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)

This pattern is surprising.

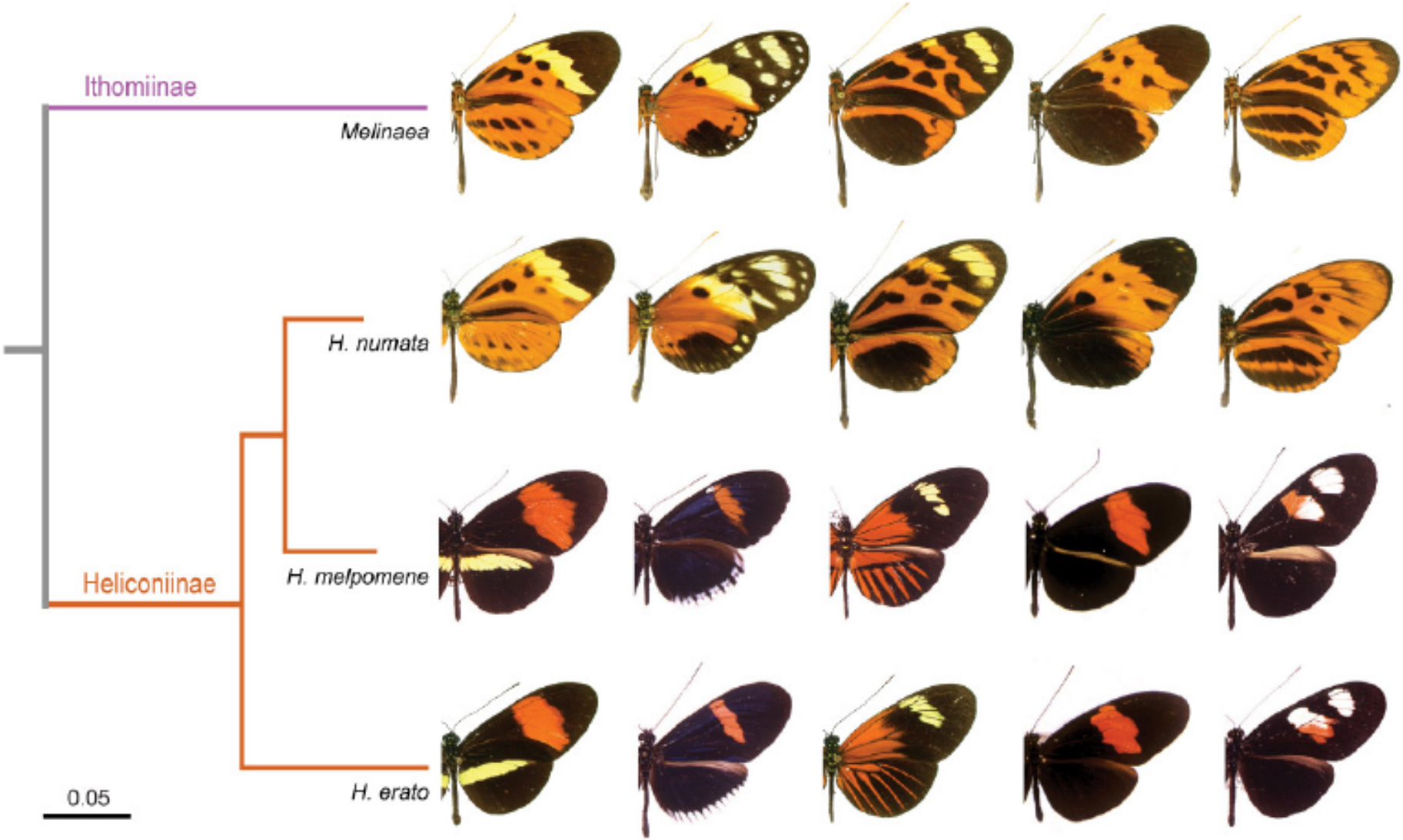
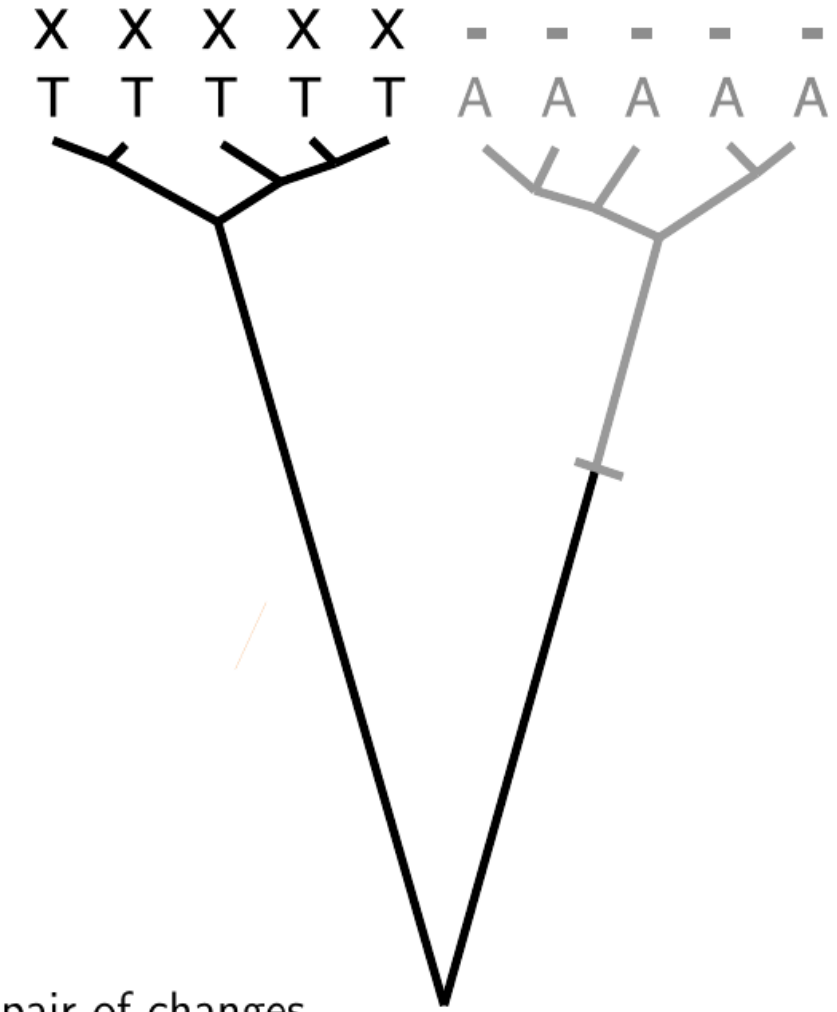
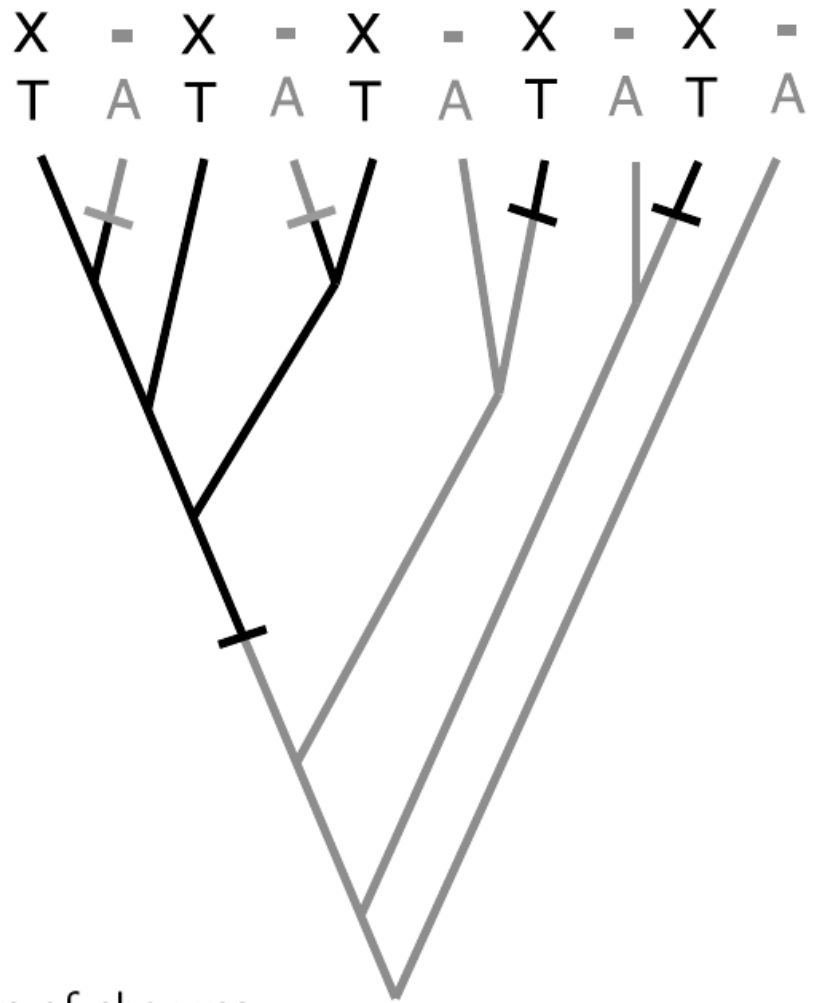


Figure by Mathieu Joron: http://xyala.cap.ed.ac.uk/joron/

These two trees tell very different evolutionary stories.



1 pair of changes.
Coincidence?



5 pairs of changes.
Much more convincing

How do we figure out what tree captures the relationships we're interested in?

Phylogenetic Inference

- We cannot see, measure, collect, or otherwise obtain a phylogeny directly from nature.

Phylogenetic Inference

- We cannot see, measure, collect, or otherwise obtain a phylogeny directly from nature.
- We have to *infer* it from data that we *can* collect from nature.

Phylogenetic Inference

- We cannot see, measure, collect, or otherwise obtain a phylogeny directly from nature.
- We have to *infer* it from data that we *can* collect from nature.
- A variety of different inferential approaches and data are used, with both increasing in sophistication and complexity over time.

Phylogenetic Inference

- We cannot see, measure, collect, or otherwise obtain a phylogeny directly from nature.
- We have to *infer* it from data that we *can* collect from nature.
- A variety of different inferential approaches and data are used, with both increasing in sophistication and complexity over time.
- Fundamental perspectives in all these approaches:

Phylogenetic Inference

- We cannot see, measure, collect, or otherwise obtain a phylogeny directly from nature.
- We have to *infer* it from data that we *can* collect from nature.
- A variety of different inferential approaches and data are used, with both increasing in sophistication and complexity over time.
- Fundamental perspectives in all these approaches:
  - Current patterns of biodiversity has been generated by processes of: (1) speciation, (2) extinction, and (3) character modification.

Phylogenetic Inference

- We cannot see, measure, collect, or otherwise obtain a phylogeny directly from nature.
- We have to *infer* it from data that we *can* collect from nature.
- A variety of different inferential approaches and data are used, with both increasing in sophistication and complexity over time.
- Fundamental perspectives in all these approaches:
  - Current patterns of biodiversity has been generated by processes of: (1) speciation, (2) extinction, and (3) character modification.
  - The phylogeny is an abstract representation ("model") of this diversification process.

Enormous numbers of topologies to consider

Taxa	Unrooted binary trees	Rooted binary trees
3	1	3
4	3	15
5	15	105
6	105	945
7	945	10,395
8	10,395	135,135
9	135,135	2,027,025
10	2,027,025	3×10^{7}
15	7×10^{12}	2×10^{14}
20	2×10^{20}	8×10^{21}
50	3×10^{74}	
100	2×10^{182}	
1,000	2×10^{2860}	
10,000	8×10^{38658}	
1,000,000	1×10^{5866723}	

Enormous numbers of topologies to consider

Taxa	Unrooted binary trees	Rooted binary trees
3	1	3
4	3	15
5	15	105
6	105	945
10	2,027,025	3×10^{7}
15	7×10^{12}	2×10^{14}
20	2×10^{20}	8×10^{21}
50	3×10^{74}	
100	2×10^{182}	
1,000	2×10^{2860}	
10,000	8×10^{38658}	
1,000,000	1×10^{5866723}	

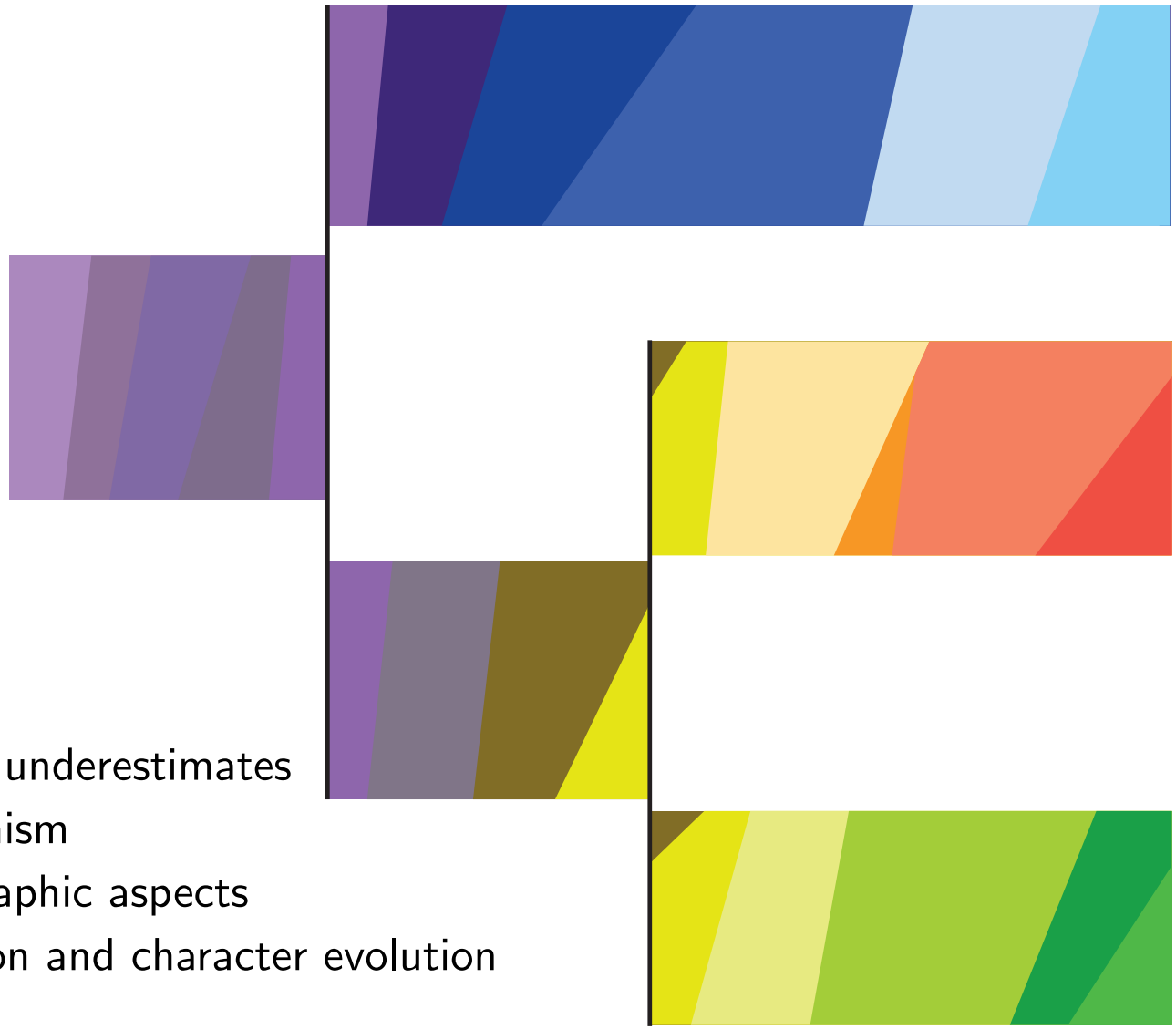
it is estimated that the there are between 10^{78} to 10^{82} atoms in the known, observable universe.

Estimating a tree from character data

Tree construction:

- strictly algorithmic approaches - use a "recipe" to construct a tree
- optimality based approaches - choose a way to "score" a trees and then search for the tree that has the best score.

Phylogeny with complete genome + "phenome" as colors:

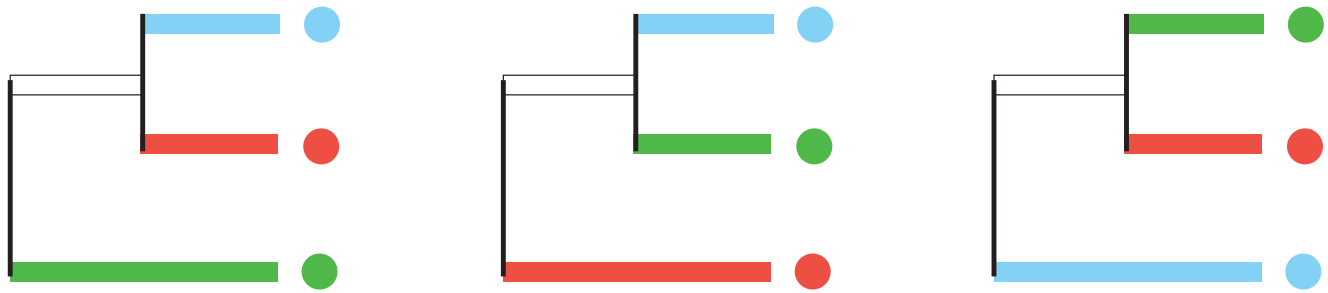


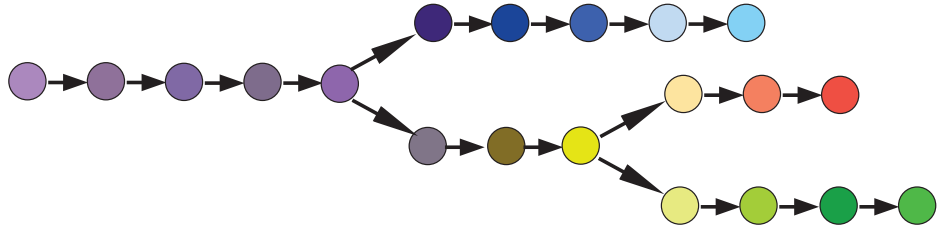
This figure:
dramatically underestimates
polymorphism
ignore geographic aspects
of speciation and character evolution

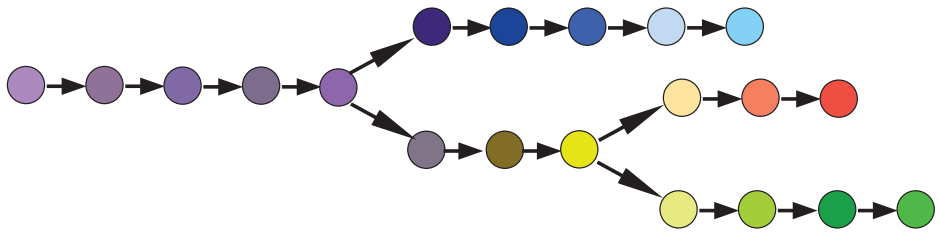
Extant species are just a thin slice of the phylogeny:

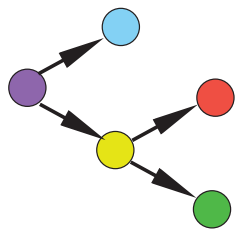
Our exemplar specimens are a subset of the current diversity:

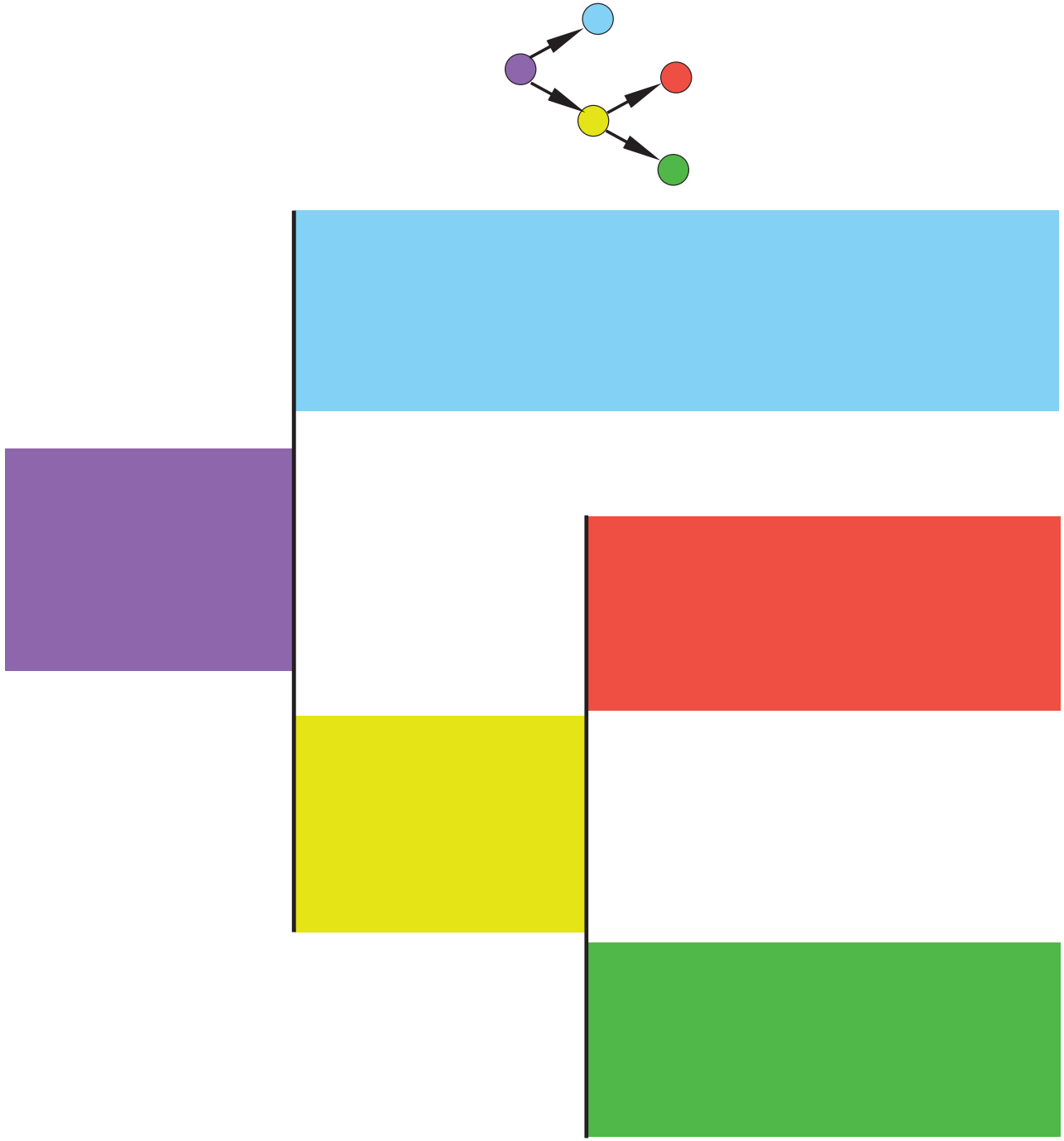
The phylogenetic inference problem:

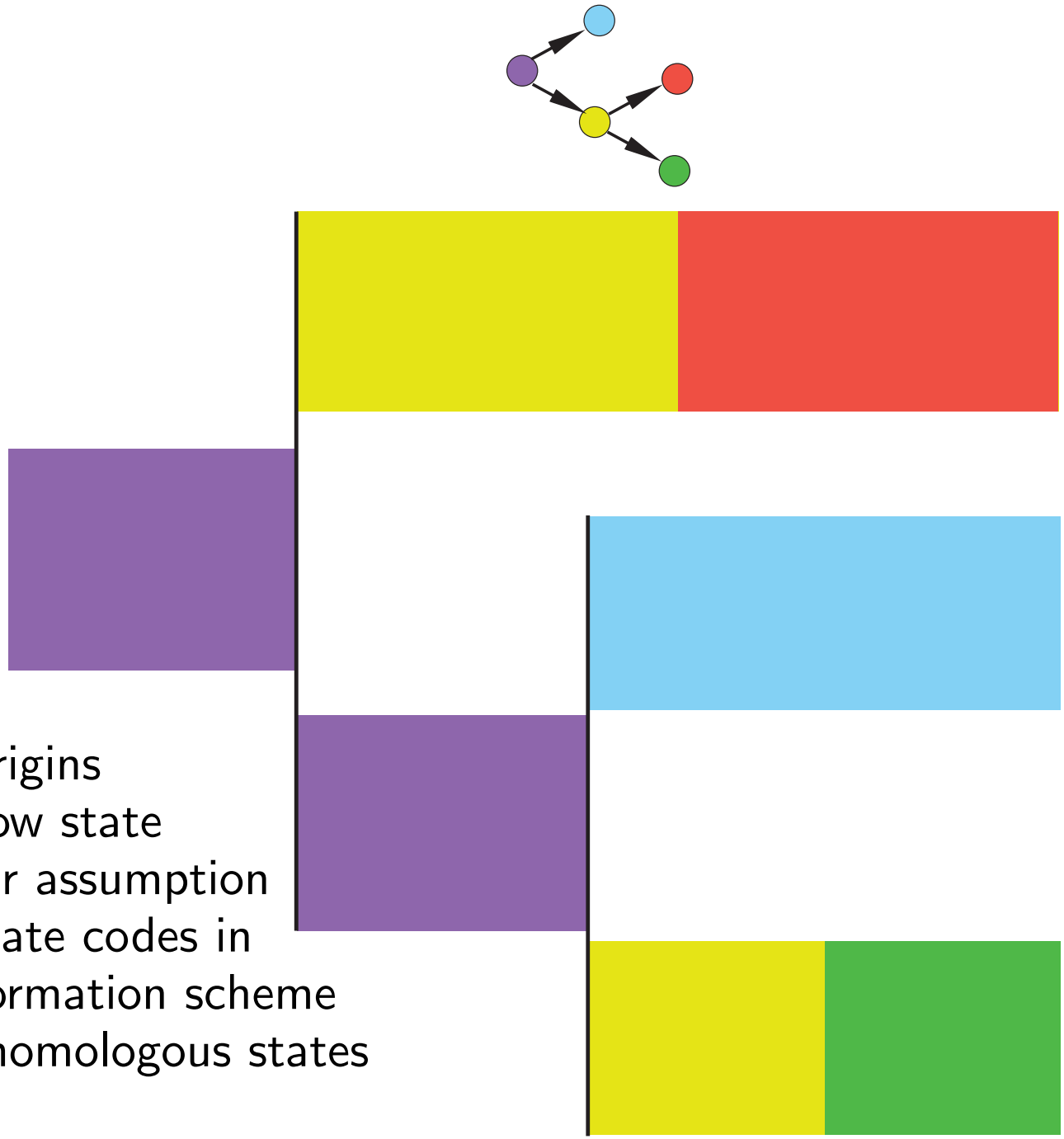




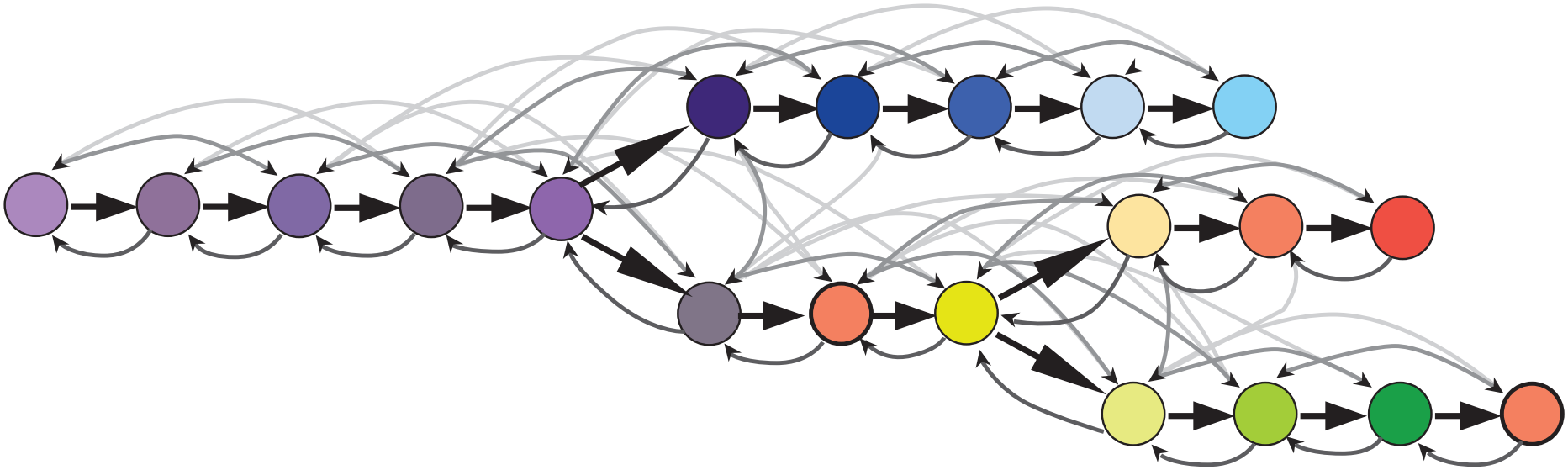








Multiple origins
of the yellow state
violates our assumption
that the state codes in
our transformation scheme
represent homologous states

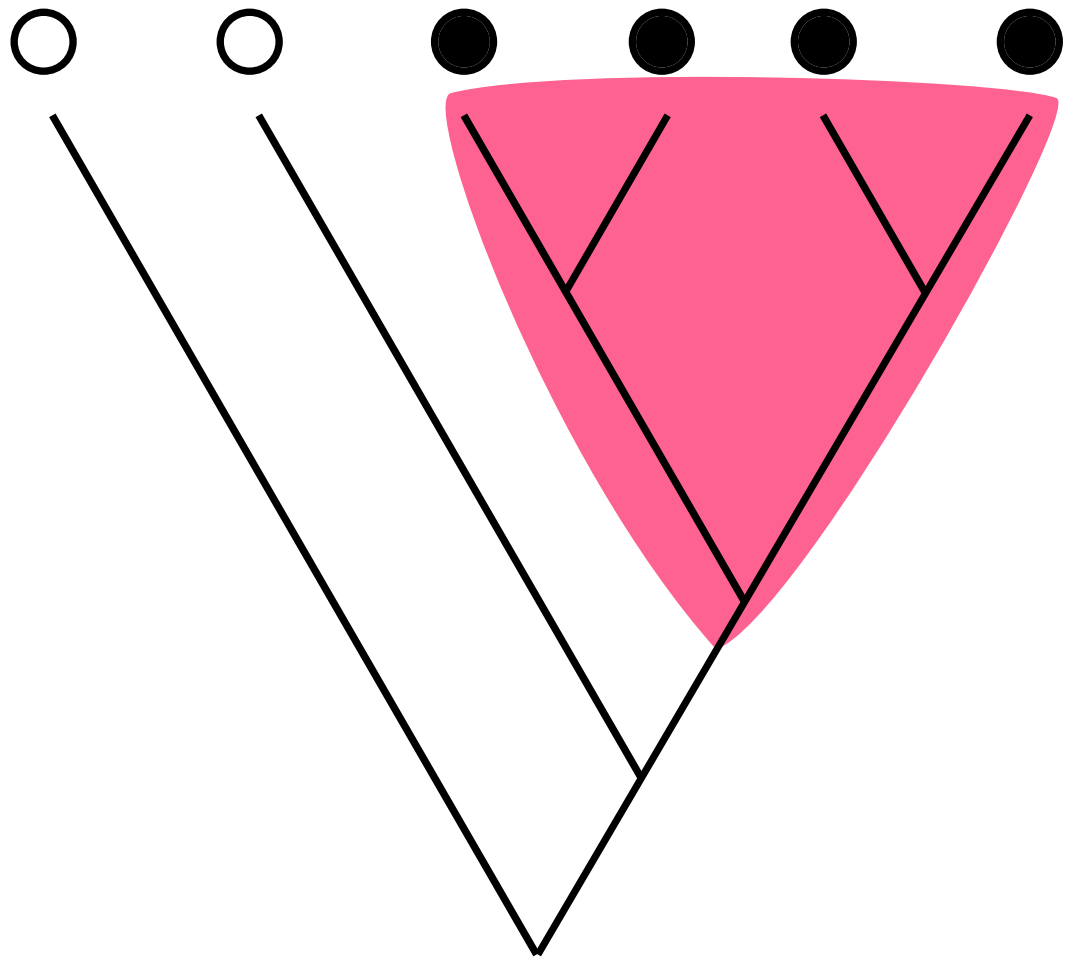


The meaning of homology (**very roughly**):

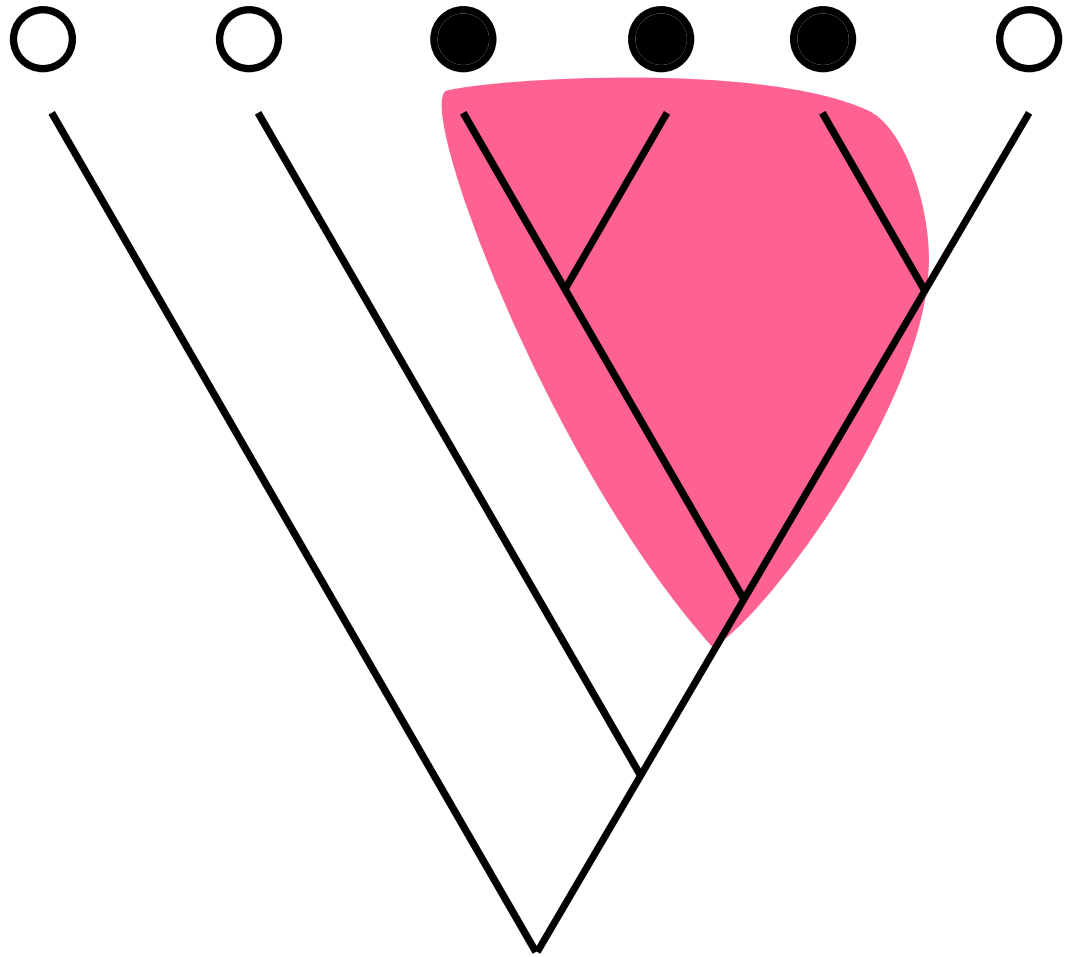
1. comparable (when applied to characters)
2. identical by descent (when applied to character states)

Ideally, each possible character state would arise once in the entire history of life on earth.

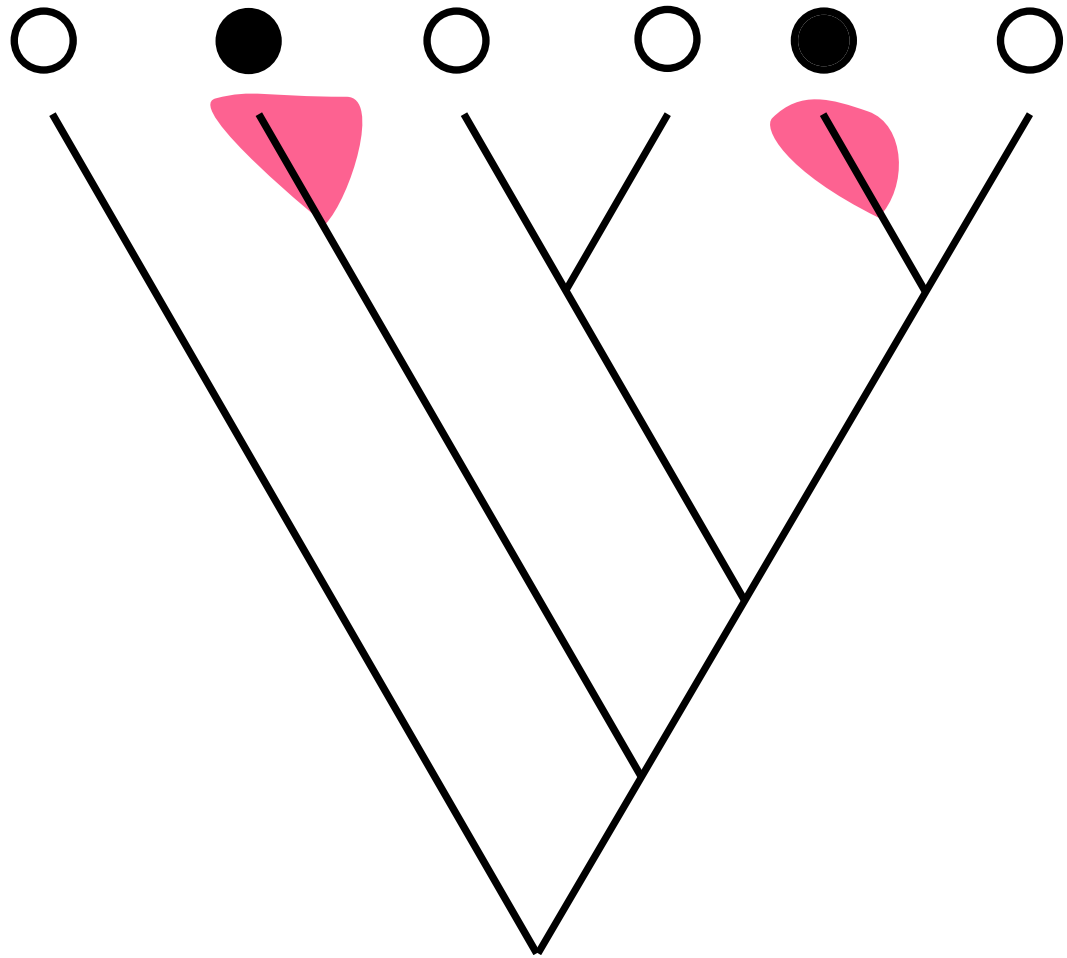
Instances of the filled character state are homologous
Instances of the hollow character state are homologous



Instances of the filled character state are homologous
Instances of the hollow character state are NOT homologous



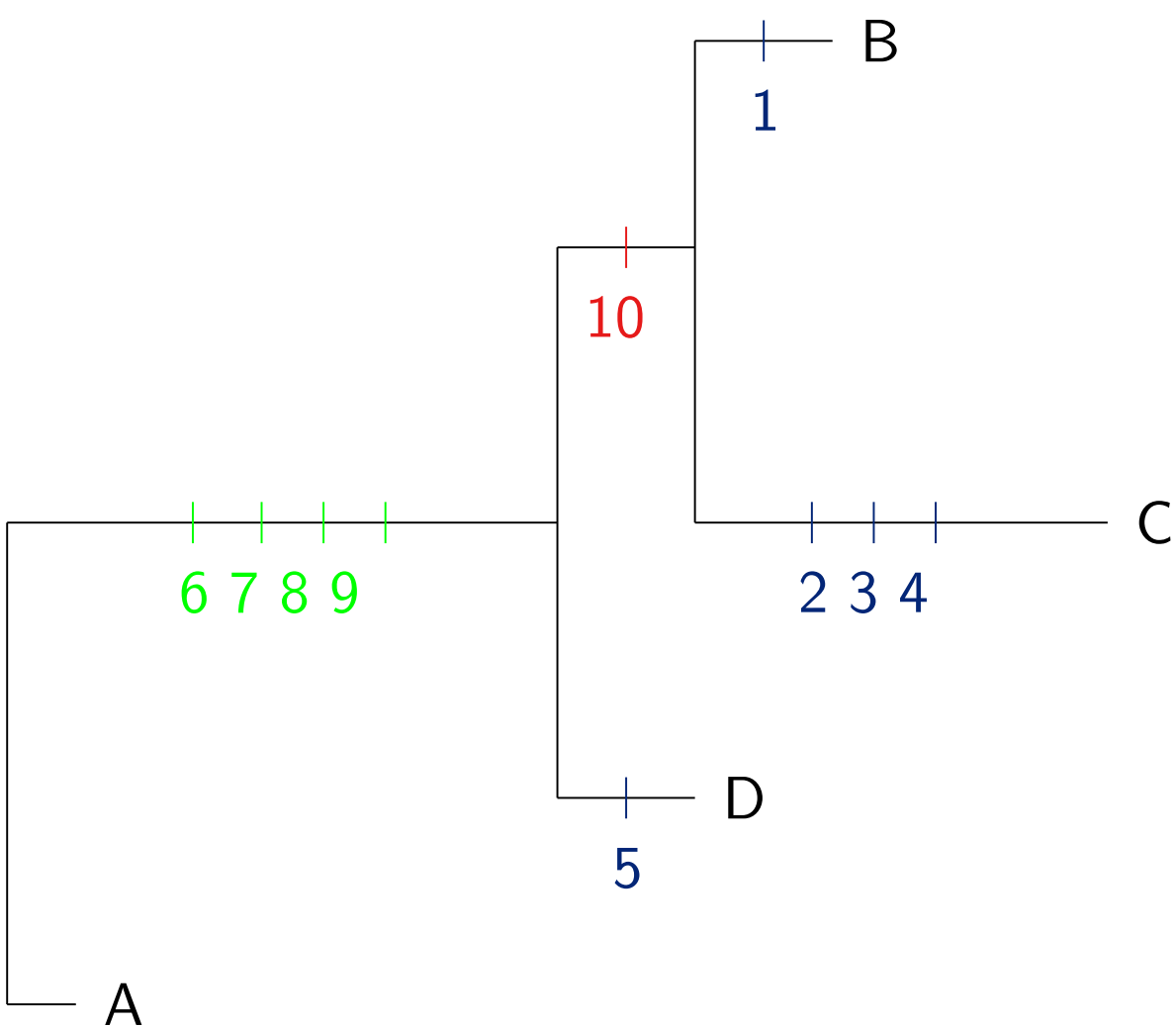
Instances of the filled character state are NOT homologous
Instances of the hollow character state are homologous

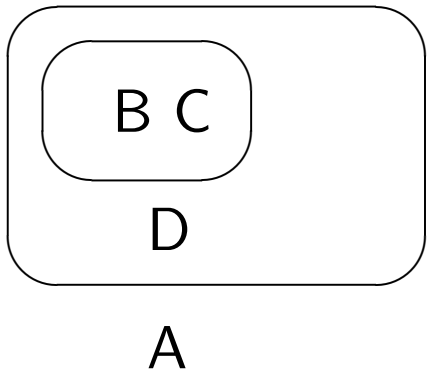


If 0 is the ancestral state, and 1 is the derived state, and there is no homoplasy, what are the relationships between these taxa?

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

Don't look at the next slide yet!!


B
1
10
C
6 7 8 9
2 3 4
D
5
A


B C
D
A

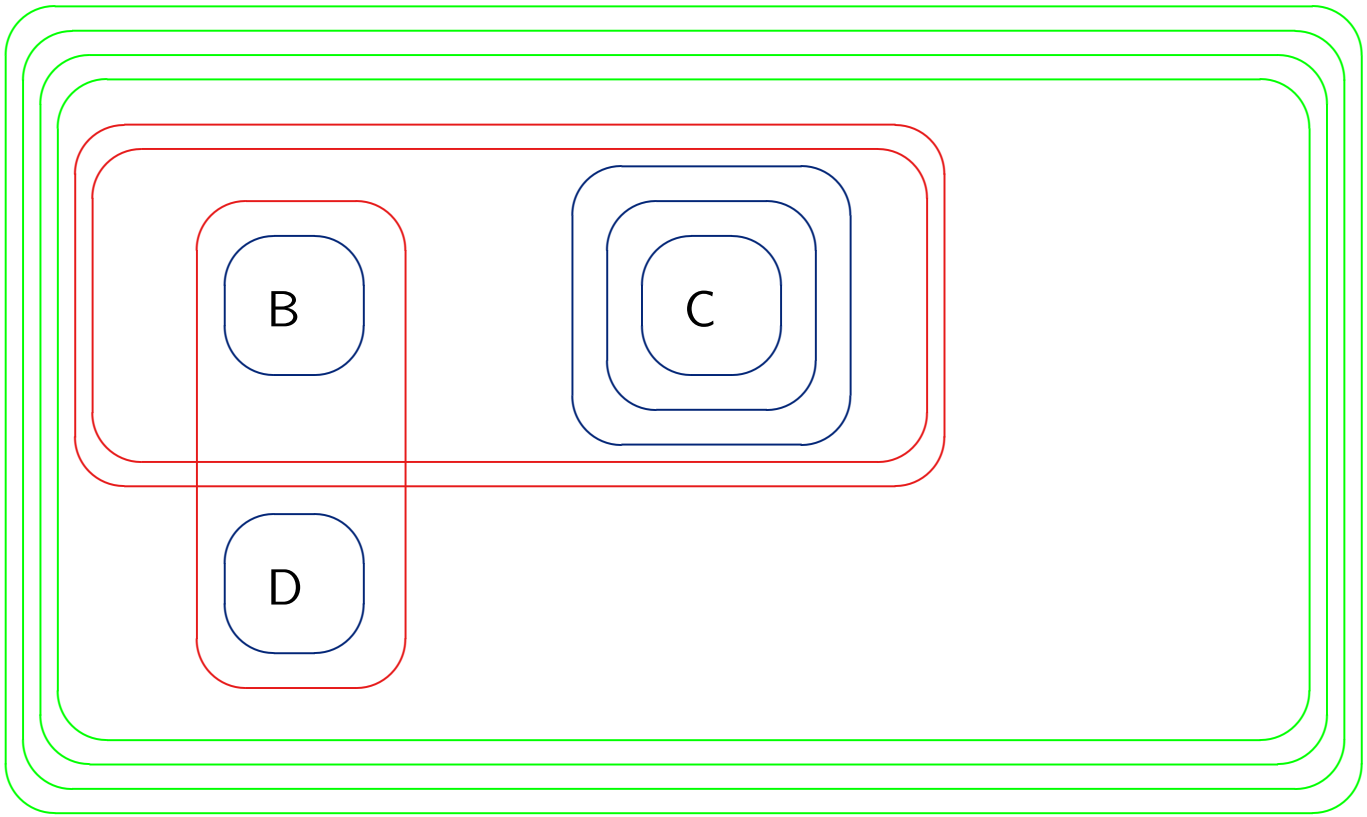
If characters are not polarized (ancestral and descendent states known) can infer unrooted trees.

Character conflict

Homo sapiens	AGTTCAAGT
Rana catesbiana	AATTCAAGT
Drosophila melanogaster	AGTTCAAGC
C. elegans	AATTCAAGC

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.

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.

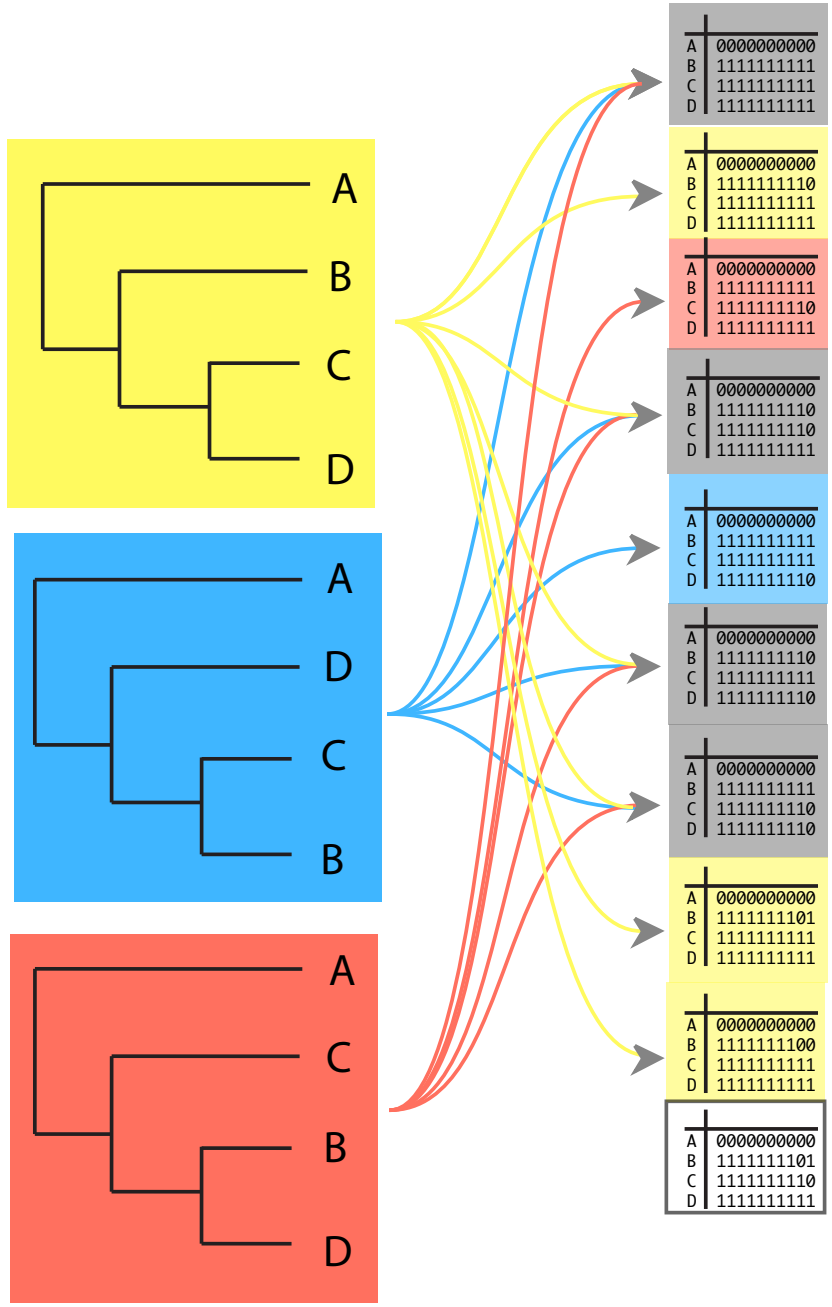
Incompatible characters are evidence of homoplasy in the data

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 Parsimony analyses.


A
B
C
D
A
D
C
B
A
C
B
D
A 0000000000
B 1111111111
C 1111111111
D 1111111111
A 0000000000
B 1111111110
C 1111111111
D 1111111111
A 0000000000
B 1111111111
C 1111111110
D 1111111111
A 0000000000
B 1111111110
C 1111111110
D 1111111111
A 0000000000
B 1111111111
C 1111111111
D 1111111110
A 0000000000
B 1111111110
C 1111111111
D 1111111110
A 0000000000
B 1111111111
C 1111111110
D 1111111110
A 0000000000
B 1111111101
C 1111111111
D 1111111111
A 0000000000
B 1111111100
C 1111111111
D 1111111111
A 0000000000
B 1111111101
C 1111111110
D 1111111111

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? No, nothing purely based on logic (and the suggestions for culling data to make matrices suitable for logical inference can lead to unsatisfyingly subjecive analyses).
- What can we do?

We must have an "error model"

In this class we will focus on Maximum Likelihood and Bayesian statistical estimates for evolutionary models.

When should we expect character conflict?

- Data type?
- Evolutionary history?

How can we deal with character conflict?

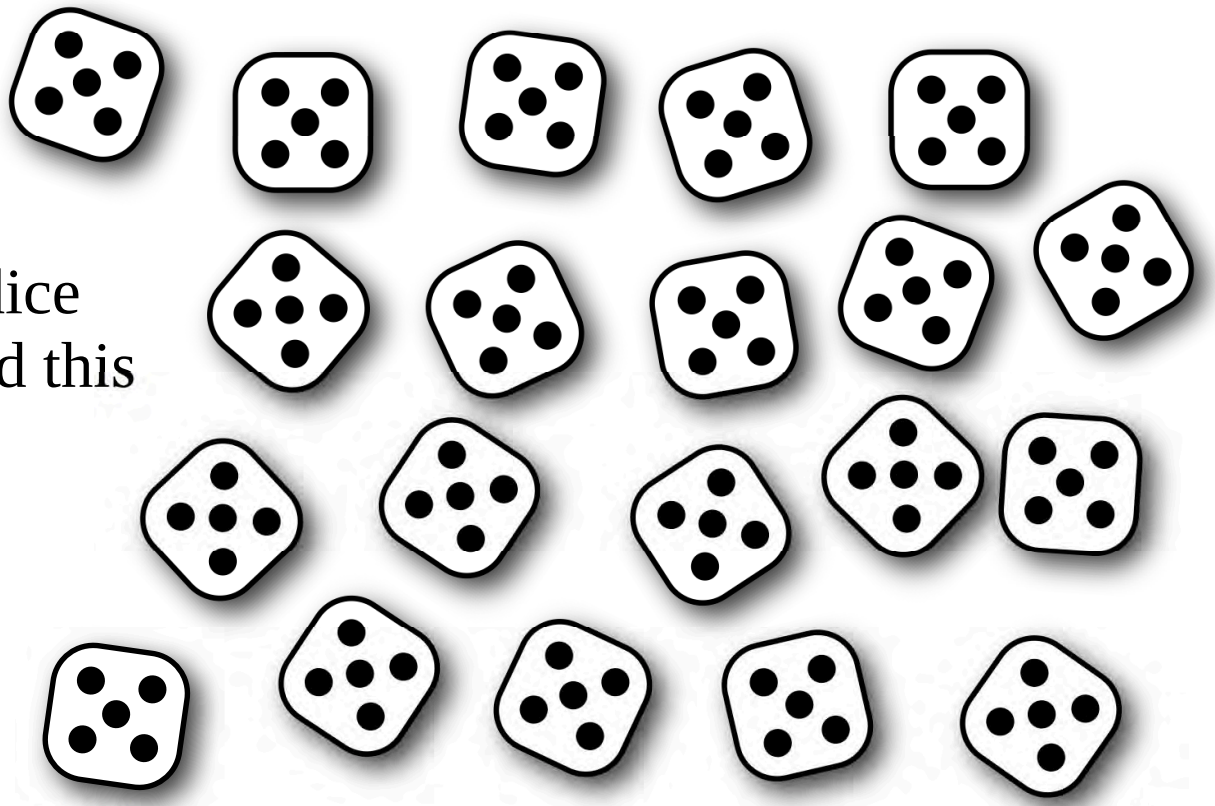
- We need to apply an error model
- Likelihood provides a measure of surprise under different models

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

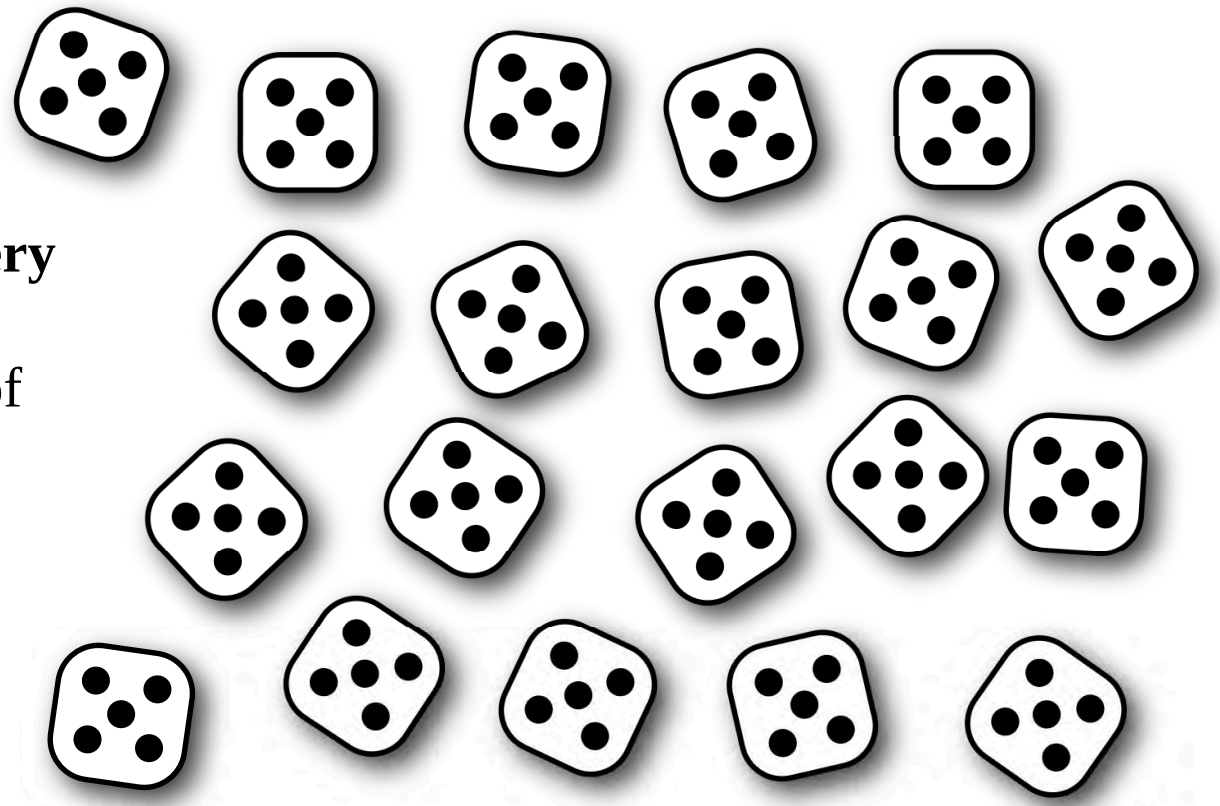
 AND ?

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

The Fair Dice model

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

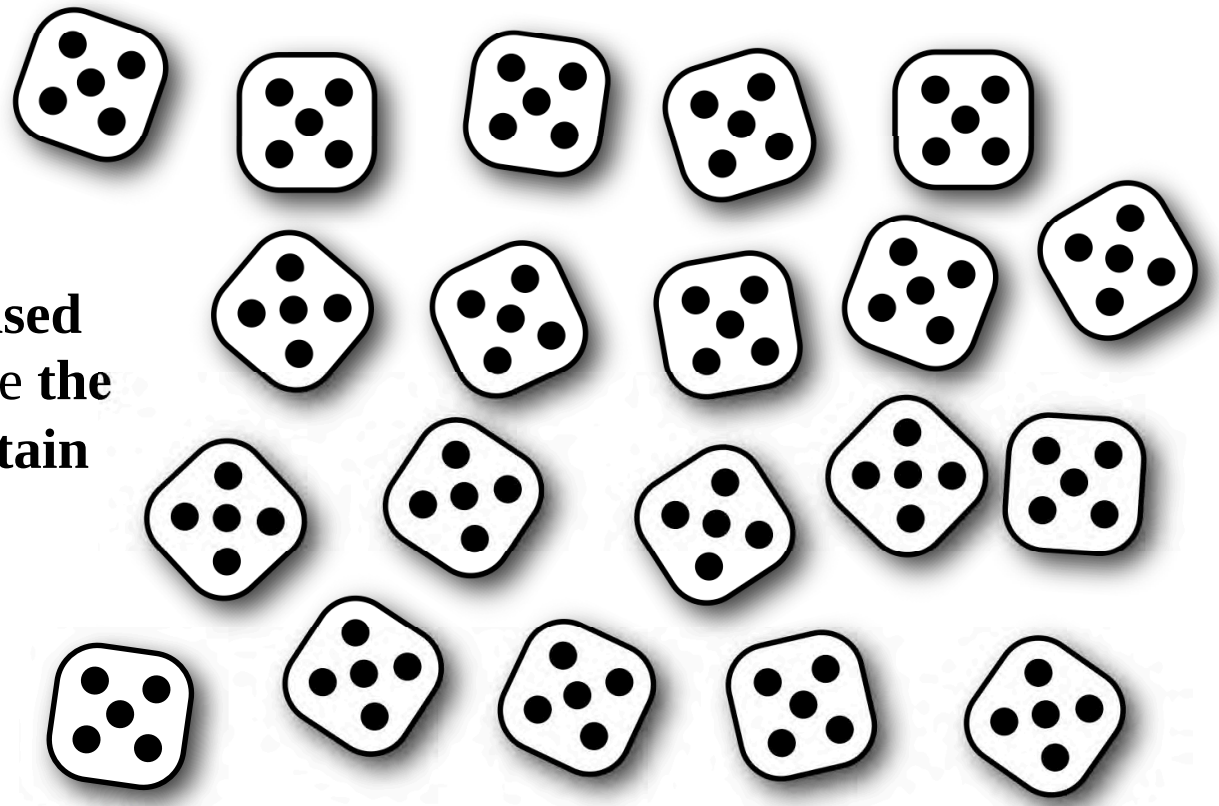
You should have been **very surprised** at this result because the probability of this event is **very small**: only 1 in 3.6 quadrillion!



The Trick Dice model

(assumes dice each have 5 on every side)

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



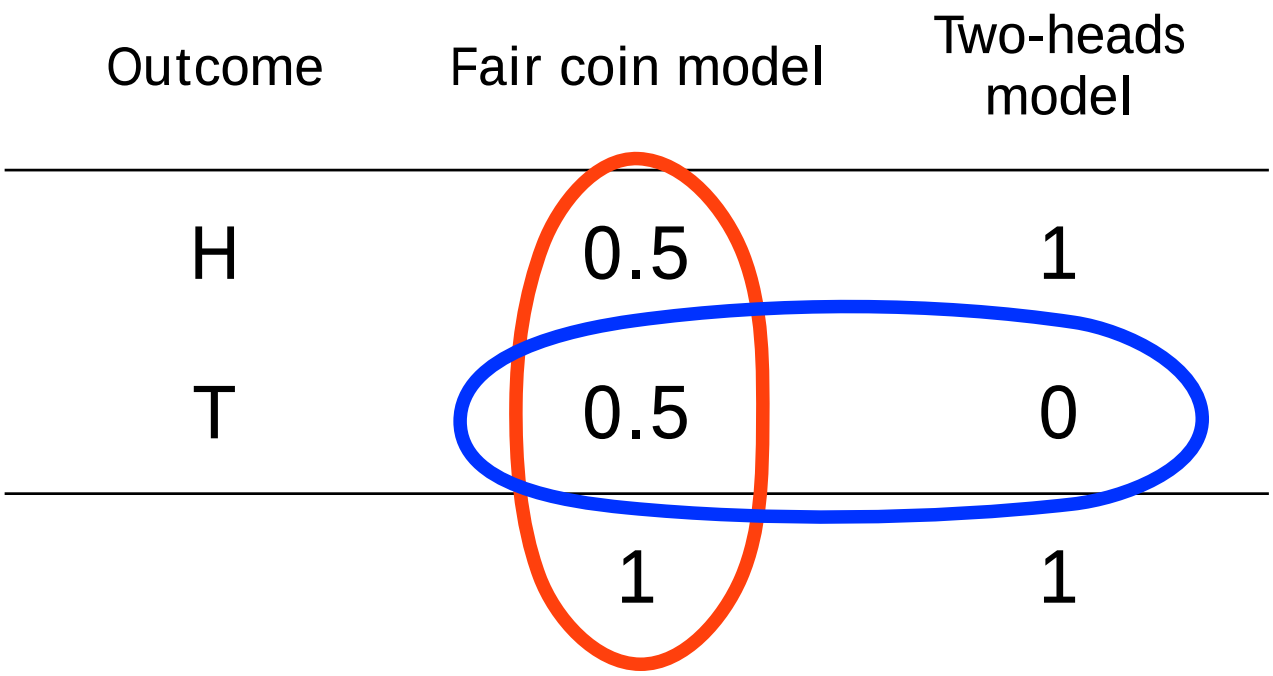
You should **not be surprised at all** at this result because **the observed outcome is certain** under this model

Results

Model	Likelihood	Surprise level
Fair Dice	$\frac{1}{3{,}656{,}158{,}440{,}062{,}976}$	Very, *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

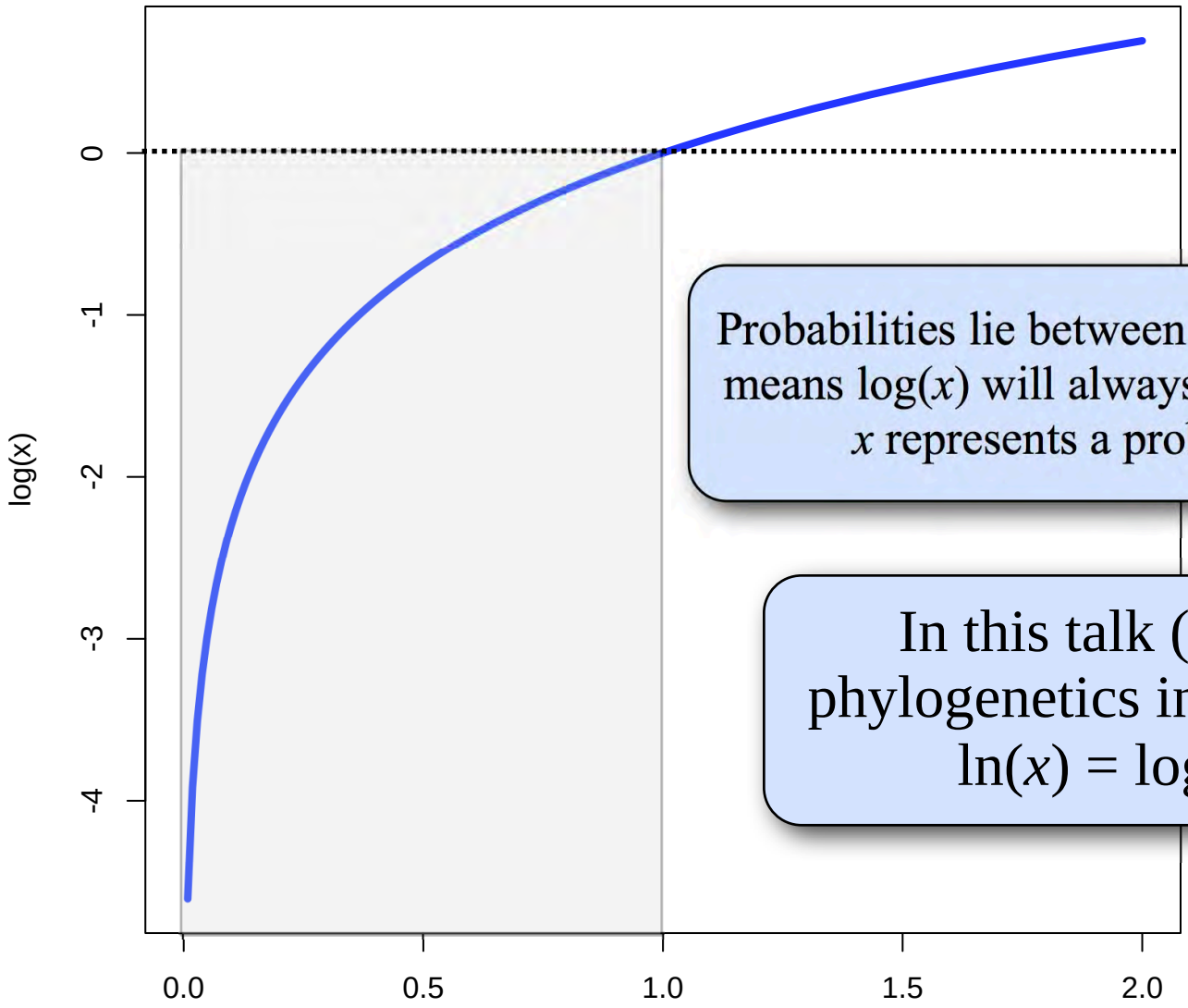
Probabilities of data outcomes given one particular model sum to 1.0

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 ψη-globin gene of gorilla:

GAAGTCCTTGAGAAATAAACTGCACACACTGG

$$L = \pi_G \pi_A \pi_A \pi_G \pi_T \pi_C \pi_C \pi_T \pi_T \pi_G \pi_A \pi_G \pi_A \pi_A \pi_A \pi_T \pi_A \pi_A \pi_A \pi_C \pi_T \pi_G \pi_C \pi_A \pi_C \pi_A \pi_C \pi_A \pi_C \pi_T \pi_G \pi_G$$
$$= \pi_A^{12} \pi_C^{7} \pi_G^{7} \pi_T^{6}$$

Note that we are assuming independence among sites here

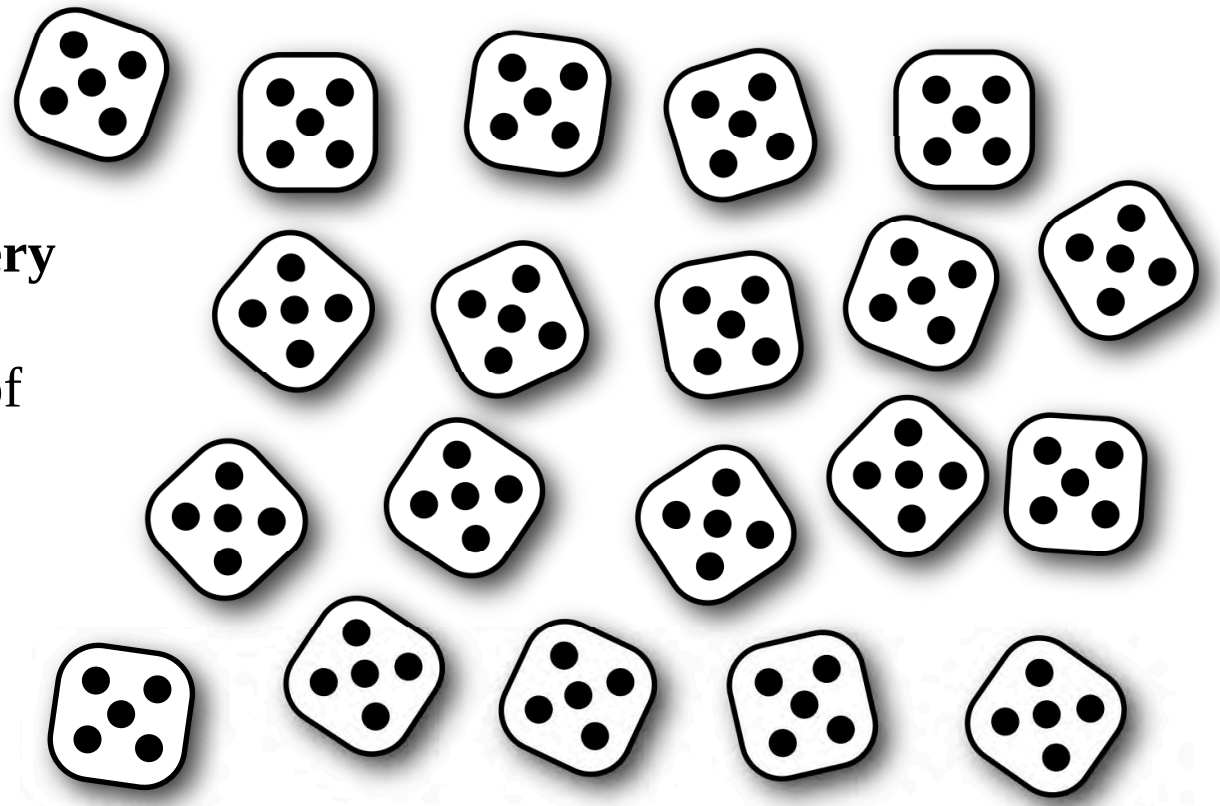
$$\log L = 12 \log(\pi_A) + 7 \log(\pi_C) + 7 \log(\pi_G) + 6 \log(\pi_T)$$

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.

The Fair Dice model

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

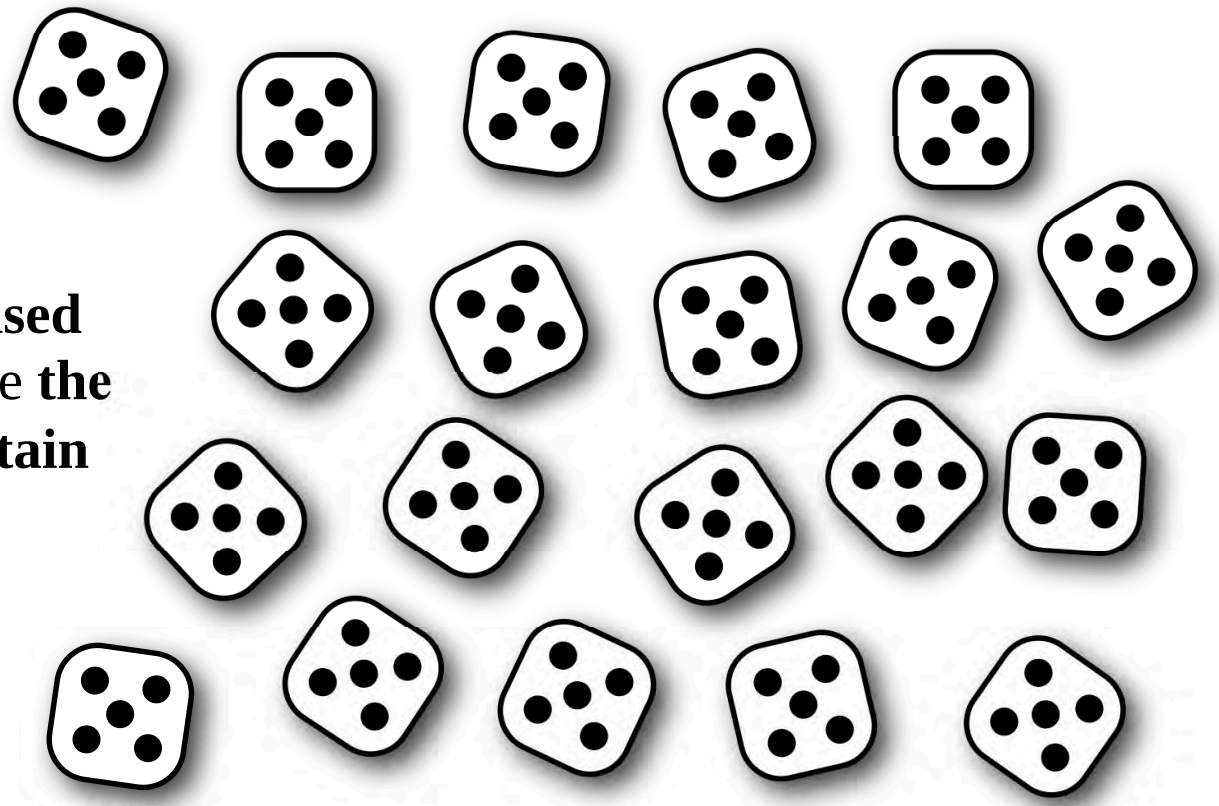
You should have been **very surprised** at this result because the probability of this event is **very small**: only 1 in 3.6 quadrillion!



The Trick Dice model

(assumes dice each have 5 on every side)

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



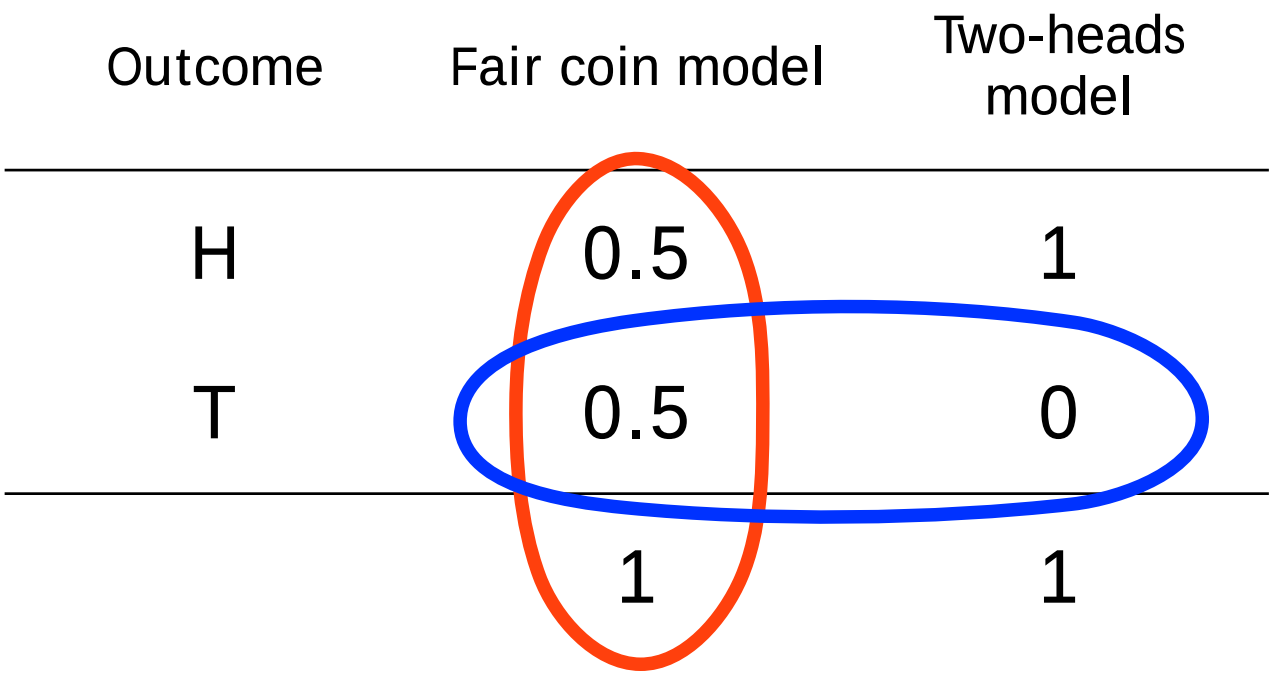
You should **not be surprised at all** at this result because **the observed outcome is certain** under this model

Results

Model	Likelihood	Surprise level
Fair Dice	$\frac{1}{3{,}656{,}158{,}440{,}062{,}976}$	Very, *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

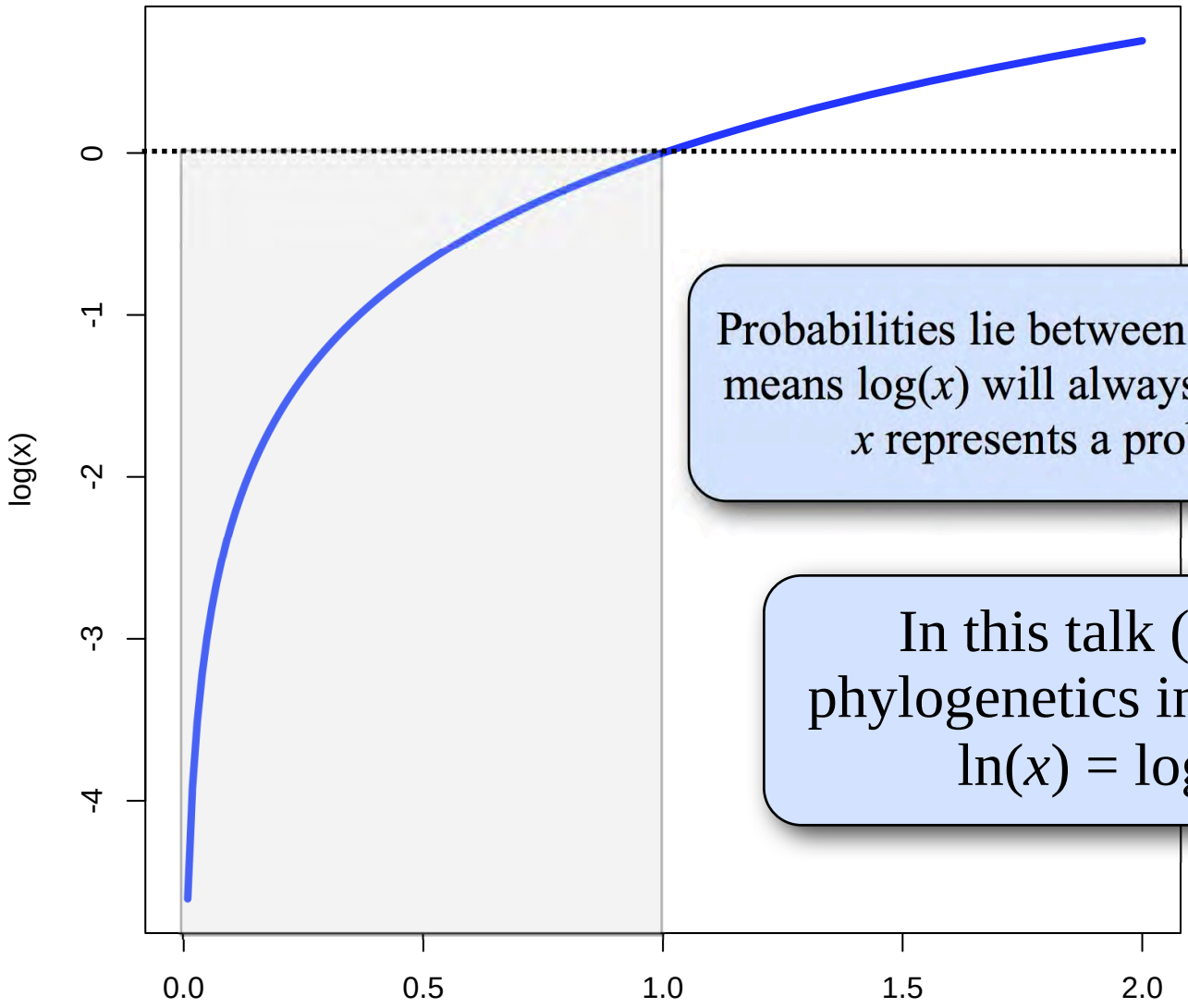
Probabilities of data outcomes given one particular model sum to 1.0

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 ψη-globin gene of gorilla:

GAAGTCCTTGAGAAATAAACTGCACACACTGG

$$L = \pi_G \pi_A \pi_A \pi_G \pi_T \pi_C \pi_C \pi_T \pi_T \pi_G \pi_A \pi_G \pi_A \pi_A \pi_A \pi_T \pi_A \pi_A \pi_A \pi_C \pi_T \pi_G \pi_C \pi_A \pi_C \pi_A \pi_C \pi_A \pi_C \pi_T \pi_G \pi_G$$
$$= \pi_A^{12} \pi_C^{7} \pi_G^{7} \pi_T^{6}$$

Note that we are assuming independence among sites here

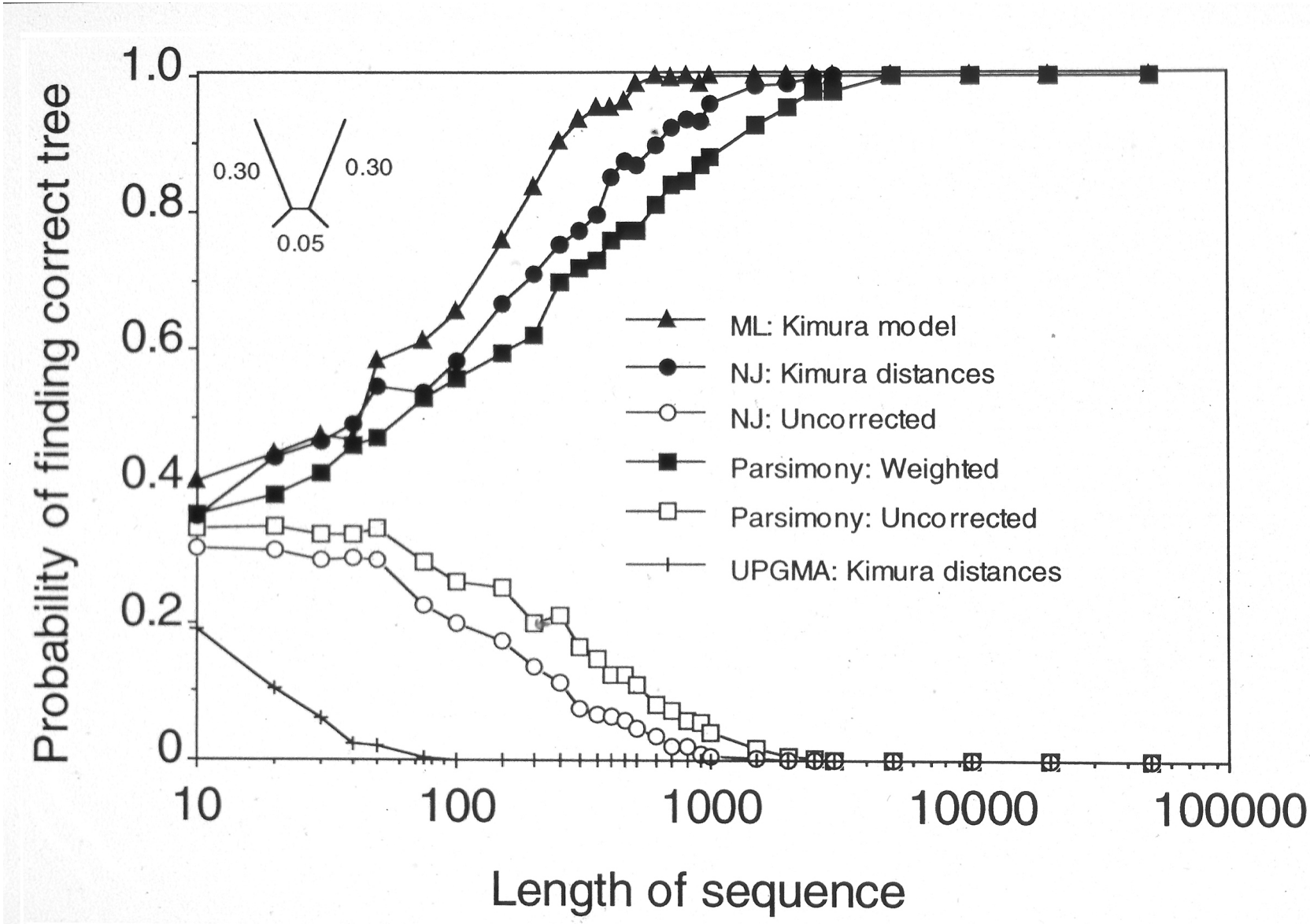
$$\log L = 12 \log(\pi_A) + 7 \log(\pi_C) + 7 \log(\pi_G) + 6 \log(\pi_T)$$

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.

Maximum Likelihood phylogenetics

- What tree makes our data the least surprising?
- Answering this question requires answering a lot of other questions too!
  - What do we expect our data to look like under different scenarios?
  - What is our model of evolution?
- We will jump into applications in the iqtree lab today, and keep building up the theory in lecture next week.