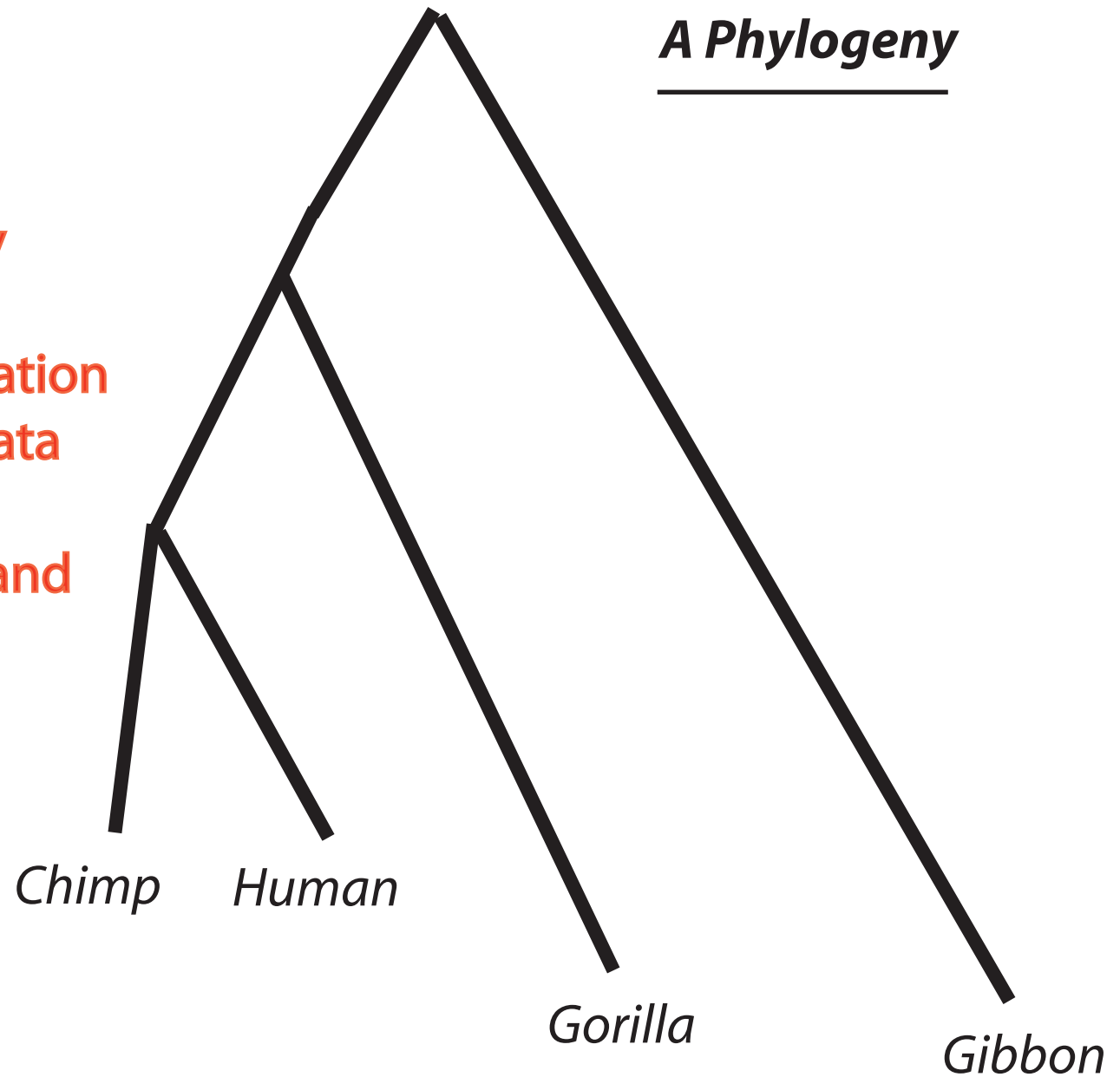
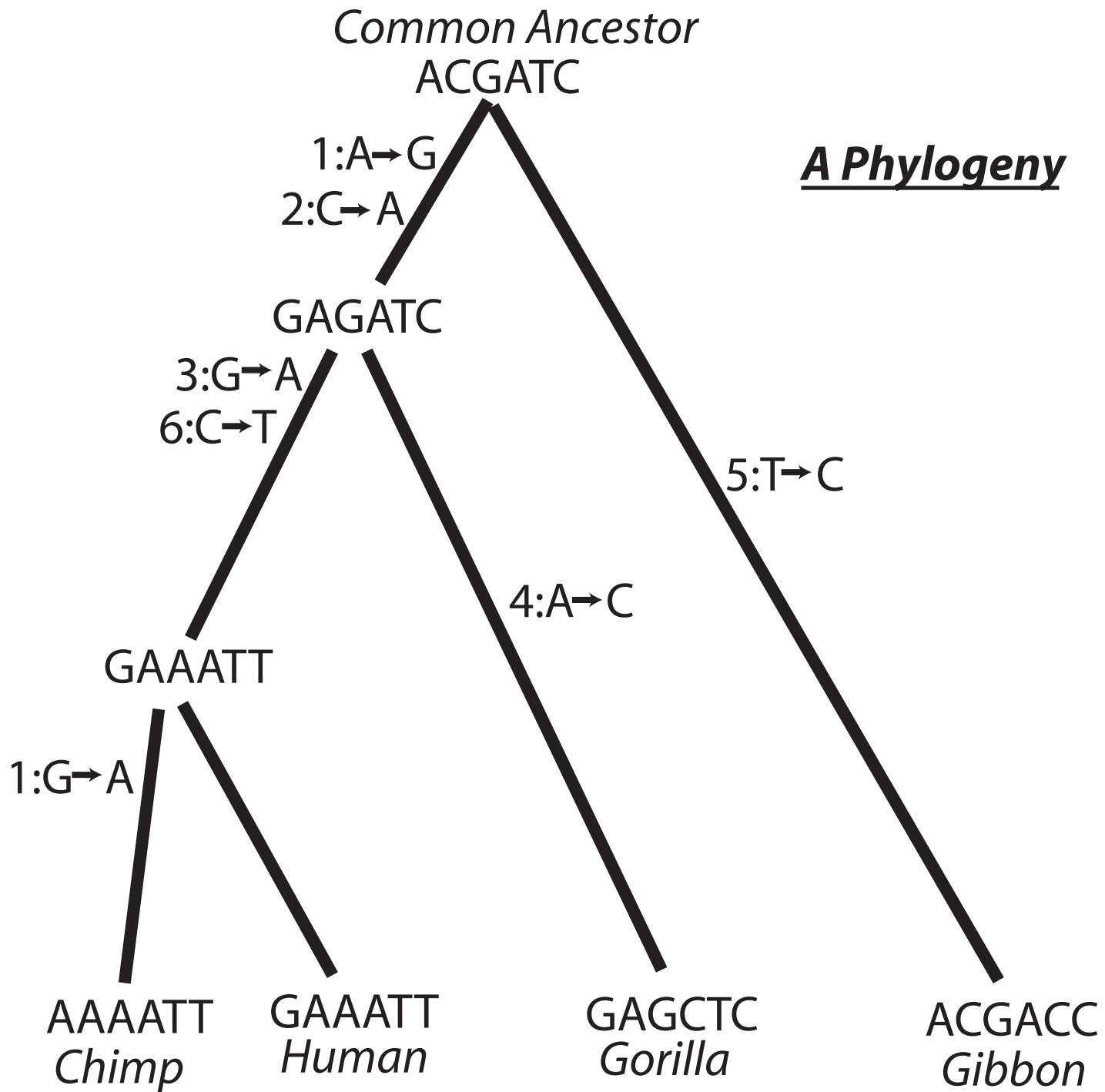


Biologists care about phylogenies because

1. They represent history
2. They represent correlation structure of biological data
3. They help us understand biological process





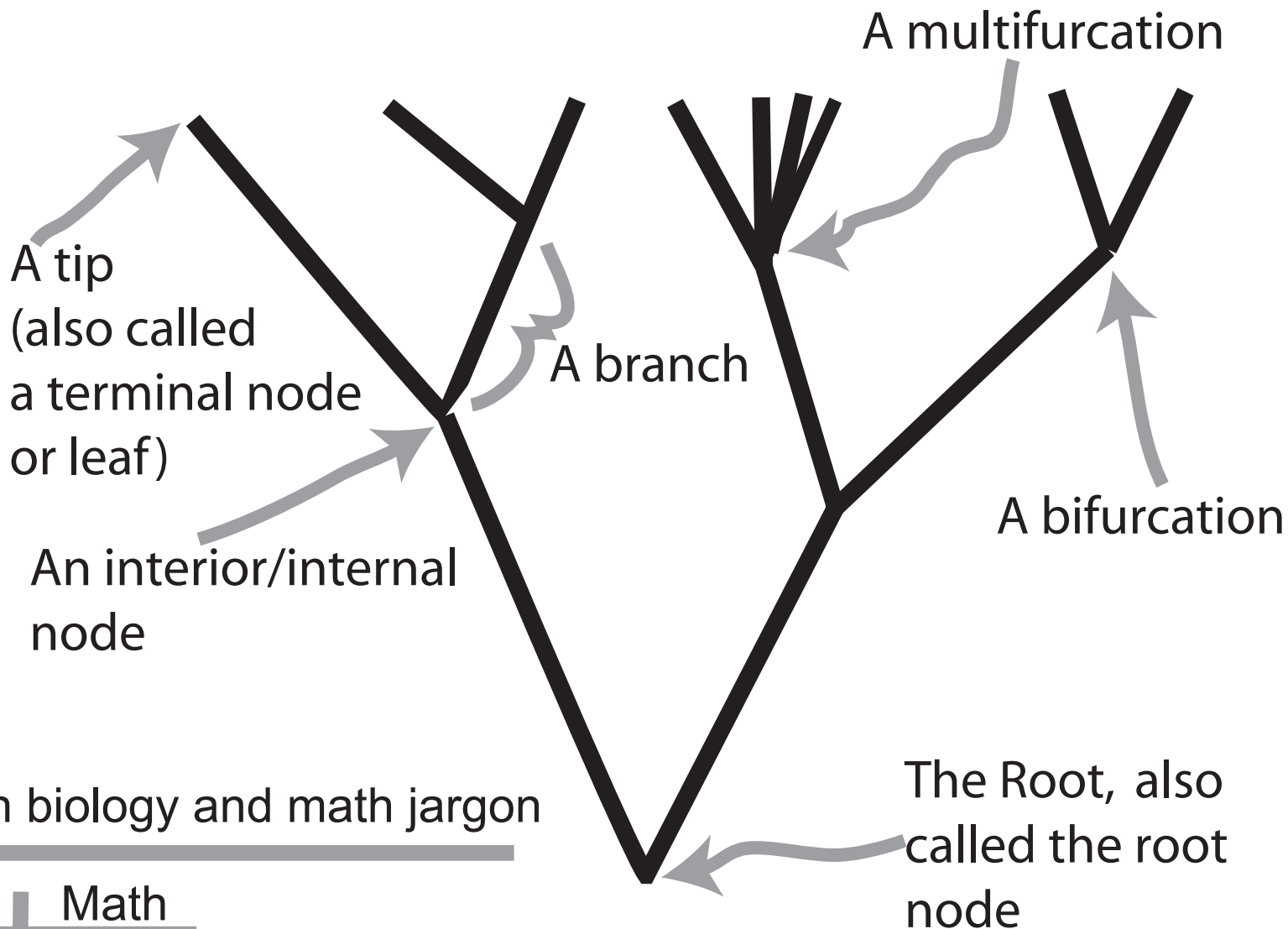
AAAATT
Chimp

GAAATT
Human

GAGCTC
Gorilla

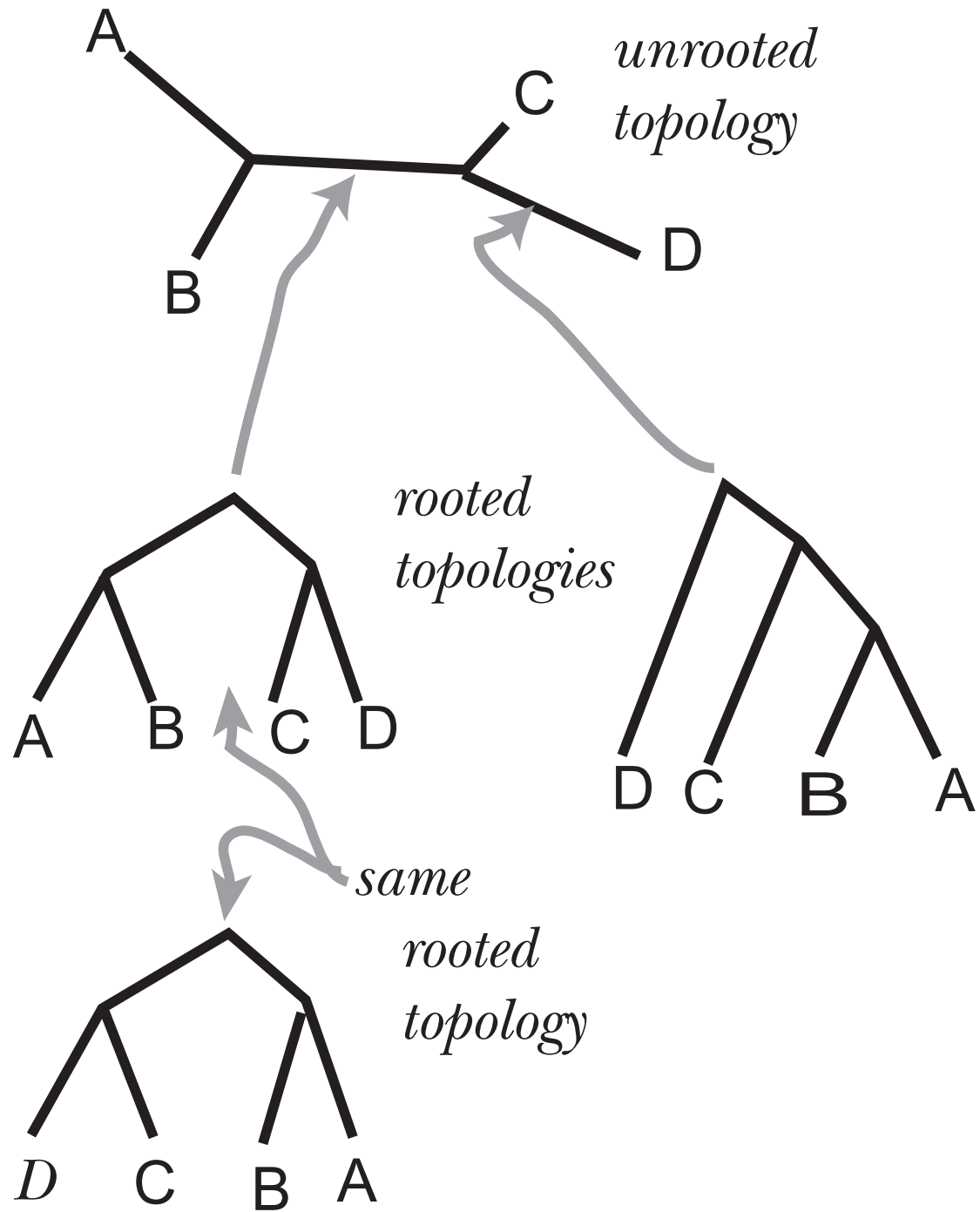
ACGACC
Gibbon

Tree Anatomy



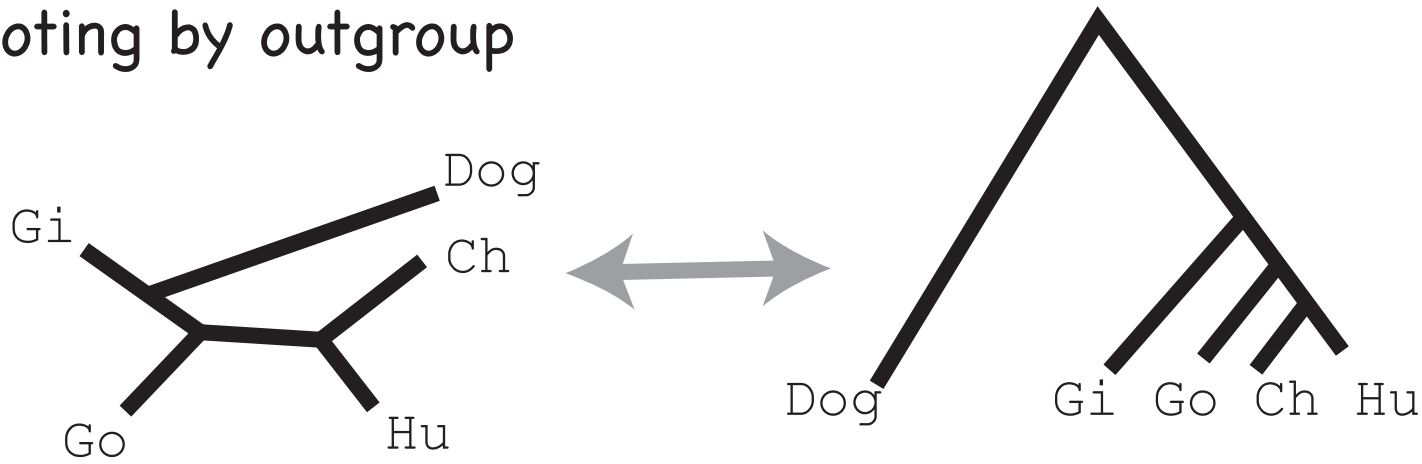
Translating between biology and math jargon

Biology	Math
Tree	Graph
Branch	Edge
Node	Vertice



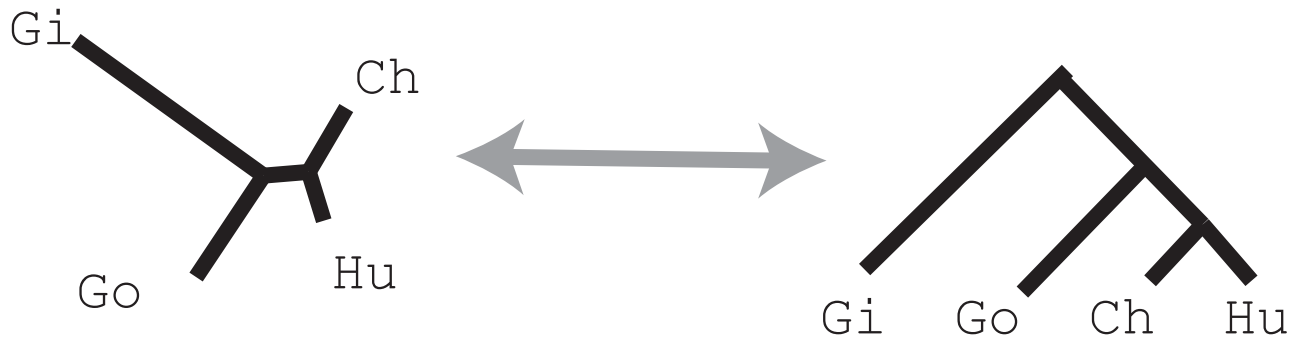
The two common ways phylogenies are rooted:

1. Rooting by outgroup

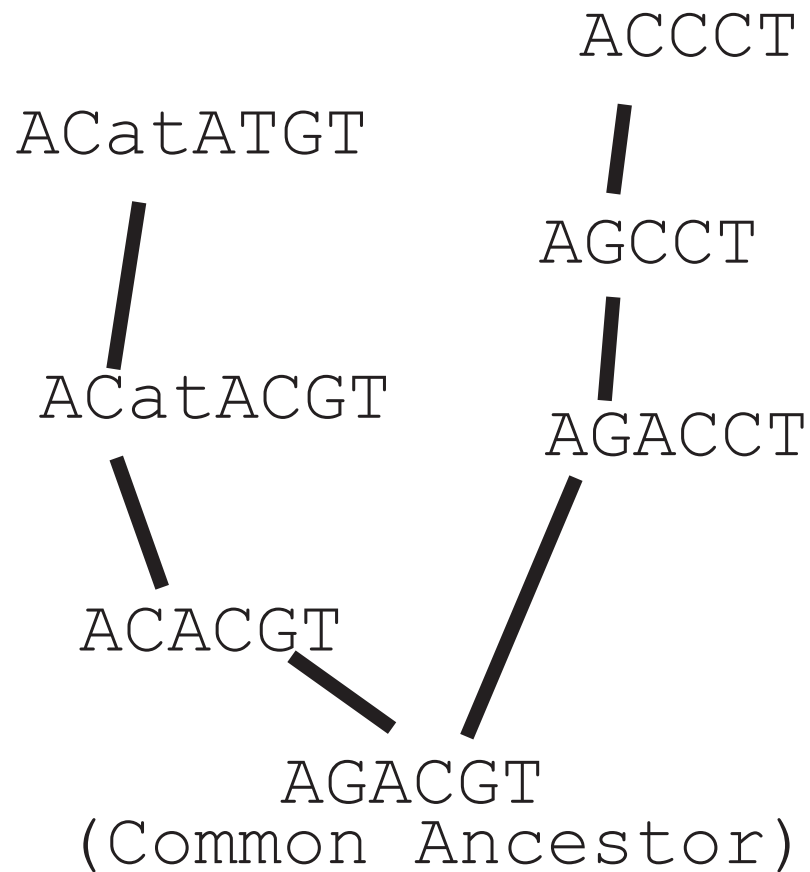


"Outgroup" = Dog "Ingroup" = Gi & Go & Ch & Hu

2. Rooting by molecular clock



All "tips" should be equally far from root

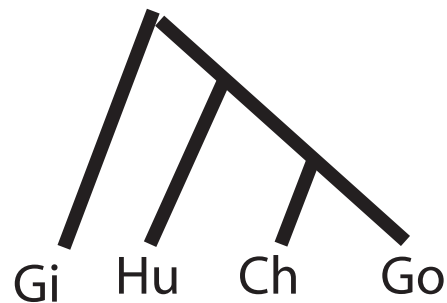
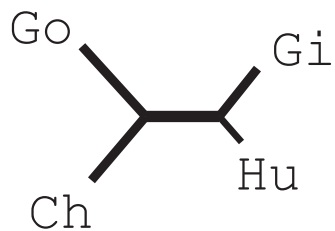
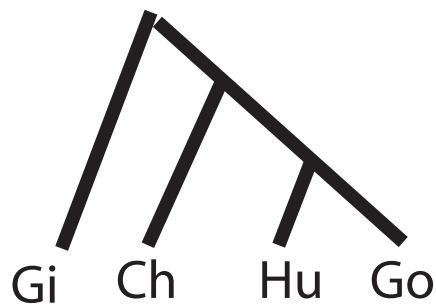
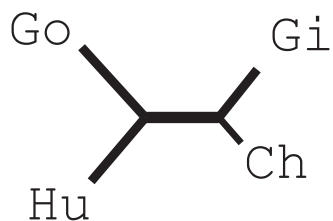
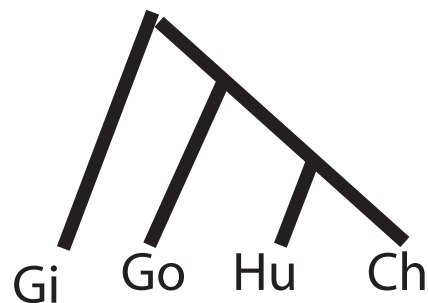
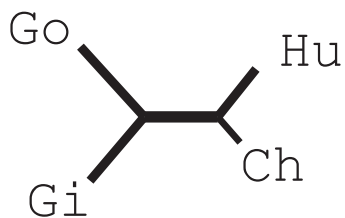


The "true" alignment:

ACATATGT

AC---CCT

Character:	123456
(Go) Gorilla:	GAGCTC
(Gi) Gibbon:	ACGACC
(Hu) Human:	GAAATT
(Ch) Chimp:	AAAATT



Maximum Parsimony Principle: The best explanation is the simplest.

Basic assumptions of parsimony as applied to phylogeny reconstruction:

1. For a particular topology and a particular character (i.e., alignment column), the most reasonable explanation of how the character evolved on the tree is the one that requires the smallest “amount” of evolutionary change.
2. The best topology is the topology that requires the smallest “amount” of evolutionary change.

the parsimony definition of ***Phylogenetically Informative Characters*** -- characters for which the most parsimonious number of changes is different among unrooted topologies.

Characters that do not vary among taxa (sequences) are not phylogenetically informative according to parsimony.

Characters where all but one taxon have same state are not phylogenetically informative **according to parsimony.**

Sequence 1: C G A
Sequence 2: C G A
Sequence 3: A G A
Sequence 4: A G T

a "parsimony
informative"

or

"phylogenetically
informative"

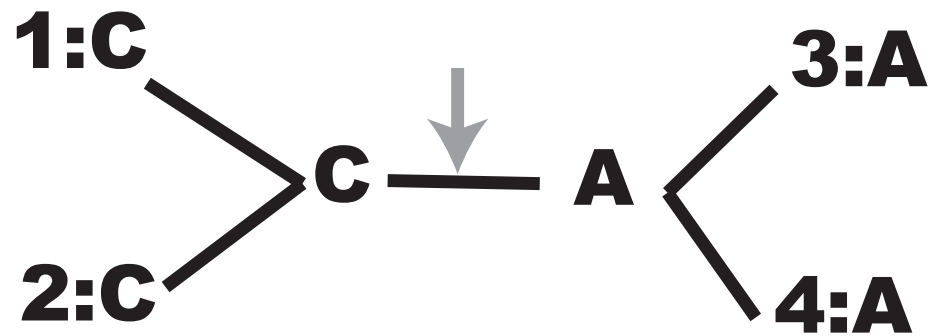
site

not

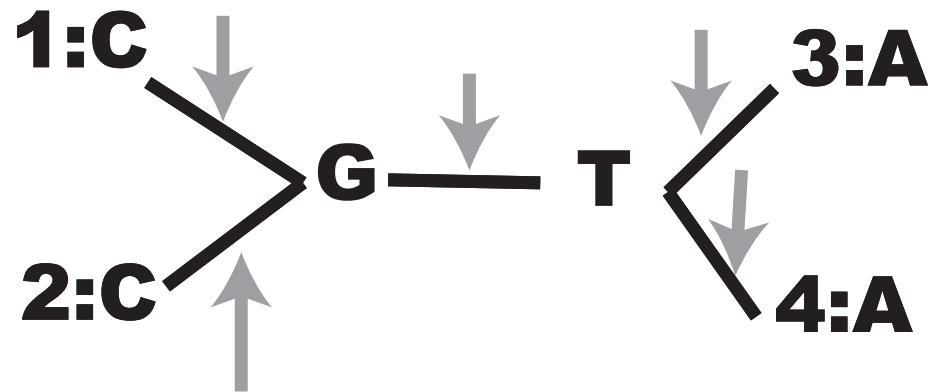
"parsimony
informative"

(column 1 will reappear on next slide)

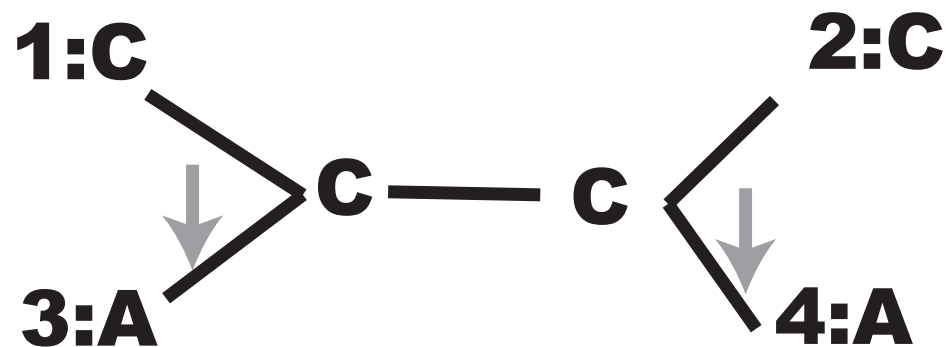
Parsimony Idea



is more reasonable than

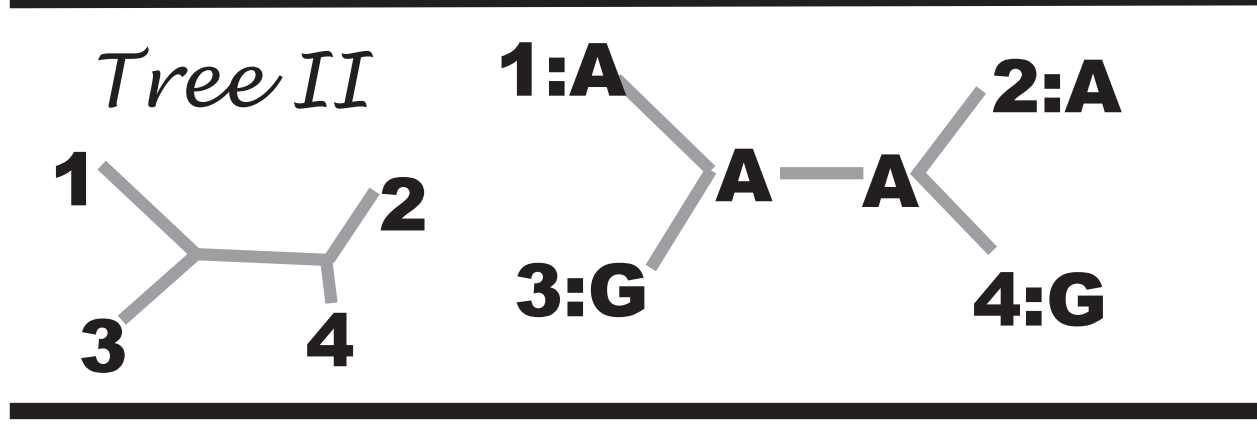
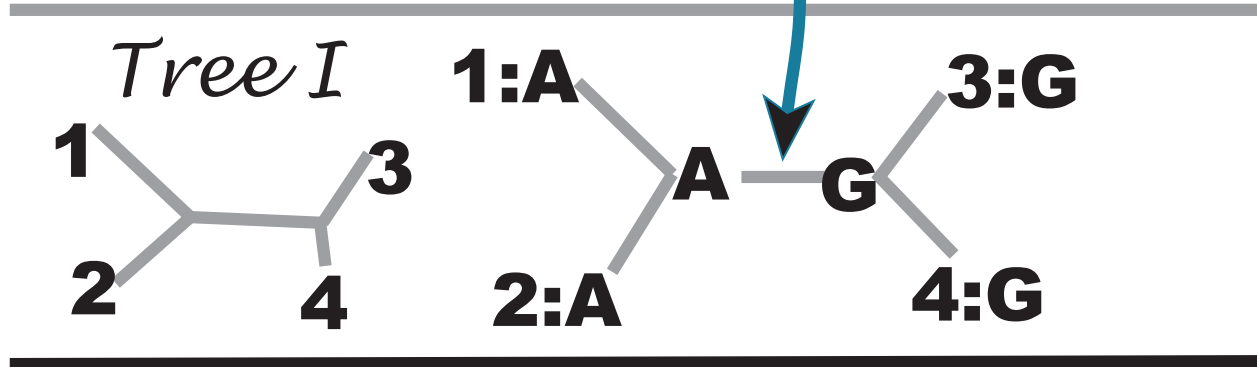


or



Taxon 1: A
Taxon 2: A
Taxon 3: G
Taxon 4: G

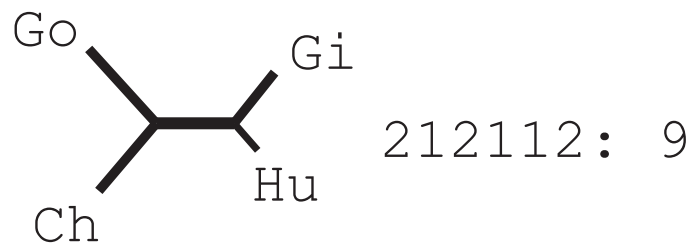
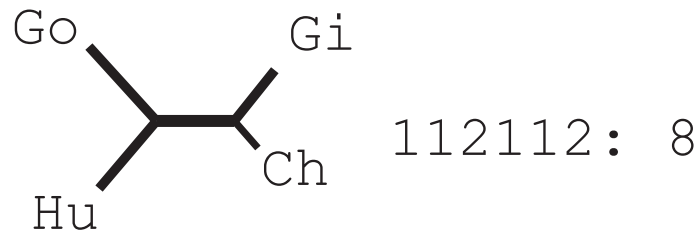
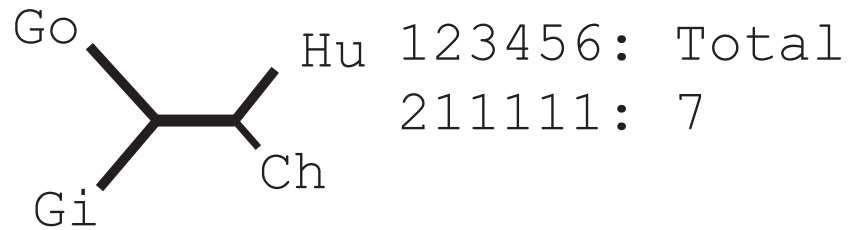
*Tree I
is most
parsimonious
here!*



Character:	123456
(Go) Gorilla:	GAGCTC
(Gi) Gibbon:	ACGACC
(Hu) Human:	GAAATT
(Ch) Chimp:	AAAATT



Parsimony Scores

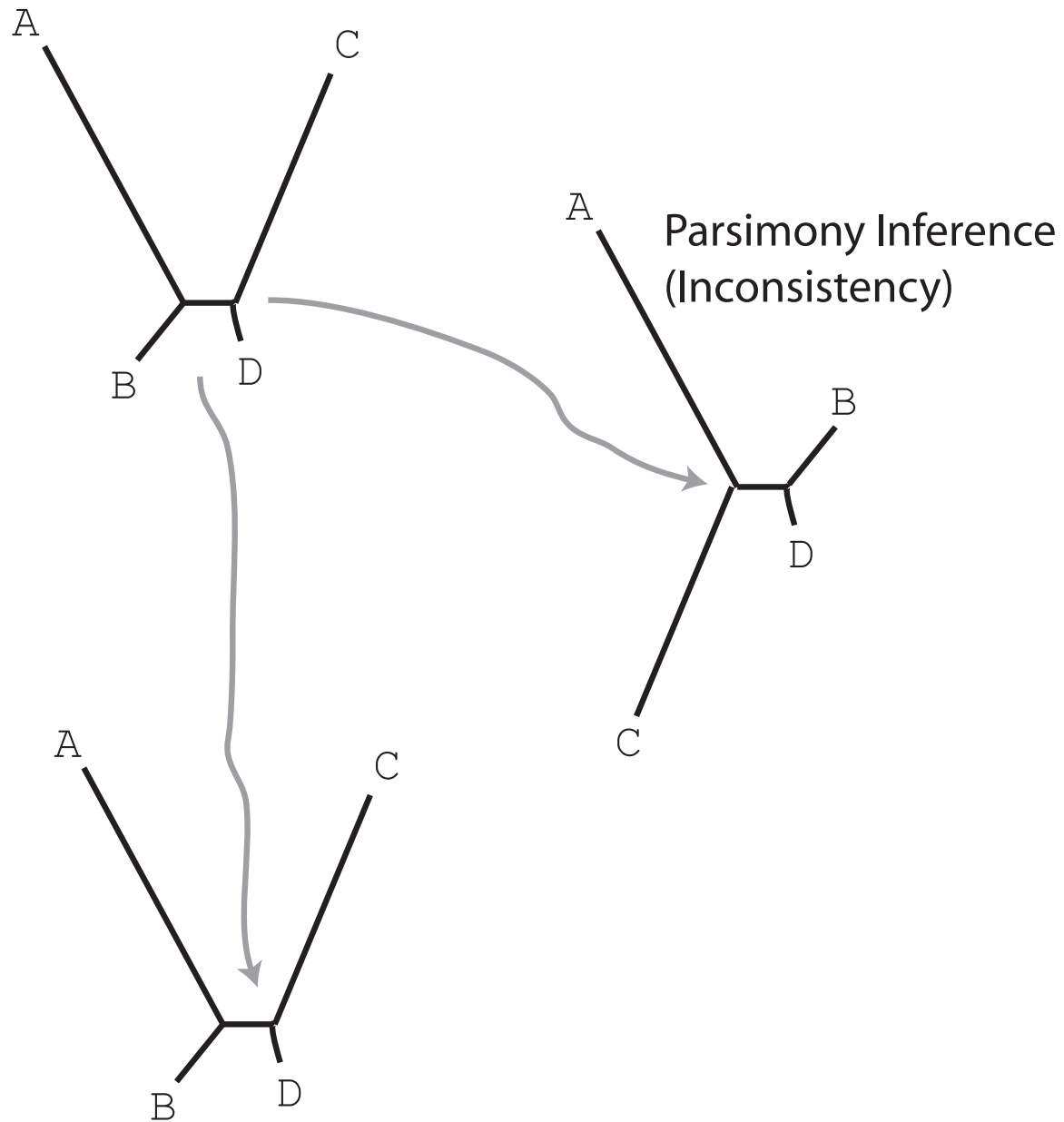


Step Matrices

		To			
		A	C	G	T
From	A	0	1	1	1
	C	1	0	1	1
	G	1	1	0	1
	T	1	1	1	0

Step matrix for Fitch parsimony

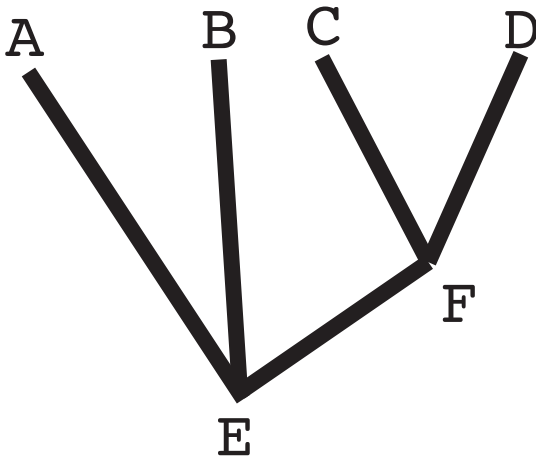
Slide borrowed with permission from
Paul Lewis



Correct model of sequence evolution with maximum likelihood or with some distance-based procedures (e.g. neighbor-joining) leads to consistent inference of topology.

Homework for Monday February 14

Name: _____



Please complete these 2 experiments before class and bring your results to class. Sorry about all of the coin flipping. But, we will rely on 2 data sets that you simulate. The idea will be that data are simulated according to an evolutionary tree as you see on the left. The letters "E" and "F" represent ancestral states of common ancestors whereas "A" and "B" and "C" and "D" are tips of the tree and will have observed states. We will discuss the value of simulation in class. But, one advantage is that you know the truth with simulation. Note that we will simulate data that are "0" or "1" rather than data that are "A" or "C" or "G" or "T" as would be the case for DNA. (Don't just copy answers or make up answers. Both are likely to be statistically detectable and will also affect the overall class experiment.)

First Case Experiment (Repeat 10 times):

- Step 1: Flip a coin.
If it is heads, Set E=0.
Otherwise, Set E=1.
- Step 2: Flip 4 coins.
If all 4 are heads, then set A=1-E.
Otherwise, set A=E.
- Step 3: Flip 4 coins. If all 4 are heads,
then set B= 1-E. Otherwise, set B=E.
- Step 4: Flip 2 coins. If both are heads,
then set F= 1-E. Otherwise, set F=E.
- Step 5: Flip 4 coins. If all 4 are heads,
then set C= 1-F. Otherwise, set C=F.
- Step 6: Flip 4 coins. If all 4 are heads,
then set D= 1-F. Otherwise, set D=F.
- Step 7: Record A&B&C&D in appropriate Count
Column Below.

My results (First Case Exp):

- Step 1: Flip a coin. T
- If it is heads, Set E=0. E=1
- Otherwise, Set E=1.
- Step 2: Flip 4 coins. TTHT
- If all 4 are heads, then set A=1-E. A=E=1
- Otherwise, set A=E.
- Step 3: Flip 4 coins. If all 4 are heads, HHTH
then set B= 1-E. Otherwise, set B=E. B=E=1
- Step 4: Flip 2 coins. If both are heads, TH
then set F= 1-E. Otherwise, set F=E. F=E=1
- Step 5: Flip 4 coins. If all 4 are heads, TTTT
then set C= 1-F. Otherwise, set C=F. C=F=1
- Step 6: Flip 4 coins. If all 4 are heads, HTTT
then set D= 1-F. Otherwise, set D=F. D=F=1
- Step 7: Record A&B&C&D in appropriate Count
Column Below.

...So record count for A=B=C=D=1

A	B	C	D		A	B	C	D	Count
0	0	0	0		1	1	1	1	1
0	0	0	1		1	1	1	0	_____
0	0	1	0		1	1	0	1	_____
0	0	1	1	OR	1	1	0	0	_____
0	1	0	0		1	0	1	1	_____
0	1	0	1		1	0	1	0	_____
0	1	1	0		1	0	0	1	_____
0	1	1	1		1	0	0	0	_____

Distance Methods for Phylogeny Inference

Most slides courtesy of ...

Dr. Mark Holder, University of Kansas

...with a few slides courtesy of ...

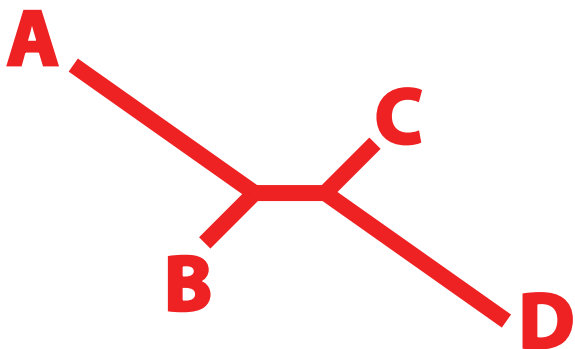
Dr. Paul Lewis, University of Connecticut

Distance-based approaches to inferring trees

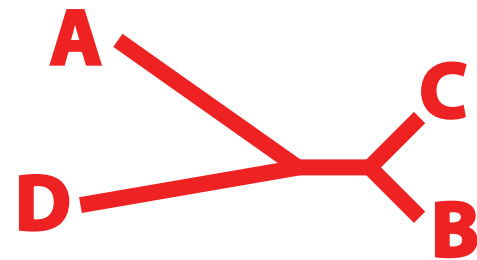
(1) Convert the raw data (sequences) to pairwise distances

(2) Find a tree that best explains these distances.

- **Do Not** simply cluster the most similar sequences



Versus



	1	2	3	4	5	6	7	8	9	10
Species 1	C	G	A	C	C	A	G	G	T	A
Species 2	C	G	A	C	C	A	G	G	T	A
Species 3	C	G	G	T	C	C	G	G	T	A
Species 4	C	G	G	C	C	A	T	G	T	A

Can be converted to a distance matrix:

	Species 1	Species 2	Species 3	Species 4
Species 1	0	0	0.3	0.2
Species 2	0	0	0.3	0.2
Species 3	0.3	0.3	0	0.3
Species 4	0.2	0.2	0.3	0

Note that the distance matrix is symmetric.

	Species 1	Species 2	Species 3	Species 4
Species 1	0	0	0.3	0.2
Species 2	0	0	0.3	0.2
Species 3	0.3	0.3	0	0.3
Species 4	0.2	0.2	0.3	0

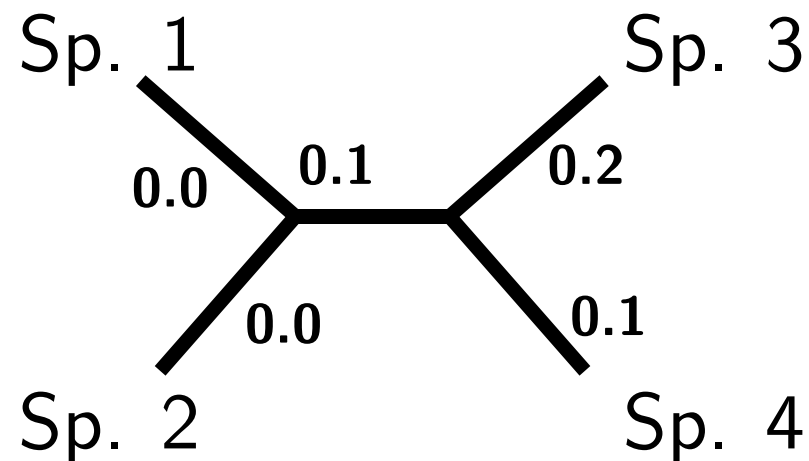
. . . so we can just use the lower triangle.

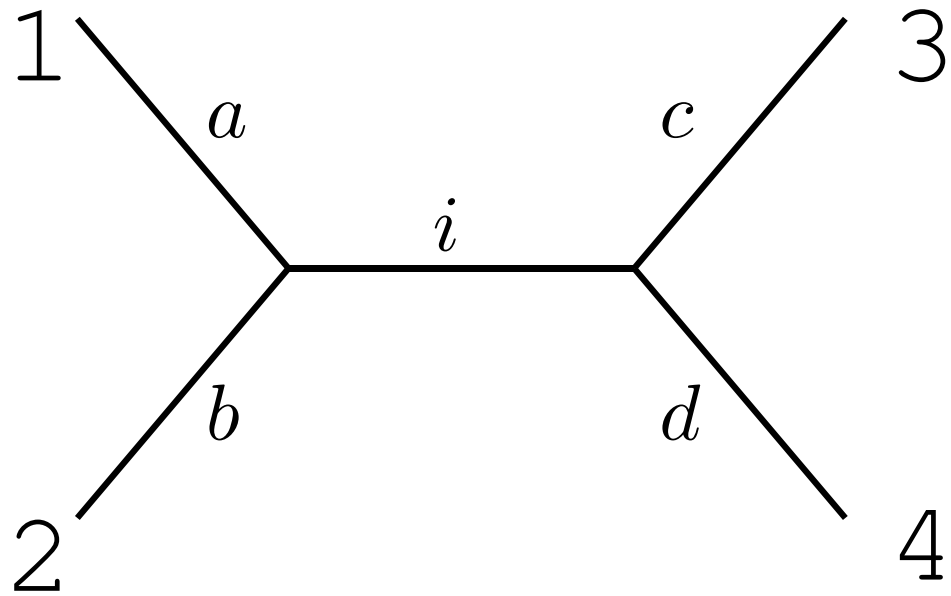
	Species 1	Species 2	Species 3
Species 2	0		
Species 3	0.3	0.3	
Species 4	0.2	0.2	0.3

Can we find a tree that would predict these observed character divergences?

	Species 1	Species 2	Species 3
Species 2	0		
Species 3	0.3	0.3	
Species 4	0.2	0.2	0.3

Can we find a tree that would predict these observed character divergences?





parameters

$$p_{12} = a + b$$

$$p_{13} = a + i + c$$

$$p_{14} = a + i + d$$

$$p_{23} = b + i + c$$

$$p_{24} = b + i + d$$

$$p_{34} = c + d$$

	data		
	1	2	3
2	d_{12}		
3	d_{13}	d_{23}	
4	d_{14}	d_{24}	d_{34}

Why doesn't simple clustering work?

Step 1: use sequences to estimate pairwise distances between taxa.

	A	B	C	D
A	-	0.2	0.5	0.4
B		-	0.46	0.4
C			-	0.7
D				-

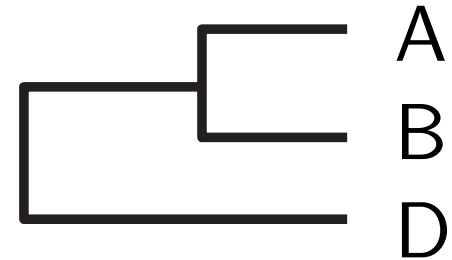
Why doesn't simple clustering work?

	A	B	C	D
A	-	0.2	0.5	0.4
B		-	0.46	0.4
C			-	0.7
D				-



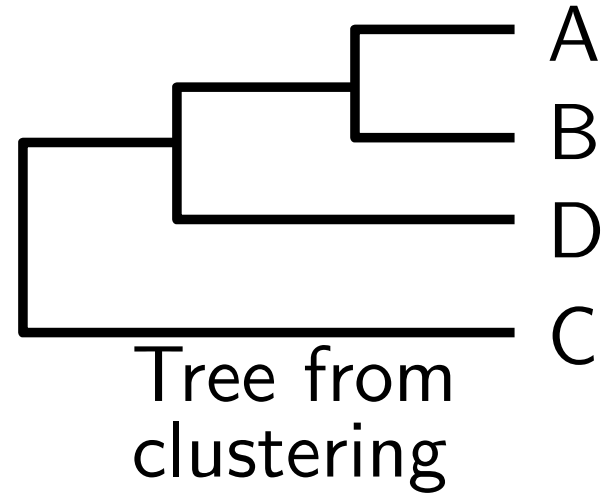
Why doesn't simple clustering work?

	A	B	C	D
A	-	0.2	0.5	0.4
B		-	0.46	0.4
C			-	0.7
D				-



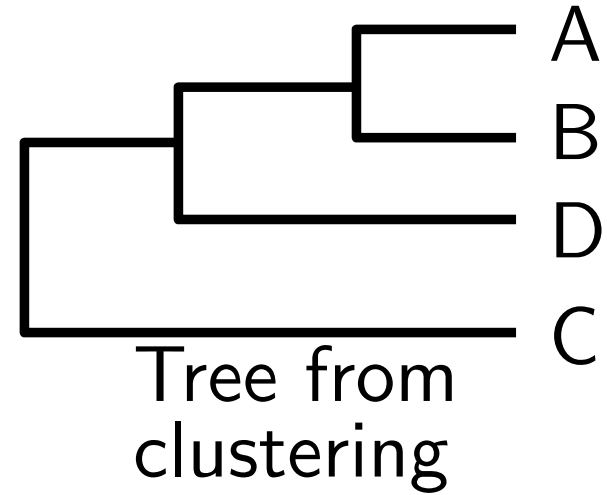
Why doesn't simple clustering work?

	A	B	C	D
A	-	0.2	0.5	0.4
B		-	0.46	0.4
C			-	0.7
D				0



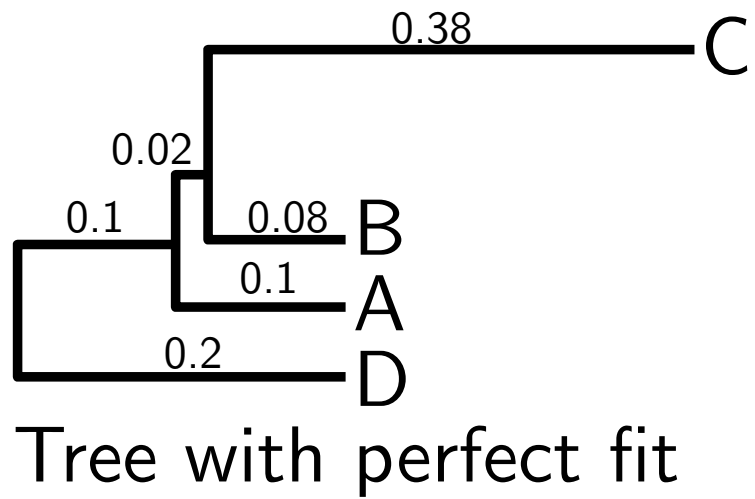
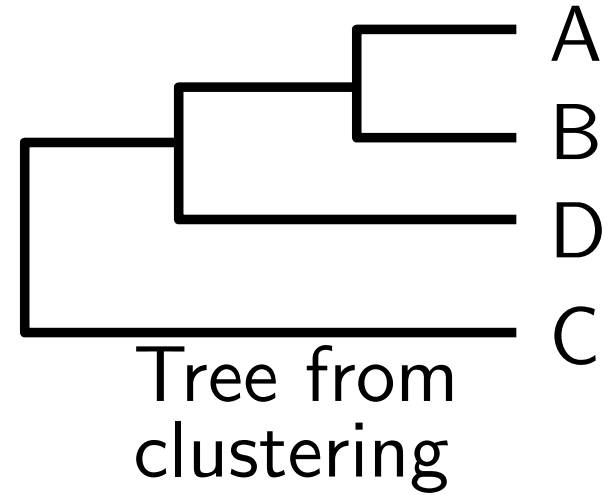
Why doesn't simple clustering work?

	A	B	C	D
A	0	0.2	0.5	0.4
B	0.2	0.2	0.46	0.4
C	0.5	0.46	0	0.7
D	0.4	0.4	0.7	0



Why doesn't simple clustering work?

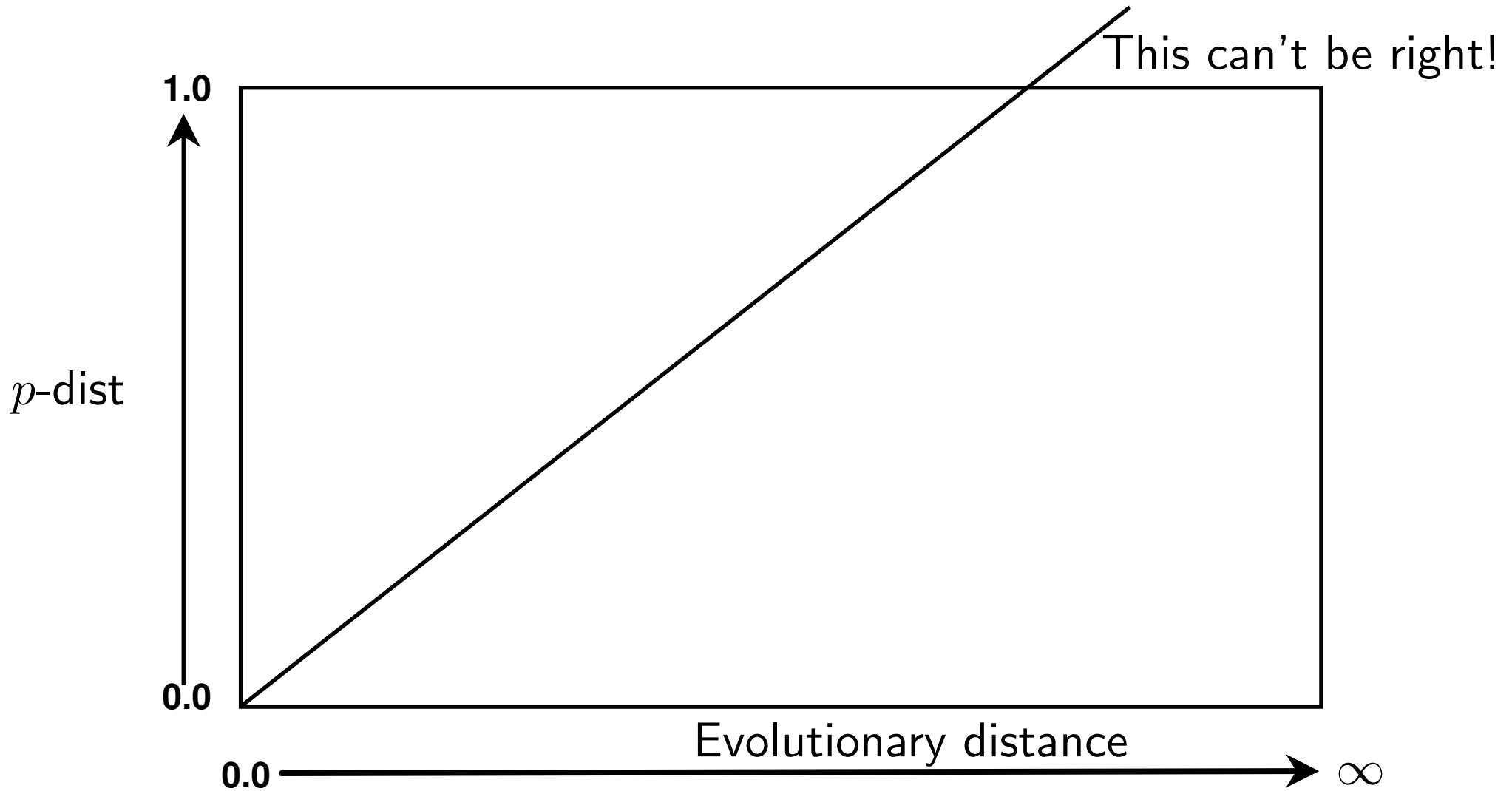
	A	B	C	D
A	0	0.2	0.5	0.4
B	0.2	0.	0.46	0.4
C	0.5	0.46	0	0.7
D	0.4	0.4	0.7	0



p-distance = proportion of positions that are different in 2 sequences.

Hamming distance = number of positions at which two sequences ("strings") differ.

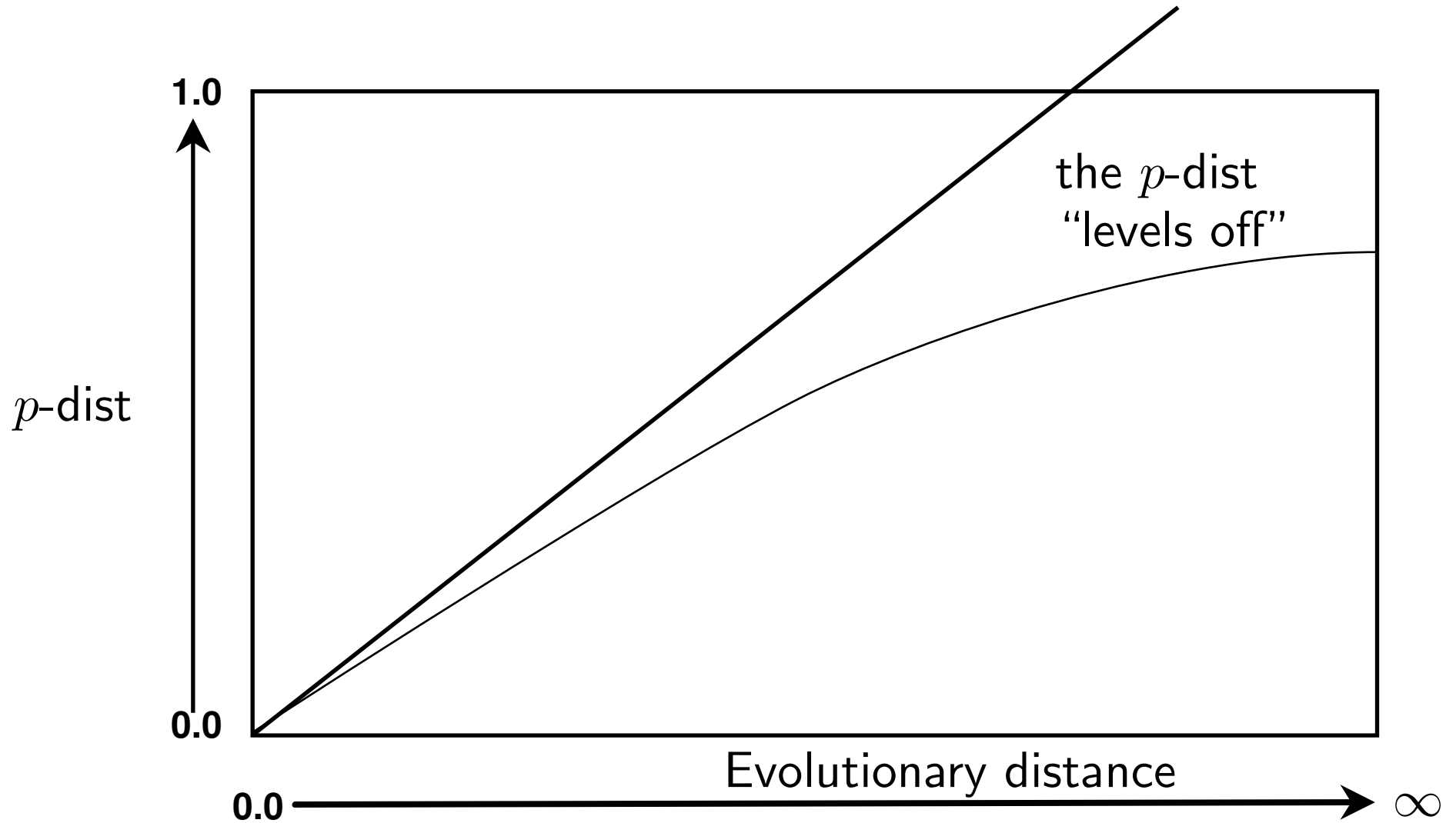
Intuition of sequence divergence vs evolutionary distance



“Multiple hits” problem (also known as saturation)

- Levelling off of sequence divergence vs time plot is caused by multiple substitutions affecting the same site in the DNA.
- At large distances the “raw” sequence divergence (also known as the p-distance or Hamming distance) is a poor estimate of the true evolutionary distance.
- Statistical models must be used to correct for unobservable substitutions
- Large p-distances respond more to model-based correction – and there is a larger error associated with the correction.

Sequence divergence vs evolutionary distance



Besides parsimony and distance-based methods for inferring evolutionary trees, there are two additional widely-used categories of methods.

Both of the other two rely on probabilistic models of sequence change and they therefore have some connection to each other.

One of these categories of method is known as maximum likelihood and the other is known as Bayesian inference.

Bayesian inference and maximum likelihood methods are more statistically sound than parsimony and distance-based methods.

Maximum likelihood uses only the probabilistic model and the data. Bayesian inference uses these plus prior information.

However, they are both computationally more demanding and so sometimes data sets are too big to use them.

Phylogeny Reconstruction is computationally difficult

Number of Tips	Number of Rooted Trees	Number of Unrooted Trees
2	1	1
3	3	1
4	15	3
5	105	15
6	945	105
7	10,395	945
8	135,135	10,395
9	2,027,025	135,135
10	34,459,425	2,027,025
...		
N	$\frac{(2N-5)!}{2^{N-3} (N-3)!}$	$\frac{(2N-3)!}{2^{N-2} (N-2)!}$