

MA3259 Lecture 10

**Molecular Phylogenetic Analysis:
Part 4. Ancestral State Reconstruction**

LX Zhang

Department of Mathematics
National University of Singapore

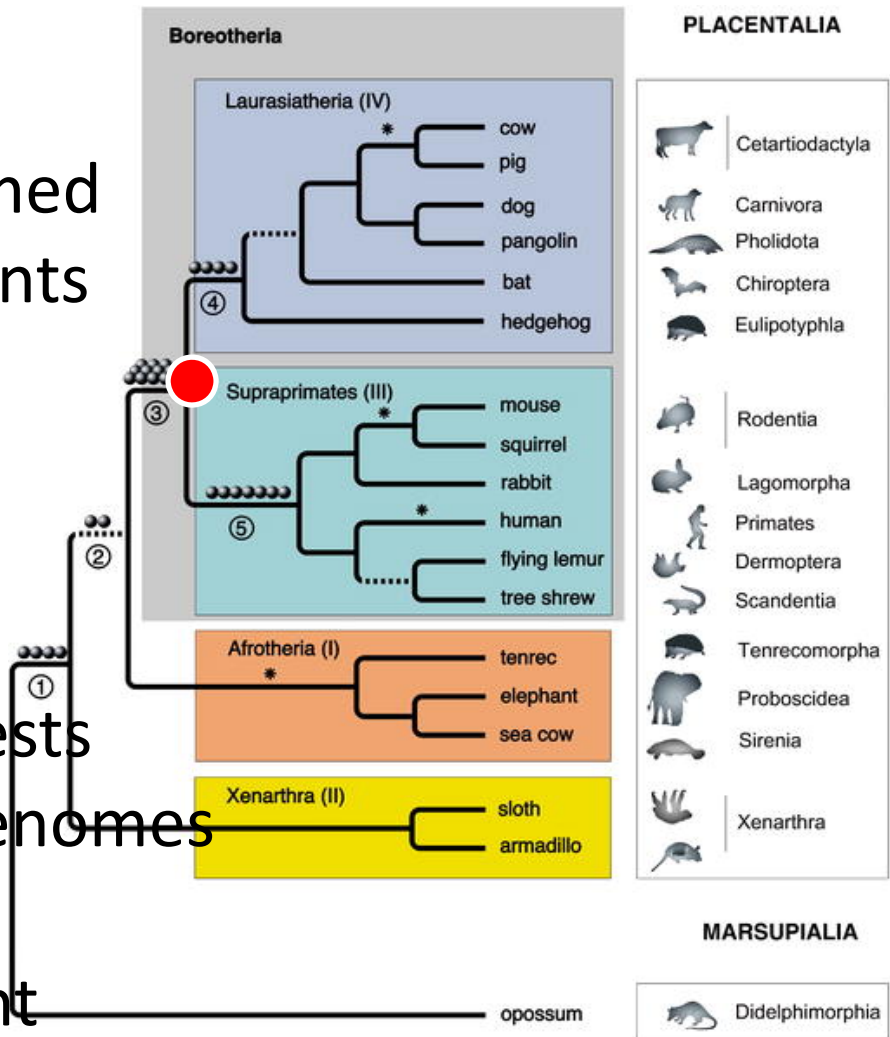
matzlx@nus.edu.sg

1. Little Background to Ancestral Sequence Reconstruction

- Ancestral sequence reconstruction incorporates sequences from modern organisms into evolutionary models to estimate the corresponding sequence of an ancestor that no longer presents on Earth.
- This approach to understanding proteins or life in general was proposed by Zuckerkandl and Pauling in 1963.
- It has become an popular approach to studying the origin, evolution, sequence-function relationship of proteins, genes and other components of life.

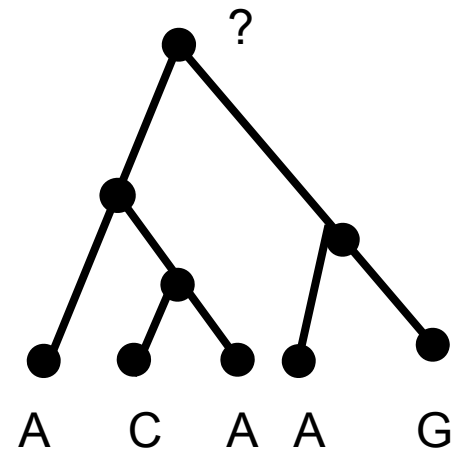
Recreate Genome of Ancient Human Ancestor

- “Boreoeutherian ancestor” lived 70 million yrs ago.
- The boreoeutheria was formed by a series of speciation events occurring rapidly after the ancestor, leading to a star-like phylogeny of boreoeutheria.
- Computer simulation suggests a small number of extant genomes can give a highly accurate reconstruction of this ancient genomes.



2. Ancestral State Reconstruction Problem

- Given a **phylogenetic tree H** with leaves labeled with letters from an alphabet, an **evolutionary model** (which gives the prior distribution of all possible letters at the root and substitution probability at each branch).
- Estimate the **true letter at the root** from the labels of leaves in the tree H.



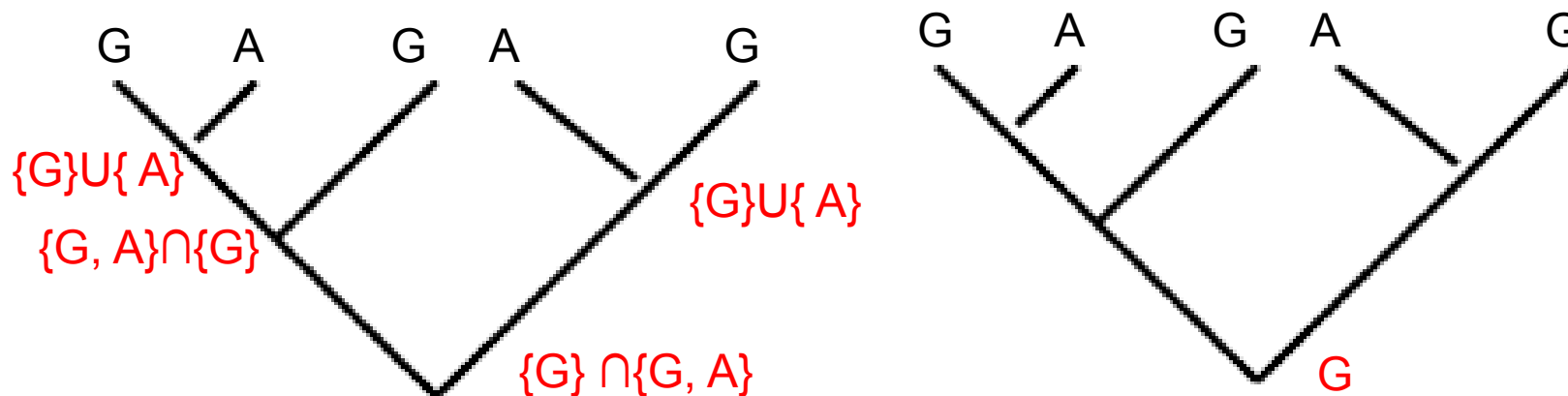
3. Methods for Ancestral State Reconstruction

Fitch Algorithm: Two-Step algorithm

- It assigns a letter to the root by minimizing the total number of substitutions placed on all branches

Step 1: Compute a subset S_x of letters for each node x .

Step 2: Select a letter from the subset obtained at the root randomly.



- It is a local and then efficient method.
- But, it ignores substitution rate on branches and is very sensitive to the topology of the tree.
- The reconstruction accuracy is not a monotone function with respect to the subset of leaves used.

Maximum Likelihood (ML) Method

-- It assigns letter a to the root that has the maximum **likelihood** defined as

$\text{Pr}[a \text{ evolves into the given letters in the leaves }]$
and a tie is broken arbitrarily.

FACT: Given the letters at leaves, the likelihood can be calculated in polynomial time.

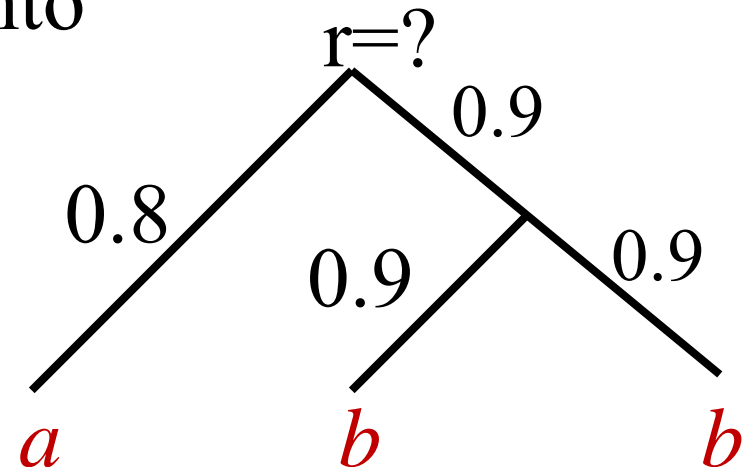
- It is a global method and so less efficient.
- But, has the largest reconstruction accuracy, over all methods.

Consider the following evolutionary model
(Tree + conservation rates + prior probabilities)
and two letter a and b .

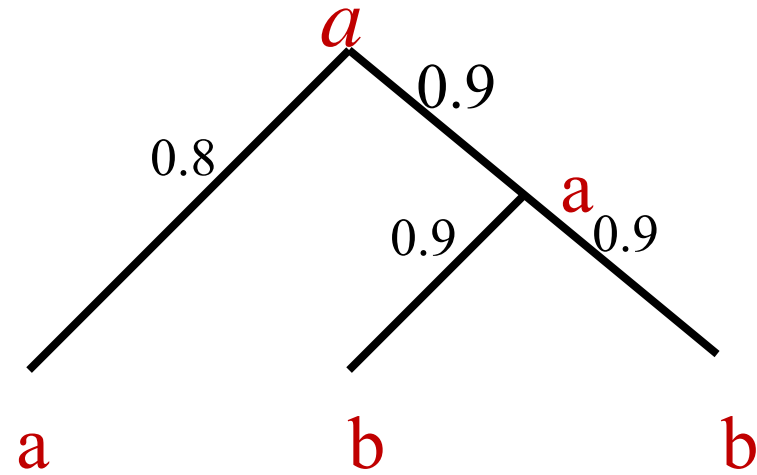
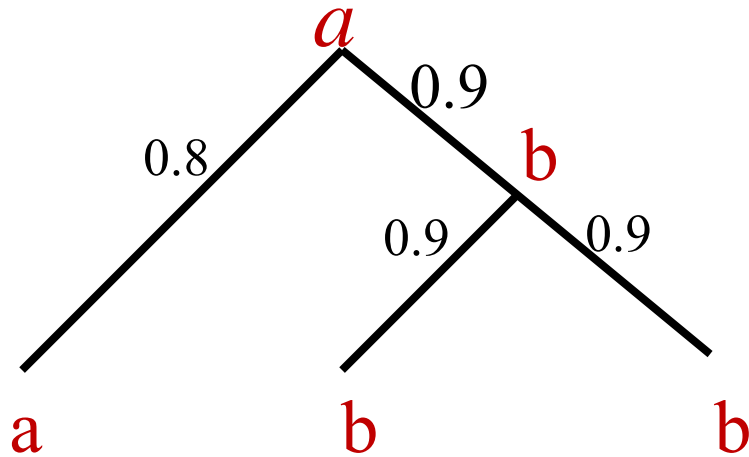
--- $p_{\text{prior}}(a) = p_{\text{prior}}(b) = 0.5$

--- Jukes-Cantor model

conservation rate 0.8 means that each letter
remains unchanged with probability 0.8
and that a letter mutates into
another letter with
probability 0.2.



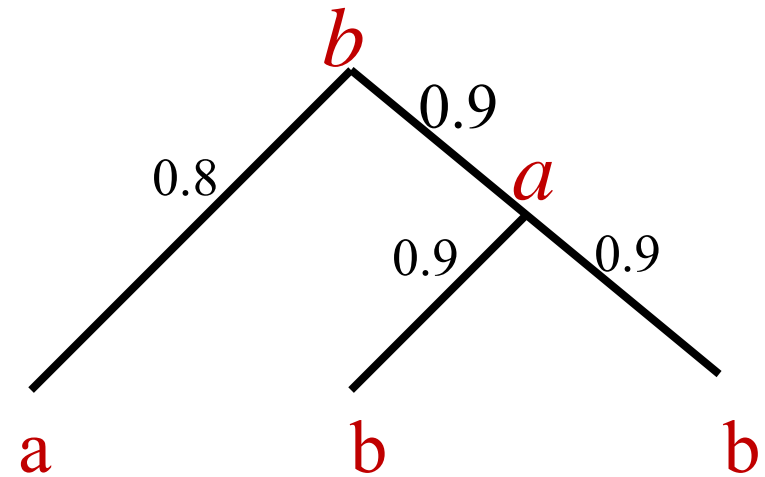
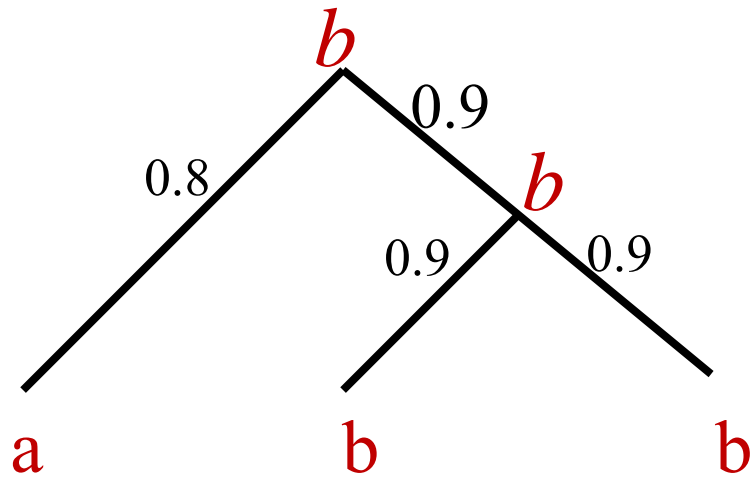
For root letter $s_r = a$, there are two possibilities



$$\Pr[a, b, b \mid a]$$

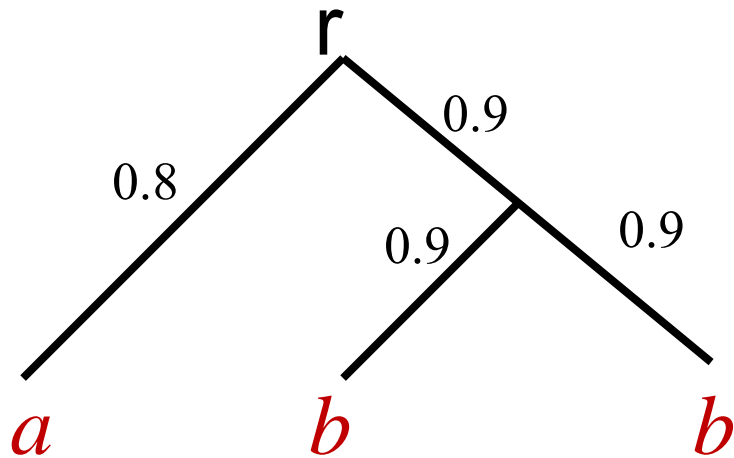
$$\begin{aligned}
 &= \Pr[a \rightarrow a] \Pr[a \rightarrow b] \Pr[b \rightarrow b] \Pr[b \rightarrow b] && \text{(left possibilities)} \\
 &\quad + \Pr[a \rightarrow a] \Pr[a \rightarrow a] \Pr[a \rightarrow b] \Pr[a \rightarrow b] && \text{(right possibilities)} \\
 &= 0.8 \times 0.1 \times 0.9 \times 0.9 + 0.8 \times 0.9 \times 0.1 \times 0.1 = 0.072
 \end{aligned}$$

For letter $s_r = b$,



$$\begin{aligned}
 & \Pr[a, b, b / b] \\
 &= \Pr[b \rightarrow a] \Pr[b \rightarrow b] \Pr[b \rightarrow b] \Pr[b \rightarrow b] \quad (\text{left case}) \\
 & \quad + \Pr[b \rightarrow a] \Pr[b \rightarrow a] \Pr[a \rightarrow b] \Pr[a \rightarrow b] \quad (\text{right case}) \\
 &= 0.2 \times 0.9 \times 0.9 \times 0.9 + 0.2 \times 0.1 \times 0.1 \times 0.1 = 0.1464
 \end{aligned}$$

In summary



$$p_{\text{prior}}(a) = p_{\text{prior}}(b) = 0.5$$

$$\Pr[a \text{ evolves into } a, b, b] = p_{\text{prior}}(a) \times 0.072$$

$$\Pr[b \text{ evolves into } a, b, b] = p_{\text{prior}}(b) \times 0.1462$$

The marginal ML method selects *b* as the root state from leaves states *a*, *b*, *b*.

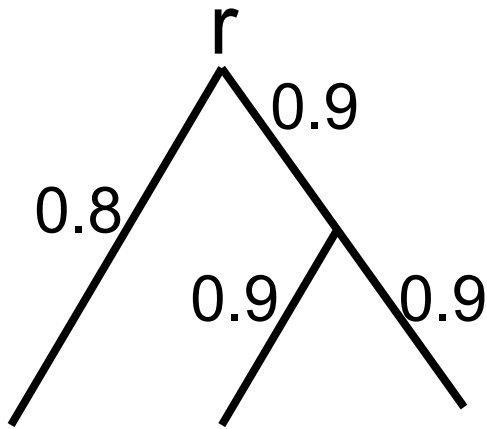
Reconstruction Accuracy

For a reconstruction method \mathcal{M} , its accuracy is the expected probability that the method reconstructs correctly the root state from possible configurations D of letters of leaves in tree H :

$$\begin{aligned} & \text{RA}_{\mathcal{M}}(H) \\ &= \sum_{c,D} \Pr[c \text{ evolves into } D] \Pr[\mathcal{M} \text{ reconstructs } c \text{ from } D] \end{aligned}$$

where $\Pr[c \text{ evolves into } D] = p_{\text{prior}}(c) \Pr[D | s_r = c]$

Fitch method' accuracy in the example with tree H.

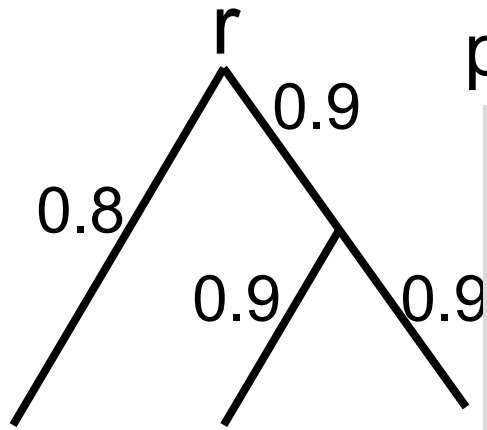


$$p_{\text{prior}}(a) = p_{\text{prior}}(b) = 0.5$$

$$\begin{aligned} \text{RA}_F(H) &= \sum_{c,D} p_{\text{prior}}(c) \Pr[D | c] \Pr[\text{reconstruct } c \text{ from } D] \\ &= \sum_D \Pr[D | a] \Pr[\text{reconstruct } a \text{ from } D] \\ &= 0.584 + 0.072 + 0.072 + 0.146 \times \frac{1}{2} + 0.072 \times \frac{1}{2} \\ &= 0.837 \end{aligned}$$

D			$\Pr[D s_r = b]$	$\Pr[D s_r = a]$	Selected root state
a	a	a	0.018	0.584	a
a	a	b	0.018	0.072	a
a	b	a	0.018	0.072	a
b	a	a	0.072	0.146	a, b (prob $\frac{1}{2}$)
a	b	b	0.146	0.072	a, b (prob $\frac{1}{2}$)
b	a	b	0.072	0.018	b
b	b	a	0.072	0.018	b
b	b	b	0.584	0.018	b

ML method' accuracy in the example.



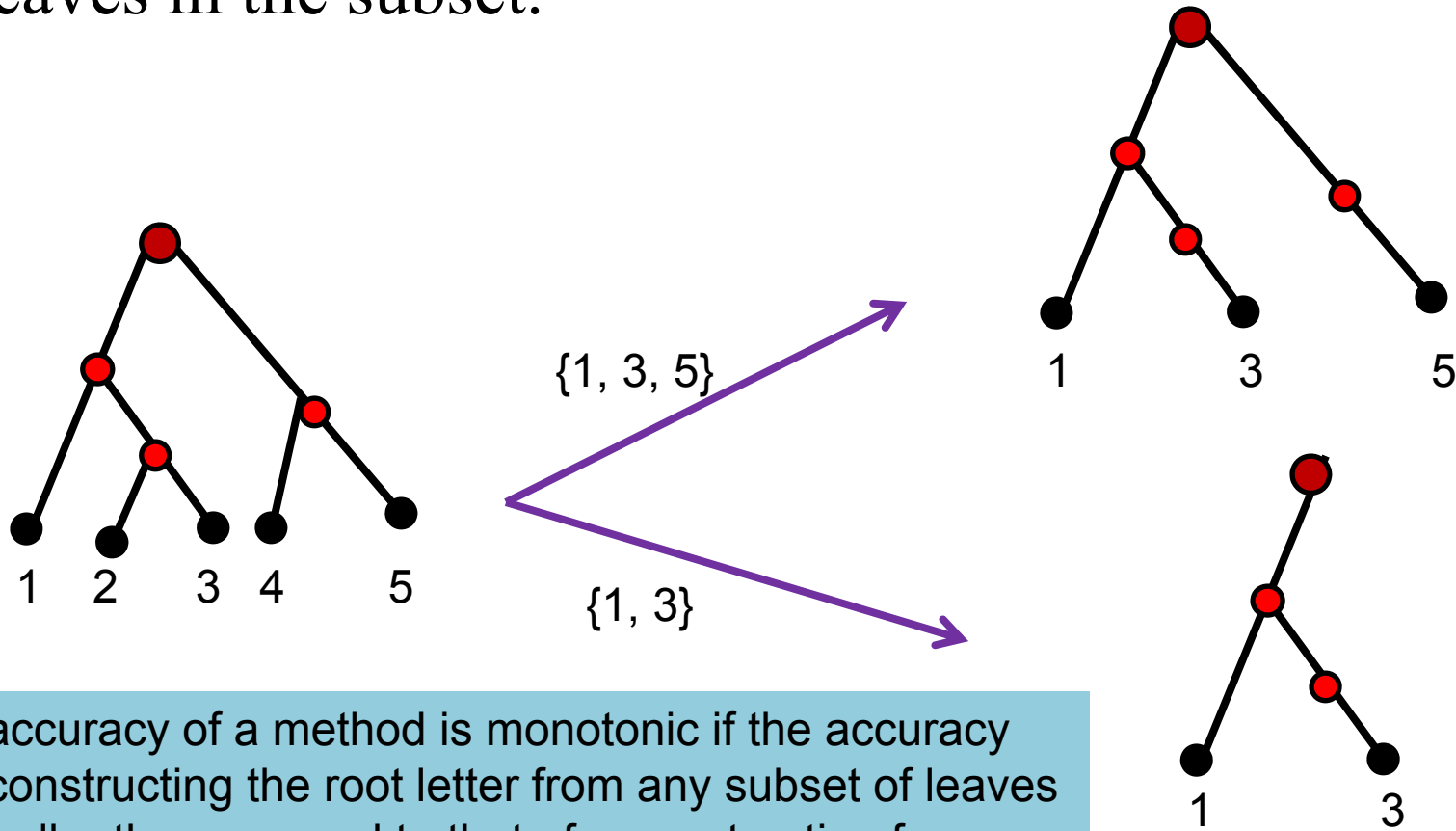
$$p_{\text{prior}}(0) = p_{\text{prior}}(1) = 0.5$$

$$\begin{aligned} \text{RA}_{\text{ML}}(H) &= \sum_{c,D} p_{\text{prior}}(c) \Pr[D | c] \Pr[\text{reconstruct } c \text{ from } D] \\ &= \sum_D \Pr[D | a] \Pr[\text{reconstruct } a \text{ from } D] \\ &= 0.584 + 0.072 + 0.072 + 0.146 \\ &= 0.874 \end{aligned}$$

D			$\Pr[D s_r = b]$	$\Pr[D s_r = a]$	Selected root state
a	a	a	0.018	0.584	a
a	a	b	0.018	0.072	a
a	b	a	0.018	0.072	a
b	a	a	0.072	0.146	a
a	b	b	0.146	0.072	b
b	a	b	0.072	0.018	b
b	b	a	0.072	0.018	b
b	b	b	0.584	0.018	b

Monotonicity of the Accuracy Function

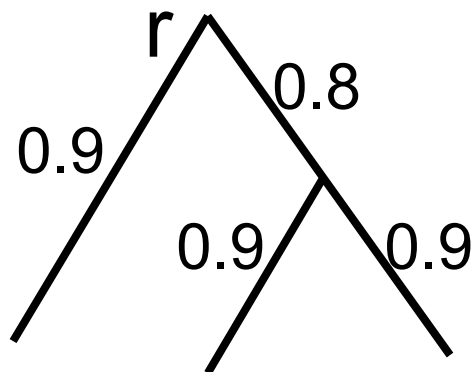
A subset of leaves induces a subtree, which is composed of the nodes and branches on all paths from the root to the leaves in the subset.



The accuracy of a method is monotonic if the accuracy of reconstructing the root letter from any subset of leaves is smaller than or equal to that of reconstruction from all leaves in any tree.

Theorem: The accuracy function of the Fitch method is not monotonic.

Proof. Consider the following tree



$$p_{\text{prior}}(a) = p_{\text{prior}}(b) = 0.5$$

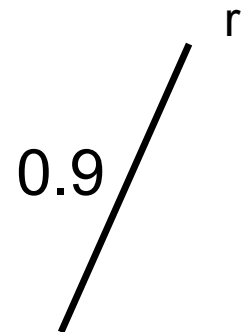
	$\Pr[D s_r = b]$	$\Pr[D s_r = a]$	Selected root state
a a a	0.017	0.595	a
a a b	0.009	0.081	a
a b a	0.009	0.081	a
b a a	0.153	0.065	a, b (prob 1/2)
a b b	0.065	0.153	a, b (prob 1/2)
b a b	0.081	0.009	b
b b a	0.081	0.009	b
b b b	0.595	0.017	b

Since the model is symmetric regarding to a and b,
the accuracy of reconstruction from all leaves is

$$\begin{aligned}
RA_F(H) &= \sum_{c,D} p_{\text{prior}}(c) \Pr[D|c] \Pr[\text{reconstructs } c \text{ from } D] \\
&= \sum_D \Pr[D|a] \Pr[\text{reconstructs } a \text{ from } D] \\
&= \Pr[a,a,a | a] + \Pr[a, a, b | a] + \Pr[a, b, a | a] + \Pr[a, b, b/a] \times \frac{1}{2} \\
&= 0.595 + 0.081 + 0.081 + 0.153 \times \frac{1}{2} + 0.065 \times \frac{1}{2} \\
&= 0.866
\end{aligned}$$

The accuracy of reconstruction from the subtree H' induced by
the left leaf is 0.9, larger than 0.866

$$\begin{aligned}
RA_H(H') &= \sum_{c,D} p_{\text{prior}}(c) \Pr[D|c] \Pr[\text{reconstruct } c \text{ from } D] \\
&= \sum_D \Pr[D|a] \Pr[\text{reconstruct } a \text{ from } D] \\
&= \Pr[a | a] \times 1.0 + \Pr[b | a] \times 0 \\
&= 0.9
\end{aligned}$$



Theorem: The maximum likelihood (ML) method has the largest reconstructing accuracy over all methods for any tree and evolution model.

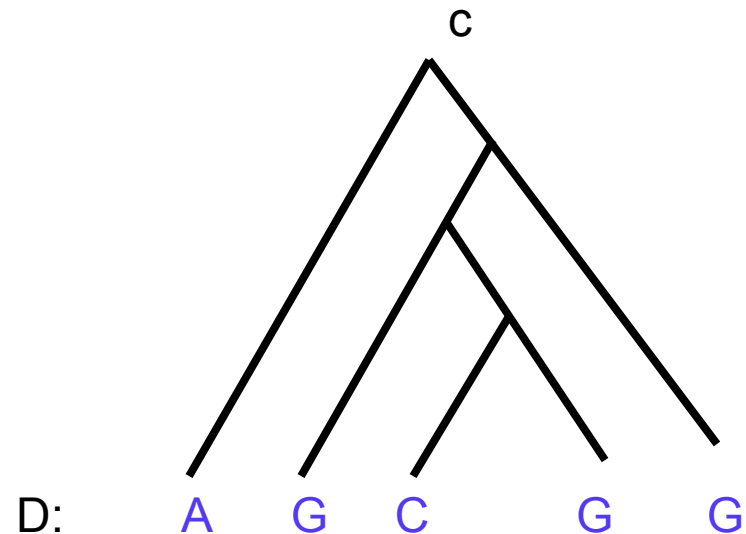
Corollary: The accuracy function of the ML method is monotonic

Proof of Corollary.

Using a subset of leaves is just a specific reconstruction method that does not **use** letter information in the other leaves and hence its accuracy is not higher than the reconstruction from all the leaves when ML is used.

Proof of Theorem: For any method \mathcal{M} and tree H ,

$$\begin{aligned} & \text{RA}_{\mathcal{M}}(H) \\ &= \sum_{c,D} \Pr[c \text{ evolves into } D] \Pr[\mathcal{M} \text{ reconstructs } c \text{ from } D] \\ &= \sum_D \sum_c \Pr[c \text{ evolves into } D] \Pr[\mathcal{M} \text{ reconstructs } c \text{ from } D] \\ &\leq \sum_D \sum_c (\max_c \Pr[c \text{ evolves into } D]) \Pr[\mathcal{M} \text{ reconstructs } c \text{ from } D] \\ &= \sum_D (\max_c \Pr[c \text{ evolves into } D]) \{\sum_c \Pr[\mathcal{M} \text{ reconstructs } c \text{ from } D]\} \\ &= \sum_D (\max_c \Pr[c \text{ evolves into } D]) \end{aligned}$$



For the ML method and tree H,

$$\begin{aligned} & \text{RA}_{\text{ML}}(H) \\ &= \sum_{c,D} \Pr[c \text{ evolves into } D] \Pr[\text{ML reconstructs } c \text{ from } D] \\ &= \sum_D \sum_c \Pr[c \text{ evolves into } D] \Pr[\text{ML reconstructs } c \text{ from } D] \\ &= \sum_D (\max_c \Pr[c \text{ evolves into } D]) \end{aligned}$$

Thus,

$$\begin{aligned} & \text{RA}_{\mathcal{M}}(H) \\ & \leq \sum_D \max_c \Pr[c \text{ evolves into } D] \\ & = \text{RA}_{\text{ML}}(H) \end{aligned}$$

