CSEP 527
Spring 2016

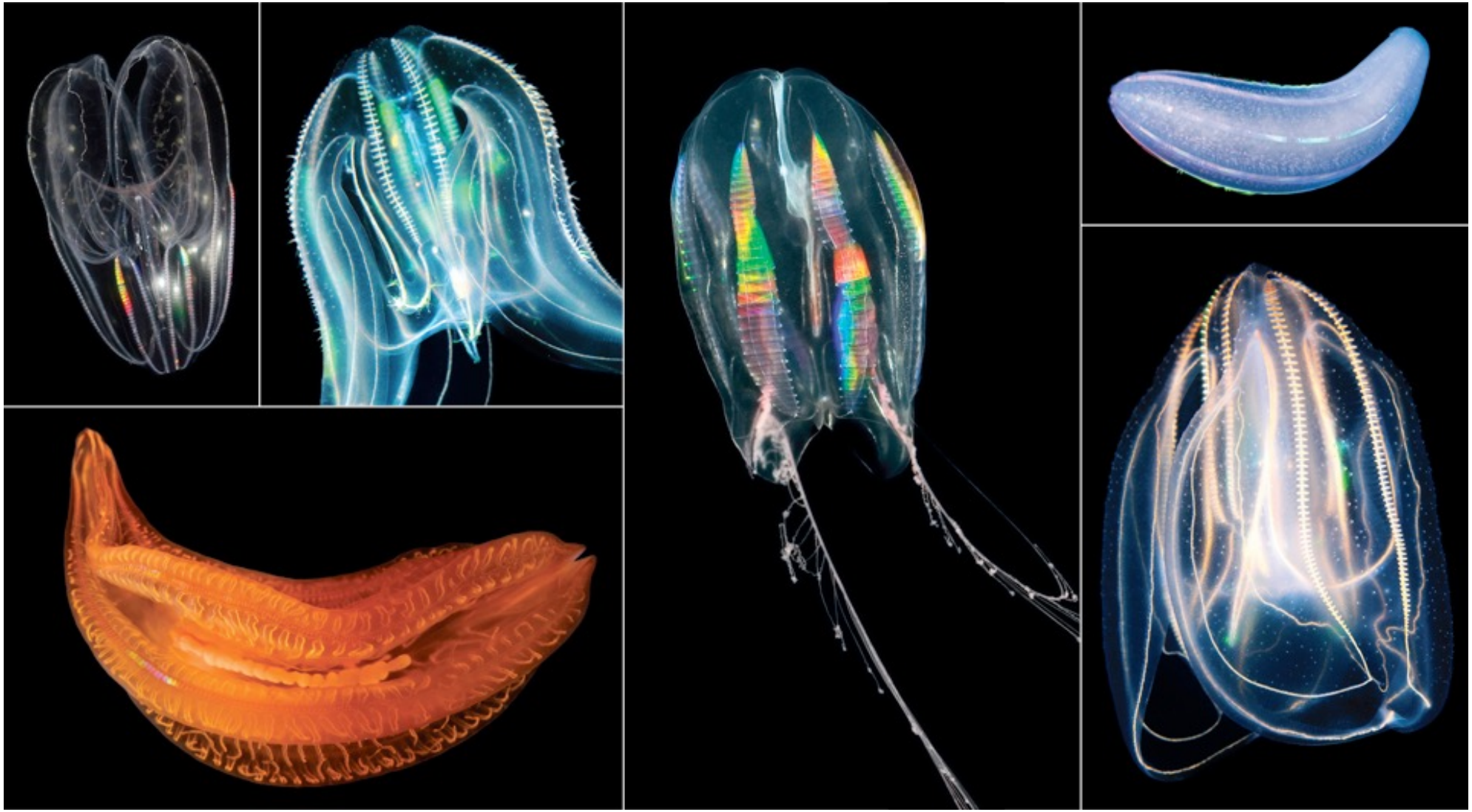
Phylogenies: Parsimony Plus a Tantalizing Taste of Likelihood

Phylogenies
(aka Evolutionary Trees)

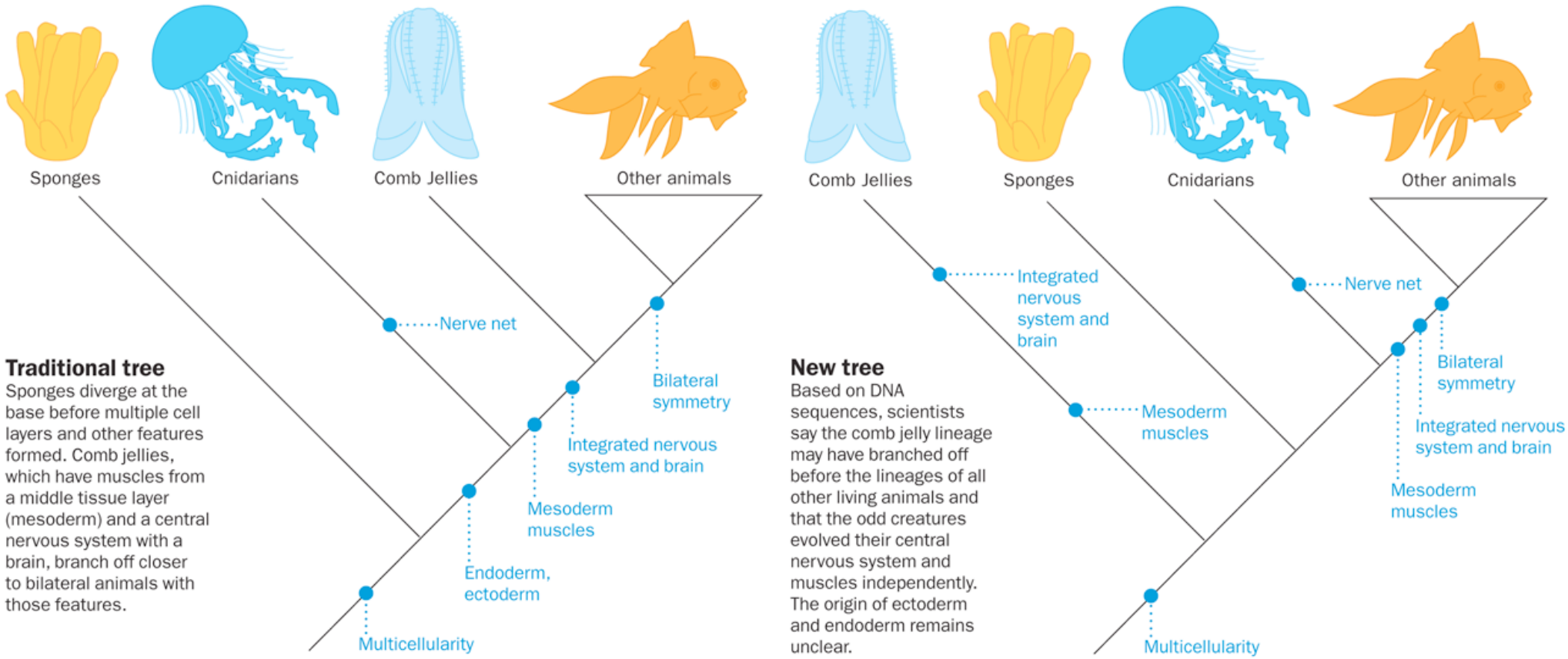
"Nothing in biology makes sense, except in the light of evolution"

-- Theodosius Dobzhansky, 1973

Comb Jellies: Evolutionary enigma



http://www.sciencenews.org/view/feature/id/350120/description/Evolutionary_enigmas



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

A Complex Question:

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

A Simpler Question:

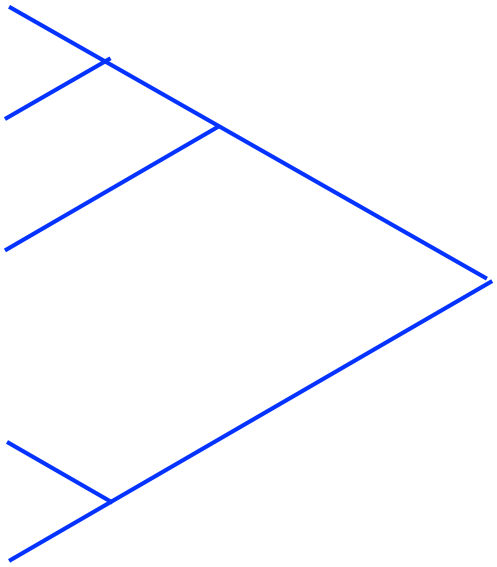
Given data *and a phylogeny*, evaluate "how much change" is needed to fit data to tree

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

Parsimony

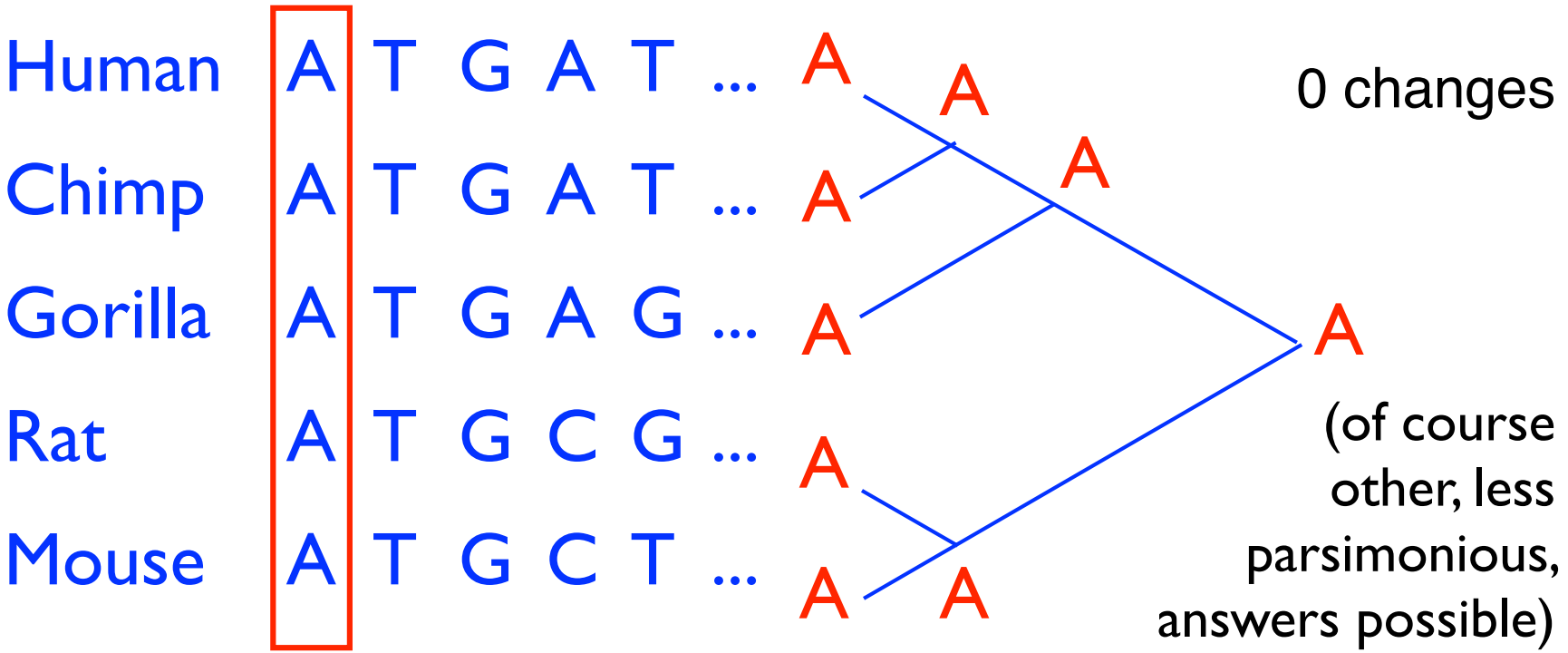
General idea ~ Occam's Razor:
Given data where change is rare, prefer an explanation that requires few events

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	...



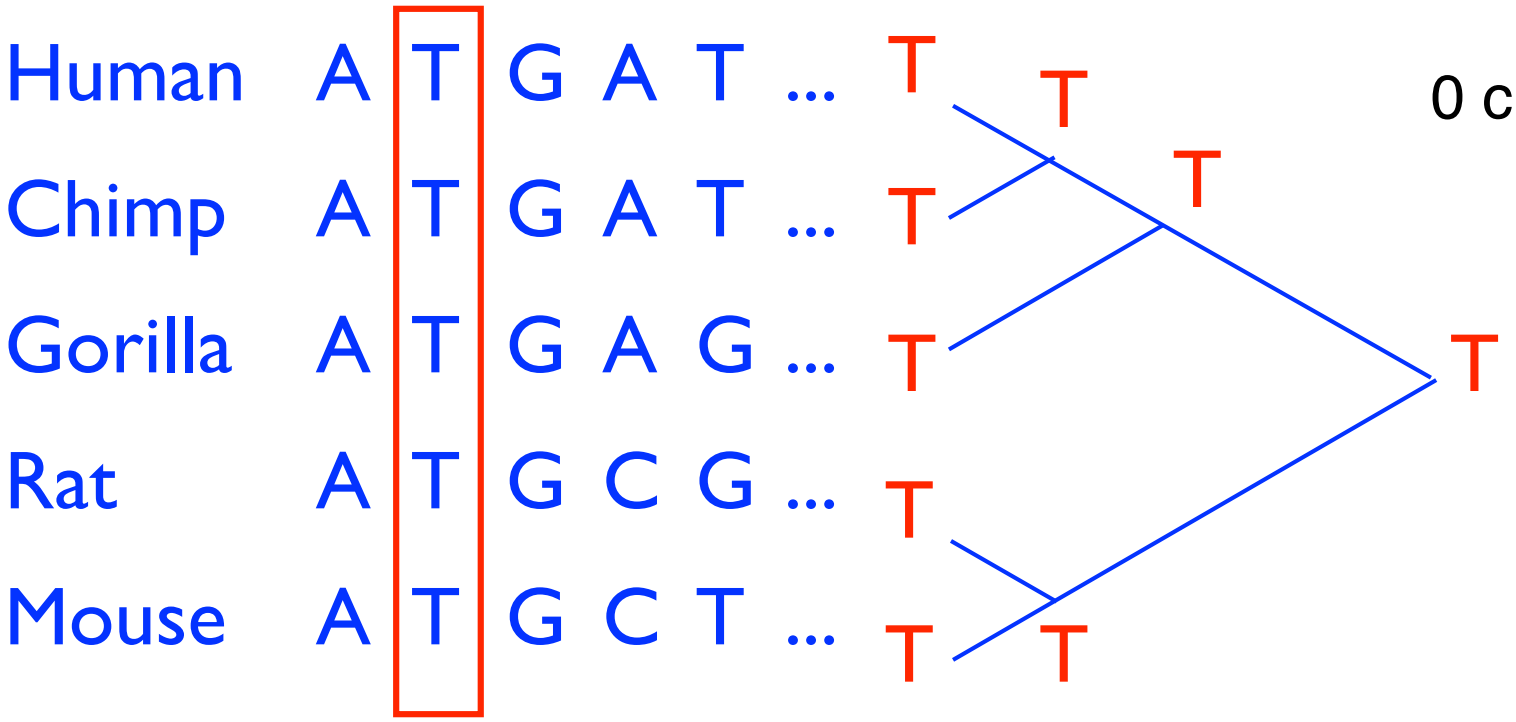
Parsimony

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



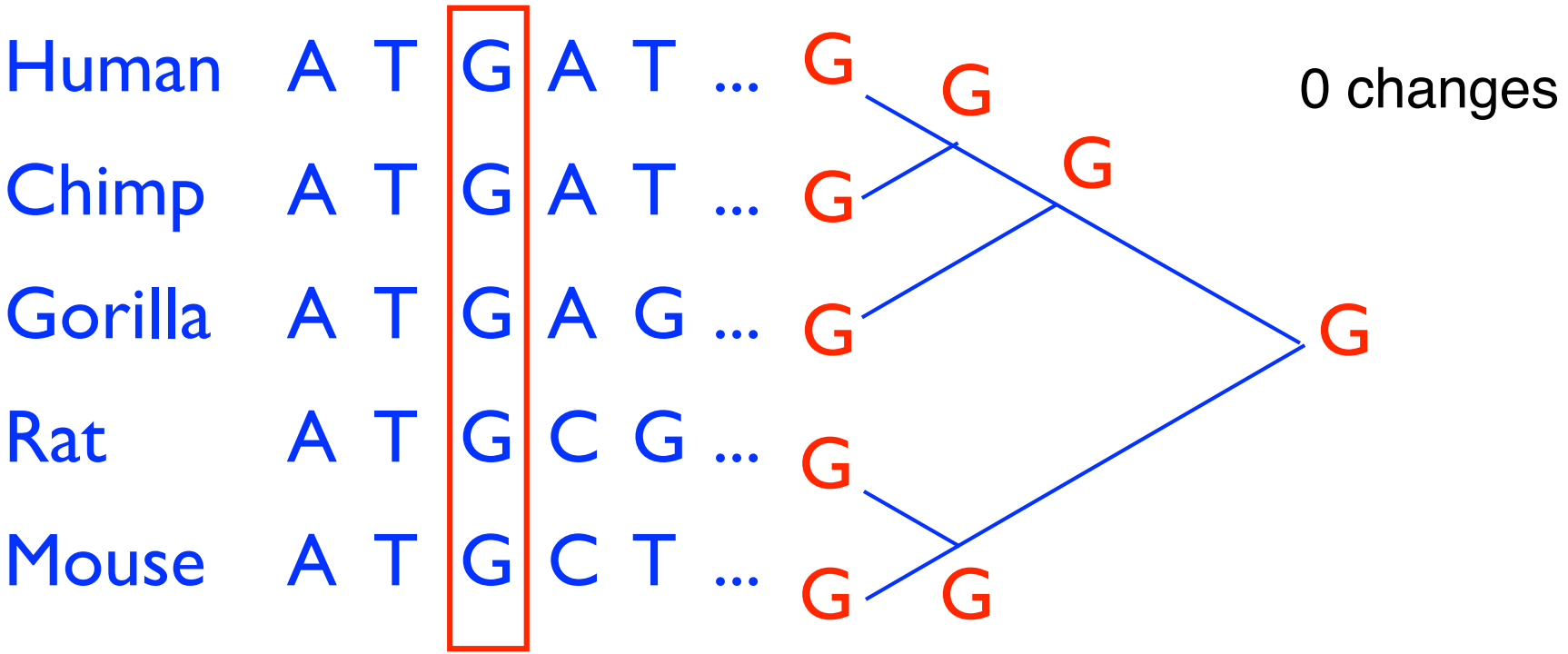
Parsimony

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



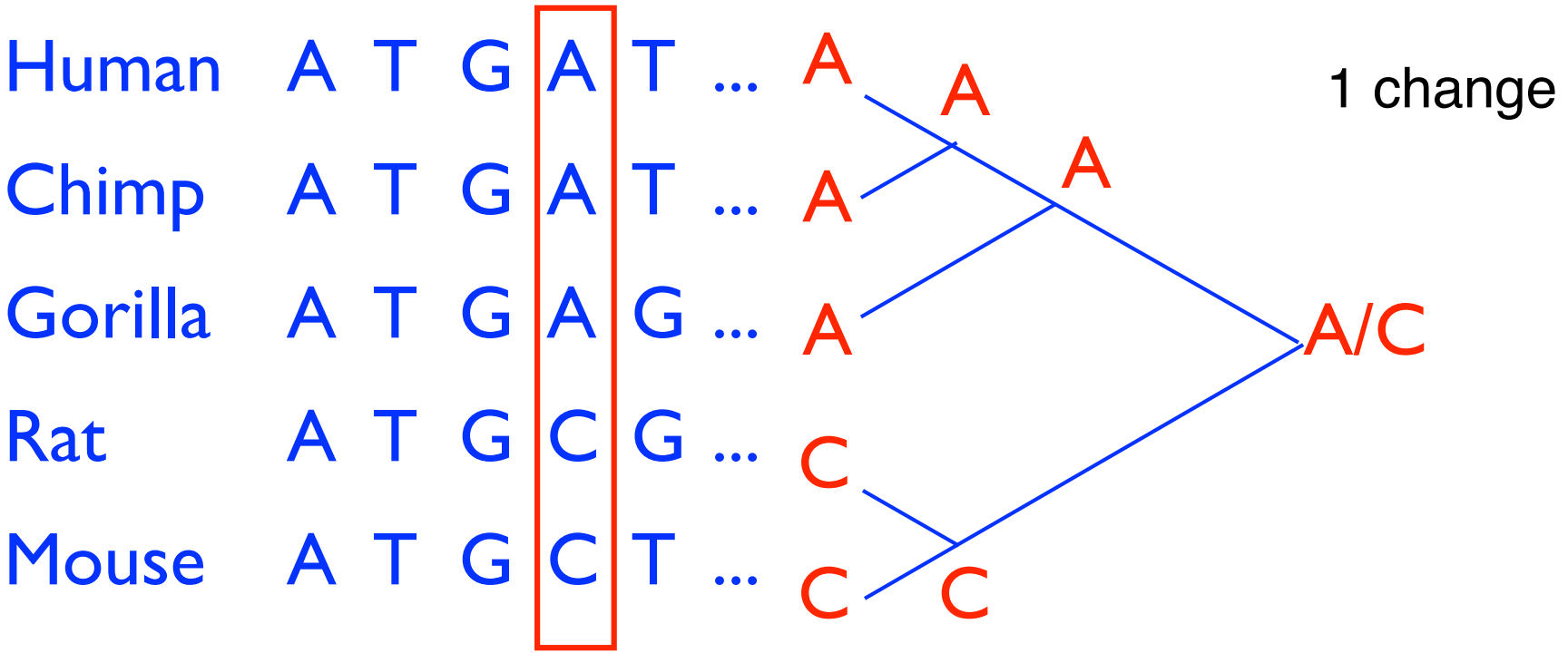
Parsimony

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



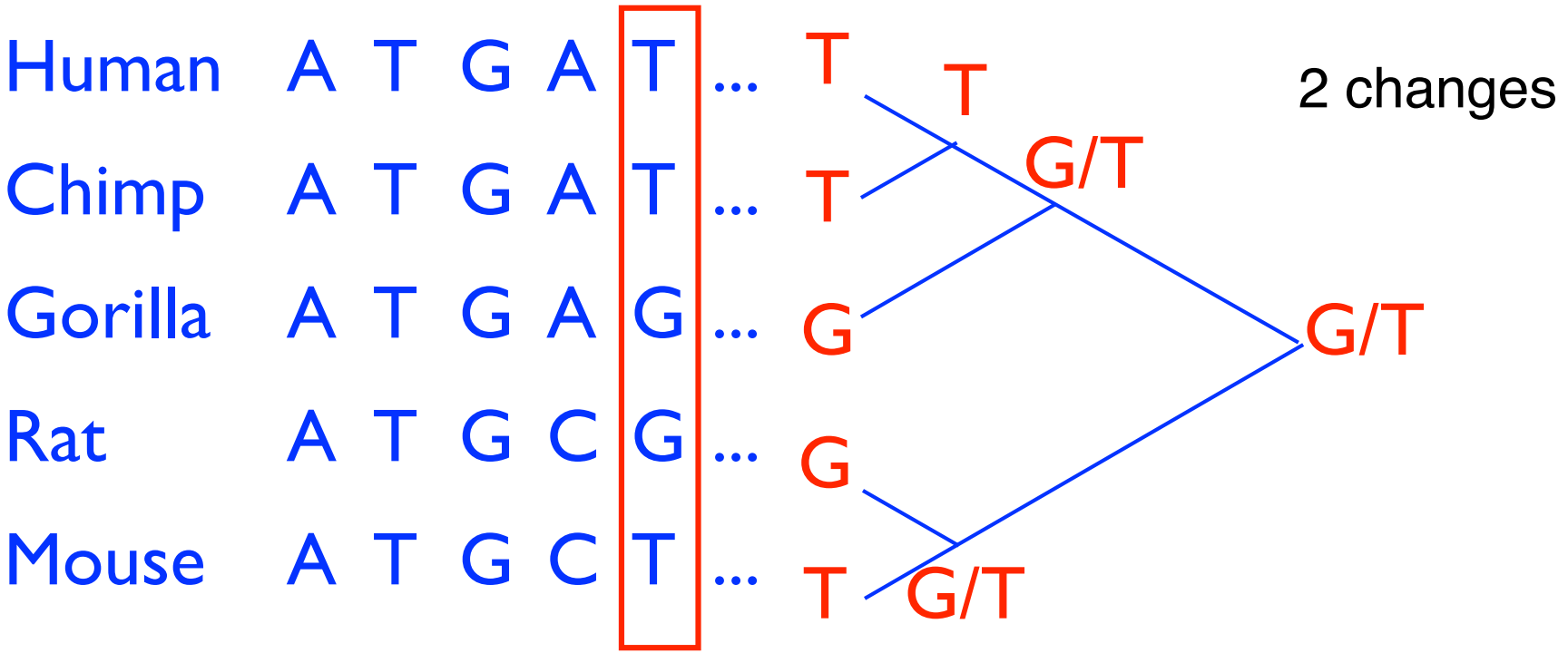
Parsimony

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

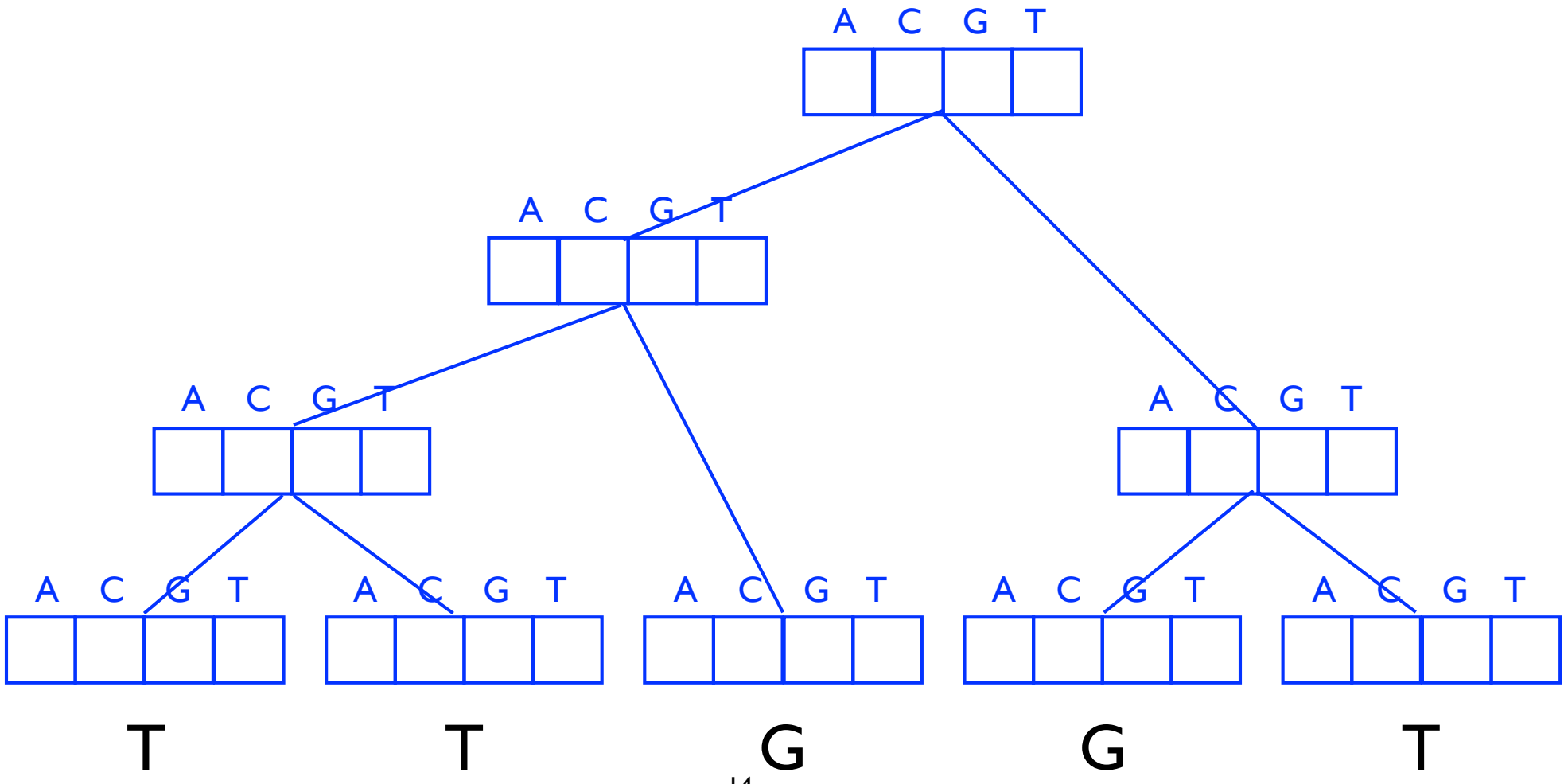
But there is a unique minimum number, of course

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

$P_u(s)$ = 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:

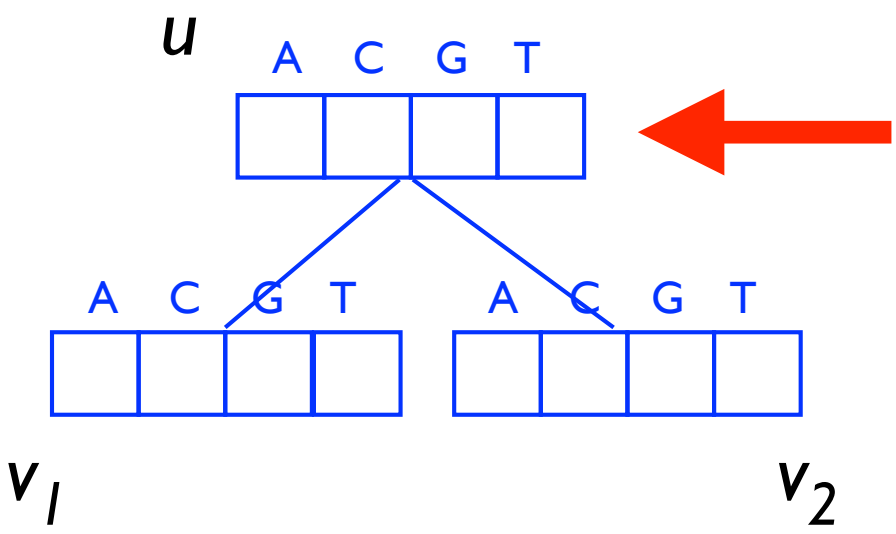
$$P_u(s) = \sum_{v \in \text{child}(u)} \min_{t \in \{A,C,G,T\}} \text{cost}(s,t) + P_v(t)$$

Time: O(alphabet^2 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

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

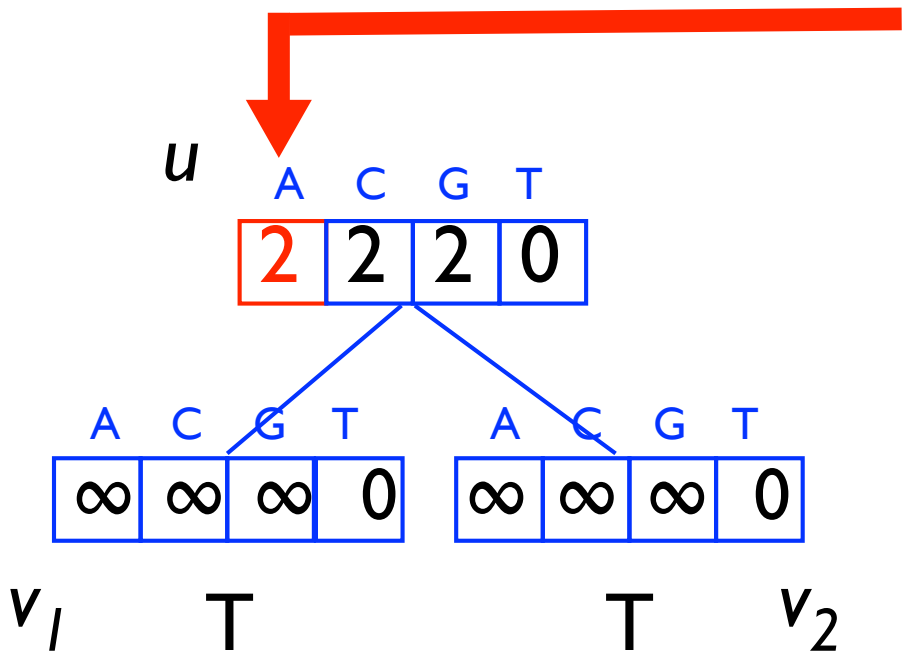


s	v	t	cost(s,t)+P_v(t)	min
	v_1	A		
		C		
		G		
		T		
	v_2	A		
		C		
		G		
		T		
			sum: $P_u(s)$ =	

Sankoff & Rousseau, '75

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

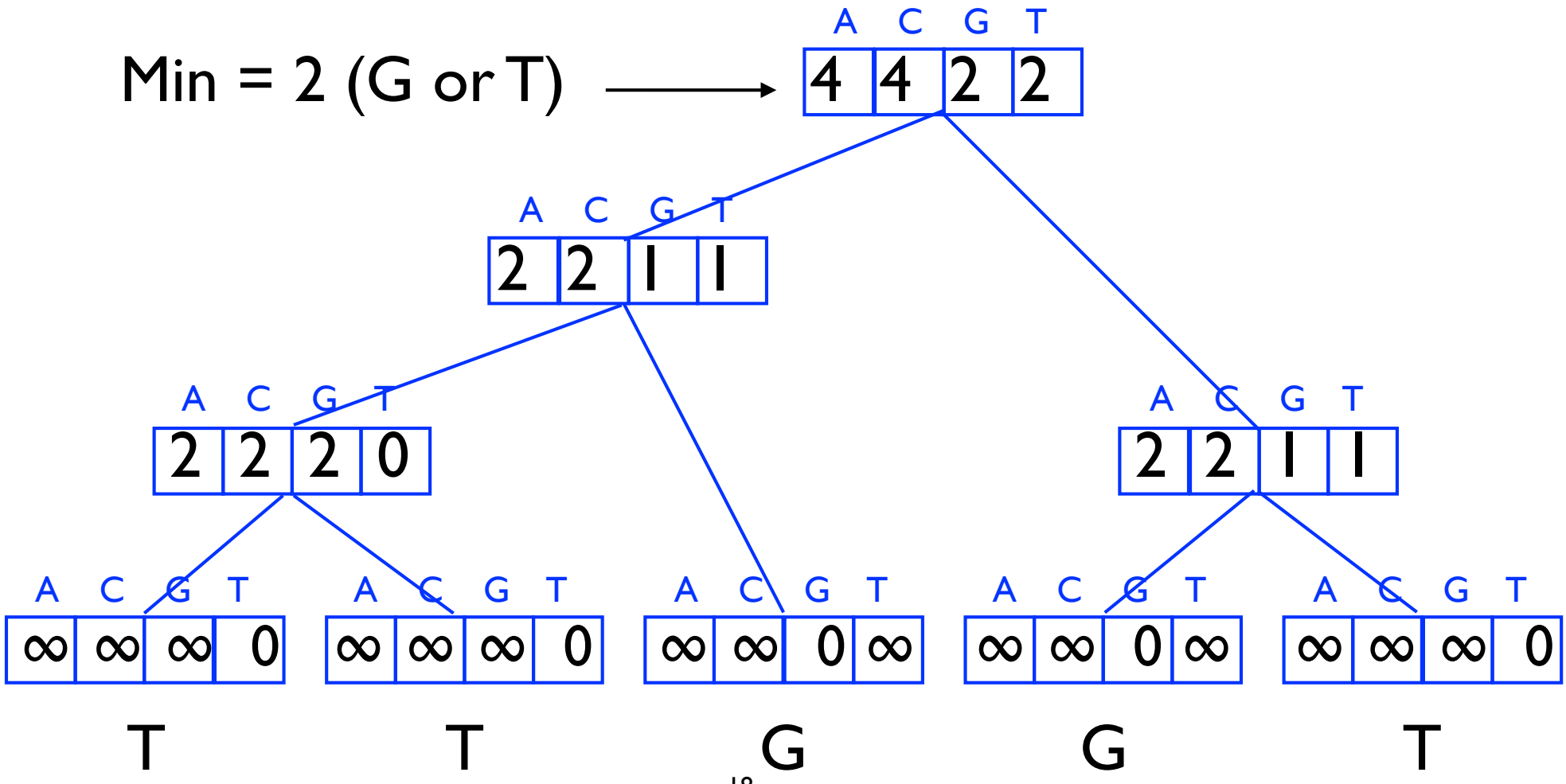
$$P_u(s) = \sum_{v \in \text{child}(u)} \min_{t \in \{A,C,G,T\}} \text{cost}(s,t) + P_v(t)$$



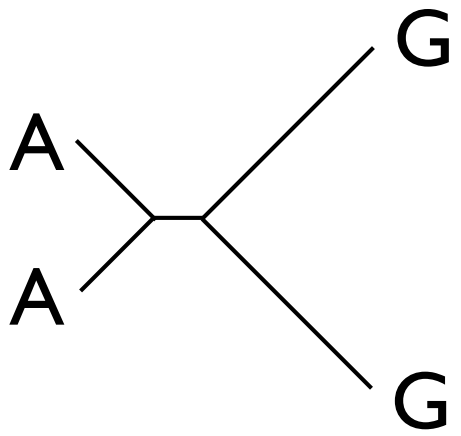
s	v	t	cost(s,t)+P_v(t)	min
A	v_1	A	0 + ∞	1
		C	1 + ∞	
		G	1 + ∞	
		T	1 + 0	
	v_2	A	0 + ∞	1
		C	1 + ∞	
		G	1 + ∞	
		T	1 + 0	
			sum: P_u(s) =	2

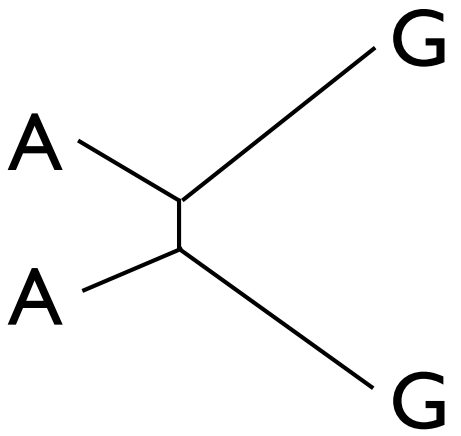
Sankoff & Rousseau, '75

$P_u(s)$ = 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

Small Example

Size of motif sought: $k = 4$

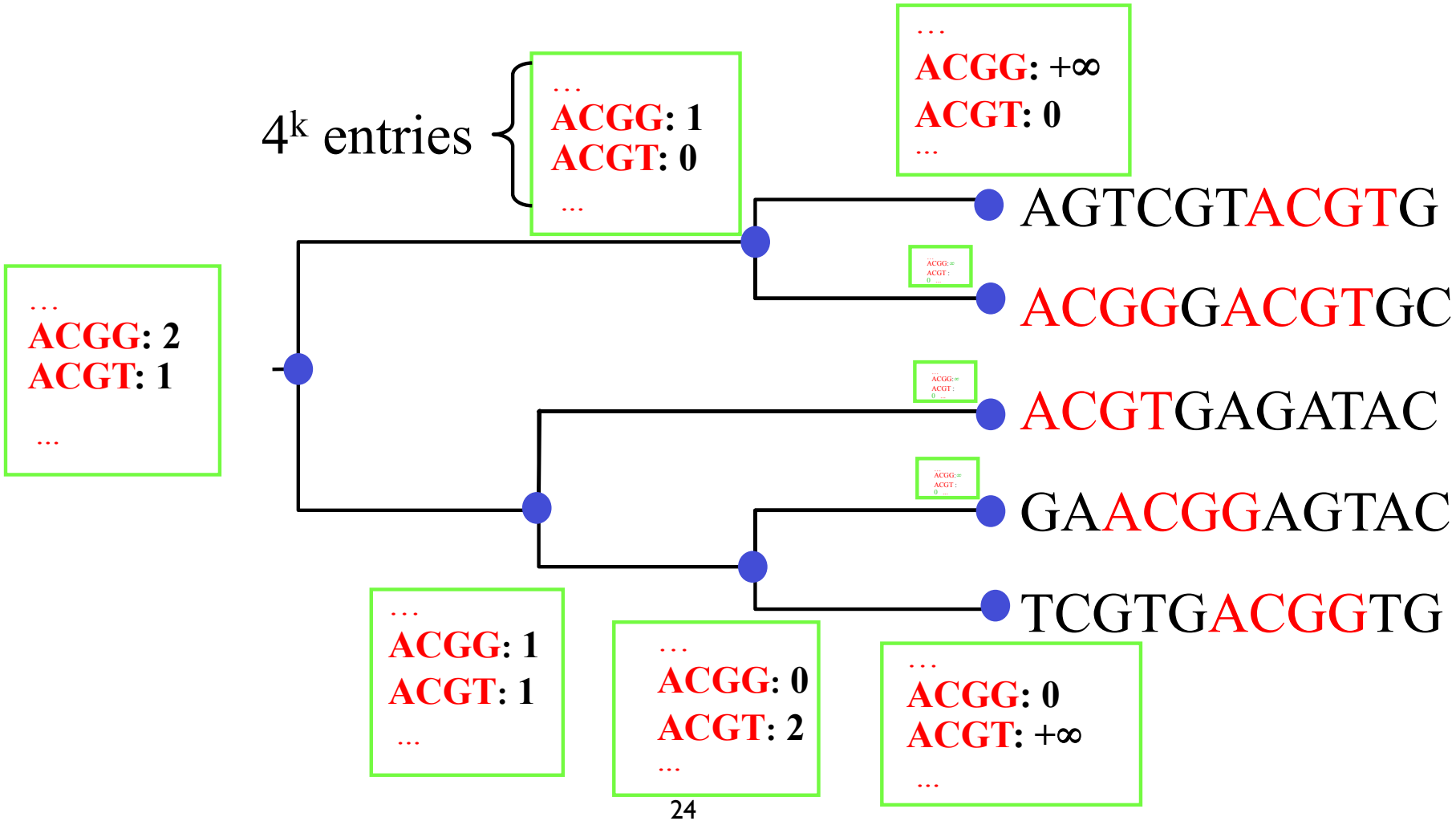
CLUSTALW multiple sequence alignment (rbcS gene)

```
Cotton      ACGGTT-TCCATTGGATGA---AATGAGATAAGAT---CACTGTGC---TTCTTCCACGTG--GCAGGTTGCCAAAGATA-------AGGCTTTACCATT
Pea         GTTTTT-TCAGTTAGCTTA---GTGGGCATCTTA----CACGTGGC---ATTATTATCCTA--TT-GGTGGCTAATGATA-------AGG--TTAGCACA
Tobacco     TAGGAT-GAGATAAGATTA---CTGAGGTGCTTTA---CACGTGGC---ACCTCCATTGTG--GT-GACTTAAATGAAGA-------ATGGCTTAGCACC
Ice-plant   TCCCAT-ACATTGACATAT---ATGGCCCGCCTGCGGCAACAAAAA---AACTAAAGGATA--GCTAGTTGCTACTACAATTC--CCATAACTCACCACC
Turnip      ATTCAT-ATAAATAGAAGG---TCCGCGAACATTG--AAATGTAGATCATGCGTCAGAATT--GTCCTCTCTTAATAGGA-------A-------GGAGC
Wheat       TATGAT-AAAATGAAATAT---TTTGCCCAGCCA-----ACTCAGTCGCATCCTCGGACAA--TTTGTTATCAAGGAACTCAC--CCAAAAACAAGCAAA
Duckweed    TCGGAT-GGGGGGGCATGAACACTTGCAATCATT-----TCATGACTCATTTCTGAACATGT-GCCCTTGGCAACGTGTAGACTGCCAACATTAATTAAA
Larch       TAACAT-ATGATATAACAC---CGGGCACACATTCCTAAACAAAGAGTGATTTCAAATATATCGTTAATTACGACTAACAAAA--TGAAAGTACAAGACC

Cotton      CAAGAAAAGTTTCCACCCTC------TTTGTGGTCATAATG-GTT-GTAATGTC-ATCTGATTT----AGGATCCAACGTCACCCTTTCTCCCA-----A
Pea         C---AAAACTTTTCAATCT-------TGTGTGGTTAATATG-ACT-GCAAAGTTTATCATTTTC----ACAATCCAACAA-ACTGGTTCT---------A
Tobacco     AAAAATAATTTTCCAACCTTT---CATGTGTGGATATTAAG-ATTTGTATAATGTATCAAGAACC-ACATAATCCAATGGTTAGCTTTATTCCAAGATGA
Ice-plant   ATCACACATTCTTCCATTTCATCCCCTTTTTCTTGGATGAG-ATAAGATATGGGTTCCTGCCAC----GTGGCACCATACCATGGTTTGTTA-ACGATAA
Turnip      CAAAAGCATTGGCTCAAGTTG-----AGACGAGTAACCATACACATTCATACGTTTTCTTACAAG-ATAAGATAAGATAATGTTATTTCT---------A
Wheat       GCTAGAAAAAGGTTGTGTGGCAGCCACCTAATGACATGAAGGACT-GAAATTTCCAGCACACACA-A-TGTATCCGACGGCAATGCTTCTTC--------
Duckweed    ATATAATATTAGAAAAAAATC-----TCCCATAGTATTTAGTATTTACCAAAAGTCACACGACCA-CTAGACTCCAATTTACCCAAATCACTAACCAATT
Larch       TTCTCGTATAAGGCCACCA-------TTGGTAGACACGTAGTATGCTAAATATGCACCACACACA-CTATCAGATATGGTAGTGGGATCTG--ACGGTCA

Cotton      ACCAATCTCT---AAATGTT----GTGAGCT---TAG-GCCAAATTT-TATGACTATA--TAT----AGGGGATTGCACC----AAGGCAGTG-ACACTA
Pea         GGCAGTGGCC---AACTAC-------------------CACAATTT-TAAGACCATAA-TAT----TGGAAATAGAA------AAATCAAT--ACATTA
Tobacco     GGGGGTTGTT---GATTTTT----GTCCGTTAGATAT-GCGAAATATGTAAAACCTTAT-CAT----TATATATAGAG------TGGTGGGCA-ACGATG
Ice-plant   GGCTCTTAATCAAAAGTTTTAGGTGTGAATTTAGTTT-GATGAGTTTTAAGGTCCTTAT-TATA---TATAGGAAGGGGG----TGCTATGGA-GCAAGG
Turnip      CACCTTTCTTTAATCCTGTGGCAGTTAACGACGATATCATGAAATCTTGATCCTTCGAT-CATTAGGGCTTCATACCTCT----TGCGCTTCTCACTATA
Wheat       CACTGATCCGGAGAAGATAAGGAAACGAGGCAACCAGCGAACGTGAGCCATCCCAACCA-CATCTGTACCAAAGAAACGG----GGCTATATATACCGTG
Duckweed    TTAGGTTGAATGGAAAATAG---AACGCAATAATGTCCGACATATTTCCTATATTTCCG-TTTTTCGAGAGAAGGCCTGTGTACCGATAAGGATGTAATC
Larch       CGCTTCTCCTCTGGAGTTATCCGATTGTAATCCTTGCAGTCCAATTTCTCTGGTCTGGC-CCA----ACCTTAGAGATTG----GGGCTTATA-TCTATA

Cotton      T-TAAGGGATCAGTGAGAC-TCTTTTGTATAACTGTAGCAT--ATAGTAC
Pea         TATAAAGCAAGTTTTAGTA-CAAGCTTTGCAATTCAACCAC--A-AGAAC
Tobacco     CATAGACCATCTTGGAAGT-TTAAAGGGAAAAAAGGAAAAG--GGAGAAA
Ice-plant   TCCTCATCAAAAGGGAAGTGTTTTTTCTCTAACTATATTACTAAGAGTAC
Larch       TCTTCTTCACAC---AATCCATTTGTGTAGAGCCGCTGGAAGGTAAATCA
Turnip      TATAGATAACCA---AAGCAATAGACAGACAAGTAAGTTAAG-AGAAAAG
Wheat       GTGACCCGGCAATGGGGTCCTCAACTGTAGCCGGCATCCTCCTCTCCTCC
Duckweed    CATGGGGCGACG---CAGTGTGTGGAGGAGCAGGCTCAGTCTCCTTCTCG
```

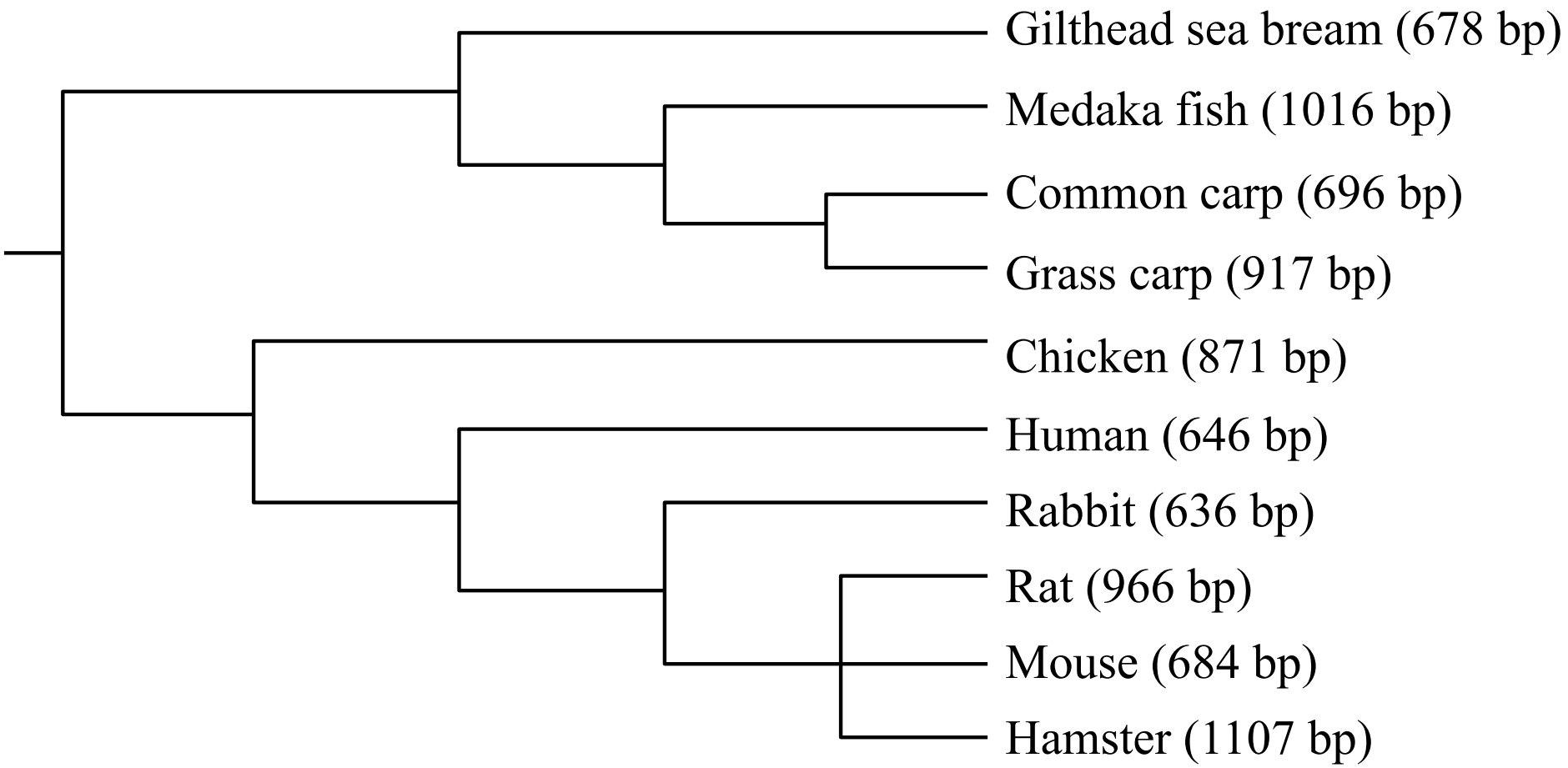
An Exact Algorithm

(generalizing Sankoff and Rousseau 1975)

$W_u[s]$ = best parsimony score for subtree rooted at node u, if u is labeled with string s.



Application to *β-actin* Gene



Common carp

ACGGACTGTTACCACTTCACGCCGACTCAACTGCGCAGAGAAAAACTTCAAACGACAACATTGGCATGGCTTTTGTTATTTTTGGCGCTTGACTCAGG
ATCTAAAAAACTGGAACGGCGAAGGTGACGGCAATGTTTTGGCAAATAAGCATCCCCGAAGTTCTACAATGCATCTGAGGACTCAATGTTTTTTTTTTTTTTTTT
CTTTAGTCATTCCAAATGTTTGTTAAATGCATTGTTCCGAAACTTATTTGCCTCTATGAAGGCTGCCCAGTAATTGGGAGCATACTTAACATTGTAGTATTGTATGTAAAT
TATGTAACAAAACAATGACTGGGTTTTTGTACTTTCAGCCTTAATCTTGGGTTTTTTTTTTTTTTGGTTCCAAAAAACTAAGCTTTACCATTCAAGATGTAAAGGTTTCATTCC
CCCTGGCATATTGAAAAAGCTGTGTGGAACGTGGCGGTGCAGACATTTGGTGGGGCCAACCTGTACACTGACTAATTCAAATAAAAGTGCACATGTAAGAC
ATCCTACTCTGTGTGATTTTTCTGTTTGTGCTGAGTGAACTTGCTATGAAGTCTTTTAGTGCACTCTTTAATAAAAGTAGTCTTCCCTTAAAGTGTCCCTTCCCTTATGGCCTTC
ACATTTCTCAACTAGCGCTTCAACTAGAAAGCACTTTAGGGACTGGGATGC

Chicken

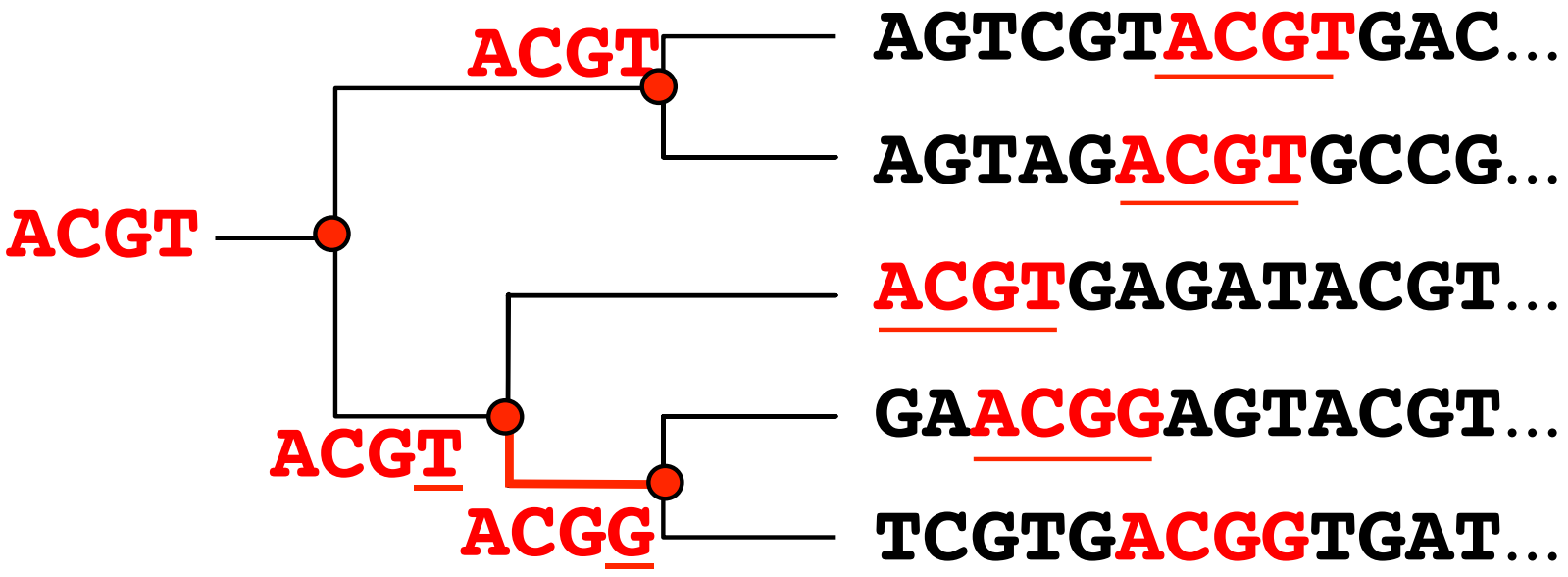
ACCGGACTGTTACCAACACCCACACCCCTGTGATGAAACAAAACCCATAAATGCGCATAAAACAAGACGAGATTGGCATGGCTTTATTTGTTTTTTCTTTTGGCGC
TTGACTCAGGATTAAAAAACTGGAATGGTGAAGGTGTCAGCAGCAGTCTTAAAATGAAACATGTTGGAGCGAACGCCCCCAAAGTTCTACAATGCAT
CTGAGGACTTTGATTGTACATTTGTTTCTTTTTTAATAGTCATTCCAAATATTGTTATAATGCATTGTTACAGGAAGTTACTCGCCTCTGTGAAGGCAACAGCCCAGCTGGG
AGGAGCCGGTACCAATTACTGGTGTTAGATGATAATTGCTTGTCTGTAAATTATGTAACCCAACAAGTGTCTTTTTGTATCTTCCGCCTTAAAAACAAAACACACTTGATCC
TTTTTGGTTTGTCAAGCAAGCGGGCTGTGTTCCCCAGTGATAGATGTGAATGAAGGCTTTACAGTCCCCCACAGTCTAGGAGTAAAGTGCCAGTATGTGGGGGAGGGAGGG
GCTACCTGTACACTGACTTAAGACCAGTTCAAATAAAAGTGCACACAATAGAGGCTTGACTGGTGTTGGTTTTTATTTCTGTGCTGCGCTGCTTGGCCGTTG
GTAGCTGTTCTCATCTAGCCTTGCCAGCCTGTGTGGGTCAGCTATCTGCATGGGCTGCGTGCTGGTGCTGTCTGGTGCAGAGGTTGGATAAACCGTGATGATATTTCAGCAA
GTGGGAGTTGGCTCTGATTCCATCCTGAGCTGCCATCAGTGTGTTCTGAAGGAAGCTGTTGGATGAGGGTGGGCTGAGTGCTGGGGGACAGCTGGGCTCAGTGGGACTG
CAGCTGTGCT

Human

GCGGACTATGACTTAGTTGCGTTACACCCTTTCTTGACAAAACCTAACTTGCGCAGAAAACAAGATGAGATTGGCATGGCTTTATTTGTTTTTTTTGTTTTGTTTTG
GTTTTTTTTTTTTTTTTGGCTTGACTCAGGATTTAAAAACTGGAACGGTGAAGGTGACAGCAGTCGGTTGGAGCGAGCATCCCCCAAAGTTCACAATG
TGGCCGAGGACTTTGATTGCATTGTTGTTTTTTTAATAGTCATTCCAAATATGAGATGCATTGTTACAGGAAGTCCCTTGCCATCCTAAAAGCCACCCCACTTCTCTCTAAG
GAGAATGGCCCAGTCCTCTCCCAAGTCCACACAGGGGAGGTGATAGCATTGCTTTCGTGTAAATTATGTAATGCAAAATTTTTTTAATCTTCGCCTTAATACTTTTTTATTTT
GTTTTATTTTGAATGATGAGCCTTCGTGCCCCCCCTTCCCCCTTTTTGTCCCCCAACTTGAGATGTATGAAGGCTTTTGGTCTCCCTGGGAGTGGGTGGAGGCAGCCAGGGC
TTACCTGTACACTGACTTGAGACCAGTTGAATAAAAGTGCACACCTTAAAAATGAGGCCAAGTGTGACTTTGTGGTGTGGCTGGGTTGGGGGCAGCAGAG
GGTG

Parsimony score over 10 vertebrates: 0 1 2

Solution



Parsimony score: 1 mutation