Likelihood and Bayesian Methods in Biology

http://mtholder.github.io/like-bayes-bio/

Mark T. Holder (Univ. Kansas, USA)
Thanks to John Kelly and Ford Ballantyne

HCMR
15-18 June, 2015.

Source code:
https://github.com/mtholder/like-bayes-bio-4-day-workshop

Goals of the course

- Cover the basic theory associated with point estimation, interval estimation, and hypothesis testing in the maximum likelihood and Bayesian paradigms.

- Use some simple computer programs in Python and R to demonstrate that it is not too difficult to apply this form of statistical theory.

- The examples I use are all "toy" simulated datasets, so we will not really get into real biology.

Schedule

- Introduction to likelihoods and basic scripts for ML-based confidence intervals (today)
- Tuesday: numerical approaches used to optimize likelihoods, and dealing with multi-parameter models.
- Bayesian Introduction: Bayesian inference and Markov chain Monte Carlo
- Using model-jumping MCMC to perform model averaging.

The two competing approaches to statistics:

- frequentist:
  - **probability**: the relative frequency of an event if you were able to repeat a trial an infinite number of times.
  - **goal**: make an argument like "This result differs significantly from what we would expect if the null model were true. Either the null is not true or we experienced an unusually large amount of sampling error." The P-value summarizes how unusual it would be to see results like yours if the null hypothesis were true.
  - **we make probability statements about**: the long-run performance of our estimation procedures.
- Bayesian:
  - **probability**: degree of belief
  - **goal**: express the uncertainty of an estimate: "Given a model and what one knew before one collected the data, one should now believe that the $\mathbb{P}(\mu > 2.3) \approx 0.87$"
  - **we make probability statements about**: the true values of parameters/models.

The frequentist "recipe" for hypothesis testing

1. Ask a scientific question.
2. State your question in terms of H_0 and H_A
3. Collect a random sample
4. Calculate a value of a test statistic
5. Determine P-value:
   (a) What values of the test statistic are expected under H_0?
   (b) How does the observed test statistic differ from these expectations?
   (c) What is the probability of observing a value of a test statistic this extreme or more extreme if the H_0 is true? – this is the P-value.
6. Make a decision about H_0 and H_A
7. Answer your question and report the results.

Choosing a test statistic can be difficult.
Deriving a null distribution (step 5a) can be really difficult.

Question: Where does "likelihood" fit in?

Answer: the choice of a test statistic in frequentist statistics. And we'll see that the likelihood is central to Bayesian inference, too.

Law of likelihood: "the extent to which the evidence supports one parameter value or hypothesis against another is equal to the ratio of their likelihoods" [1]

Using a likelihood ratio as a test statistic $\rightarrow$ powerful test.

Using maximum likelihood as an estimator: $\rightarrow$ powerful, and statistically consistent (for well-behaved models) estimator.

[1] definition from Wikipedia – everyone's favorite source of assertions

P-values for likelihood ratios

Sometimes it is analytically tractable to calculate a null distribution.

More commonly, we test null models that are nested within a richer model, so we use the following trick:

For large sample sizes,
if you calculate a likelihood under a model with x extra parameters,
$2 \times$ the natural logarithm of
the ratio of likelihoods between the larger model and the true model
will be distributed according to $\chi^2_{df=x}$

This is often referred to as the "likelihood ratio test."

Definition of likelihood

In common English usage: "likelihood" $=$ "probability."

In statistics: The likelihood of a model/parameter θ based on observing data X is:

$$\begin{aligned} \ell(\theta) &= \mathbb{P}(X \mid \theta) \\ \ell(\theta) &= f(X \mid \theta) \end{aligned}$$

Because we use likelihood *ratios* for estimation, it is acceptable to use any function proportional to the probability of the data.

$$\begin{aligned} \ell(\theta) &\propto \mathbb{P}(X \mid \theta) \\ \ell(\theta) &\propto f(X \mid \theta) \end{aligned}$$

Only compare likelihoods when they are calculated on the same data set X.

The sum of likelihoods over all parameter values is...

some number, but nothing you can use.

$$\sum_i \mathbb{P}(X \mid \theta_i) \text{ is not necssarily } 1$$

By the laws of probability:

$$\sum \mathbb{P}(X_i \mid \theta) = 1$$

but we know what our data is, so we do not sum over all possible data sets!

Modeling

Modeling to perform likelihood calculations is the art of moving from a scientific question to abstract representations of the question which allow you to calculate the probability of a particular data outcome.

Modeling require understanding the rules of probability and often also requires:

- knowledge of what statistical distributions are natural fits for different processes, and
- knowledge of some form of stochastic process theory (often Markov processes).

Expertise in modeling requires a lot of training and is beyond the scope of this workshop.

But a few rules, go a long way...

rules of probability

1. $\mathbb{P}(A) = 1 - \mathbb{P}(\text{not } A)$
2. $\mathbb{P}(A, B) = \mathbb{P}(A)\mathbb{P}(B \mid A) = \mathbb{P}(B)\mathbb{P}(A \mid B)$
3. $\mathbb{P}(A, B) = \mathbb{P}(A)\mathbb{P}(B)$ if A and B are independent
4. $\mathbb{P}(A \text{ or } B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A, B)$
5. $\mathbb{P}(A \text{ or } B) = \mathbb{P}(A) + \mathbb{P}(B)$ if A and B are mutually exclusive.
6. $\mathbb{P}(A) = \sum_{i=1}^{n} [\mathbb{P}(A \mid B = b_i)\mathbb{P}(B = b_i)]$ if $B \in \{b_1, b_2, \ldots, b_n\}$.
7. Bayes' rule:

$$\mathbb{P}(A \mid B) = \frac{\mathbb{P}(B \mid A)\mathbb{P}(A)}{\mathbb{P}(B)}$$

A widely used result from continuous time Markov processes

If Q is a matrix of instantaneous rates where q_{ij} is the rate of transitioning from state i to state j, then:

$$\mathbb{P}(t) = e^{tQ}$$

where t is the time.

This lets us extrapolate about the effects of a constant process over time.

Working some examples

Now we'll switch to 3 examples.

There is python and R code for 2 of them.

The code is not the most robust in the world - you should try using a number of optimizers before you publish.

But it should help us demonstrate the concepts.

A toy example: estimating the mean from a sample

X: a set of random, independent continuous measurements sampled from the same population.

Assume that we know the population standard deviation, and want to estimate the population mean, μ.

$$\begin{aligned}
\ell(\mu) = \mathbb{P}(X \mid \mu, \sigma) &= \prod_{i=1}^{n} \mathbb{P}(x_i \mid \mu, \sigma) \\
&= \prod_{i=1}^{n} \left[\frac{1}{\sqrt{2\pi\sigma^2}} \, e^{-\frac{(x_i-\mu)^2}{2\sigma^2}} \right]
\end{aligned}$$

Note that μ that minimizes the squared error (the least squares estimator) maximizes the likelihood.

$$
\begin{aligned}
\ell(\mu) = \mathbb{P}(X \mid \mu, \sigma) &= \prod_{i=1}^{n} \mathbb{P}(x_i \mid \mu, \sigma) \\
&= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}}\, e^{-\frac{(x_i-\mu)^2}{2\sigma^2}} \\
\log L(\mu) &= \sum_{i=1}^{n} \log \left[\frac{1}{\sqrt{2\pi\sigma^2}}\, e^{-\frac{(x_i-\mu)^2}{2\sigma^2}} \right] \\
&= n \log \left[\frac{1}{\sqrt{2\pi\sigma^2}} \right] + \sum_{i=1}^{n} \log \left[e^{-\frac{(x_i-\mu)^2}{2\sigma^2}} \right] \\
&= n \log \left[\frac{1}{\sqrt{2\pi\sigma^2}} \right] - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2
\end{aligned}
$$

Note that μ that minimizes the squared error (the least squares estimator) maximizes the likelihood.

An example from phylogenetics

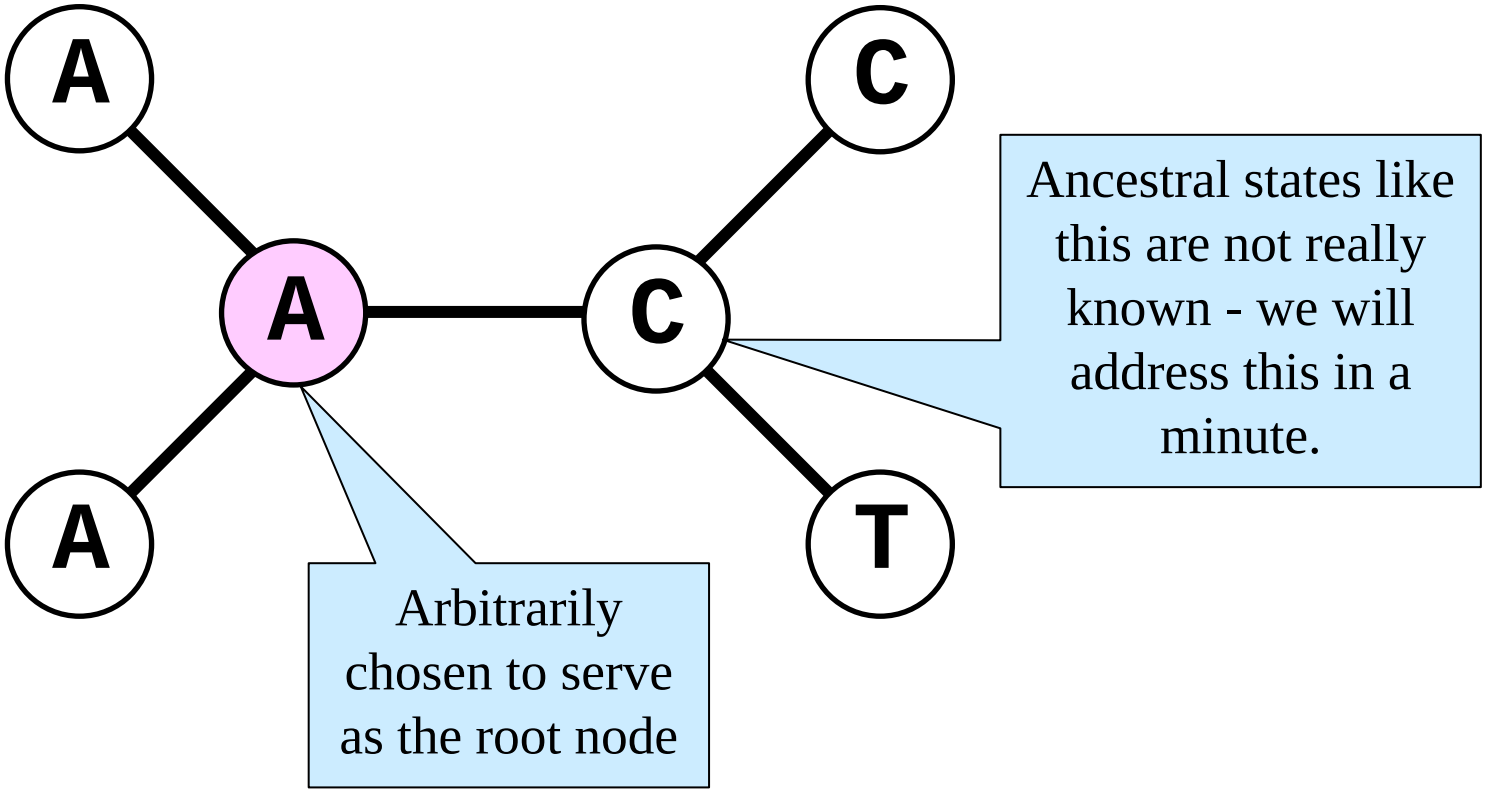
If we have a tree T and branch lengths ν as our model, what is the probability of an alignment of DNA sequences?

We can think of a branch length being the product of time t and rate of sequences evolution α:

$$\nu = \alpha t$$

Likelihood of a tree

(data for only one site shown)



JC69 model

- Bases are assumed to be equally frequent (all 0.25)
- Assumes rate of substitution (α) is the same for all possible substitutions
- Usually described as a 1-parameter model (the parameter being α)
- Remember, however, that each edge in a tree can have its own α, so there are really as many parameters in the model as there are edges in the tree!

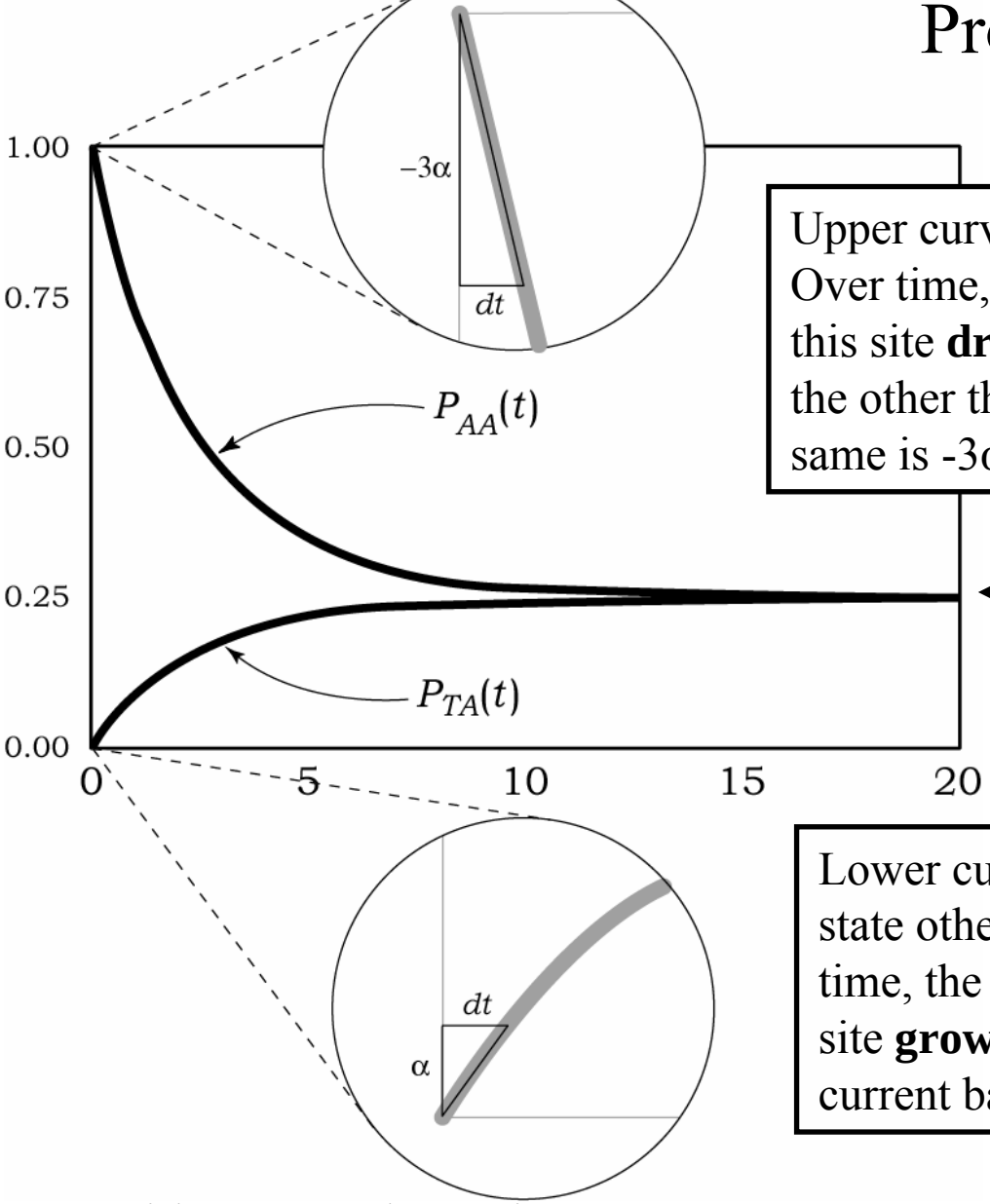
Jukes, T. H., and C. R. Cantor. 1969. Evolution of protein molecules. Pages 21-132 *in* H. N. Munro (ed.), *Mammalian Protein Metabolism*. Academic Press, New York.

JC instantaneous rate matrix - the Q matrix for JC

The 1 parameter is α (sometimes parameterized in terms of μ). This is the rate of replacements ("disruptions" that change the state):

		To State			
		A	C	G	T
From State	A	-3α	α	α	α
	C	α	-3α	α	α
	G	α	α	-3α	α
	T	α	α	α	-3α

Probability of "A present" as a function of time

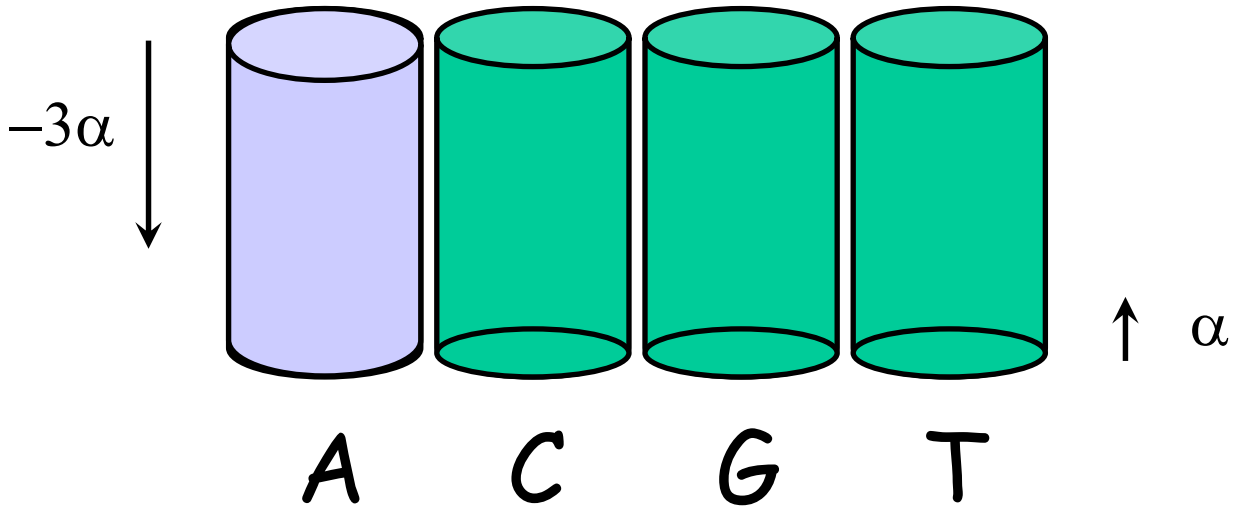


Upper curve assumes we started with A at time 0. Over time, the probability of still seeing an A at this site **drops** because rate of changing to one of the other three bases is 3α (so rate of staying the same is -3α).

The equilibrium relative frequency of A is 0.25

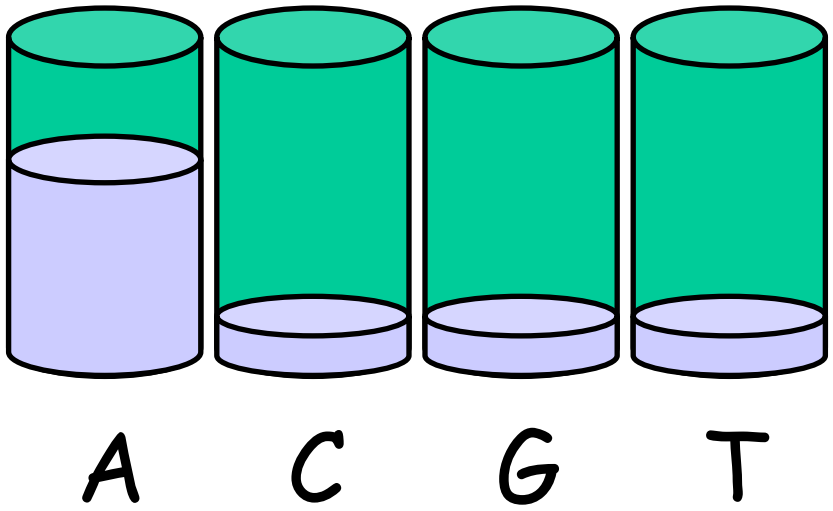
Lower curve assumes we started with some state other than A (T is used here). Over time, the probability of seeing an A at this site **grows** because the rate at which the current base will change into an A is α.

Water analogy (time 0)



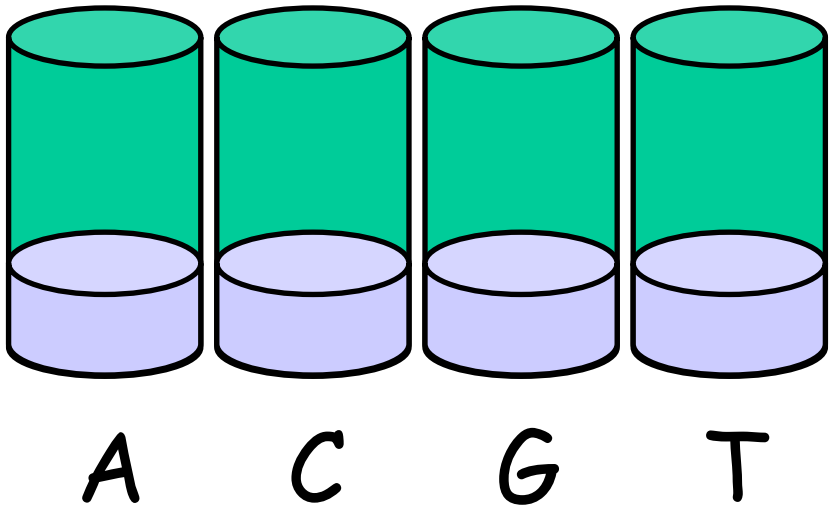
- Start with container A completely full and others empty
- Imagine that all containers are connected by tubes that allow same rate of flow between any two
- Initially, A will be losing water at 3 times the rate that C (or G or T) gains water

Water analogy (after some time)



A's level is not dropping as fast now because it is now also *receiving* water from C, G and T

Water analogy (after a very long time)



Eventually, all containers are one fourth full and there is zero *net* volume change – **stationarity** (equilibrium) has been achieved

(Thanks to Kent Holsinger for this analogy)

Change probabilities

We can calculate a transition probability matrix as a function of time by:

$$\mathbf{P}(t) = e^{\mathbf{Q}t}$$

The important thing to note is the rates ($\mathbf{Q}$ matrix) is multiplied by the time.

We can't separate rates and times since we always see the effect of their product.

Is a medium level of character divergence:

1. medium rate of change and medium amount of time,
2. high rate, but short time period,
3. low rate, but a long time period?

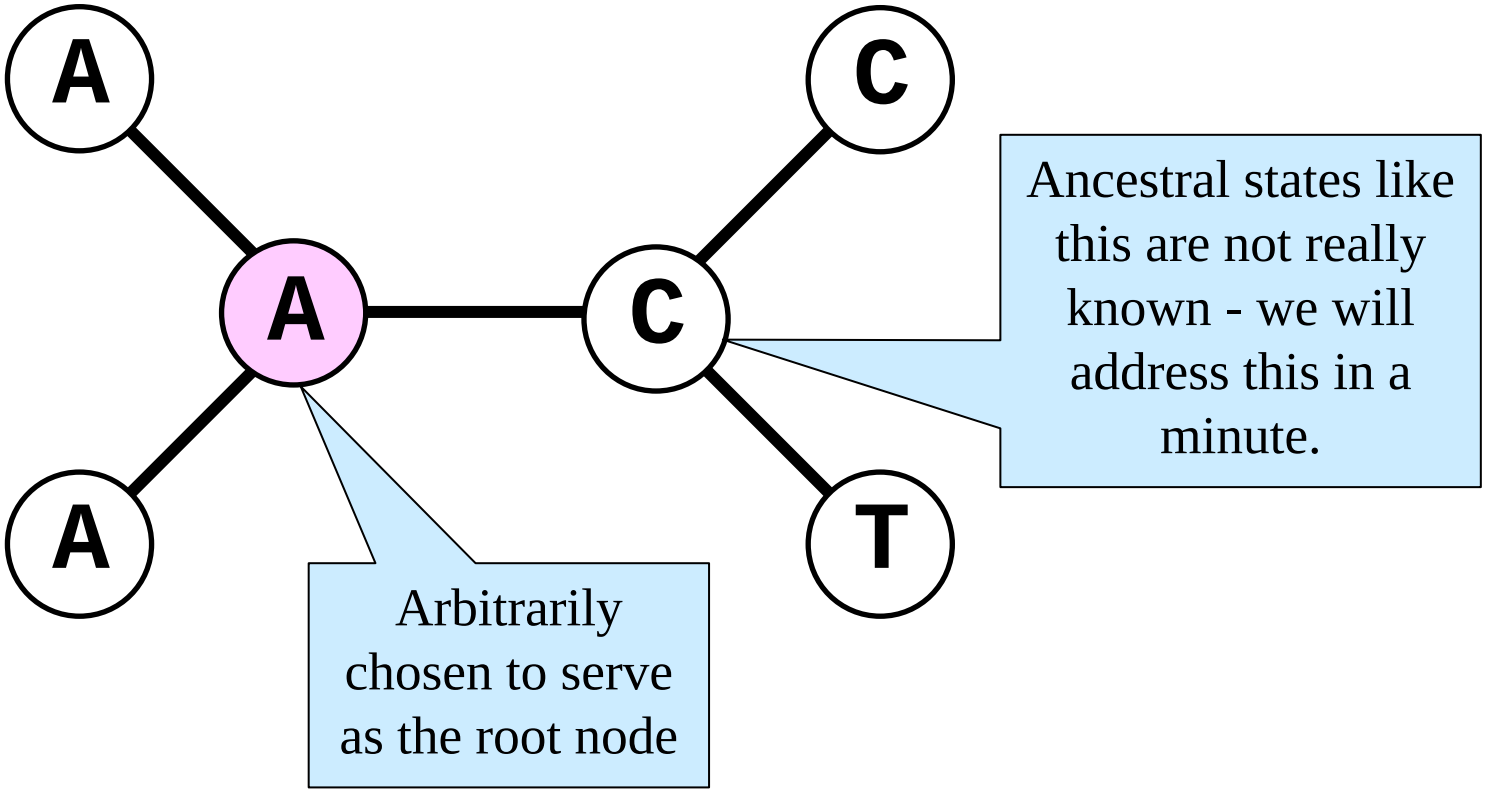
JC transition probabilities

$$\mathbb{P}_{ii}(\nu) = \mathbb{P}(\mathsf{end} = i \mid \mathsf{start} = i, \nu) = \frac{1}{4} + \frac{3}{4}e^{-4\nu/3}$$

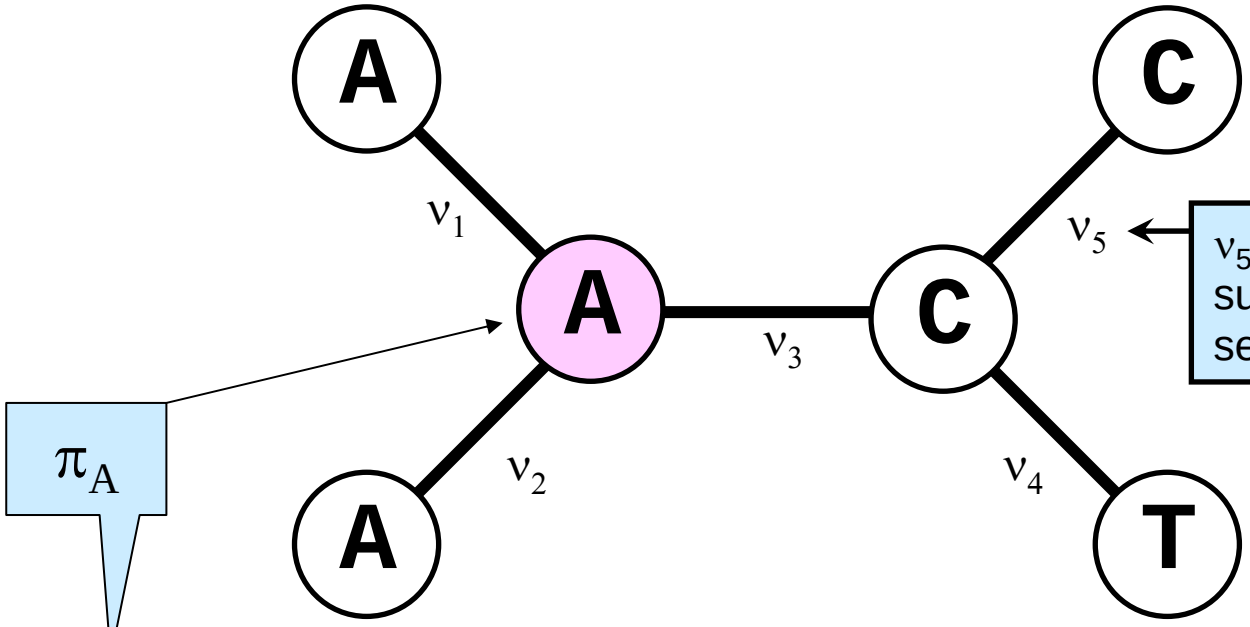
$$\mathbb{P}_{ij}(\nu) = \mathbb{P}(\mathsf{end} = j \mid \mathsf{start} = i, \nu) = \frac{1}{4} - \frac{1}{4}e^{-4\nu/3}$$

Likelihood of a tree

(data for only one site shown)



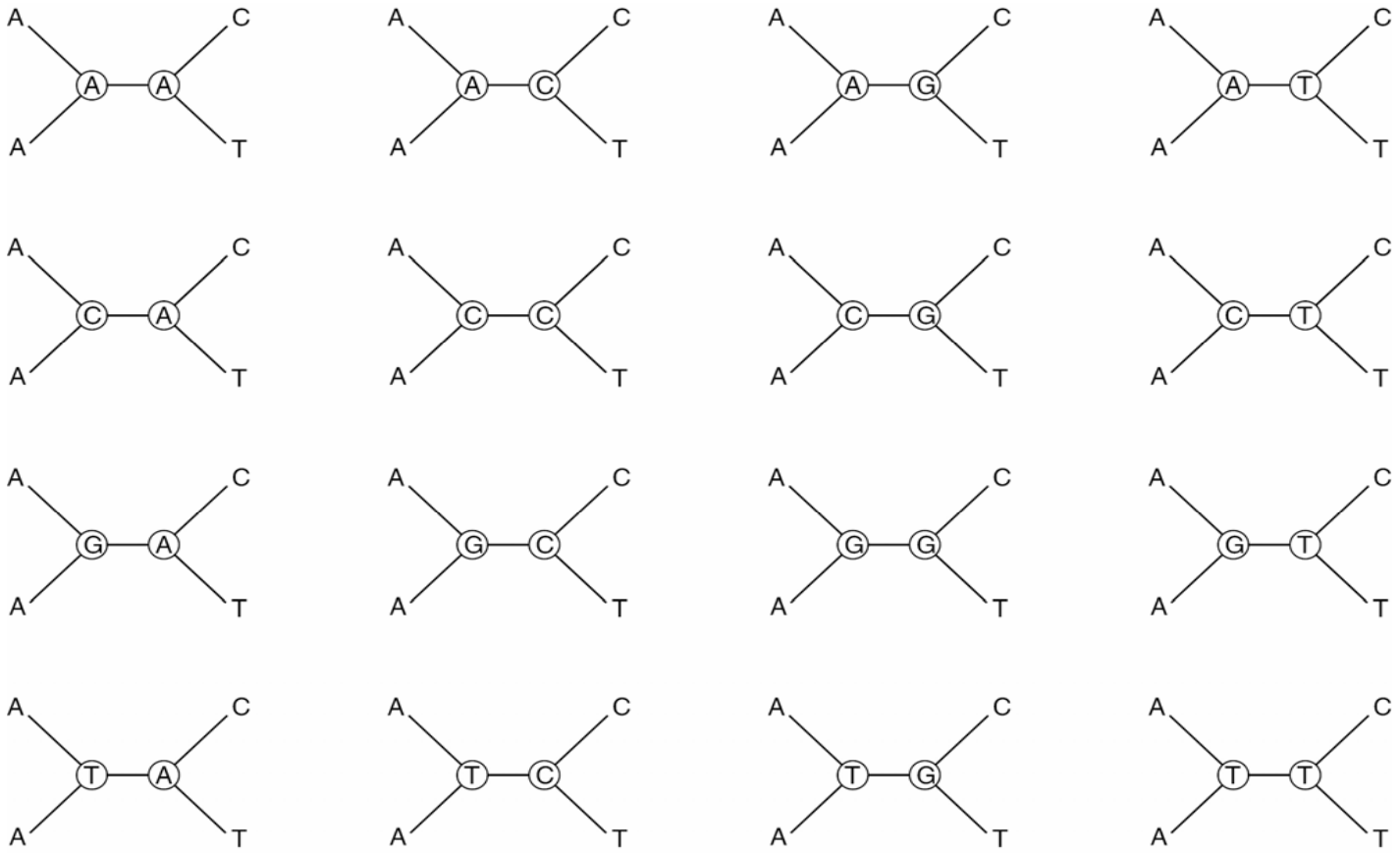
Likelihood for site k



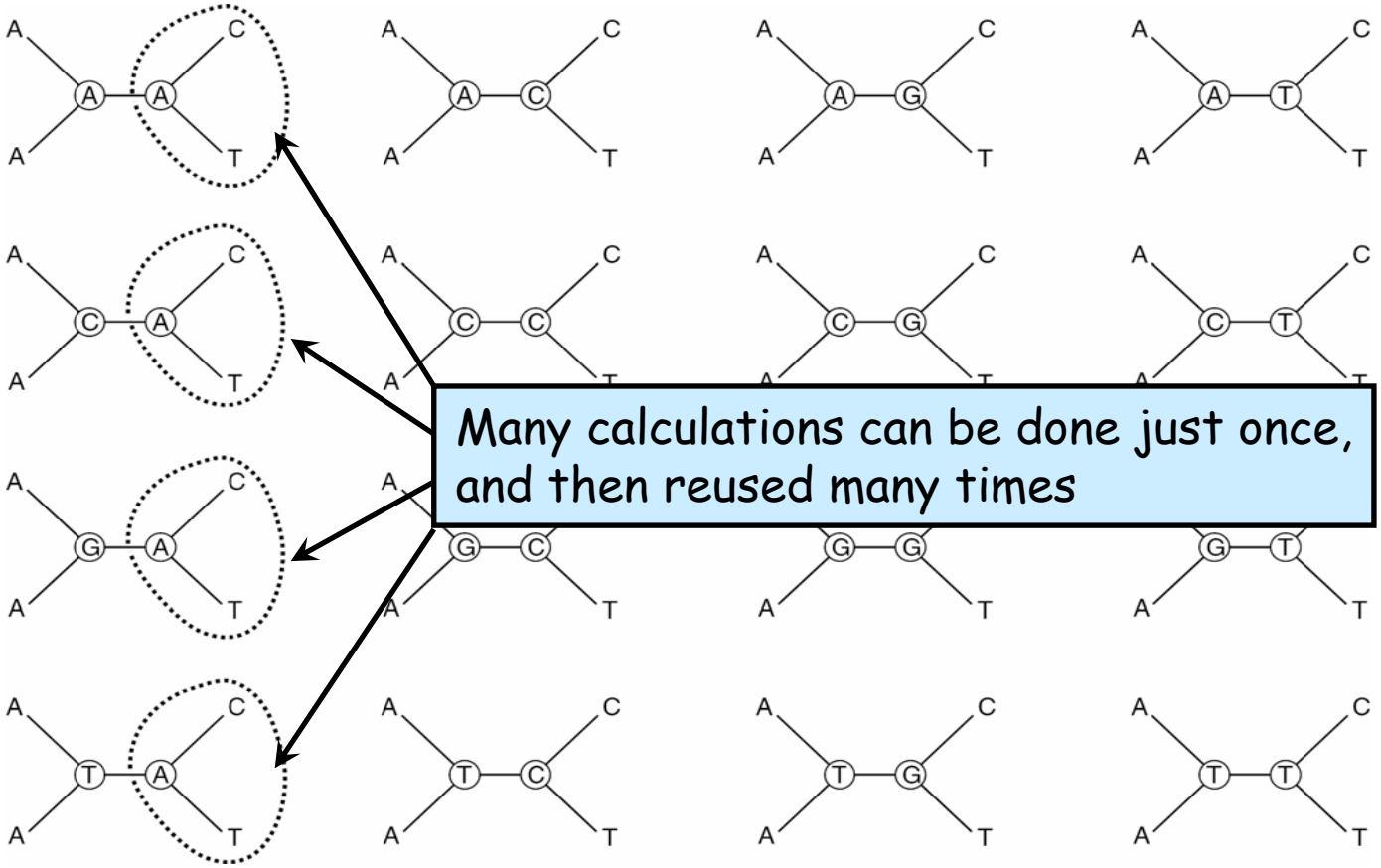
$$L_k = \tfrac{1}{4}\left[\tfrac{1}{4}+\tfrac{3}{4}e^{-4\nu_1/3}\right]\left[\tfrac{1}{4}+\tfrac{3}{4}e^{-4\nu_2/3}\right]\left[\tfrac{1}{4}-\tfrac{1}{4}e^{-4\nu_3/3}\right]\left[\tfrac{1}{4}-\tfrac{1}{4}e^{-4\nu_4/3}\right]\left[\tfrac{1}{4}+\tfrac{3}{4}e^{-4\nu_5/3}\right]$$

$P_{AA}(\nu_1)$ $P_{AA}(\nu_2)$ $P_{AC}(\nu_3)$ $P_{CT}(\nu_4)$ $P_{CC}(\nu_5)$

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



Pruning algorithm*
(same result, much less time)



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

Another toy example

You tag 1000 territorial animals with transmitting tags.

Every month you survey the area. Assume that you can detect every tag attached to a living organism.

You know (from other studies) that the probability of a tag falling off are: 0.10 in the first month, 0.15 in the second month, 0.2 in the third month, and 0.25 for every month after that.

Can you estimate the per-month probability of death?