Phylogenetic inference and likelihood 2

Emily Jane McTavish

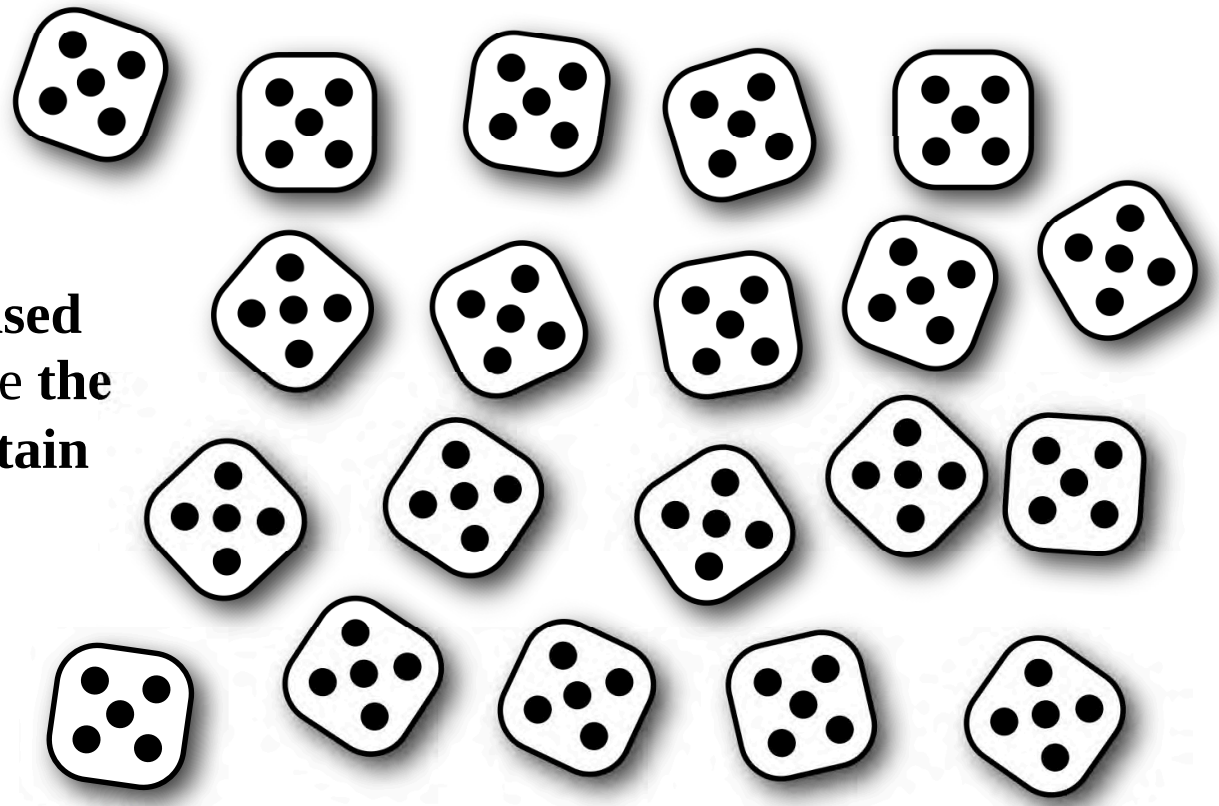
Life and Environmental Sciences
University of California, Merced
ejmctavish@ucmerced.edu

(With thanks to Mark Holder and Paul Lewis for slides)

The Trick Dice model

(assumes dice each have 5 on every side)

$$\Pr(\text{obs.}|\text{trick dice model}) = 1^{20} = 1$$



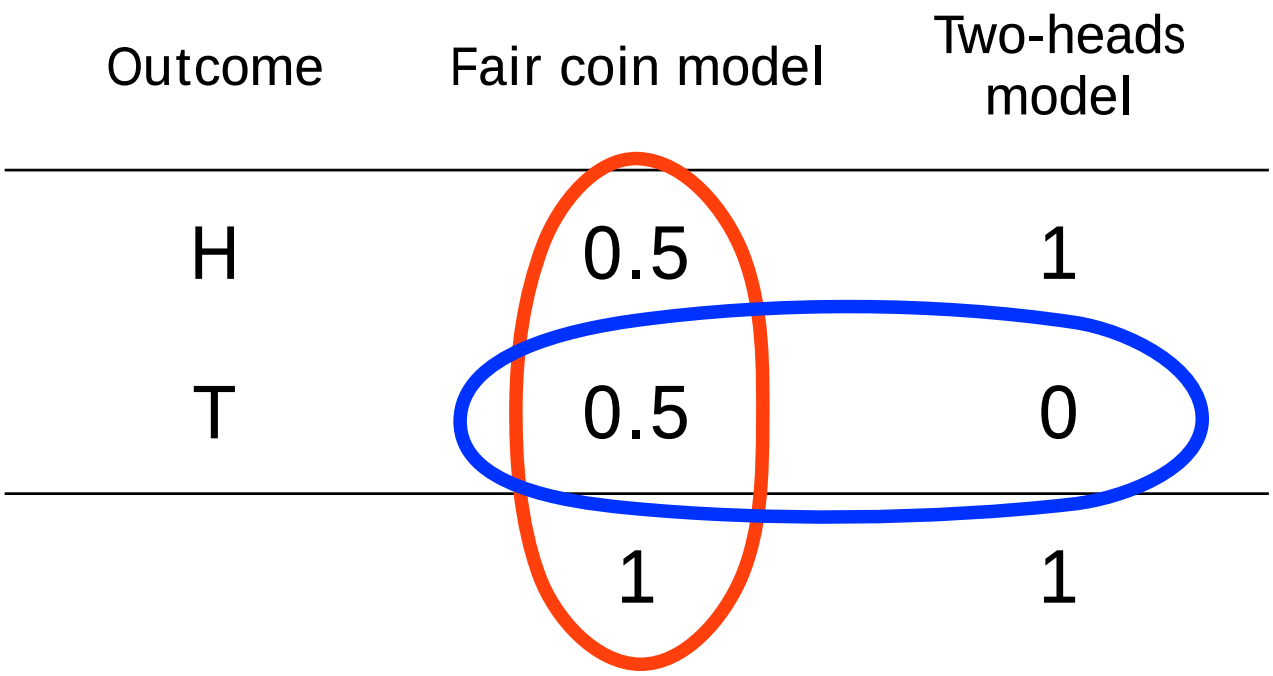
You should **not be surprised at all** at this result because **the observed outcome is certain** under this model

Results

Model	Likelihood	Surprise level
Fair Dice	$\frac{1}{3{,}656{,}158{,}440{,}062{,}976}$	Very, *very*, ***very*** surprised
Trick Dice	1	Not surprised at all

winning model maximizes likelihood (and thus minimizes surprise)

Likelihood: why a new term?



Likelihoods of models given one particular data outcome are *not* expected to sum to 1.0

Probabilities of data outcomes given one particular model sum to 1.0

What is the probability of getting a 2 tails for two coin flips under each of these models:

- Fair coin
- Two heads
- Two tails

What is the likelihood of each of these models given that you got 2 tails:

- Fair coin
- Two heads
- Two tails

Does that mean that this coin likely has 2 tails?

What is the likelihood of this sequence?
AAGGATC

What is the likelihood of this sequence?

AAGGATC

Depends on a lot of stuff!!

What is the likelihood of this sequence, if sites are independent, and all bases are equally common in the genome?

AAGGATC

What is the likelihood of this sequence, if sites are independent, and all bases are equally common in the genome?

AAGGATC

What is the log likelihood?

What is the likelihood of this sequence, if sites are independent, and all bases are as common in the genome as they are in this sequence?
AAGGATC

What is the likelihood of this sequence, if sites are independent, and all bases are as common in the genome as they are in this sequence?

AAGGATC

What is the log likelihood under this model?

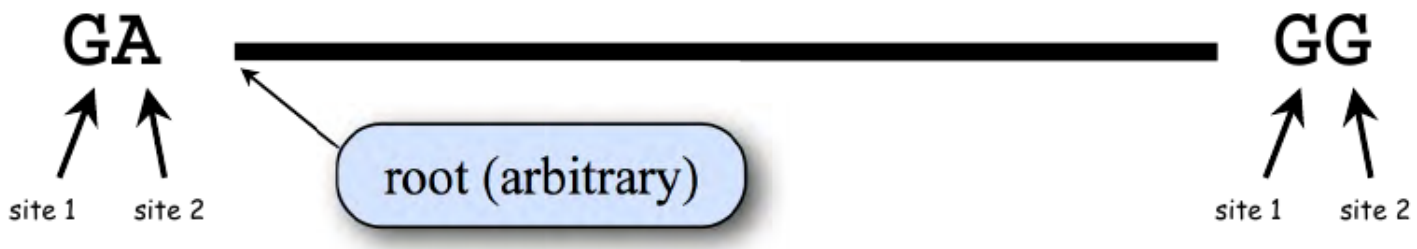
Which model is a better fit for the data? Is it possible for the other model to be better?

Does the likelihood ratio test statistic support the flexible model or the equal rates model?

Comparing models in phylogenetics can be challenging, as topologies are not nested within one another.
We will discuss appropriate statistical approaches later in the course.

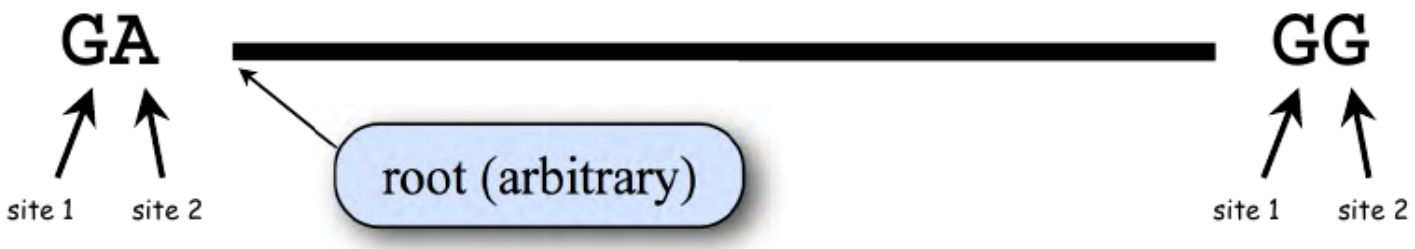
What do we need to know to calculate the likelihood of this very simple tree?

To keep things simple, assume that the sequences are only 2 nucleotides long:



What do we need to know to calculate the likelihood of this very simple tree?

To keep things simple, assume that the sequences are only 2 nucleotides long:

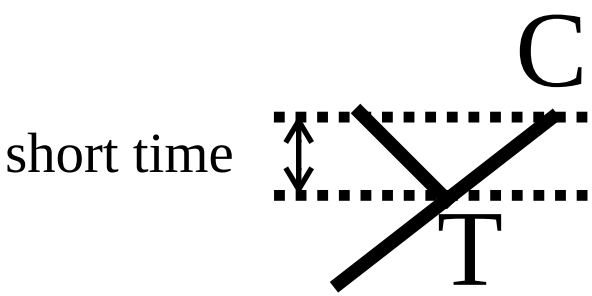


We could calculate the probability of each sequence.

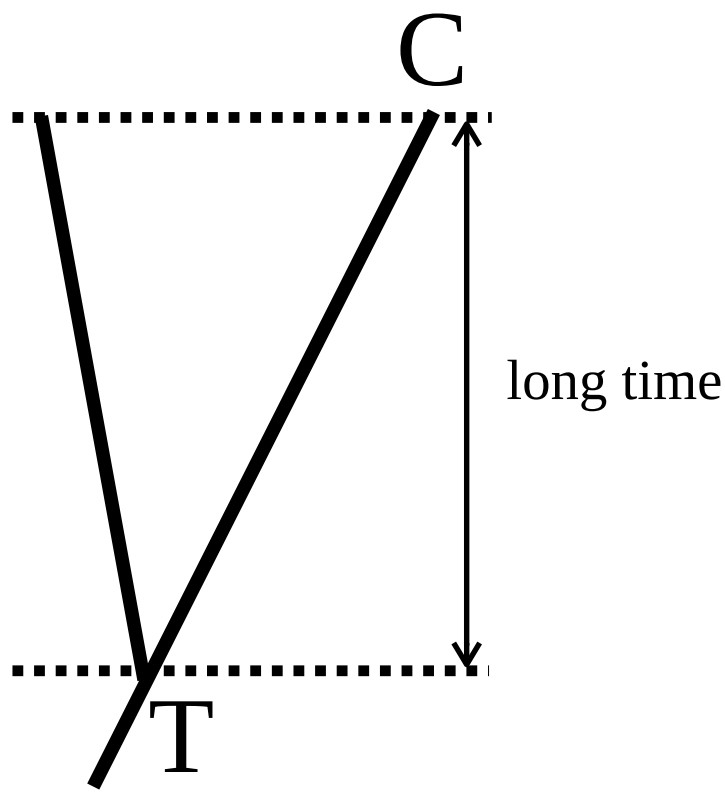
Are they independent of each other?

Non-independence in molecular evolution

The state present in the descendant is **not independent** of the state in the ancestor



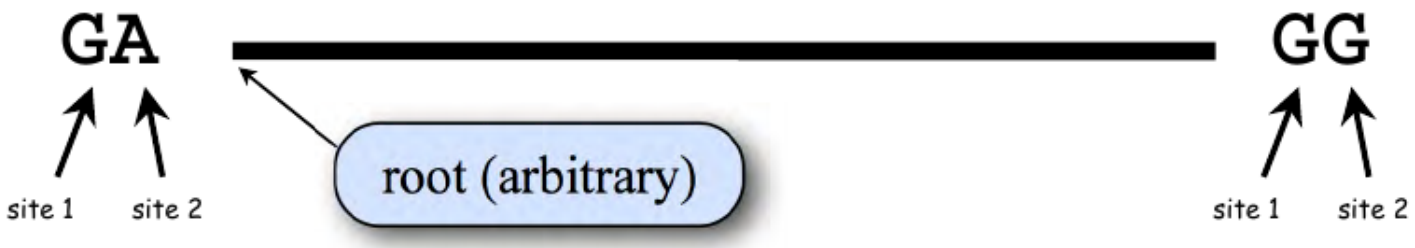
less probable



more probable

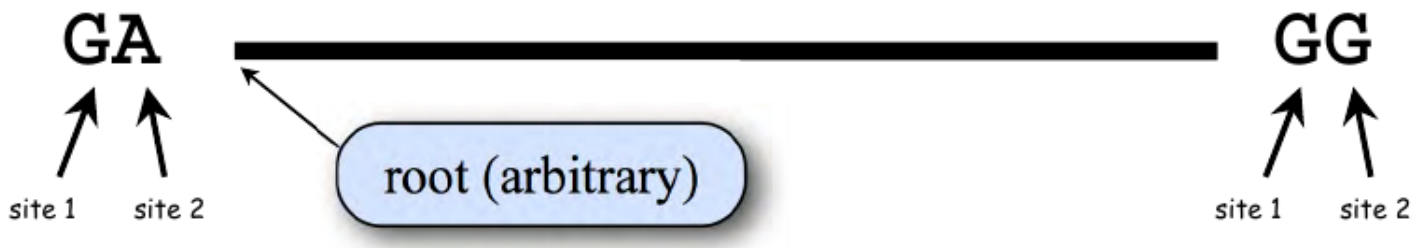
What do we need to know to calculate the likelihood of this very simple tree?

To keep things simple, assume that the sequences are only 2 nucleotides long:



What do we need to know to calculate the likelihood of this very simple tree?

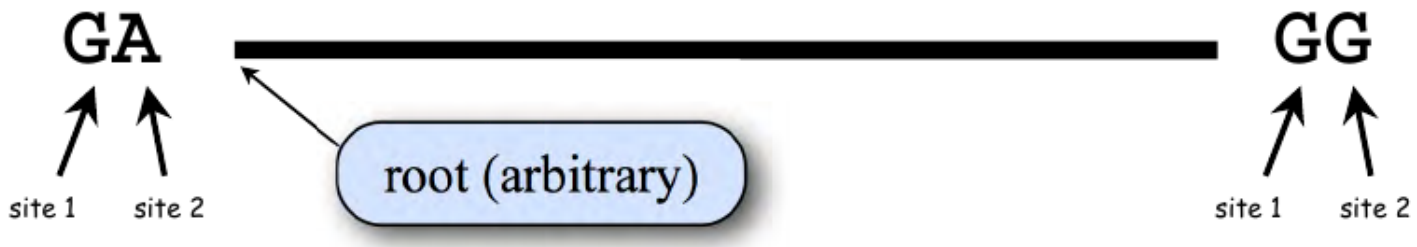
To keep things simple, assume that the sequences are only 2 nucleotides long:



What is the probability a base has changed?
What is the probability a base has stayed the same?

What do we need to know to calculate the likelihood of this very simple tree?

To keep things simple, assume that the sequences are only 2 nucleotides long:



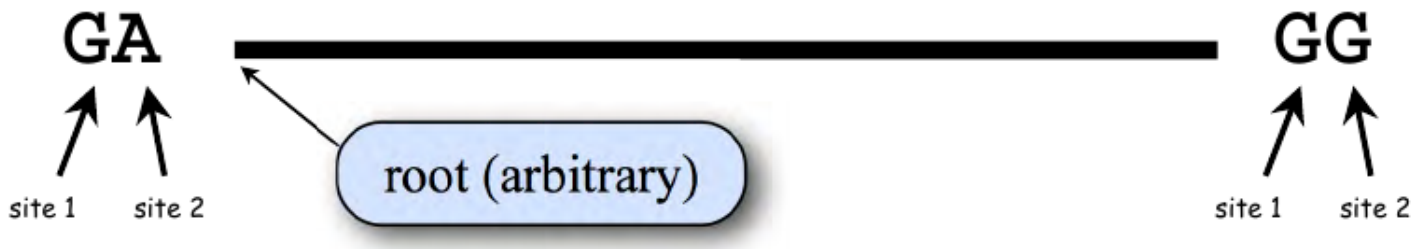
If the branch length is 0:

What is the probability a base has changed?

What is the probability a base has stayed the same?

What do we need to know to calculate the likelihood of this very simple tree?

To keep things simple, assume that the sequences are only 2 nucleotides long:



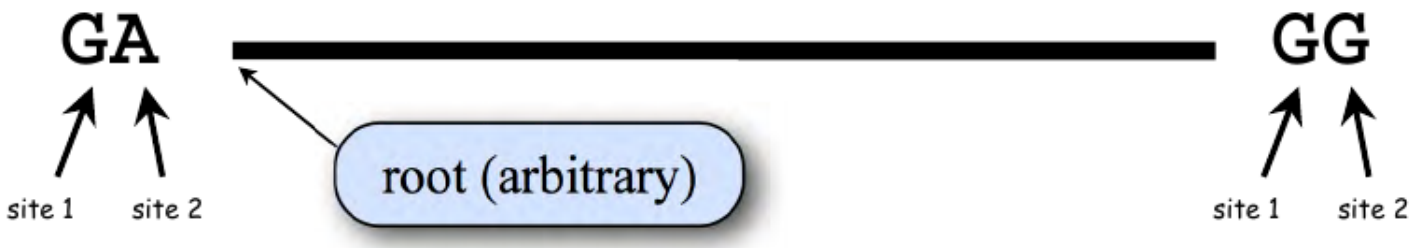
If the branch length is infinitely long:

What is the probability a base has changed?

What is the probability a base has stayed the same?

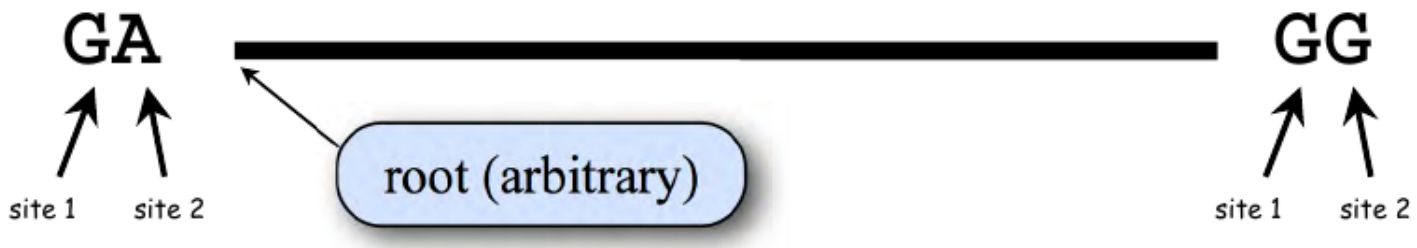
What do we need to know to calculate the likelihood of this very simple tree?

To keep things simple, assume that the sequences are only 2 nucleotides long:



What do we need to know to calculate the likelihood of this very simple tree?

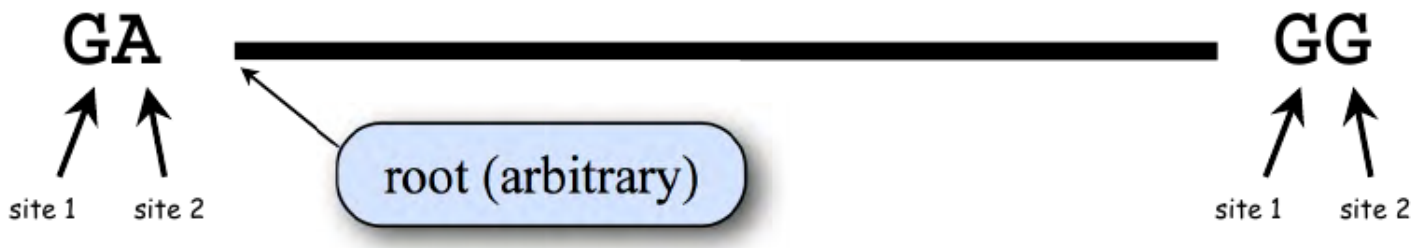
To keep things simple, assume that the sequences are only 2 nucleotides long:



The branch length is probably between 0 and infinity....

What do we need to know to calculate the likelihood of this very simple tree?

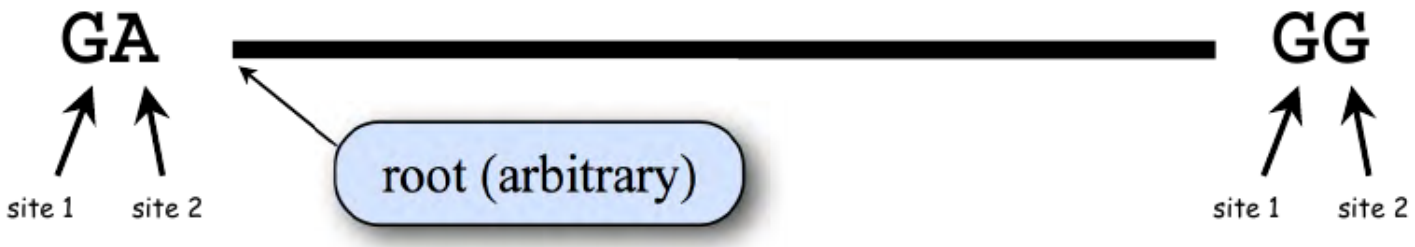
To keep things simple, assume that the sequences are only 2 nucleotides long:



What is the probability we start with a G at the first base?
What is the probability we start with an A at the second base?

What do we need to know to calculate the likelihood of this very simple tree?

To keep things simple, assume that the sequences are only 2 nucleotides long:

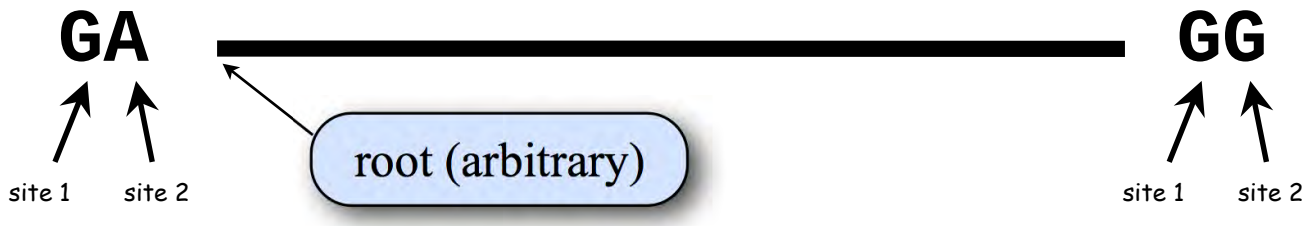


What is the probability there is no change (G->G) at the first base?
What is the probability there is a change (A->G) at the second base?

Likelihood of the simplest tree

sequence 1 ——— sequence 2

To keep things simple, assume that the sequences are only 2 nucleotides long:



$$
\begin{aligned}
L &= L_1 \, L_2 \\
&= \left[\left(\frac{1}{4}\right) \left(\frac{1}{4} + \frac{3}{4}e^{-4\alpha t}\right)\right] \left[\left(\frac{1}{4}\right) \left(\frac{1}{4} - \frac{1}{4}e^{-4\alpha t}\right)\right]
\end{aligned}
$$

Pr(G) Pr(G|G, αt) Pr(A) Pr(G|A, αt)

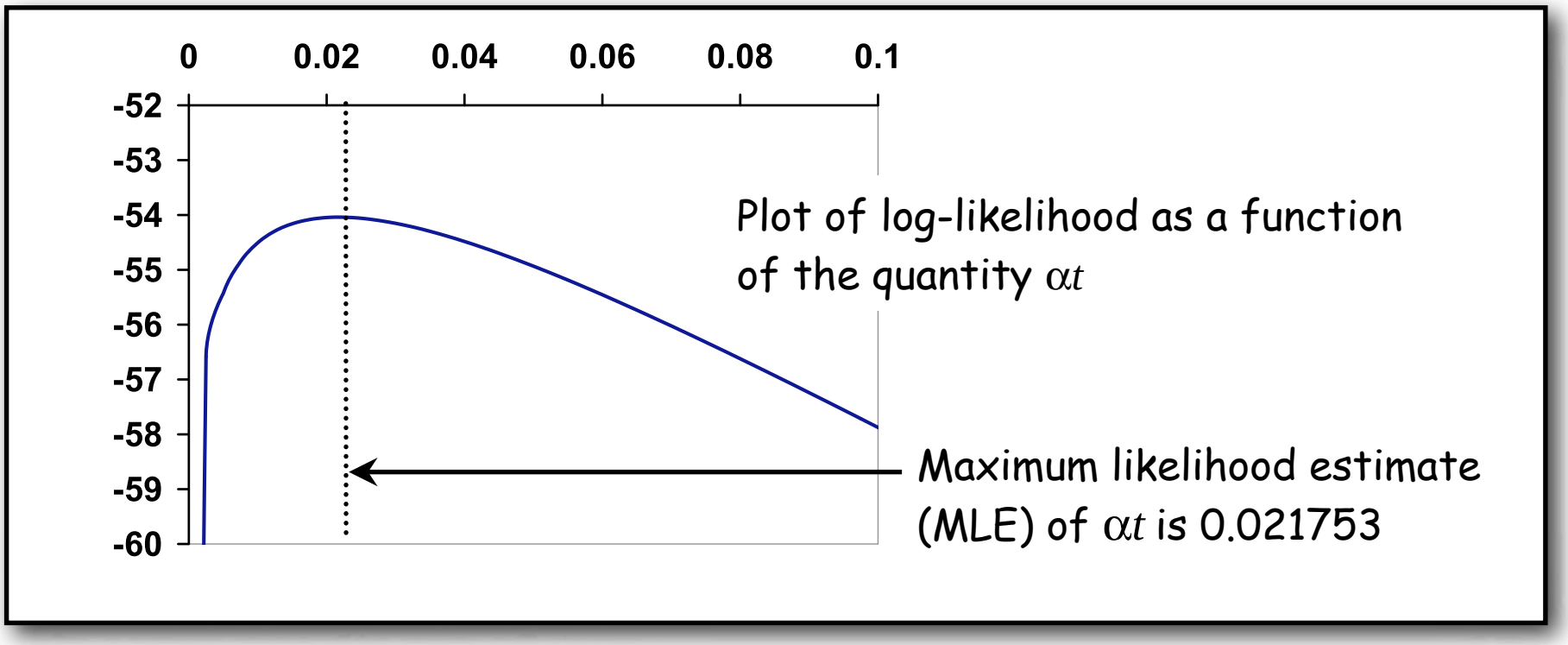
Note that we are NOT assuming independence here

Maximum likelihood estimation

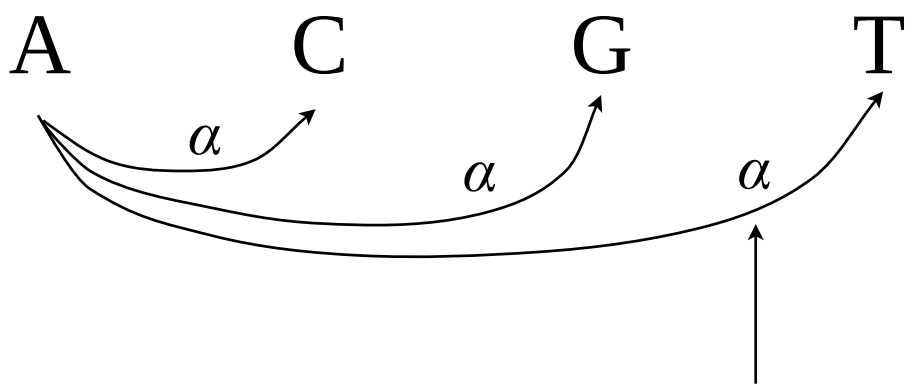
First 32 nucleotides of the ψη-globin gene of gorilla and orangutan:

```
gorilla   GAAGTCCTTGAGAAATAAACTGCACACACTGG
orangutan GGACTCCTTGAGAAATAAACTGCACACACTGG
```

$$L = \left[\left(\tfrac{1}{4}\right)\ \left(\tfrac{1}{4}+\tfrac{3}{4}e^{-4\alpha t}\right)\right]^{30}\left[\left(\tfrac{1}{4}\right)\ \left(\tfrac{1}{4}-\tfrac{1}{4}e^{-4\alpha t}\right)\right]^{2}$$



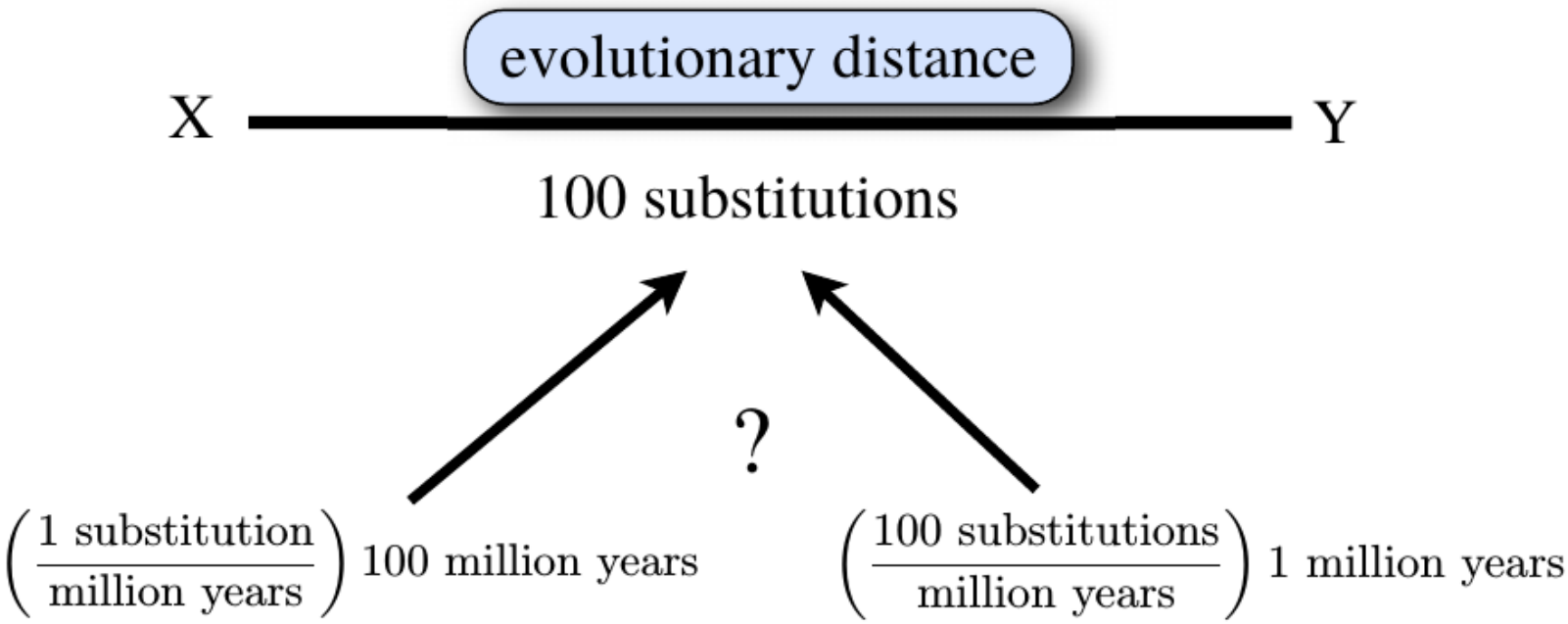
number of substitutions = rate × time



Overall substitution rate is 3α, so the expected number of substitutions (v) is

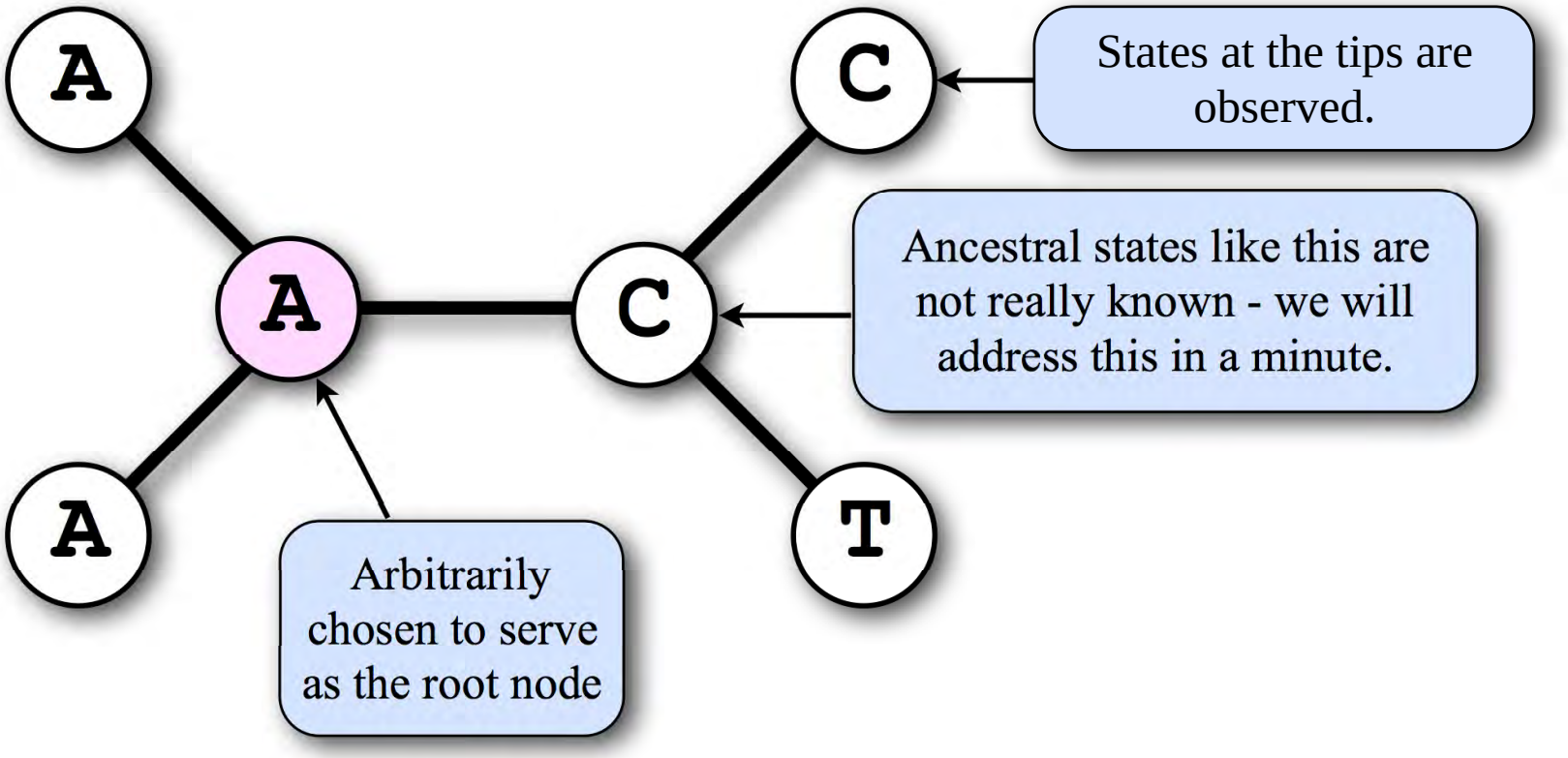
$$v = 3\alpha t$$

Rate and time are confounded

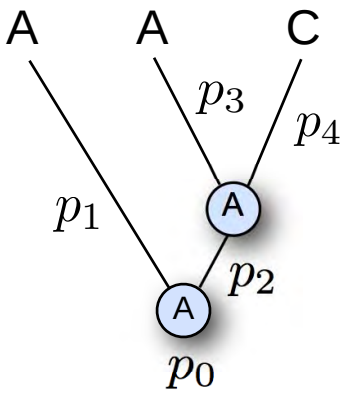


Likelihood of an unrooted tree

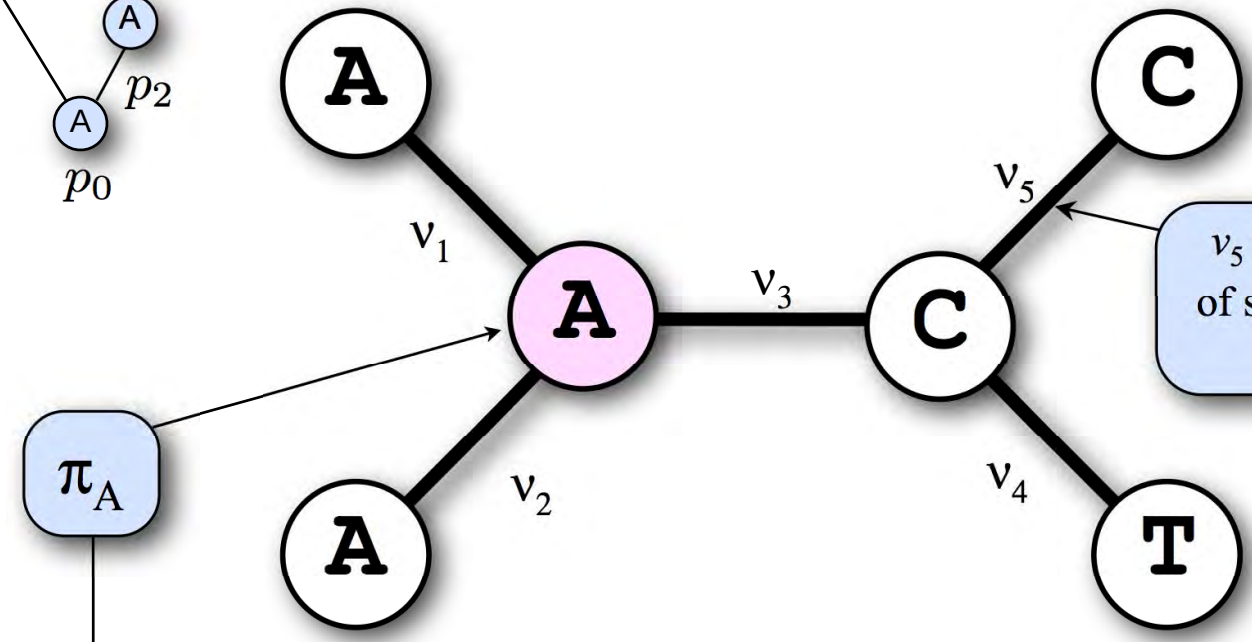
(data shown for only one site)



From slide 6



Likelihood for site *k*



v_5 is the expected number of substitutions for just this one branch

$$L_k = \frac{1}{4}\left[\frac{1}{4}+\frac{3}{4}e^{-4\nu_1/3}\right]\left[\frac{1}{4}+\frac{3}{4}e^{-4\nu_2/3}\right]\left[\frac{1}{4}-\frac{1}{4}e^{-4\nu_3/3}\right]\left[\frac{1}{4}-\frac{1}{4}e^{-4\nu_4/3}\right]\left[\frac{1}{4}+\frac{3}{4}e^{-4\nu_5/3}\right]$$

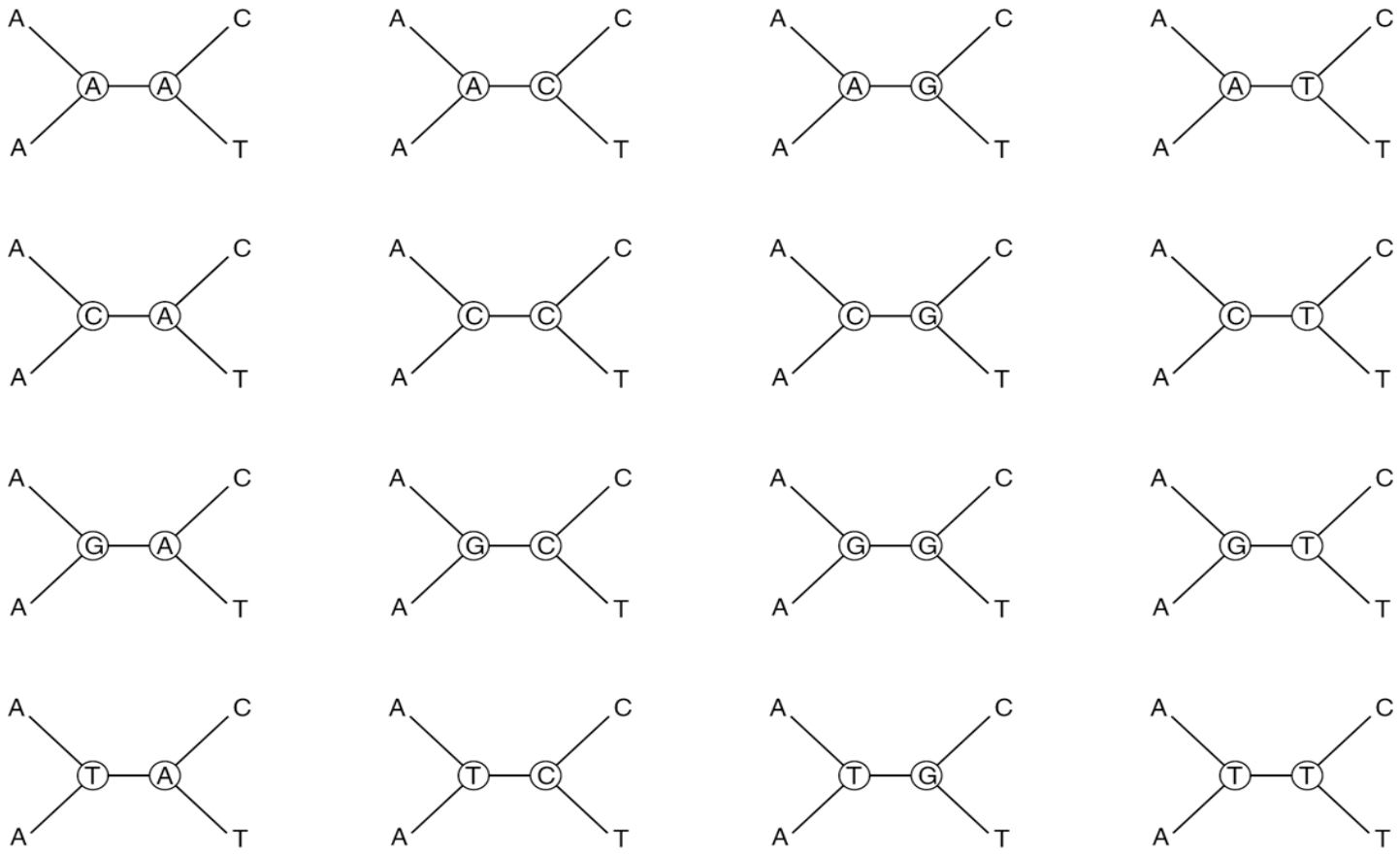
$P_{AA}(v_1)$ $P_{AA}(v_2)$ $P_{AC}(v_3)$ $P_{CT}(v_4)$ $P_{CC}(v_5)$

Note use of the AND probability rule

How do we know the states at the internal nodes?

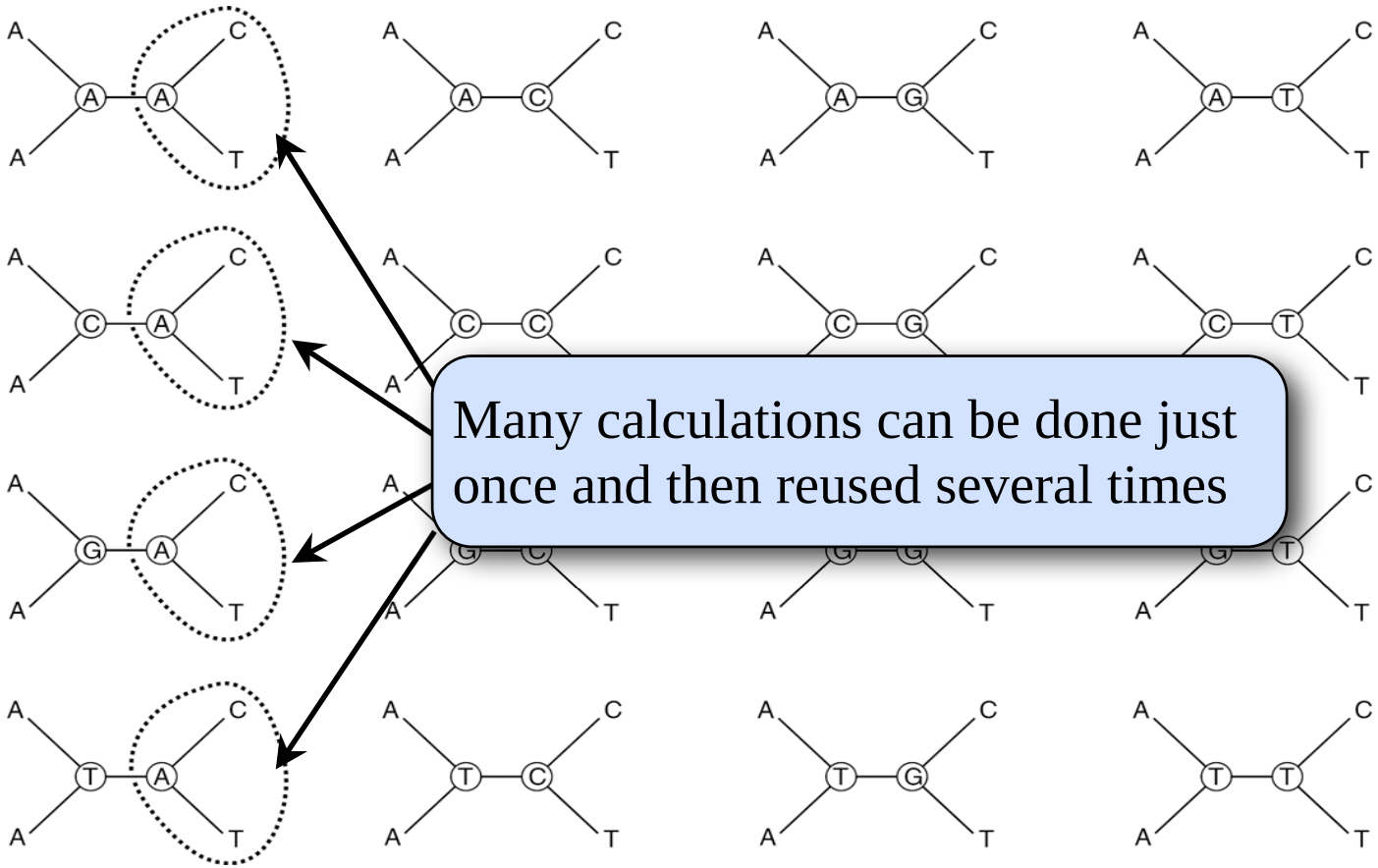
How do we know the states at the internal nodes?
We don't! We must consider all possible paths.

Brute force approach would be to calculate L_k for all 16 combinations of ancestral states and sum them



Note use of the OR probability rule

Pruning algorithm
(same result, less time)

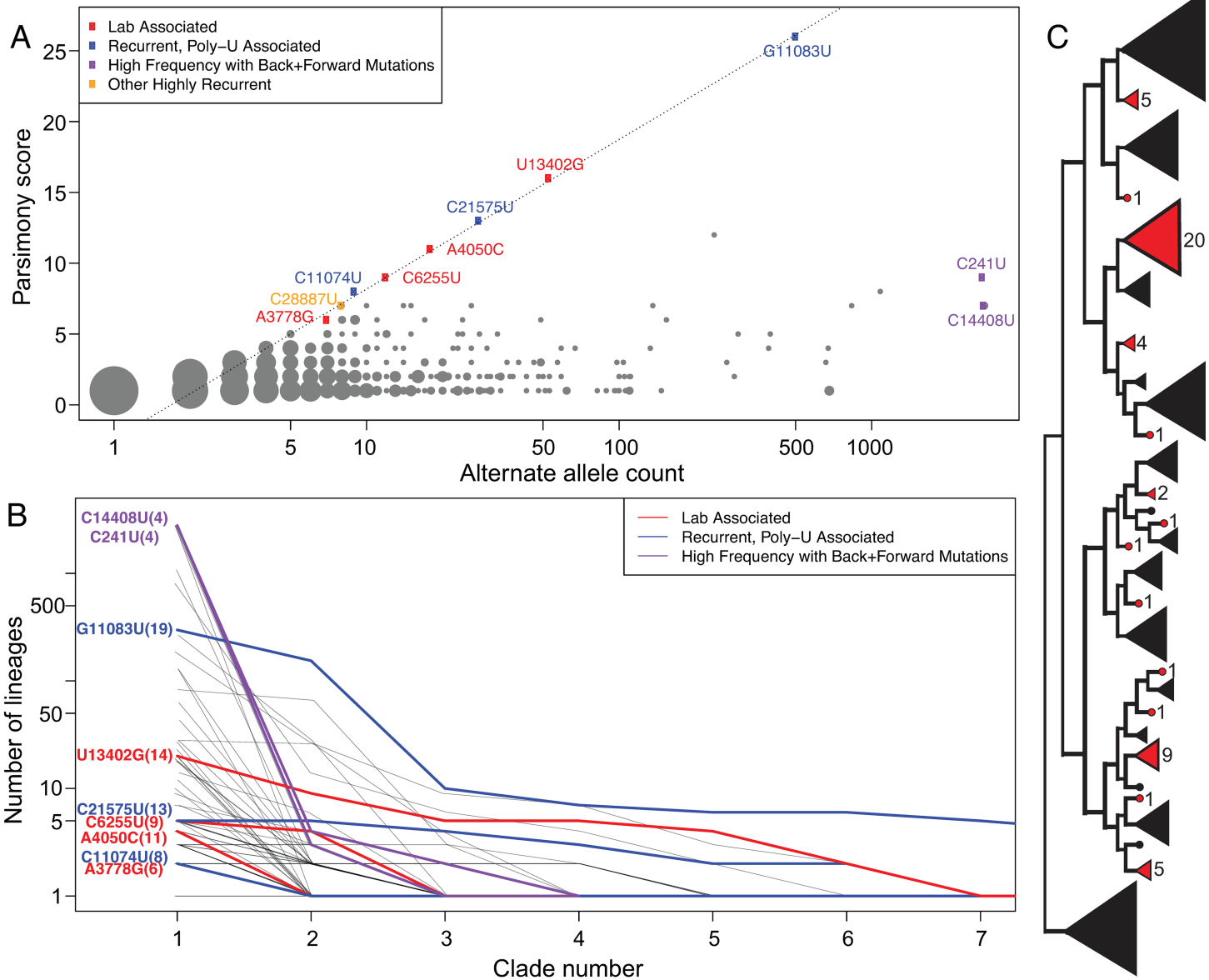


Felsenstein, J. 1981. Evolutionary trees from DNA sequences: a maximum likelihood approach. *Journal of Molecular Evolution* **17**:368-376

We will explore likelihood of different trees using these super cool widgets developed by Mark Holder:

- `http://phylo.bio.ku.edu/mephytis/barcharts.html`
- `http://phylo.bio.ku.edu/mephytis/brlen-opt.html`
- `http://phylo.bio.ku.edu/mephytis/tree-opt.html`

Discussion: How does the likelihood for the data observed at different site types vary, conditioned on the tree and branch lengths?



Stability of SARS-CoV-2 phylogenies Turakhia Y, De Maio N, Thornlow B, Gozashti L, Lanfear R, et al. (2020) Stability of SARS-CoV-2 phylogenies. PLOS Genetics 16(11): e1009175. https://doi.org/10.1371/journal.pgen.1009175

Bootstrapping

- Draw new data sets from your original data by sampling with replacement
- For each pseudoreplicate dataset estimate the best tree
- Calculate how often each split is recovered
- If not close to 100% may be sampling error
- If close to 100% likely not sampling error... but could still be a lot of other kinds of error!! Bootstraps are usually very high in very large (genomic) data sets.

https://phylo.bio.ku.edu/mephytis/boot-sample.html