

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)

Rule: Two taxa that share a character state must be more closely related to each other than either is to a taxon that displays a different state.(method suggested by Hennig)

Is this a valid rule?

Hennigian logical analysis

The German entomologist Willi Hennig (in addition to providing strong arguments for phylogenetic classifications) clarified the logic of phylogenetic inference.

Hennig's correction to our rule: Two taxa that share a **derived** character state must be more closely related to each other than either is to a taxon that displays the **primitive** state.

Hennig's logic is valid

Here we will use 0 for the primitive state, and 1 for the derived state.

	placenta	vertebra
<i>Homo sapiens</i>	1	1
<i>Rana catesbiana</i>	0	1
<i>Drosophila melanogaster</i>	0	0

Now the character “placenta” does not provide a grouping, but “vertebra” groups human and frog as sister taxa.

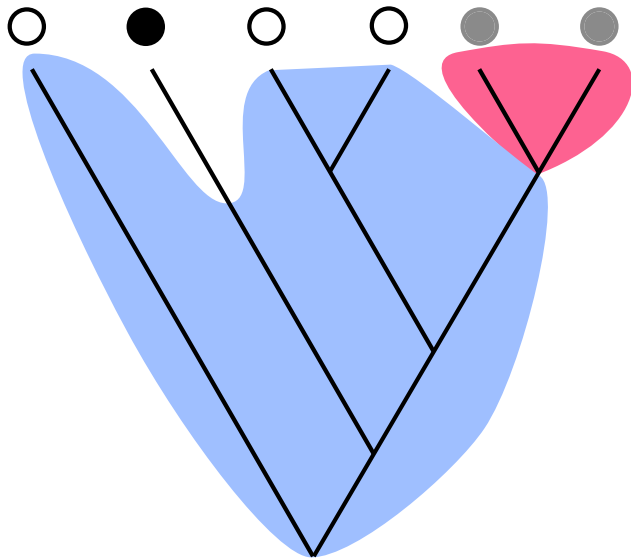
Hennigian terminology

prefixes:

- “apo” - refers to the new or derived state
- “plesio” - refers to the primitive state
- “syn” or “sym” - used to indicate shared between taxa
- “aut” - used to indicate a state being unique to one taxon

Hennigian rules

- synapomorphy - shared, derived states. Used to diagnose monophyletic groups.
- symplesiomorphy - shared, primitive states. Diagnose icky, unwanted paraphyletic groups.
- autapomorphy – a unique derived state. **No** evidence of phylogenetic relationships.
- constant characters – columns in a matrix with no variability between taxa. **No** evidence of phylogenetic relationships.



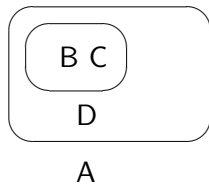
Hennigian inference

When we create a character matrix for Hennig's system, it is crucial that:

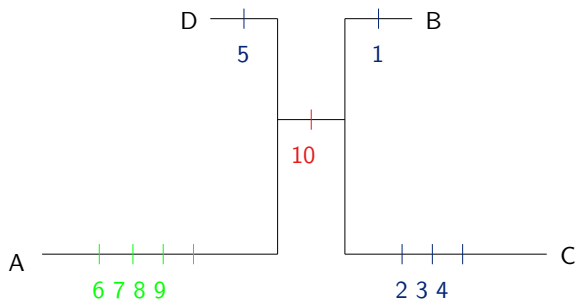
- traits assigned the same state represent homologous states (trace back to the MRCA)
- we correctly identify the directionality of the transformations (which state is plesiomorphic and which is apomorphic).
The process of identifying the direction of change is called polarization.

Polarization could be done based on developmental considerations, paleontological evidence, or biogeographic considerations, but the most common technique is outgroup polarization.

Taxon	Character #									
	1	2	3	4	5	6	7	8	9	10
A	0	0	0	0	0	0	0	0	0	0
B	1	0	0	0	0	1	1	1	1	1
C	0	1	1	1	0	1	1	1	1	1
D	0	0	0	0	1	1	1	1	1	0



- ▶ If characters are not polarized (ancestral and descendent states known) method can infer unrooted trees.
- ▶ We can infer tree topology, but be unable to tell paraphyletic from monophyletic groups.
- ▶ The outgroup method amounts to inferring an unrooted tree and then rooting the tree on the branch that leads to an outgroup.



B	C
A	D

problems with this approach

- ▶ We don't know polarization
- ▶ We observe character conflict in real data sets

Character conflict

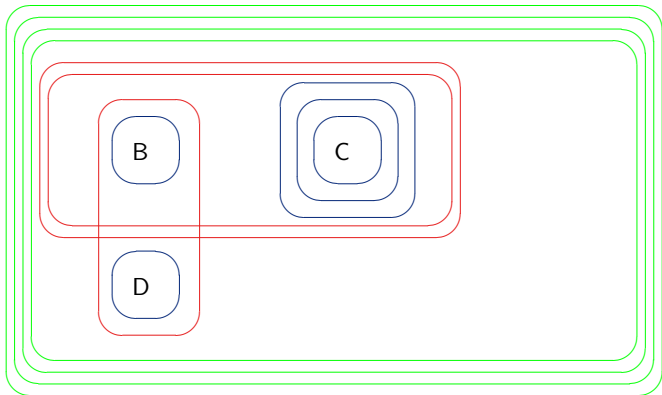
<i>Homo sapiens</i>	A G TTCAAG T
<i>Rana catesbiana</i>	A A TTCAAG T
<i>Drosophila melanogaster</i>	A G TTCAAG C
<i>C. elegans</i>	A A TTCAAG C

The red character implies that either (*Homo* + *Drosophila*) is a group (if G is derived) and/or (*Rana* + *C. elegans*) is a group.

The green character implies that either (*Homo* + *Rana*) is a group (if T is derived) and/or (*Drosophila* + *C. elegans*) is a group.

The green and red character cannot both be correct.

Taxon	Character #											
	1	2	3	4	5	6	7	8	9	10	11	12
A	0	0	0	0	0	0	0	0	0	0	0	0
B	1	0	0	0	0	1	1	1	1	1	1	1
C	0	1	1	1	0	1	1	1	1	1	1	0
D	0	0	0	0	1	1	1	1	1	0	0	1



A

Character conflict

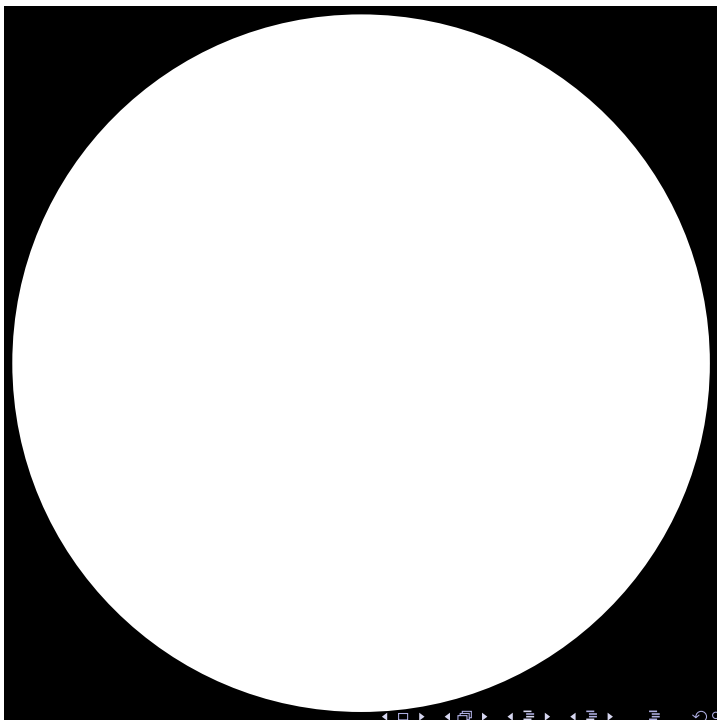
Two characters are compatible if they can both be mapped on the same tree so that all of the character states displayed could be homologous.

Incompatible characters are evidence of *homoplasy* in the data

Homoplasy literally means the “same change” has occurred more than once in the evolutionary history of the group.

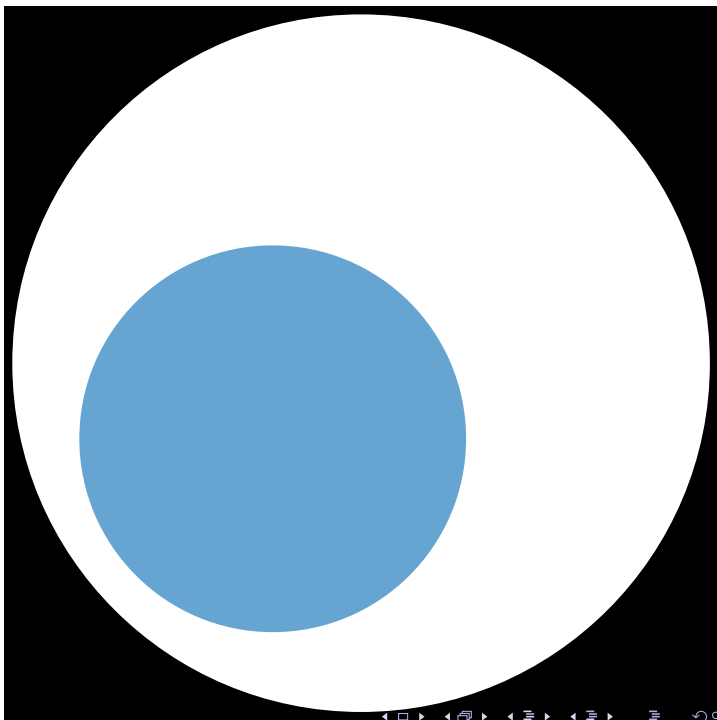
The presence of homoplasy undermines Hennigian analyses.

white = space
of all possible
matrices



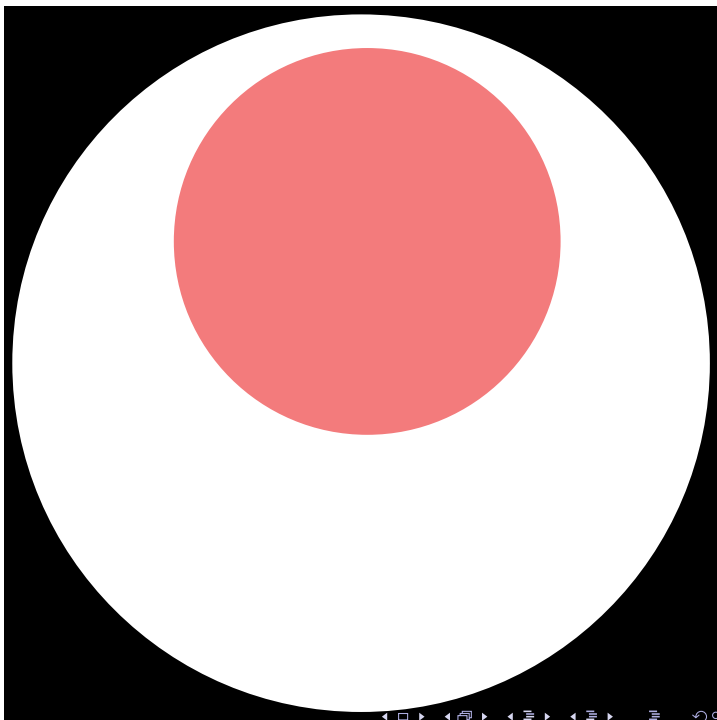
blue = space
of matrices with
the pattern:

A	B	C	D
-	*	*	-



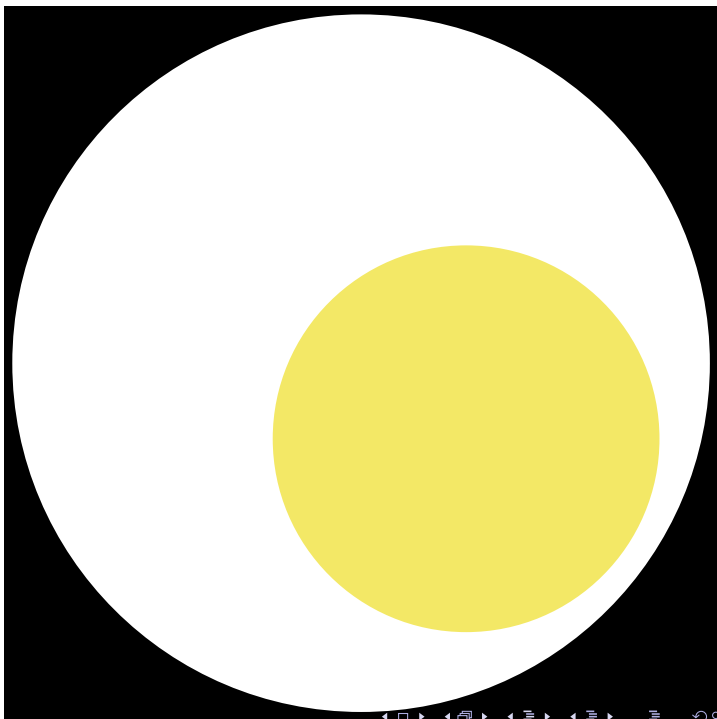
red = space
of matrices with
the pattern:

A	B	C	D
-	*	-	*



yellow = space
of matrices with
the pattern:

A	B	C	D
-	-	*	*



all eight
categories of
matrices



blue = space
of matrices
compatible
with tree:

$(A, (B, C), D)$



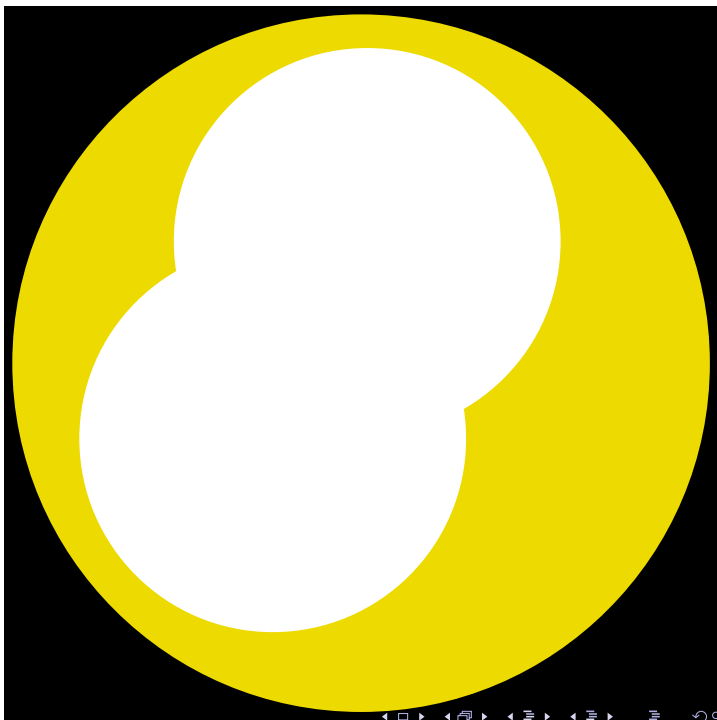
blue = space
of matrices
compatible
with tree:

$(A, C, (B, D))$



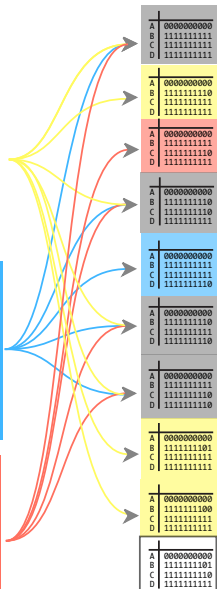
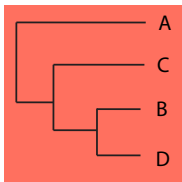
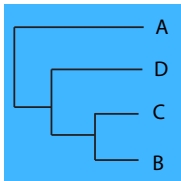
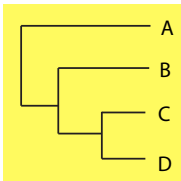
blue = space
of matrices
compatible
with tree:

$(A,B,(C,D))$



Hennigian:
grey = any tree
blue = B+C
red = B+D
yellow = C+D
white = no tree
(conflicting
characters)





What can we do if our data end up in the white (character conflict) or grey (uninformative characters only) zone?

- can we detect character conflict?
- is there a logic-based solution to the problem of character conflict?

Detecting character conflict in binary characters

Consider the four possible combinations of states in a two-character matrix.

The characters are incompatible *iff* (when you look across all taxa) you see all four state combinations.

		Char 1	
		0	1
Char 2	0	×	×
	1	×	×

What can we do if our data end up in the white (character conflict) or grey (uninformative characters only) zone?

- Can we detect character conflict? Yes
- Is there a logic-based solution to the problem of character conflict?
 - recoding characters?
 - “reciprocal illumination”?

What can we do if our data end up in the white (character conflict) or grey (uninformative characters only) zone?

- Can we detect character conflict? Yes
- Is there a logic-based solution to the problem of character conflict? No, nothing purely based on logic (and the suggestions for culling data to make matrices suitable for logical inference can lead to unsatisfyingly subjective analyses).
- What can we do?

We must have an “error model”

Should we expect character conflict?

- ▶ Data type?
- ▶ Evolutionary history?

How can we deal with character conflict?

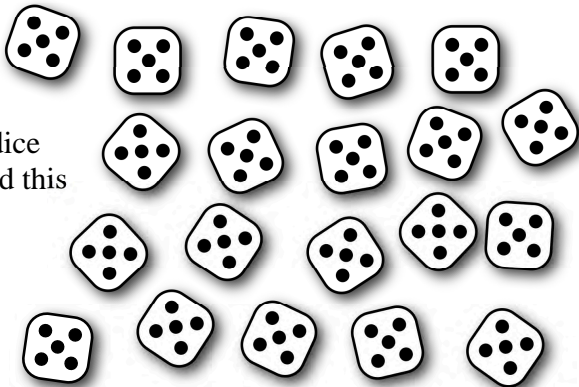
- ▶ We need to apply an error model
- ▶ Likelihood provides a measure of surprise under different models

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

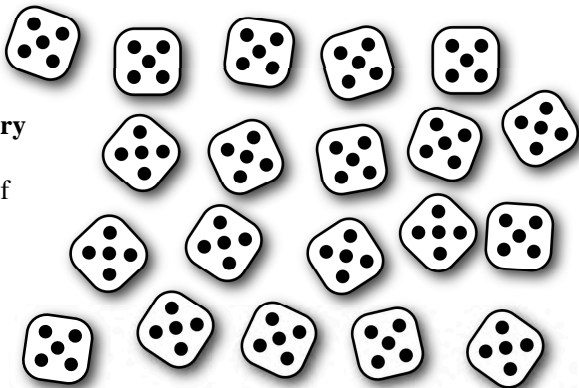


$$(1/6) \times (1/6) = 1/36$$

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!

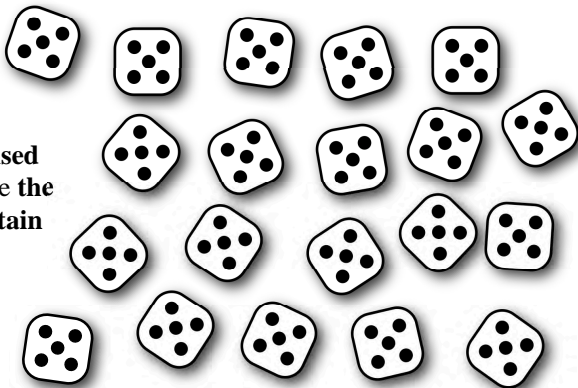


The Trick Dice model

(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> , very 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

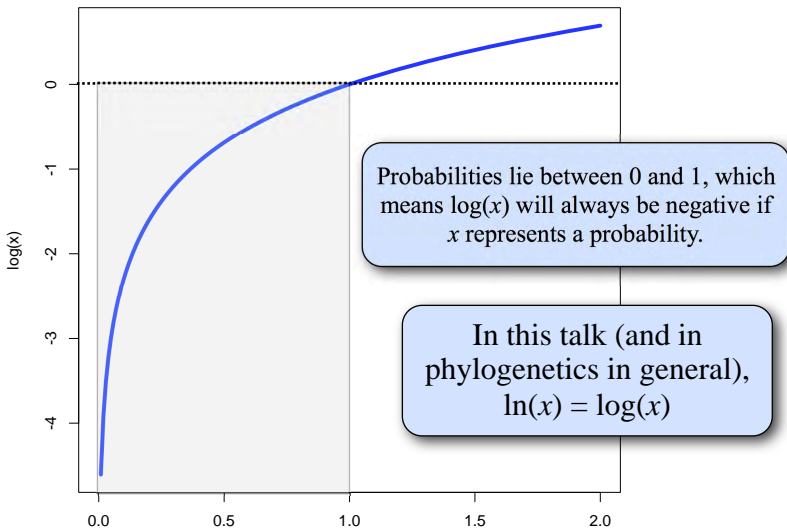
Probabilities of data outcomes given one particular model sum to 1.0

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



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

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

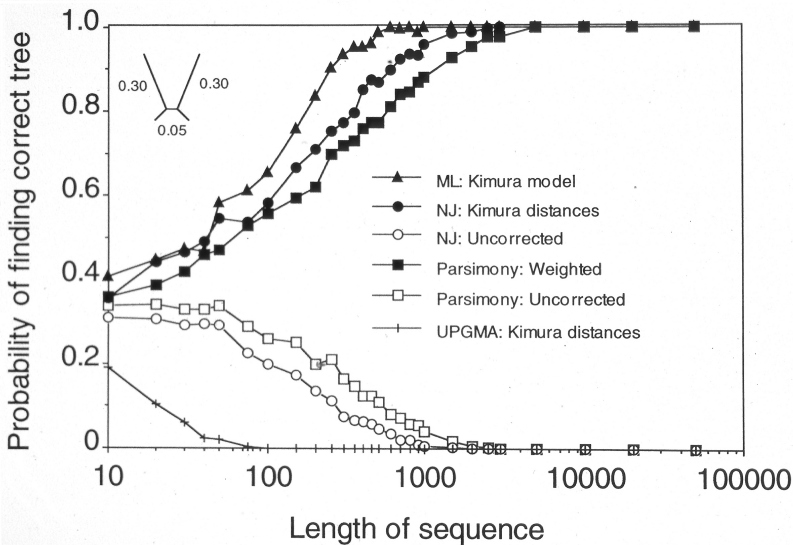
We can already see by eye-balling this that a model allowing **unequal** base frequencies will **fit better** than a model that assumes **equal** base frequencies because there are about twice as many As as there are Cs, Gs and Ts.

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.