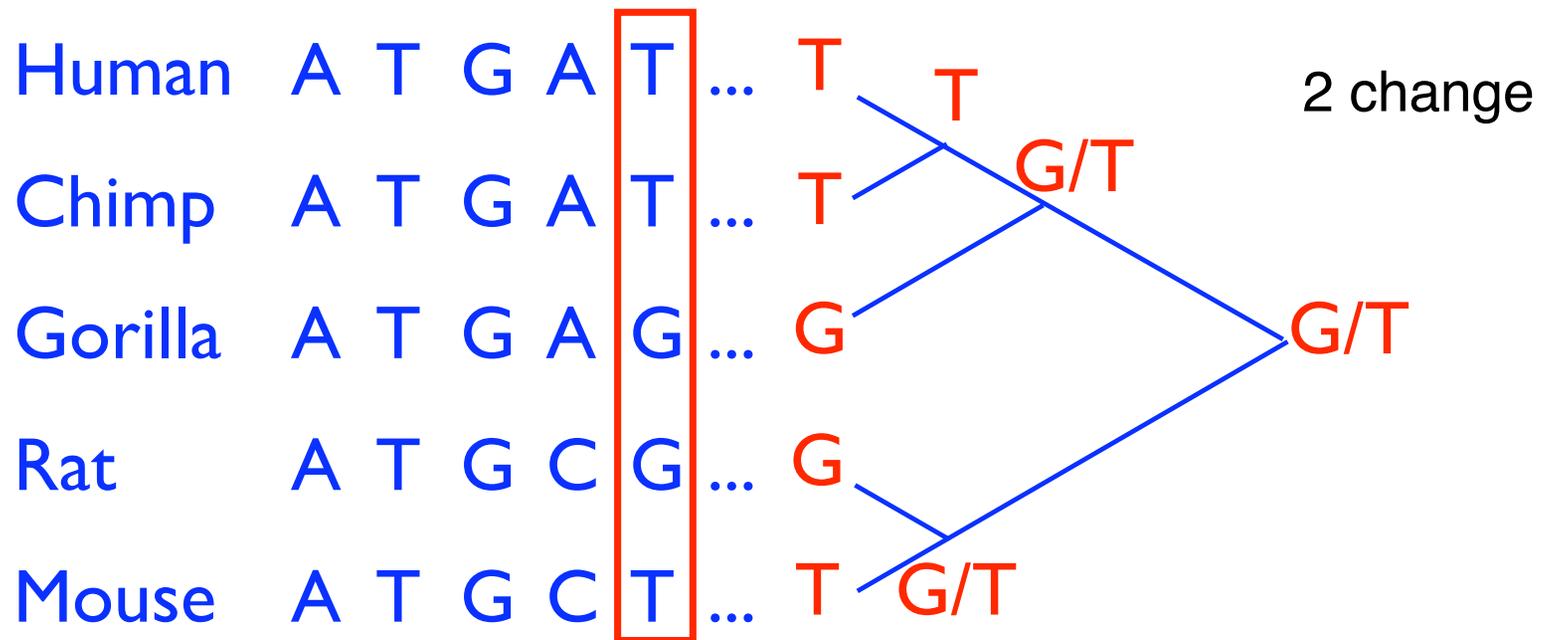
Parsimony

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

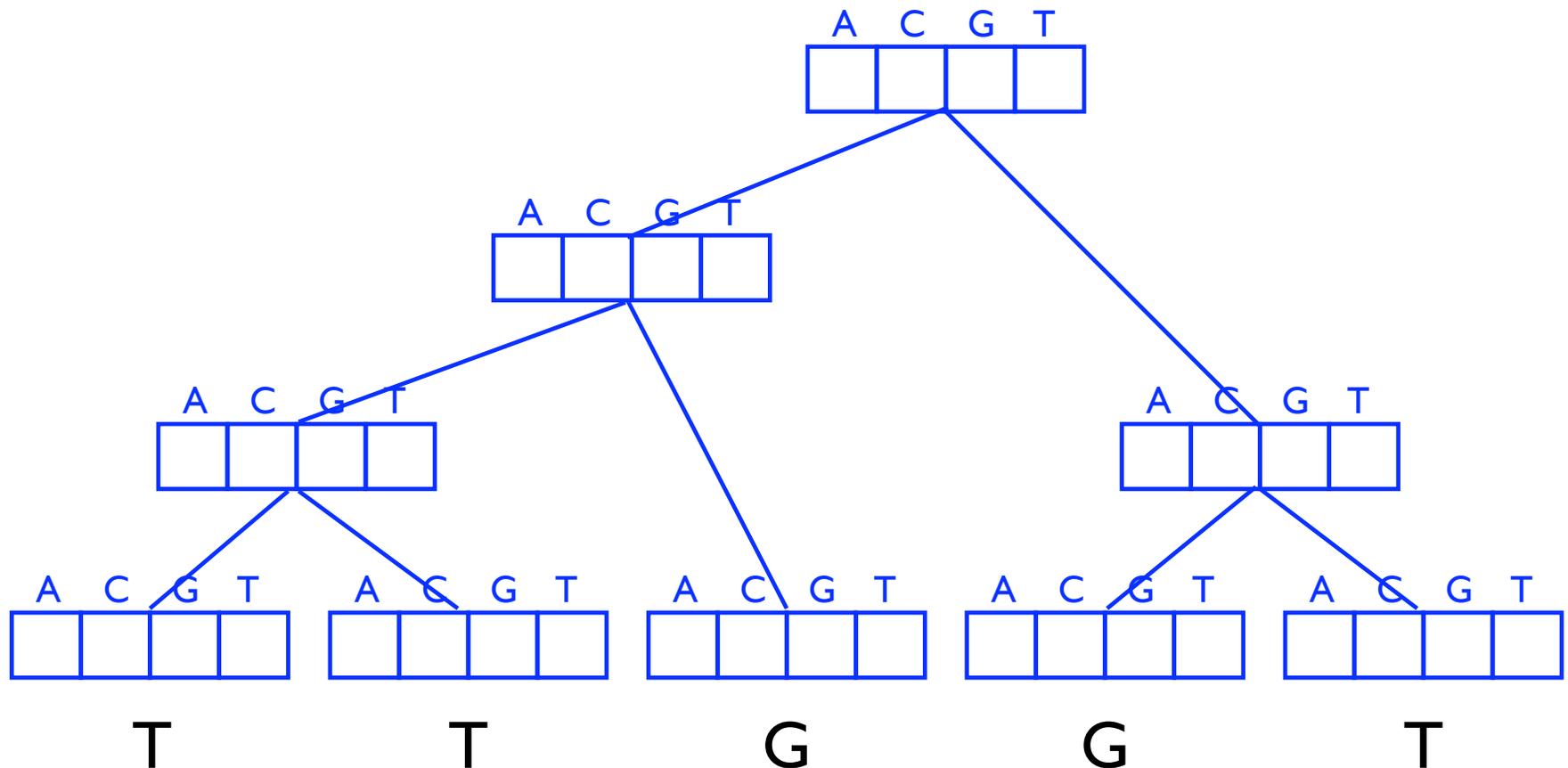


Counting Events Parsimoniously

- 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

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 :

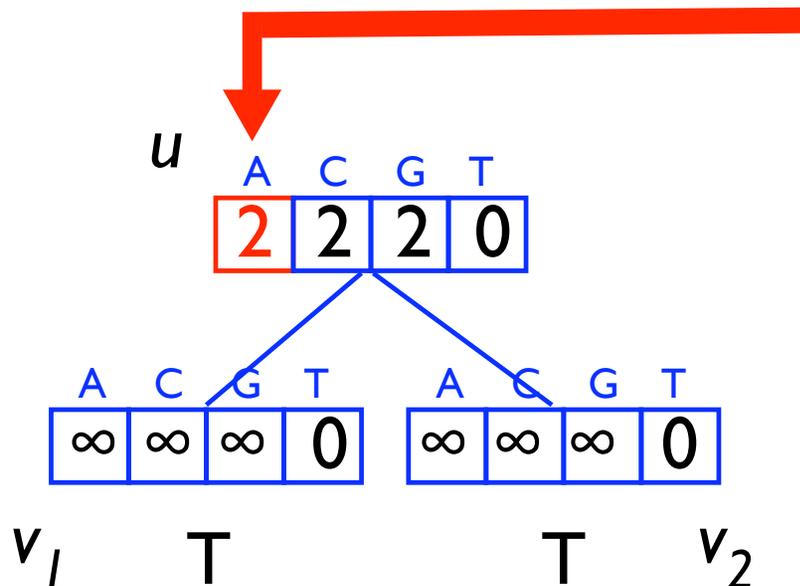
$$P_u(s) = \sum_{v \in \text{child}(u)} \min_{t \in \{A, C, G, T\}} \text{cost}(s, t) + P_v(t)$$

Time: linear in alphabet x tree size

Sankoff & Rousseau, '75

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

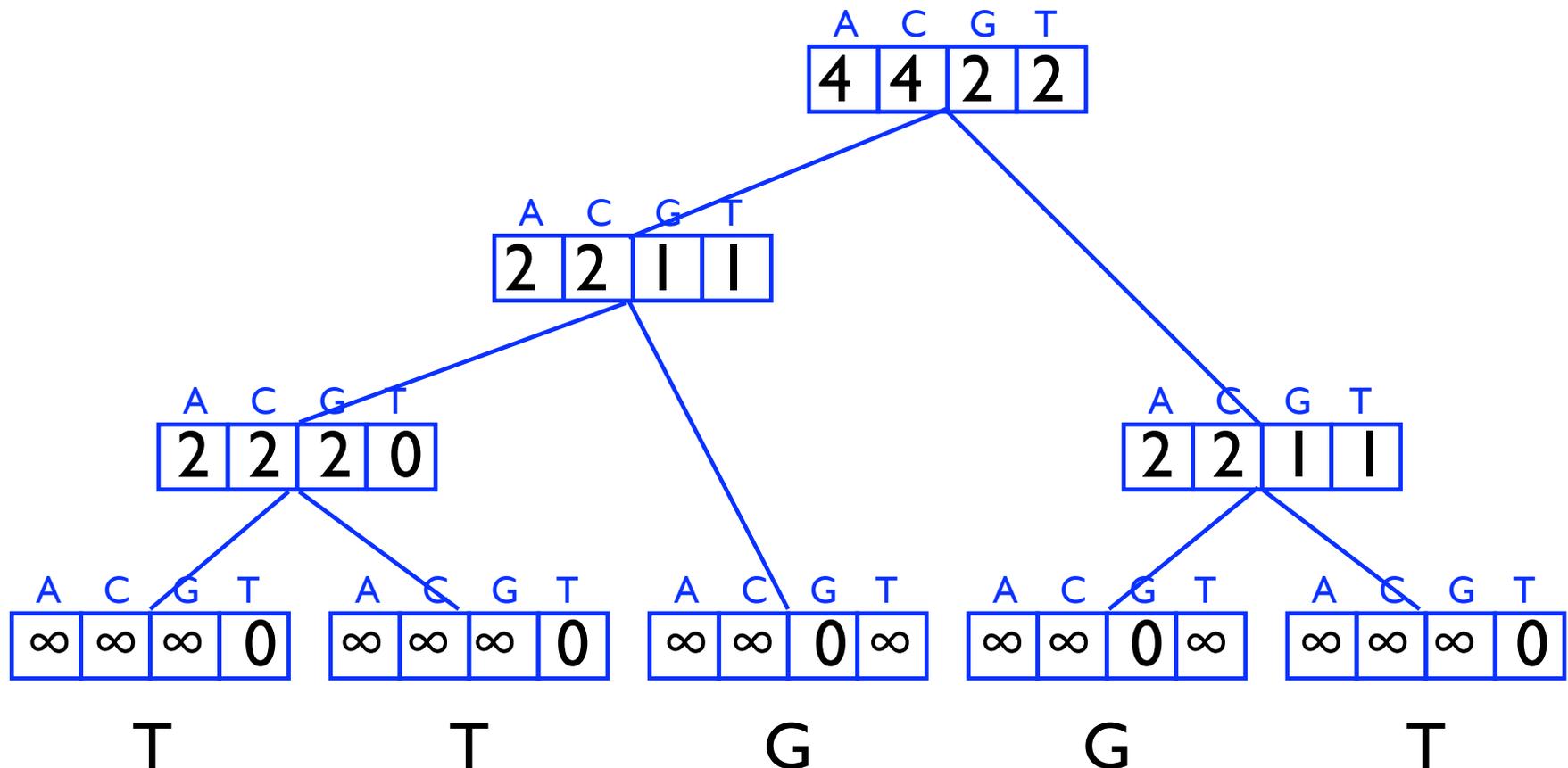
$$P_u(s) = \sum_{v \in \text{child}(u)} \min_{t \in \{A,C,G,T\}} \text{cost}(s, t) + P_v(t)$$



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



Parsimony – Generalities

- Parsimony is not necessarily the best way to evaluate a phylogeny (maximum likelihood generally preferred)
- But it is a natural approach, & 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

See link to Tompa's slides on course web page
<http://www.cs.washington.edu/homes/tompa/papers/ortho.ppt>