

# Phylogenetic inference and likelihood

Emily Jane McTavish

Life and Environmental Sciences  
University of California, Merced

`ejmctavish@ucmerced.edu`, `twitter:snacktavish`

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

## Should we expect character conflict?

- ▶ Data type?
- ▶ Evolutionary history?

## How can we deal with character conflict?

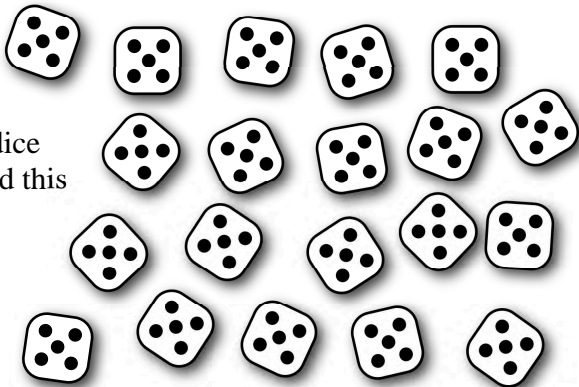
- ▶ We need to apply an error model
- ▶ Likelihood provides a measure of surprise under different models

# The Likelihood Criterion

The probability of the observations computed using a model tells us how surprised we should be.

*The preferred model is the one that surprises us least.*

Suppose I threw 20 dice down on the table and this was the result...



# Combining probabilities

- *Multiply* probabilities if the component events must happen **simultaneously** (i.e. where you would naturally use the word AND when describing the problem)

Using 2 dice, what is the probability of

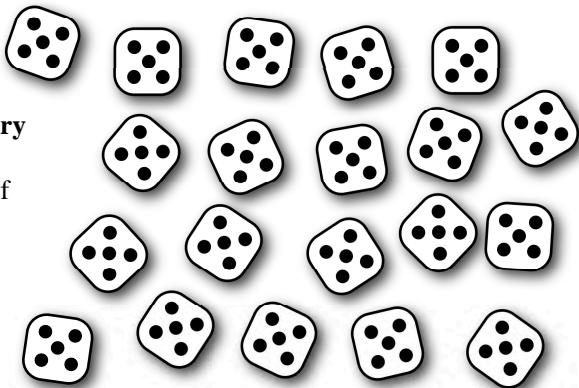


$$(1/6) \times (1/6) = 1/36$$

# The Fair Dice model

$$\Pr(\text{obs.} | \text{fair dice model}) = \left(\frac{1}{6}\right)^{20} = \frac{1}{3,656,158,440,062,976}$$

You should have been **very surprised** at this result because the probability of this event is **very small**: only 1 in 3.6 quadrillion!

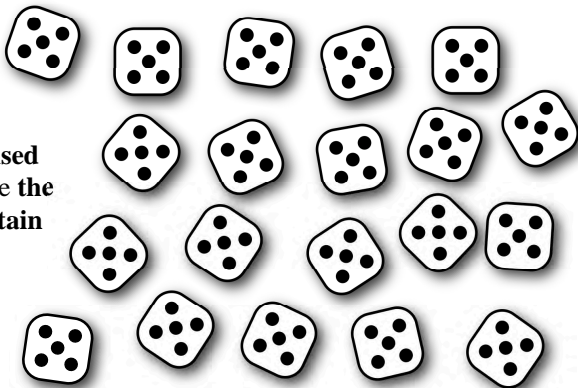


# The Trick Dice model

(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, <i>very</i> , <b>very</b> surprised
Trick Dice	1	Not surprised at all

winning model maximizes likelihood  
(and thus minimizes surprise)



# Likelihood: why a new term?

Outcome	Fair coin model	Two-heads model
H	0.5	1
T	0.5	0
	1	1

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

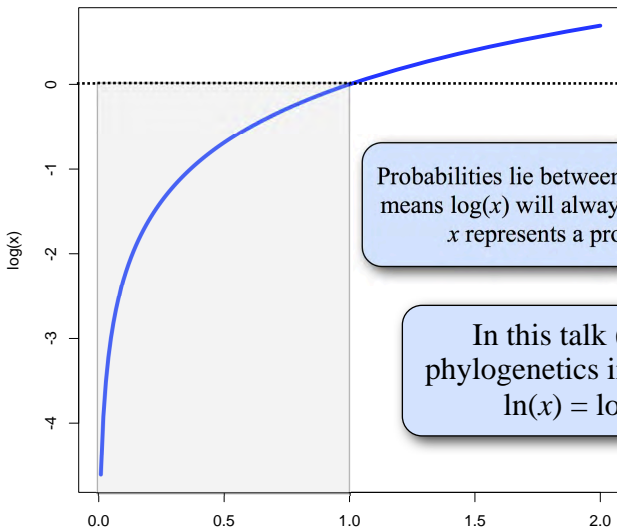
Probabilities of data outcomes given one particular model sum to 1.0

# Likelihood and model comparison

- Analyses using likelihoods ultimately involve **model comparison**
- The models compared can be **discrete** (as in the fair vs. trick dice example)
- More often the models compared differ **continuously**:
  - Model 1: branch length is 0.01
  - Model 2: branch length is 0.02
  - Model 3: branch length is 0.03

Rather than having an infinity of models, we instead think of the branch length as a **parameter** within one model

# Likelihoods vs. log-likelihoods



Probabilities lie between 0 and 1, which means  $\log(x)$  will always be negative if  $x$  represents a probability.

In this talk (and in phylogenetics in general),  
 $\ln(x) = \log(x)$

# Likelihood calculated from a single sequence

$$\Pr(A) = \pi_A$$

$$\Pr(C) = \pi_C$$

$$\Pr(G) = \pi_G$$

$$\Pr(T) = \pi_T$$

First 32 nucleotides of the  $\psi\eta$ -globin gene of gorilla:

**GAAGTCCTTGAGAAATAAACTGCACACACTGG**

$$\begin{aligned} L &= \pi_G \pi_A \pi_A \pi_G \pi_T \pi_C \pi_C \pi_T \pi_T \pi_G \pi_A \pi_G \pi_A \pi_A \pi_A \pi_T \pi_A \pi_A \pi_A \pi_C \pi_T \pi_G \pi_C \pi_A \pi_C \pi_A \pi_C \pi_A \pi_C \pi_T \pi_G \pi_G \\ &= \pi_A^{12} \pi_C^7 \pi_G^7 \pi_T^6 \end{aligned}$$

Note that we are assuming independence among sites here

$$\log L = 12 \log(\pi_A) + 7 \log(\pi_C) + 7 \log(\pi_G) + 6 \log(\pi_T)$$

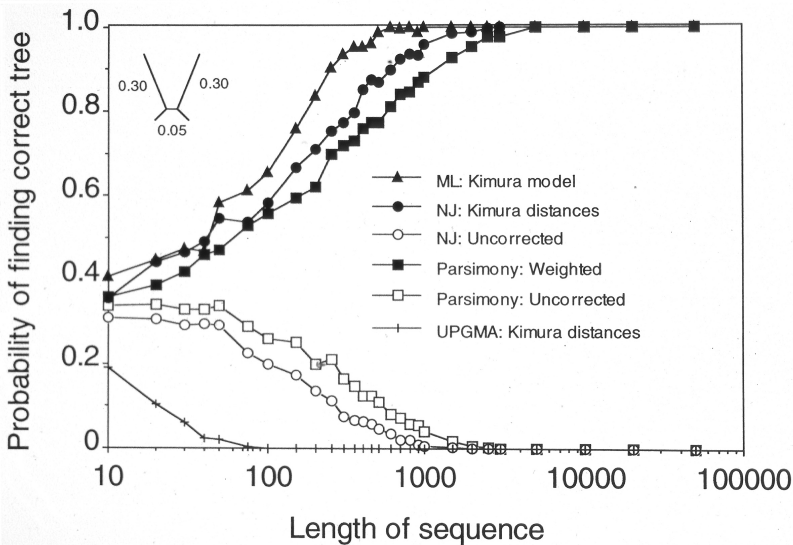
We can already see by eye-balling this that a model allowing **unequal** base frequencies will **fit better** than a model that assumes **equal** base frequencies because there are about twice as many As as there are Cs, Gs and Ts.

## Discussion Question

Is it possible for the EQUAL model to fit a data set better (using the likelihood to measure model fit) than the FLEXIBLE model? Why or why not?

## Historical aside





Hillis, D. M., J. P. Huelsenbeck, and D. L. Swofford. 1994. Hobgoblin of Phylogenetics? Nature 369:363-364.

# Likelihood calculated from a single sequence

$$\Pr(A) = \pi_A$$

$$\Pr(C) = \pi_C$$

$$\Pr(G) = \pi_G$$

$$\Pr(T) = \pi_T$$

First 32 nucleotides of the  $\psi\eta$ -globin gene of gorilla:

**GAAGTCCTTGAGAAATAAACTGCACACACTGG**

$$\begin{aligned} L &= \pi_G \pi_A \pi_A \pi_G \pi_T \pi_C \pi_C \pi_T \pi_T \pi_G \pi_A \pi_G \pi_A \pi_A \pi_A \pi_T \pi_A \pi_A \pi_A \pi_C \pi_T \pi_G \pi_C \pi_A \pi_C \pi_A \pi_C \pi_A \pi_C \pi_T \pi_G \pi_G \\ &= \pi_A^{12} \pi_C^7 \pi_G^7 \pi_T^6 \end{aligned}$$

Note that we are assuming independence among sites here

$$\log L = 12 \log(\pi_A) + 7 \log(\pi_C) + 7 \log(\pi_G) + 6 \log(\pi_T)$$

We can already see by eye-balling this that a model allowing **unequal** base frequencies will **fit better** than a model that assumes **equal** base frequencies because there are about twice as many As as there are Cs, Gs and Ts.



# Model ranking using LRT or AIC

Likelihood Ratio Tests (LRT) and the Akaike Information Criterion (AIC) provide two ways to evaluate whether an **unconstrained** model fits the data significantly better than a **constrained** version of the same model.

Find *maximum*  $\log L$  under the *unconstrained* model:

$$\begin{aligned}\log L_{\text{unconstrained}} &= 12 \log(\pi_A) + 7 \log(\pi_C) + 7 \log(\pi_G) + 6 \log(\pi_T) \\ &= 12 \log(0.375) + 7 \log(0.219) + 7 \log(0.219) + 6 \log(0.187) \\ &= -43.1\end{aligned}$$

This model has 3 estimated parameters

Find *maximum*  $\log L$  under the *constrained* model:

$$\begin{aligned}\log L_{\text{constrained}} &= 12 \log(\pi_A) + 7 \log(\pi_C) + 7 \log(\pi_G) + 6 \log(\pi_T) \\ &= 12 \log(0.25) + 7 \log(0.25) + 7 \log(0.25) + 6 \log(0.25) \\ &= -44.4\end{aligned}$$

This model has 0 estimated parameters

# Likelihood Ratio Test (LRT)

Calculate the likelihood ratio test statistic:

$$\begin{aligned} R &= -2 [\log(L_{\text{constrained}}) - \log(L_{\text{unconstrained}})] \\ &= -2 [-44.4 - (-43.1)] \\ &= 2.6 \end{aligned}$$

(Note that the log-likelihoods used in the test statistic have been *maximized* under each model separately)

“unconstrained” does fit better than “constrained” ( $-43.1 > -44.4$ ), but not significantly better ( $P = 0.457$ , chi-squared with 3 d.f.\*)

\*The number of degrees of freedom equals the difference between the two models in the number of estimated parameters. In this case, unconstrained has 3 parameters and constrained has 0, so d.f. =  $3 - 0 = 3$

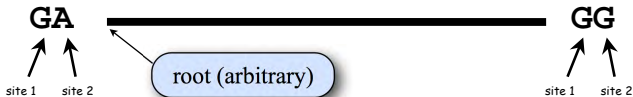
Comparing models in phylogenetics can be challenging, as topologies are not nested within one another.

We will discuss appropriate statistical approaches later in the course.

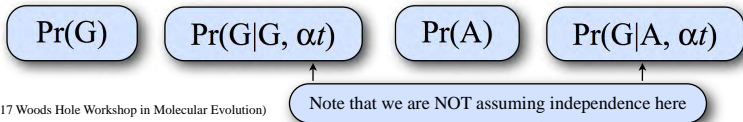
# Likelihood of the simplest tree

sequence 1  sequence 2

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



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



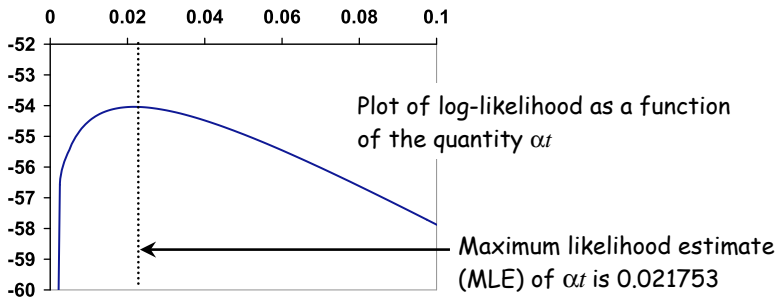
# Maximum likelihood estimation

First 32 nucleotides of the  $\psi\eta$ -globin gene of gorilla and orangutan:

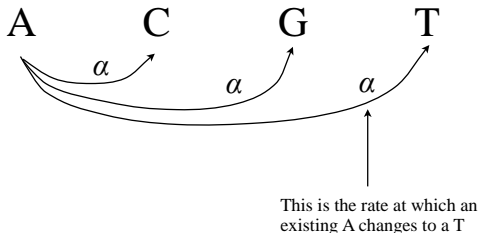
gorilla **GAAG**TCCTTGAGAAATAAACTGCACACACTGG

orangutan **GAAC**TCCTTGAGAAATAAACTGCACACACTGG

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



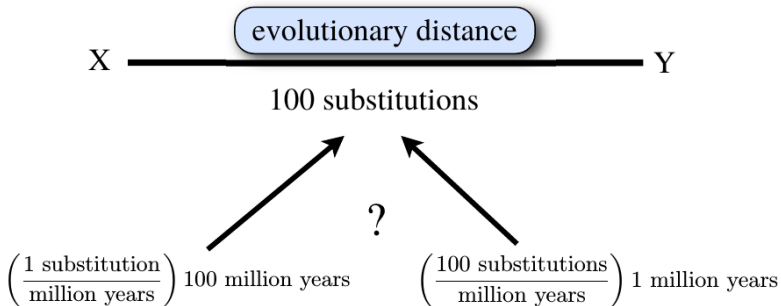
number of substitutions = rate  $\times$  time



Overall substitution rate is  $3\alpha$ , so the expected number of substitutions ( $v$ ) is

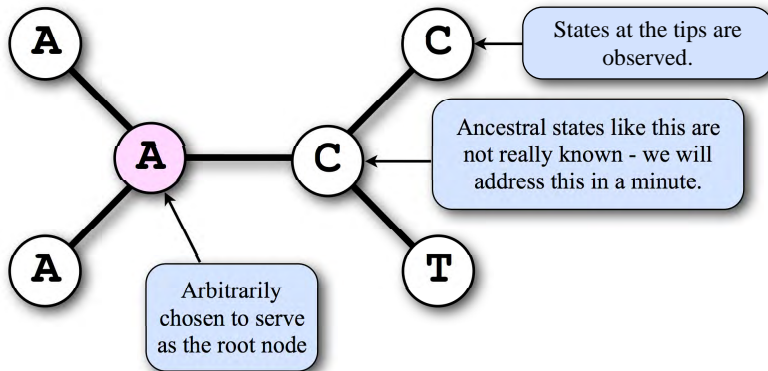
$$v = 3\alpha t$$

## Rate and time are confounded



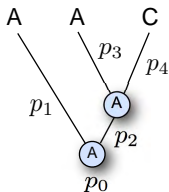
# Likelihood of an unrooted tree

(data shown for only one site)

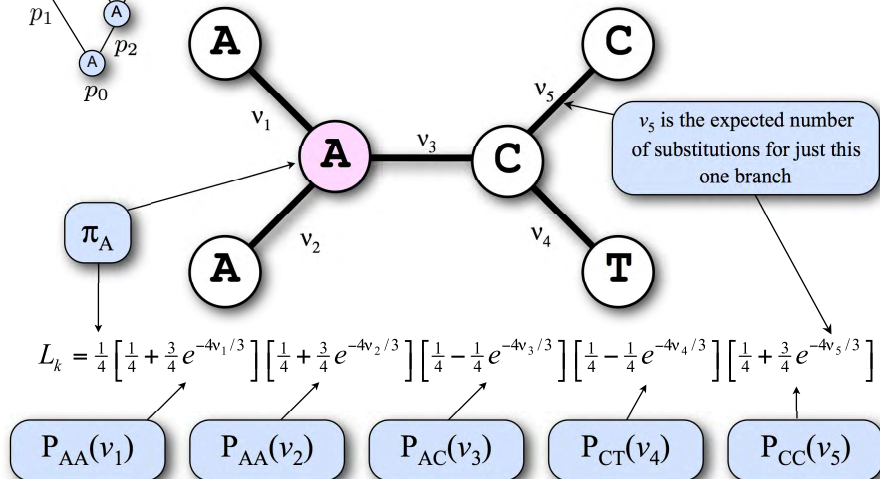




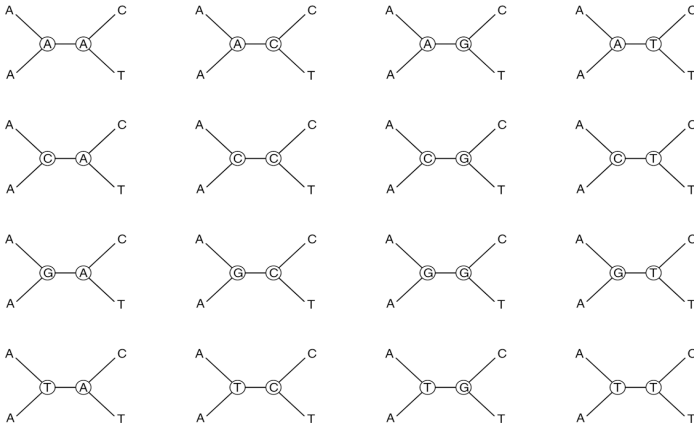
From slide 6



# Likelihood for site $k$



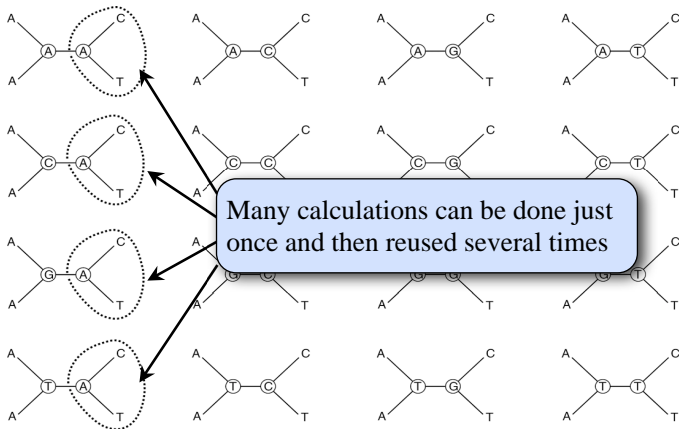
Brute force approach would be to calculate  $L_k$  for all 16 combinations of ancestral states and sum them



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 uper 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>

and using sequence simulation and analyses using seq-gen and paup

## 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>