Mon, Oct 19

Lecture 06A:

Molecular phylogenetics

Lonicera flava
© Kathy Melton/
Missouri Botanical Garden

Practical Bioinformatics (Biol 4220)
Instructor: Michael Landis
Email: michael.landis@wustl.edu

Lecture 6A outline

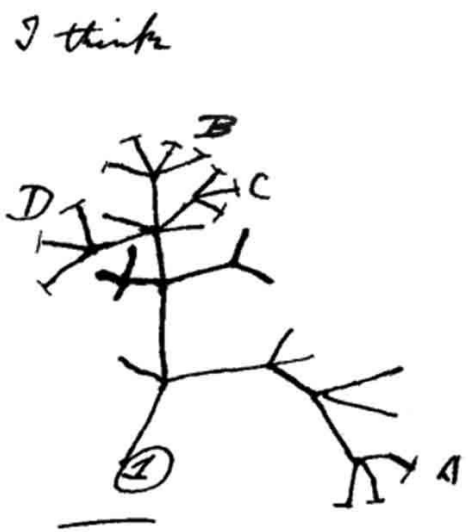
1. Intro to phylogenetics
2. Phylogenetic inference
3. Lab 6A overview

Phylogenetics

Phylogenetics studies the relationships among heritable, biological units (often called ***taxa***)

Phylogenies are useful for

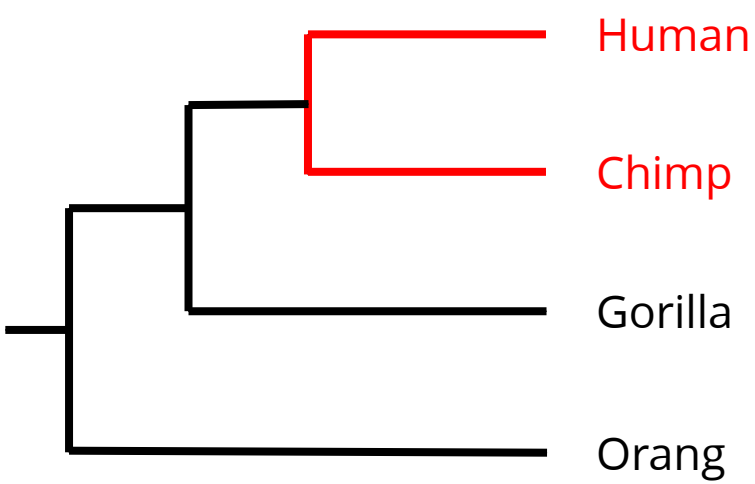
- gene annotation
- tracking the spread of a virus
- identifying zoonotic events
- reconstructing tumorogenesis
- conservation biology assays
- inferring species relationships



a phylogeny by Darwin

Reading a phylogeny

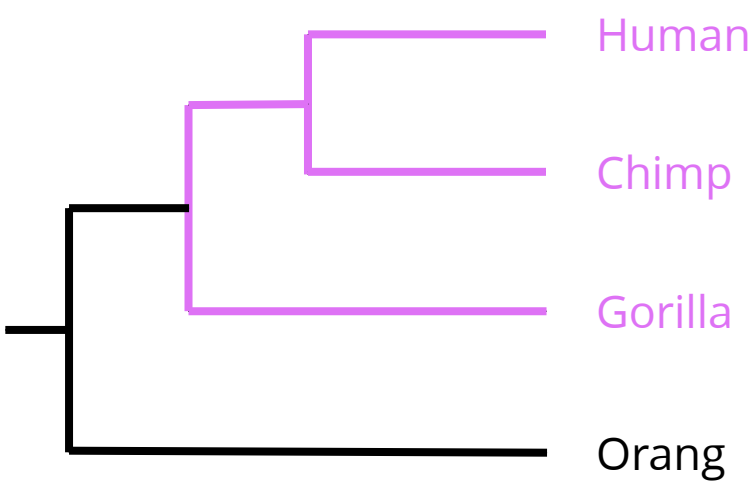
Phylogenetic relationships are hierarchical, and most often represented as ***trees***



Human and Chimp are more closely related to each other than to Gorilla or Orang

Reading a phylogeny

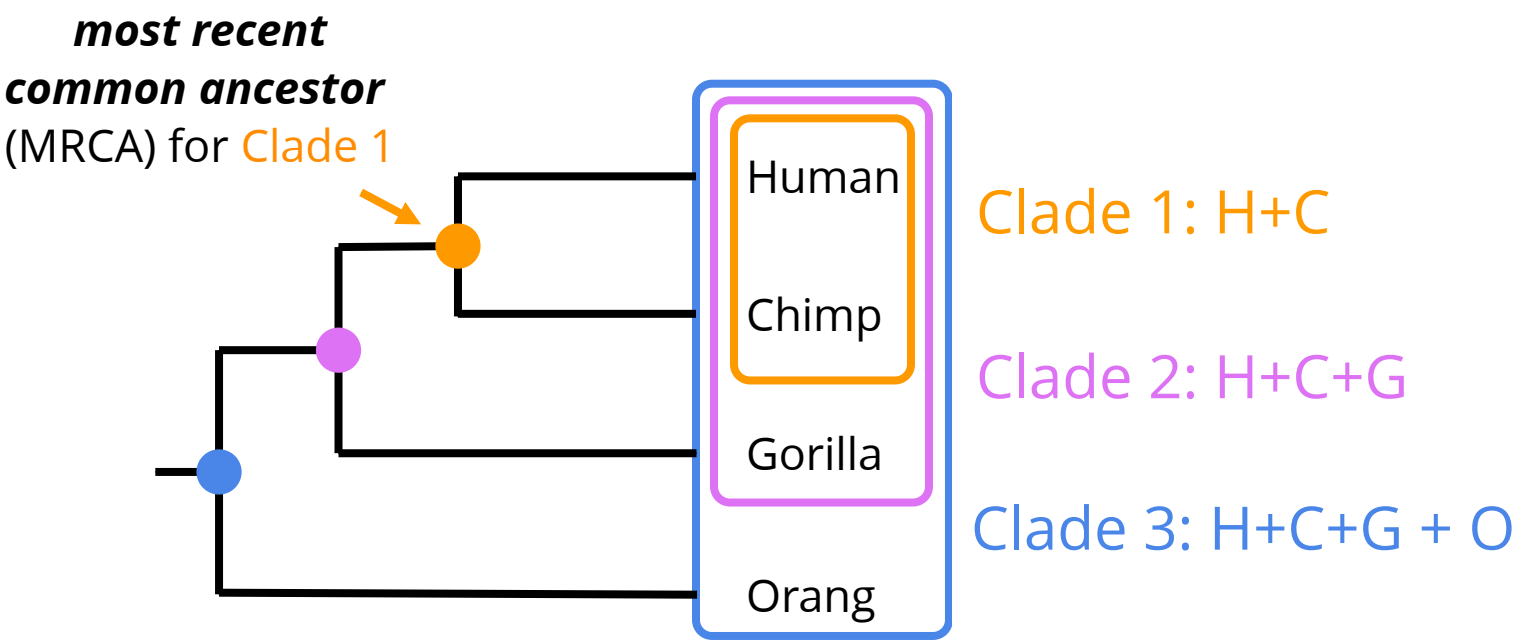
Phylogenetic relationships are hierarchical, and most often represented as ***trees***



Human, Chimp, and Gorilla are more closely related to each other than to Orang

Reading a phylogeny

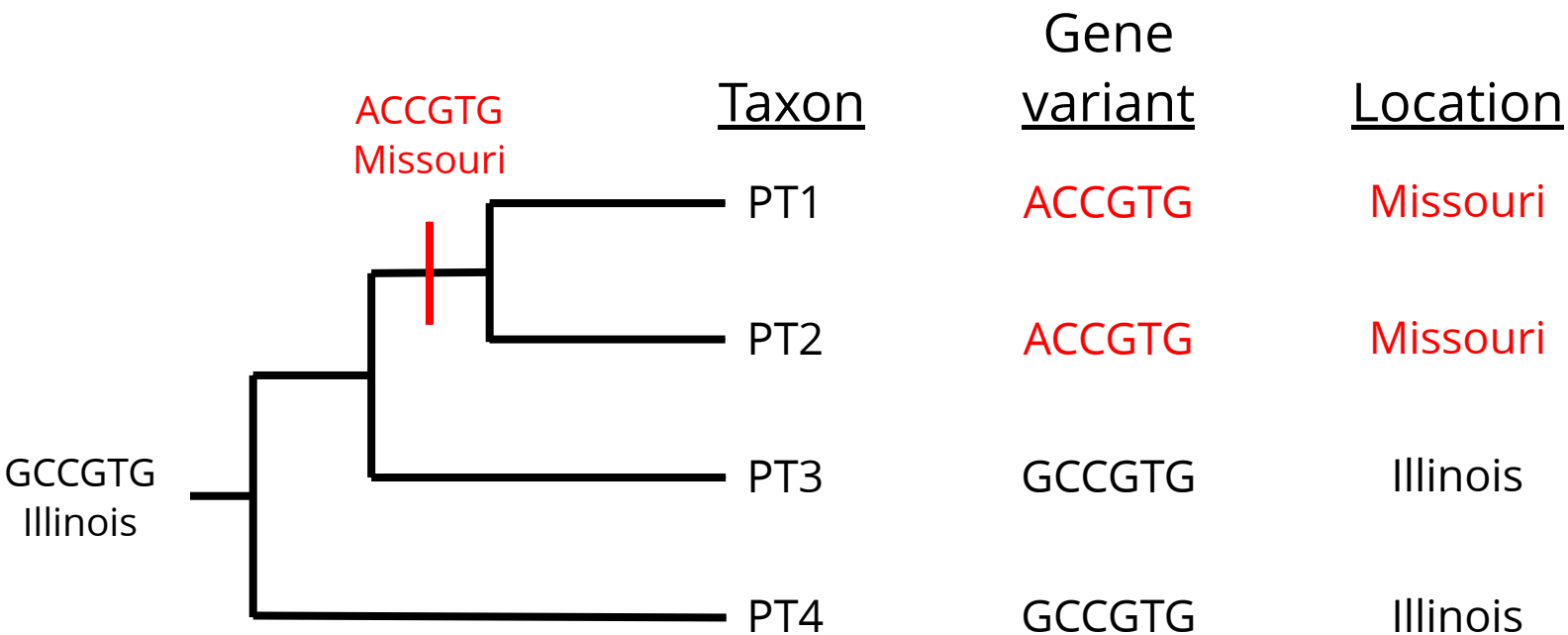
Taxa that are more closely related to one another, over any other taxa, are called ***clades***



"Tree-thinking"

Taxon	Gene variant	Location
PT1	ACCGTG	Missouri
PT2	ACCGTG	Missouri
PT3	GCCGTG	Illinois
PT4	GCCGTG	Illinois

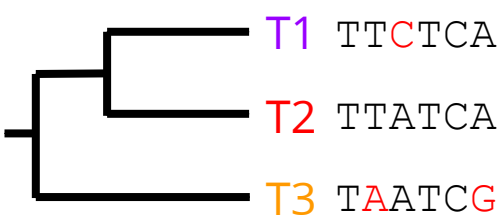
"Tree-thinking"

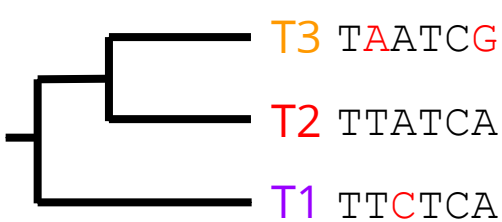


Phylogeny informs when and where variation arose, guiding biological research

Inferring phylogeny

How are taxa T1, T2, and T3 related?





Which phylogeny generated the observed pattern of molecular variation?

Inferring phylogeny

Phylogenetic inference methods take a matrix of characters (*e.g. DNA alignment*) as input

Measure how well any possible phylogenetic estimate explains the data matrix pattern by assigning a ***cost*** to each considered estimate

Methods generally ***optimize*** to estimate the phylogeny with the lowest cost for provided data matrix

Tree-space is large

# taxa	# rooted trees
3	3
4	15
5	105
6	945
7	10395
8	135135
9	2027025
10	34459425

Major challenge is efficiently exploring trees with optimal scores

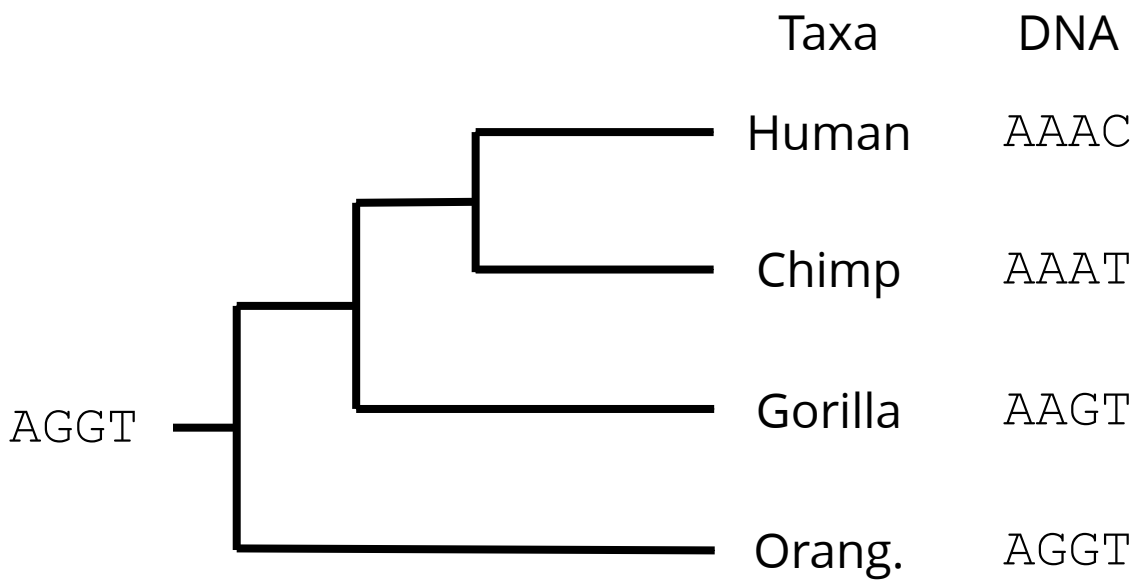
Phylogenetic method types

Most methods used to infer phylogenies compute scores based

1. event counting (***parsimony***)
2. event probabilities (***likelihood***)
3. pattern distances (***neighbor joining***)

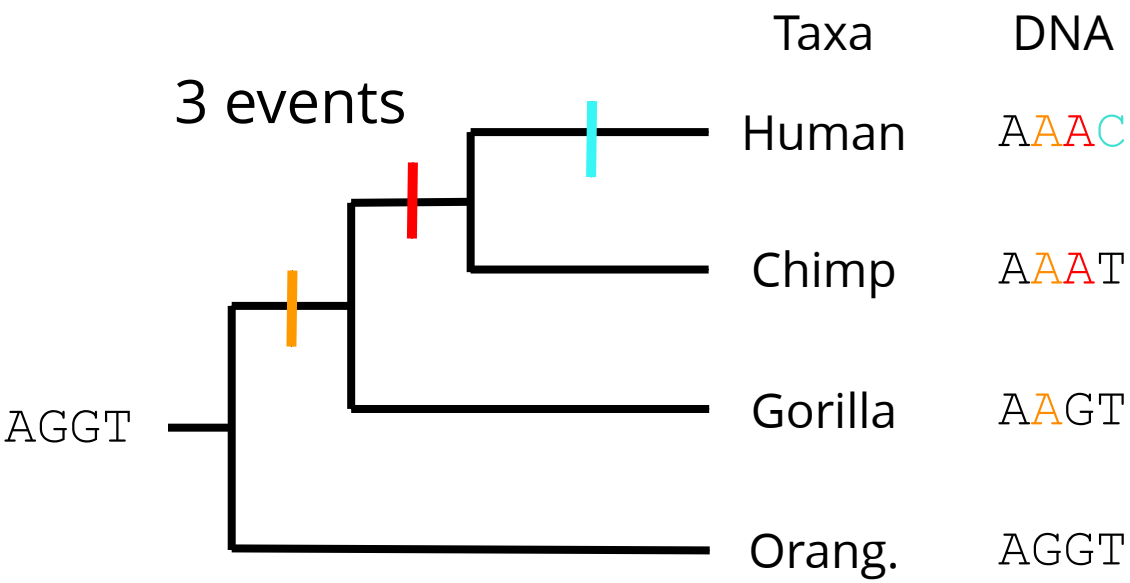
Method choice often relates to concerns regarding accuracy, speed, scalability, etc.

Parsimony



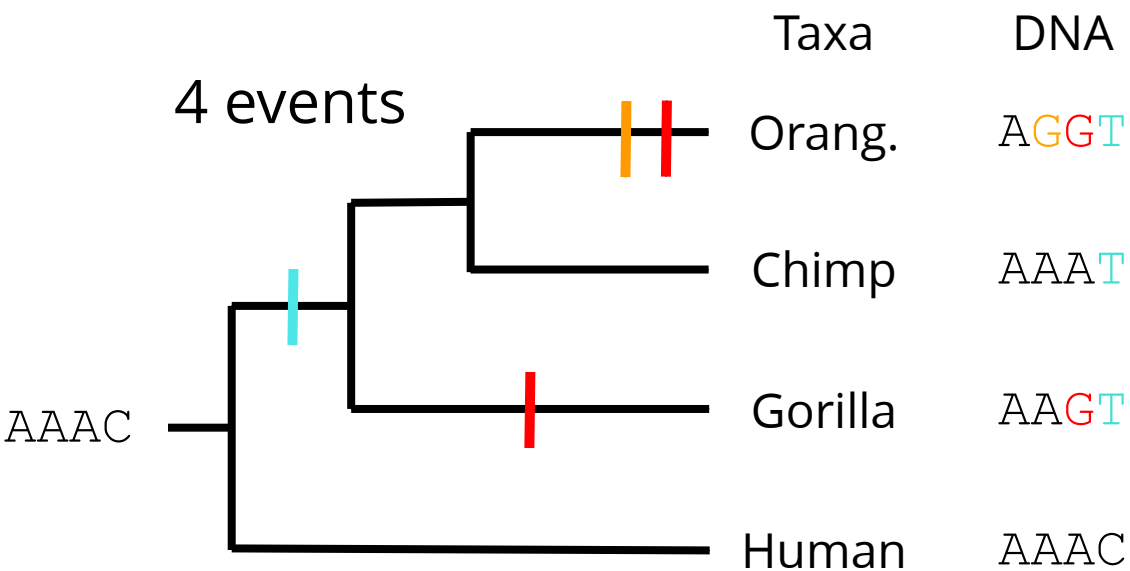
What phylogeny requires the fewest character substitution events?

Parsimony



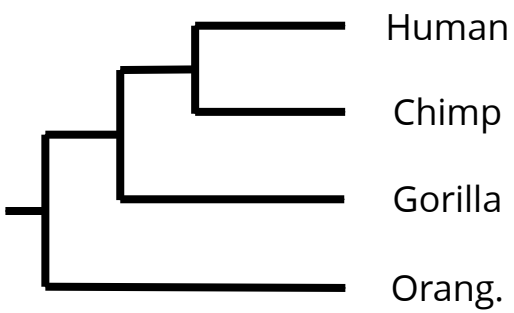
What phylogeny requires the fewest character substitution events?

Parsimony

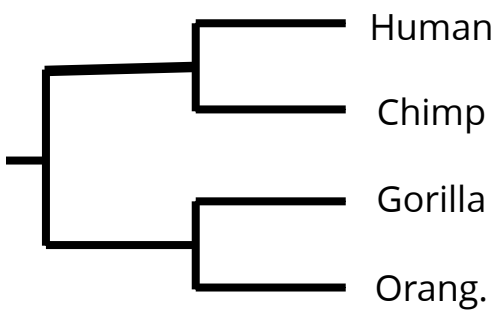


What phylogeny requires the fewest character substitution events?

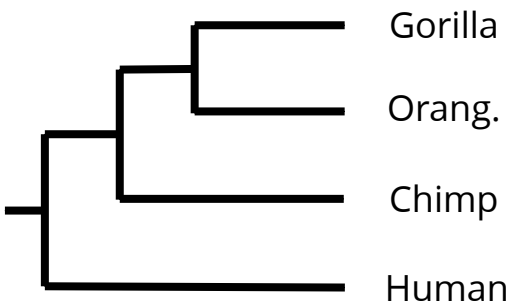
Parsimony



3 events



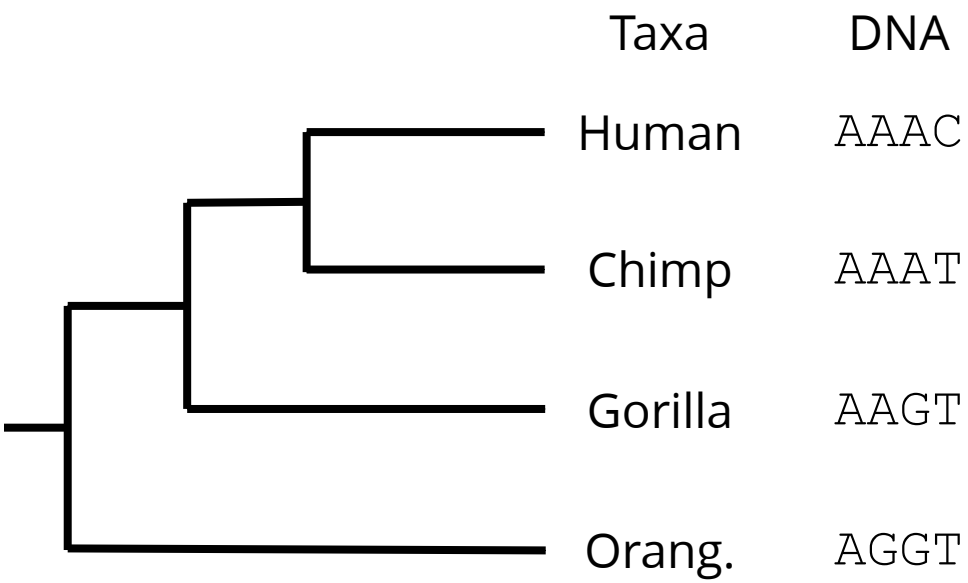
4 events



5 events

What phylogeny requires the fewest character substitution events?

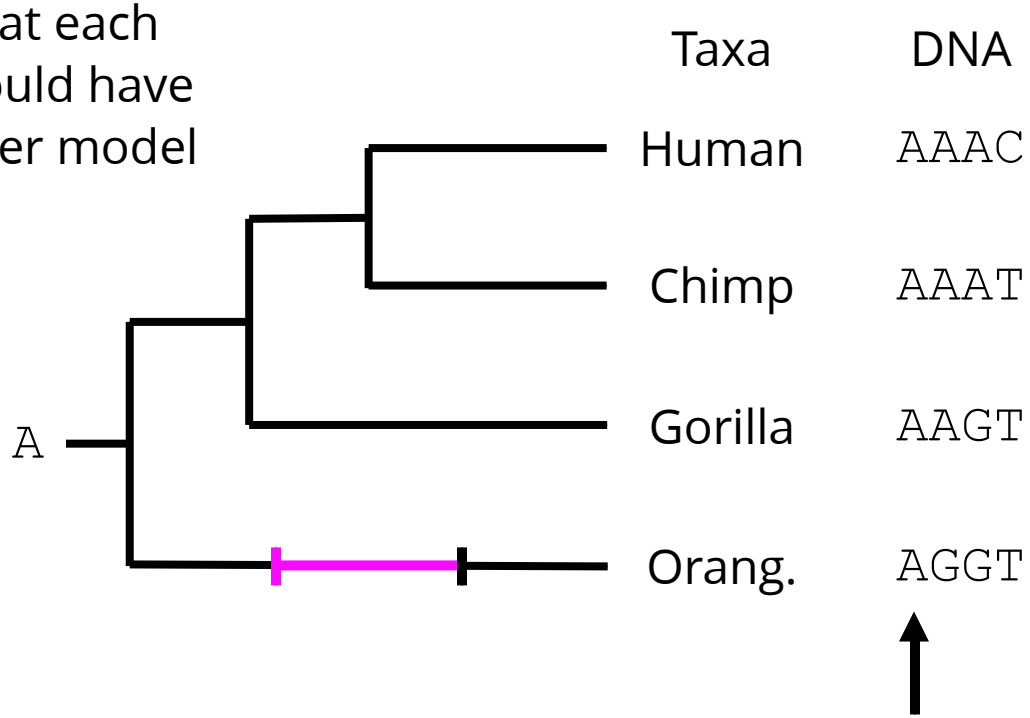
Likelihood



What phylogeny and model of evolution
is *most likely* to generate the character data?

Likelihood

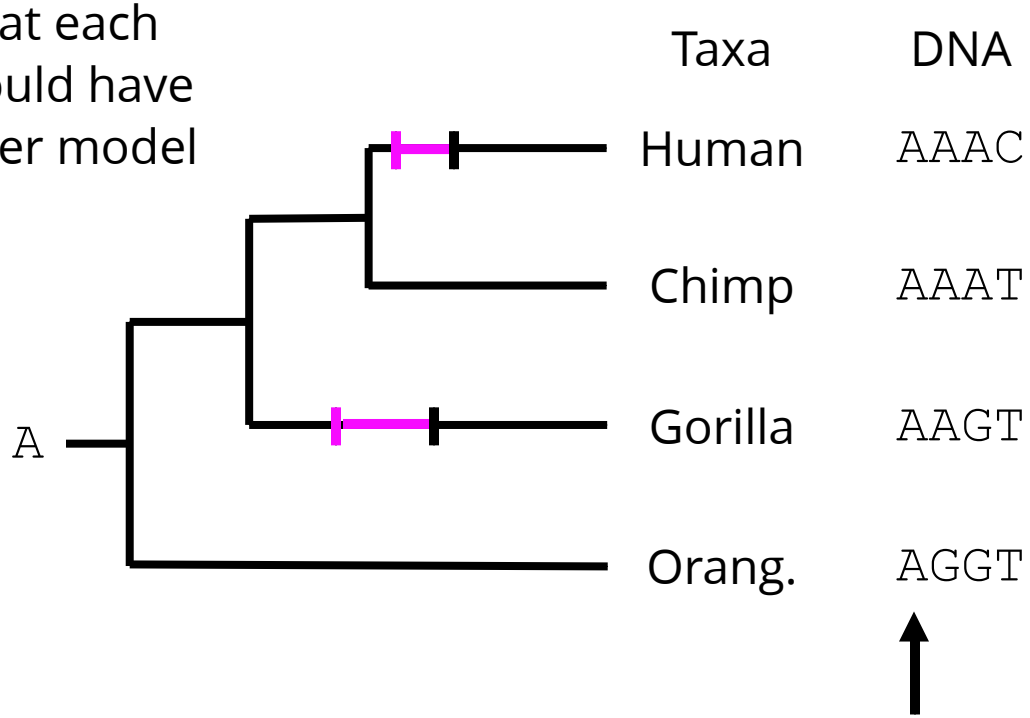
Compute probability for all ways that each character could have evolved under model



What phylogeny and model of evolution is *most likely* to generate the character data?

Likelihood

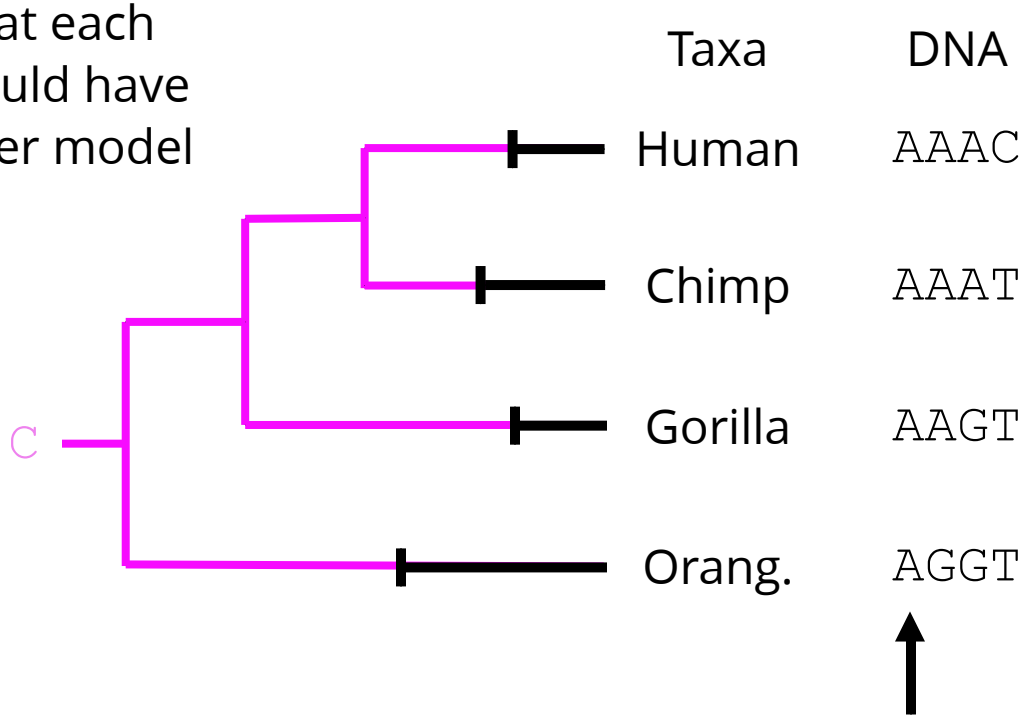
Compute probability for all ways that each character could have evolved under model



What phylogeny and model of evolution is *most likely* to generate the character data?

Likelihood

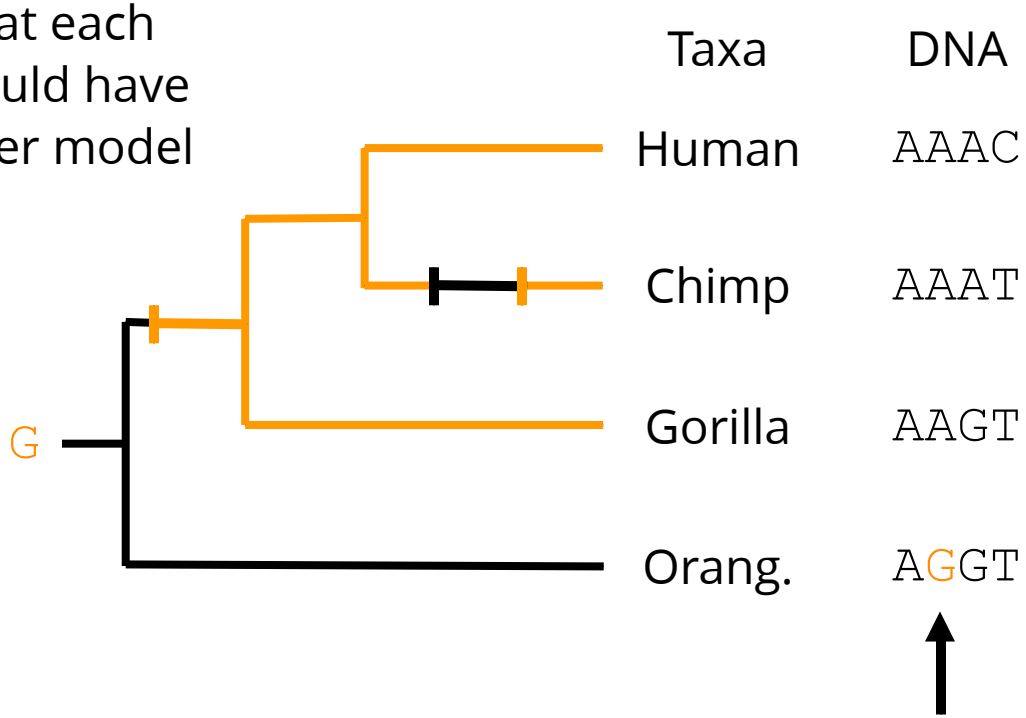
Compute probability for all ways that each character could have evolved under model



What phylogeny and model of evolution is *most likely* to generate the character data?

Likelihood

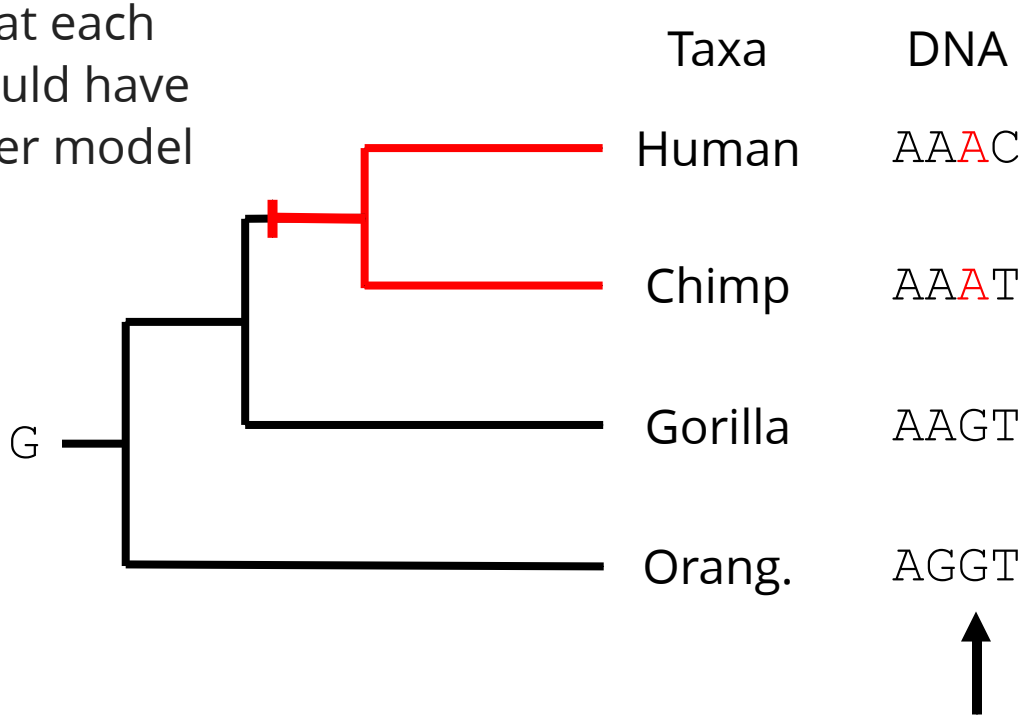
Compute probability for all ways that each character could have evolved under model



What phylogeny and model of evolution is *most likely* to generate the character data?

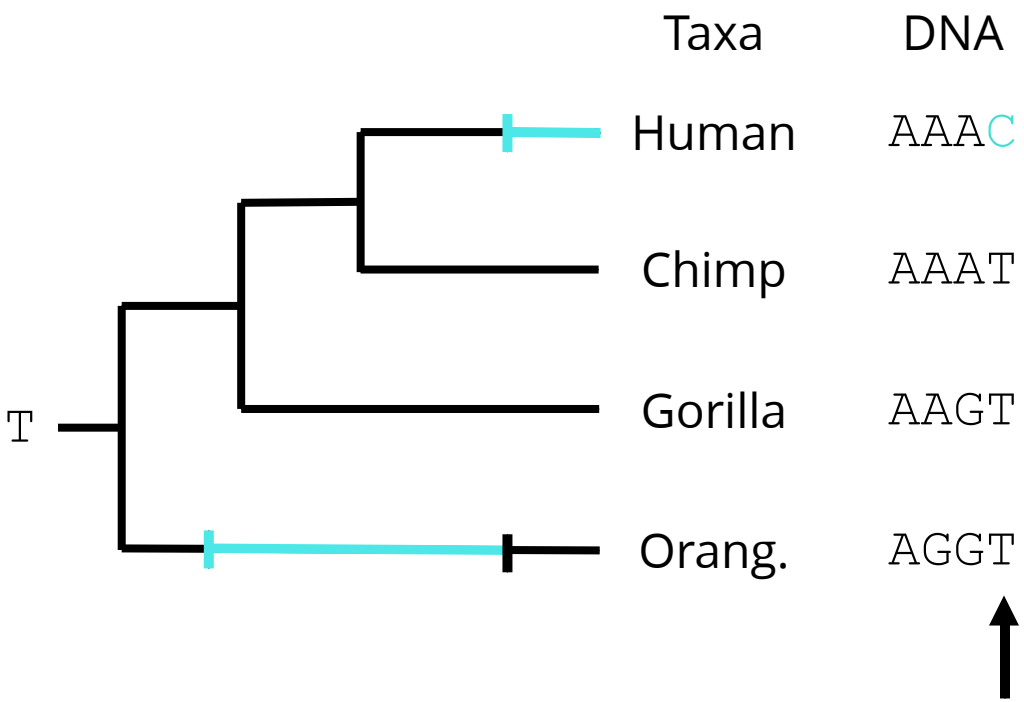
Likelihood

Compute probability for all ways that each character could have evolved under model



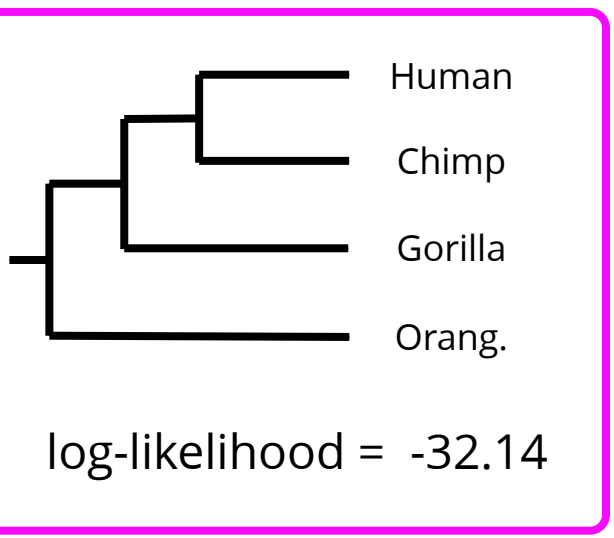
What phylogeny and model of evolution is *most likely* to generate the character data?

Likelihood

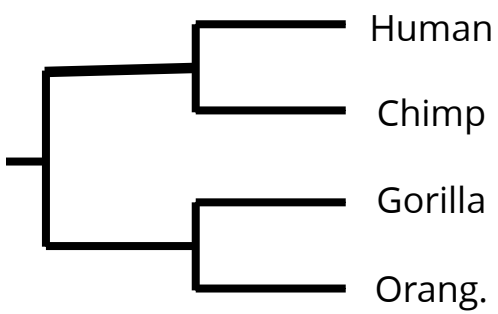


What phylogeny and model of evolution
is *most likely* to generate the character data?

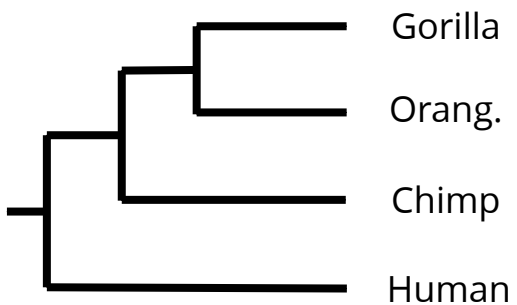
Likelihood



log-likelihood = -32.14



log-likelihood = -42.77



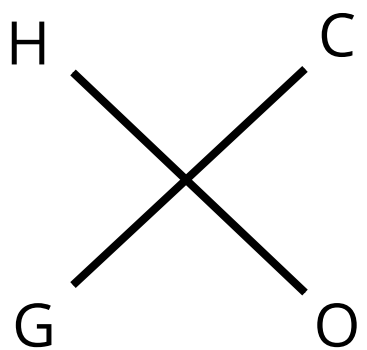
log-likelihood = -39.08

What phylogeny and model of evolution
is *most likely* to generate the character data?

Neighbor-joining

	H	C	G	O
H	0	1	3	5
C	1	0	3	5
G	3	3	0	2
O	5	5	2	0

distances matrix
for sequence pairs

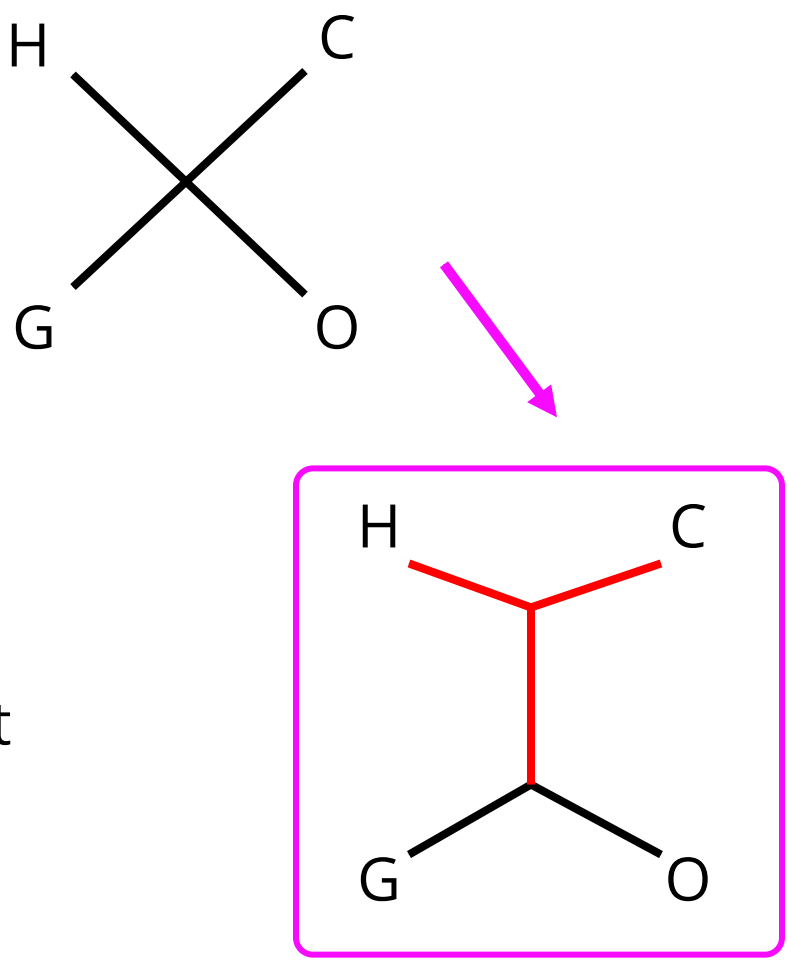


Select pairs of taxa with short sequence distances, and join them as neighbors in tree

Neighbor-joining

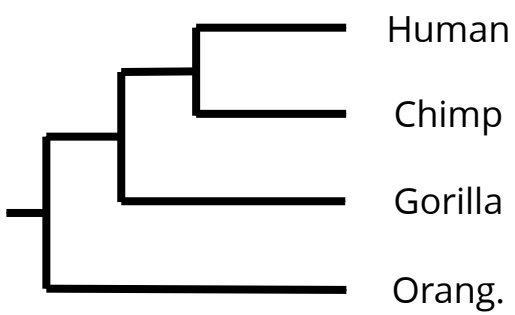
	H	C	G	O
H	0	1	3	5
C	1	0	3	5
G	3	3	0	2
O	5	5	2	0

distances matrix
for sequence pairs

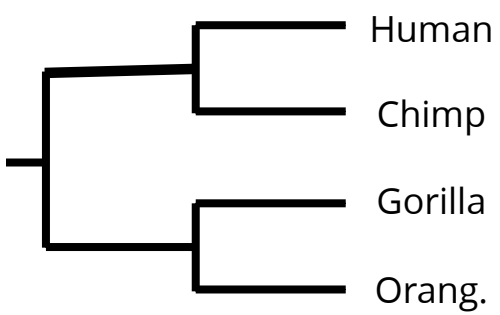


Select pairs of taxa with short sequence distances, and join them as neighbors in tree

Newick strings



(((Human,Chimp),Gorilla),Orang);



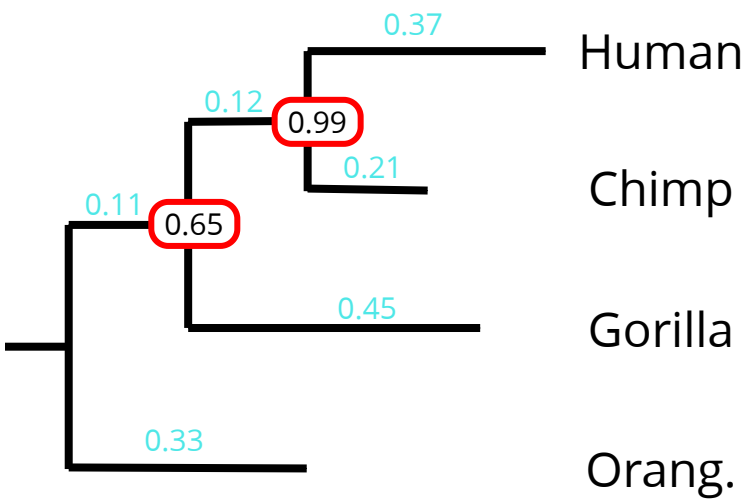
(((Human,Chimp),(Gorilla,Orang));

Taxa in parentheses define clades;
commas define divergences

Newick strings

Branch lengths
measured in expected
substitutions per site

Clade support
measures reliability of
clade in tree estimate



(((Human:0.37,Chimp:0.21)0.99:0.12,
Gorilla:0.45)0.65:0.11,Orang:0.33);

Lab 6A

github.com/WUSTL-Biol4220/home/labs/lab_06A.md