

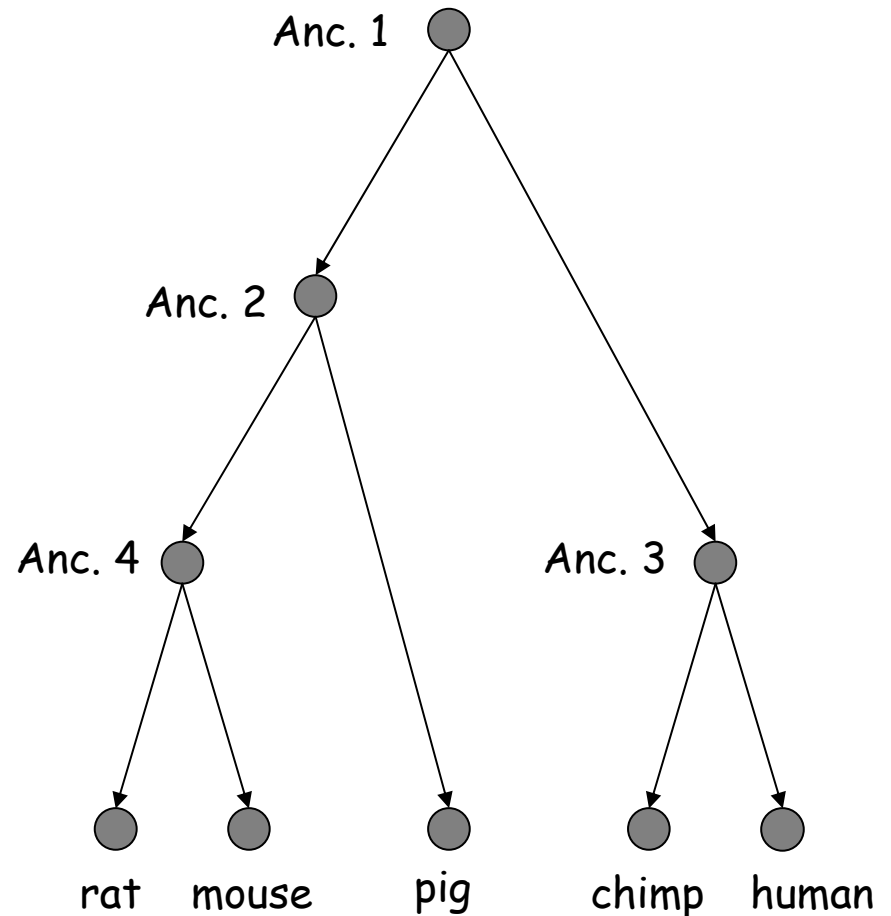
Inferring Phylogenies by Likelihood Methods: Computational Hardness

Sébastien Roch
sroch@stat.berkeley.edu
Department of Statistics
UC Berkeley

Phylogenetic Reconstruction






- *Phylogenetic Model:*

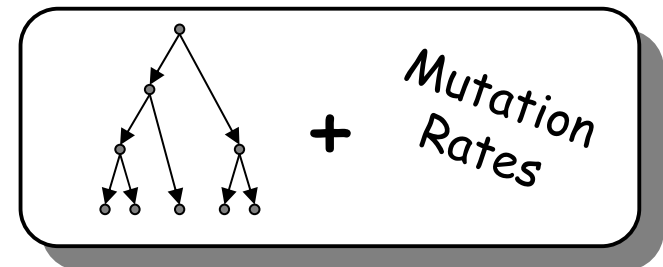
- Ancestry relationships: tree
- Leaves: known species
- Internal nodes: ancestral (unknown) species
- Data: molecular sequences at nodes
- Evolution mechanism: mutations, deletions, insertions, etc.



Phylogenetic Reconstruction (cont'd)

- *Reconstruction:*
 - Given: aligned DNA sequences of known species (leaves)
 - Task: infer evolutionary history of known species (ancestors, tree, mutation rates) **efficiently**

				
rat	mouse	pig	chimp	human
A	A	A	G	T
A	A	T	T	T
T	G	C	C	C
A	A	A	C	C
C	C	G	C	C

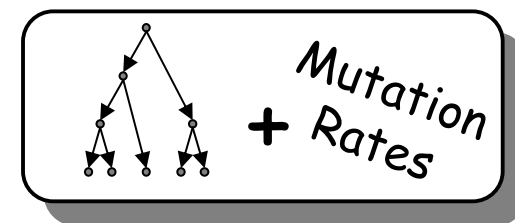


Our Result: ML is "Hard"

- *Most Popular Techniques:*

- **Parsimony:** not efficient -> Graham & Foulds, *Math. Biosc.* 1982
- **Bayesian methods:** probably not efficient -> see e.g. Mossel & Vigoda, *Science* 2005
- **Likelihood:** NOT EFFICIENT -> two independent proofs
 - 1) Chor & Tuller, RECOMB 2005
 - 2) Roch, IEEE TCBB 2005

	●	●	●	●	●
	↓	↓	↓	↓	↓
	rat	mouse	pig	chimp	human
A	A	A	A	G	T
A	A	A	T	T	T
T	G	C	C	C	C
A	A	A	C	C	C
C	C	G	C	C	C



Model of Evolution

- *CFN Model:*

- Tree: $T = (V, E)$

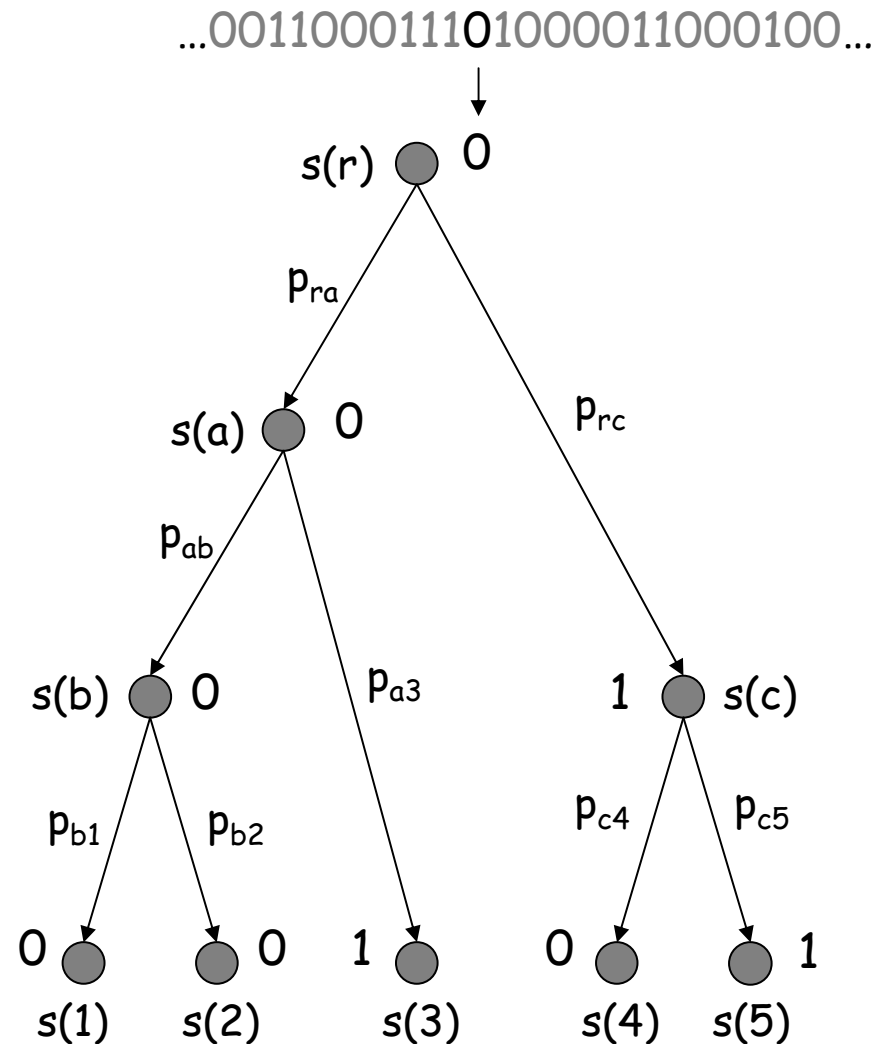
- Node states:

$$\{s(v) \in \{0,1\} : v \in V\}$$

- Mutation probabilities:

$$\{0 < p_e < 1/2 : e \in E\}$$

0: Purines (A,G)
1: Pyrimidines (C,T)

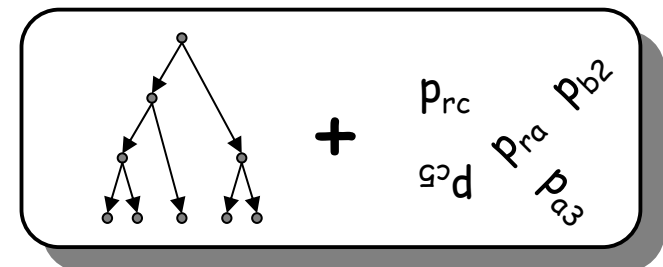


Phylogenetic Reconstruction

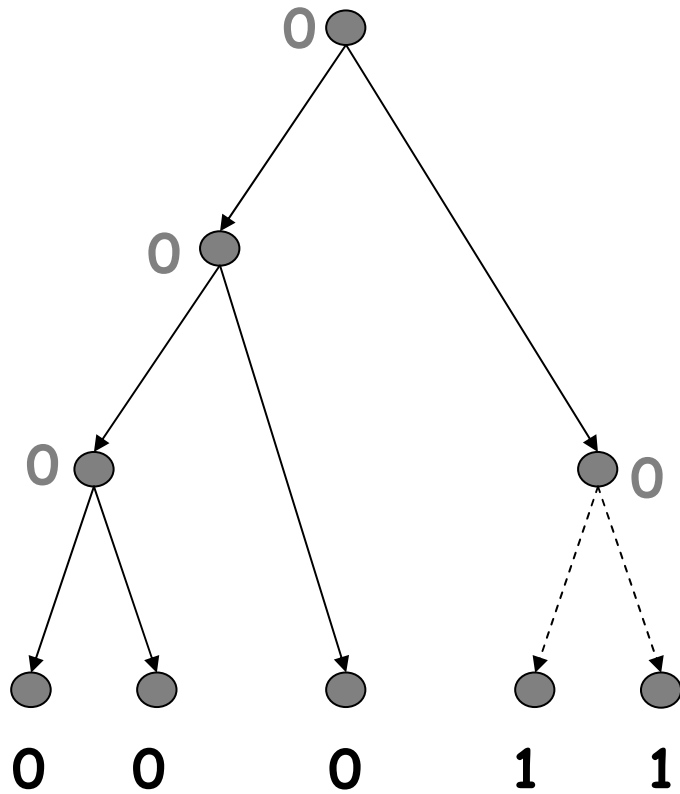
- *Reconstruction:*

- Given: i.i.d. samples at the leaves
- Task: fully reconstruct the model, i.e. find **tree and mutation probabilities** (and, *if possible*, do so **efficiently**)

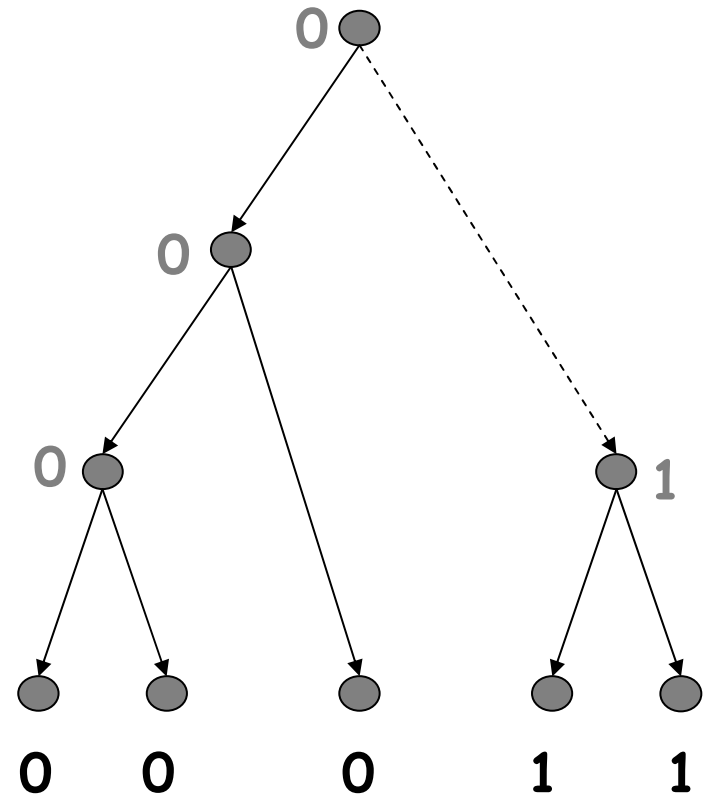
↓	↓	↓	↓	↓
●	●	●	●	●
$s(1)$	$s(2)$	$s(3)$	$s(4)$	$s(5)$
0	0	1	1	1
0	0	0	1	1
1	1	0	0	1
0	0	1	1	1
1	0	0	1	1



Reminder: Parsimony



Suboptimal Reconstruction



Most parsimonious
(Parsimony Score = 1)

- In general, sum the parsimony scores of all sites

Reminder: Maximum Likelihood

- *Example [Biased Coin Tossing]:*

- p-Biased Coin:
$$X = \begin{cases} 0, & \text{prob. } 1-p \\ 1, & \text{prob. } p \end{cases}$$

- Toss coin N times (independently):

$$(X_1, X_2, X_3, \dots, X_N) = (0, 1, 0, \dots, 0)$$

- Likelihood:

$$\Lambda(p; X_1, \dots, X_N) = \prod_{i=1}^N p^{X_i} (1-p)^{1-X_i} = (1-p)p(1-p)\cdots(1-p)$$

- MLE: value of $p(=p^*)$ minimizing (N_0 : # of 0's, N_1 : # of 1's)

$$-\ln \Lambda(p; X_1, \dots, X_N) = N_0(-\ln[1-p]) + N_1(-\ln p)$$

- Easy to see (differentiate):

$$p^* = \frac{N_1}{N}$$

Phylogenetics by Max Likelihood

- Data: n $\{0,1\}$ -sequences of length k (at leaves)

$$\{S(j) = (s_1(j), \dots, s_k(j)) \in \{0,1\}^k : 1 \leq j \leq n\}$$

- Likelihood:

$$\Lambda(T, \{p_e\}; S(1), \dots, S(n)) = \prod_{i=1}^k \sum_{s^* \in \text{Ext}(s_i)} \prod_{e=(u,v) \in E} p_e^{\langle s^*(u) \neq s^*(v) \rangle} (1 - p_e)^{\langle s^*(u) = s^*(v) \rangle}$$

- MLE:

$$(T^*, \{p_e^*\}) = \arg \min_{(T, \{p_e\})} [-\ln \Lambda(T, \{p_e\}; S(1), \dots, S(n))]$$

- **PROBLEM:** Large discrete optimization space -> **NO** simple formula for MLE

Tree Space

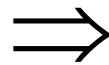
- **Complicated structure** (vs. 1D space of biased coin tossing)
- **Very large** -> number of rooted trees on n labelled leaves (from Felsenstein, 2004):

<u>n</u>	<u># of trees</u>
4	26
8	660,032
12	188,666,182,784
16	238,513,970,965,257,728
20	887,094,711,304,119,347,388,416

Reminder: NP reductions

- *NP-complete problems:*
 - "If **any** has an efficient algorithm, **all** do."
 - Example in Phylogeny: Parsimony
 - Conjecture ($P \neq NP$): No efficient algorithm for NP
- **Our goal: show that ML is NP-hard**

Efficient Algo. for ML



Efficient Algo. for NP

- **Technique: reduction** (i.e. NPC problem is "special case" of ML)

Instance of NPC problem



Instance of ML

Open Questions

- (1) Computing MLE for mutation probabilities on given tree: NP-hard?
- (2) Information-theoretic efficiency of ML?

THANK YOU...