Lecture 12
Molecular phylogenetics

Lonicera flava
© Kathy Melton/
Missouri Botanical Garden

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

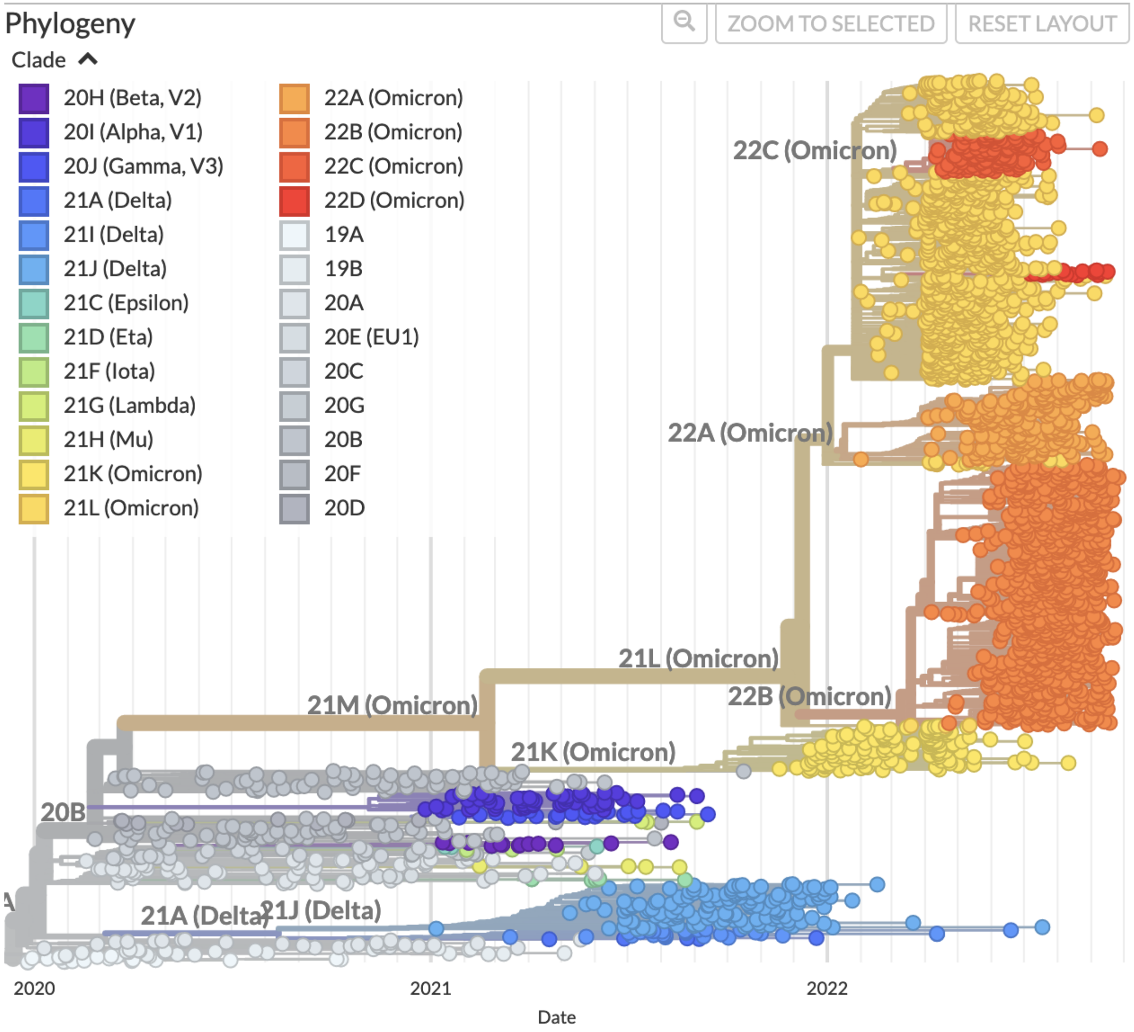
Lecture 12 outline

Last time: sequencing & clusters

This time: phylogenetics

- interpreting trees
- tree-thinking
- inferring trees
- inference methods

Phylogenetics



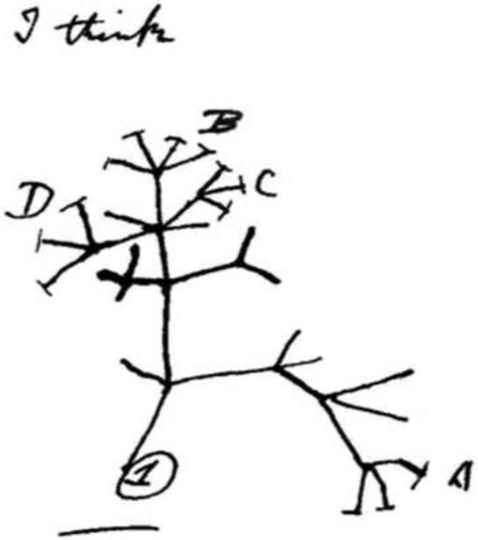
SARS-CoV-2 phylogeny from nextstrain.org

Phylogenetics

Phylogenetics studies the relationships among evolutionary lineages (often called ***taxa***)

Phylogenies are useful for

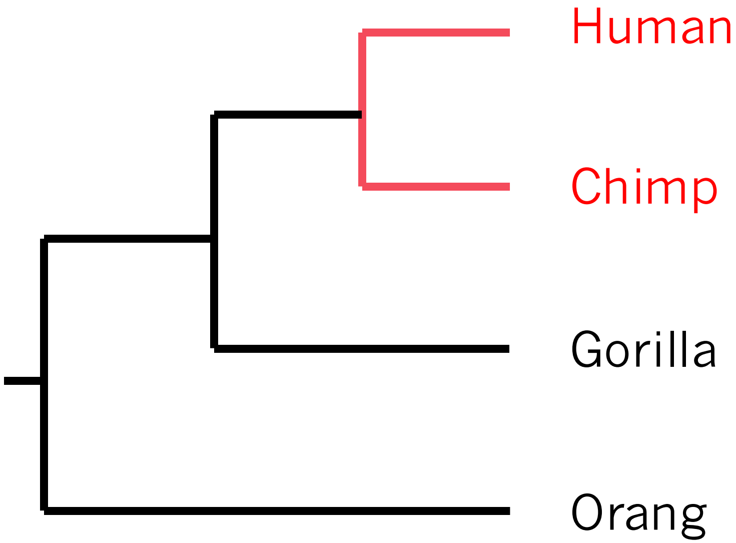
- gene annotation
- tracking viral spread
- identifying zoonosis
- reconstructing tumorogenesis
- conservation biology assays
- inferring species relationships



phylogeny sketch by Darwin

Reading a phylogeny

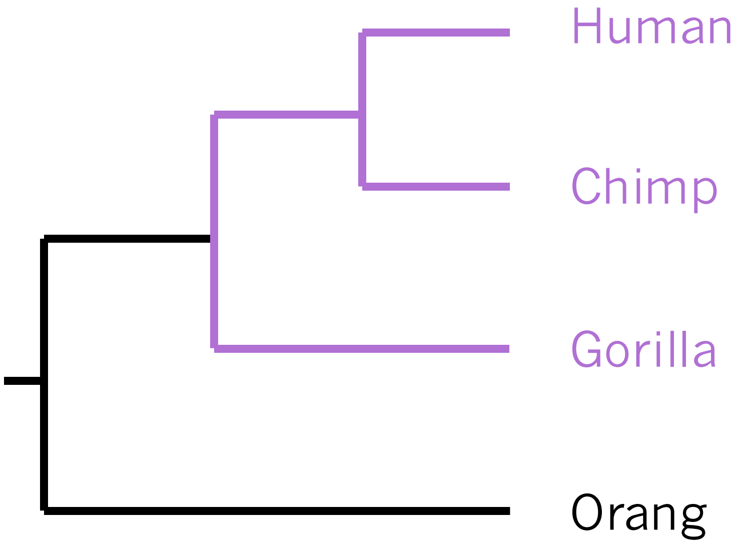
Phylogenetic relationships are hierarchical, and most often represented as bifurcating ***trees***



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

Reading a phylogeny

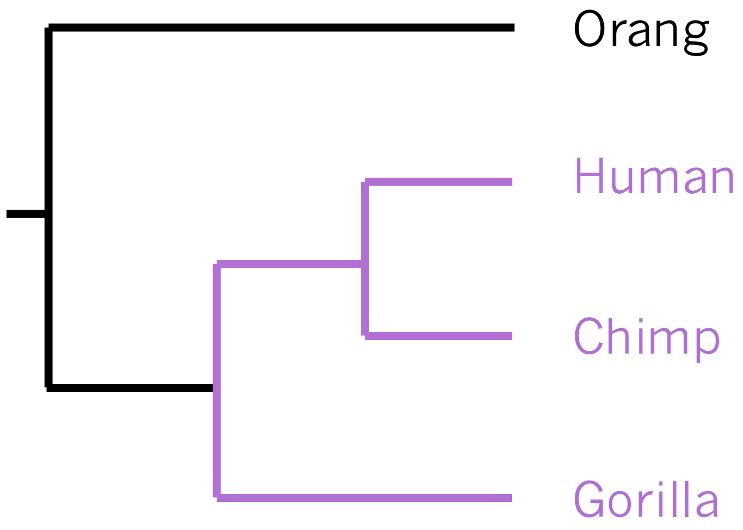
Phylogenetic relationships are hierarchical, and most often represented as bifurcating ***trees***



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

Reading a phylogeny

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

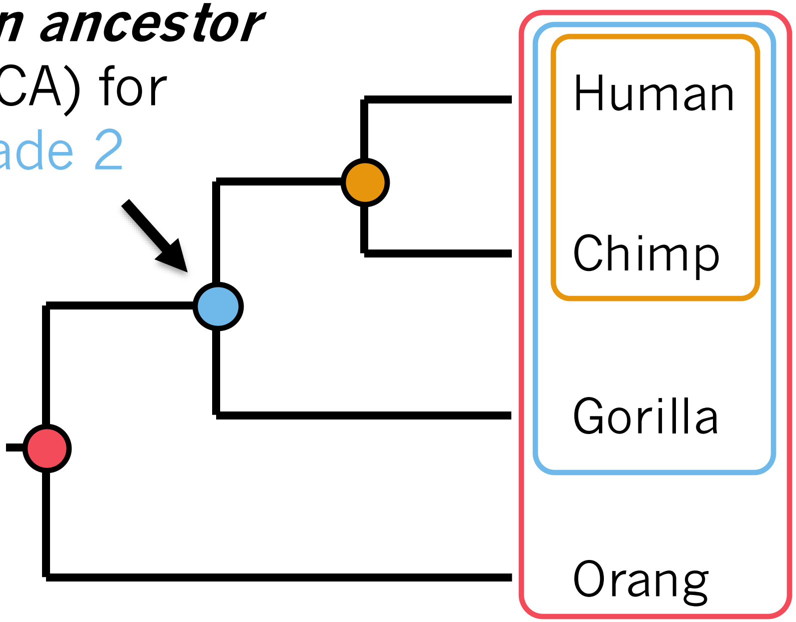


Human, Chimp and Gorilla are more closely related to each other than to Orang

Reading a phylogeny

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

most recent common ancestor (MRCA) for Clade 2



Clade 1: H+C

Clade 2: H+C+G

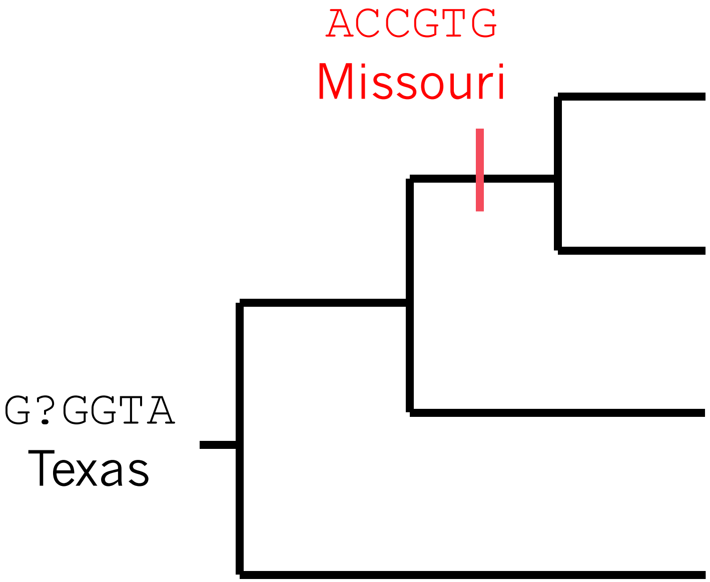
Clade 3: H+C+G+O

“Tree-thinking”

Taxon	Gene variant	Location
PT1	ACCGTG	Missouri
PT2	ACCGTG	Missouri
PT3	GAGGTA	Texas
PT4	GCGGTA	Texas

Four sequences,
but no historical context

“Tree-thinking”

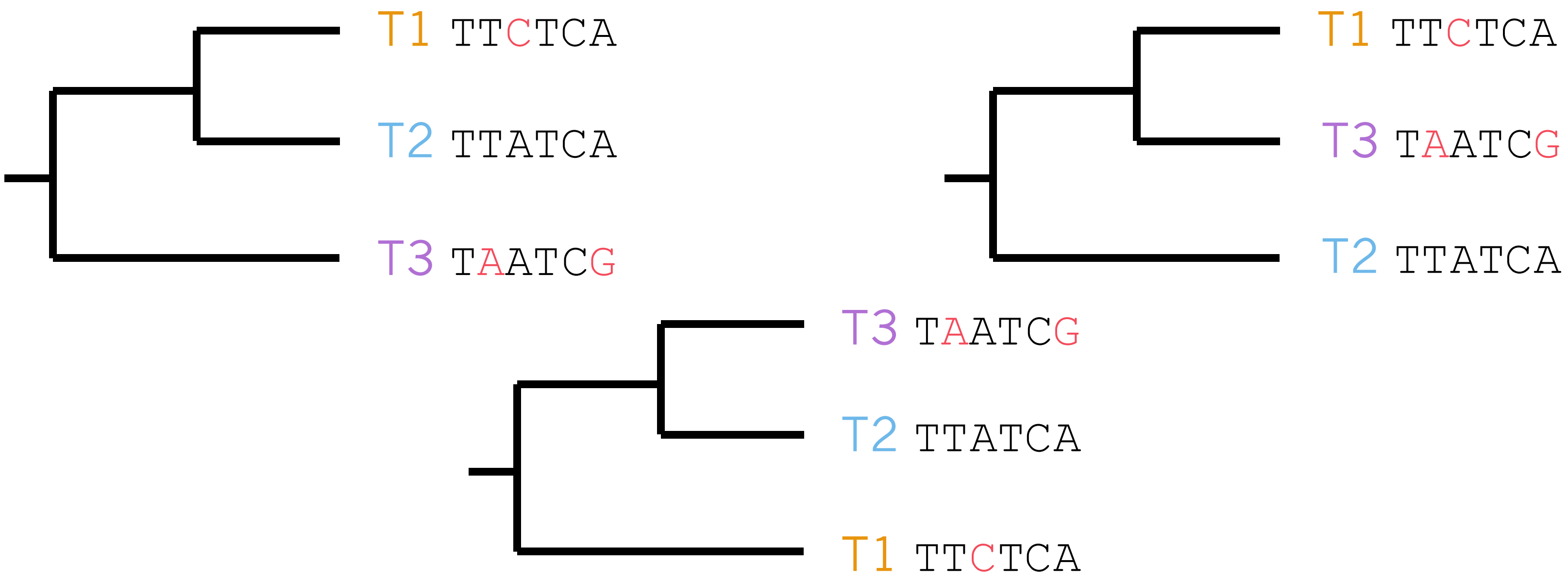


Taxon	Gene variant	Location
PT1	ACCGTG	Missouri
PT2	ACCGTG	Missouri
PT3	GAGGTA	Texas
PT4	GCGGTA	Texas

Phylogeny informs when and where variation arose, which can guide future 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*** the cost to estimate the phylogeny with the lowest cost for the provided data matrix

Phylogenetic method types

Most methods used to infer phylogenies compute scores based on

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

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

Tree-space is large

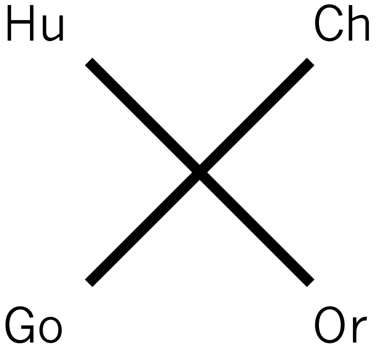
# taxa	# rooted trees
3	3
4	15
5	105
6	945
7	10395
8	135135
9	2027025
10	34459425

A major challenge: how to efficiently search
for trees with optimal scores?

Neighbor-joining

	Hu	Ch	Go	Or
Hu	0	1	3	5
Ch	1	0	3	5
Go	3	3	0	2
Or	5	5	2	0

divergence matrix
for sequence pairs



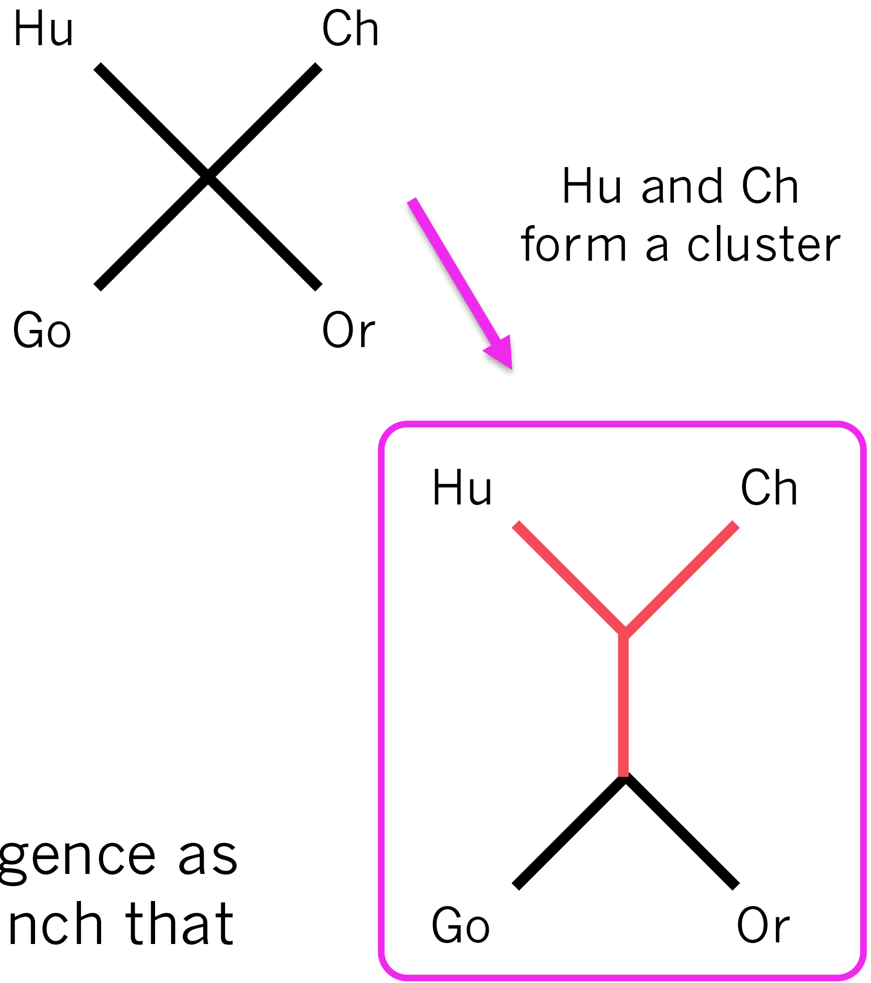
"minimum evolution"
method

Divergence matrix is
based on relative distances
among sequences

Neighbor-joining

	Hu	Ch	Go	Or
Hu	0	1	3	5
Ch	1	0	3	5
Go	3	3	0	2
Or	5	5	2	0

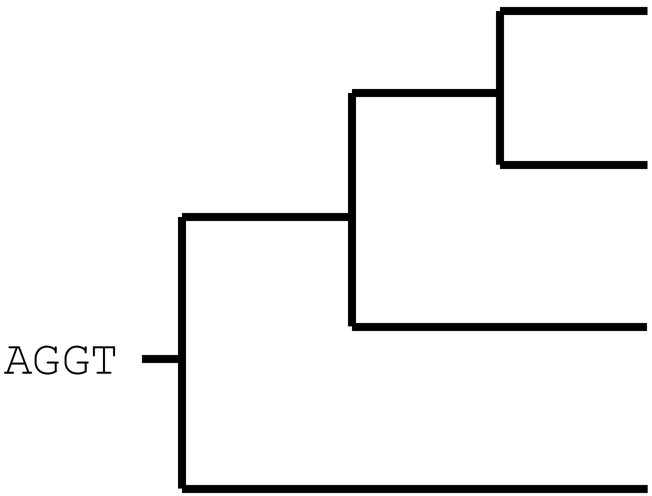
divergence matrix
for sequence pairs



Hu and Ch
form a cluster

Join sequences with least divergence as clade, assign length to new branch that ***minimizes*** other divergence scores

Parsimony



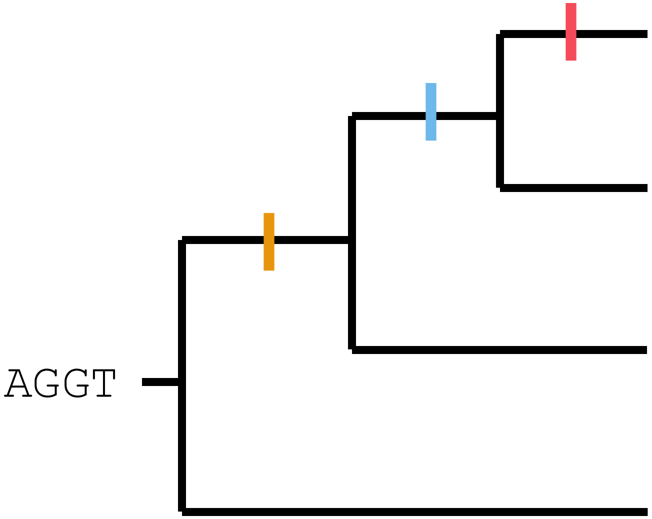
AGGT

Taxa	DNA
Human	AAAC
Chimp	AAAT
Gorilla	AAGT
Orang	AGGT

What phylogeny requires the fewest character change events?

Parsimony

3 events
(most parsimonious)

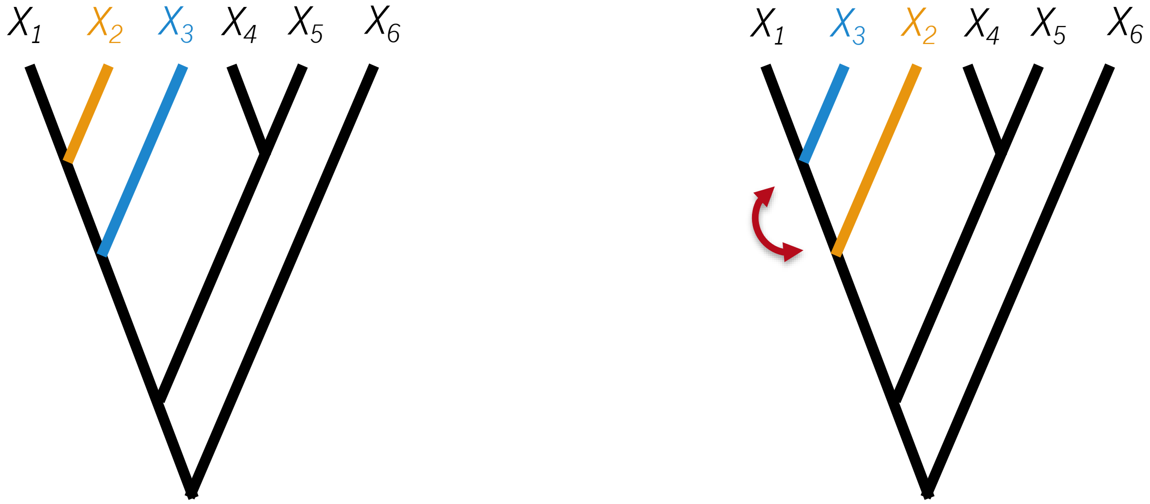


Taxa	DNA
Human	AAAC
Chimp	AAAT
Gorilla	AAGT
Orang	AGGT

What phylogeny requires the fewest character change events?

Exploring tree space

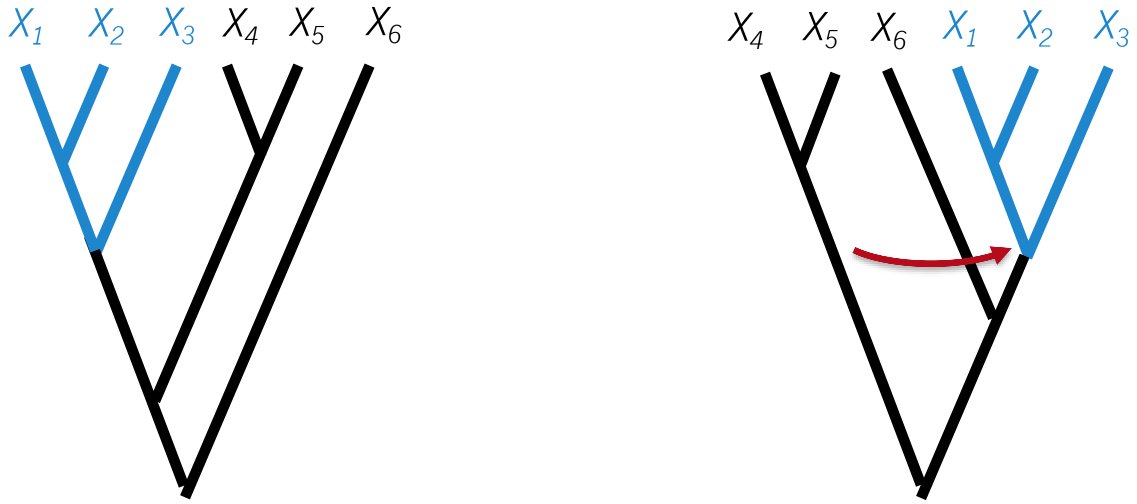
Define stochastic "moves" that modify topology,
prefer moves that improve tree score



Nearest neighbor interchange (NNI)

Exploring tree space

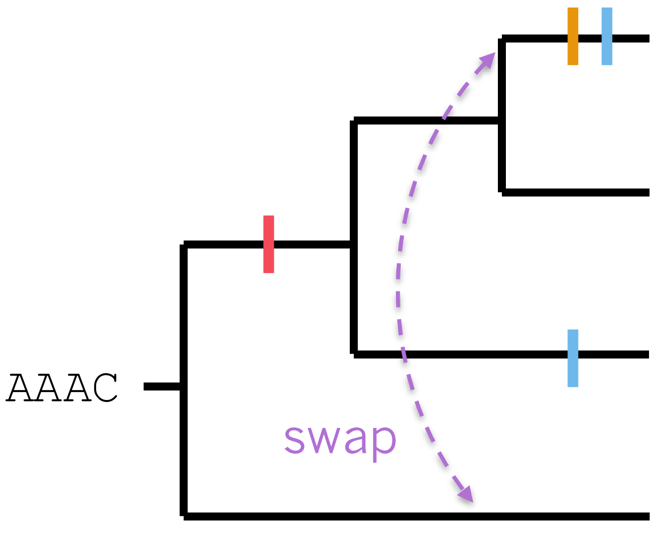
Define stochastic “moves” that modify topology, prefer moves that improve tree score



Subtree-prune-regraft (SPR)

Parsimony

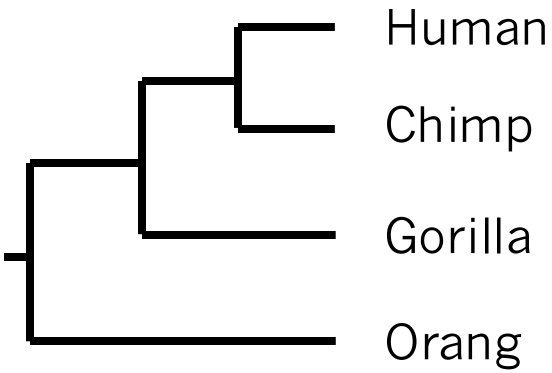
4 events
(less parsimonious)



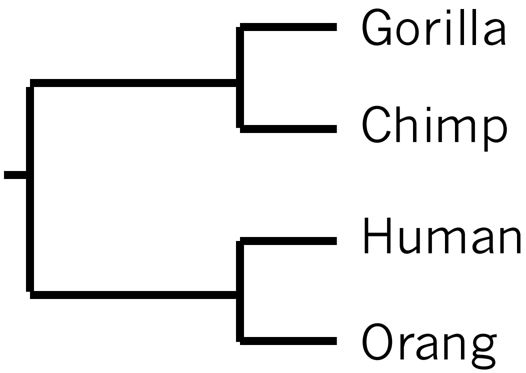
Taxa	DNA
Orang	AGGT
Chimp	AAAT
Gorilla	AAGT
Human	AAAC

What phylogeny requires the fewest character change events?

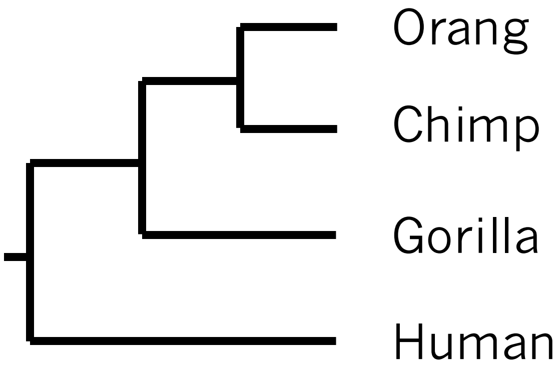
Parsimony



3 events



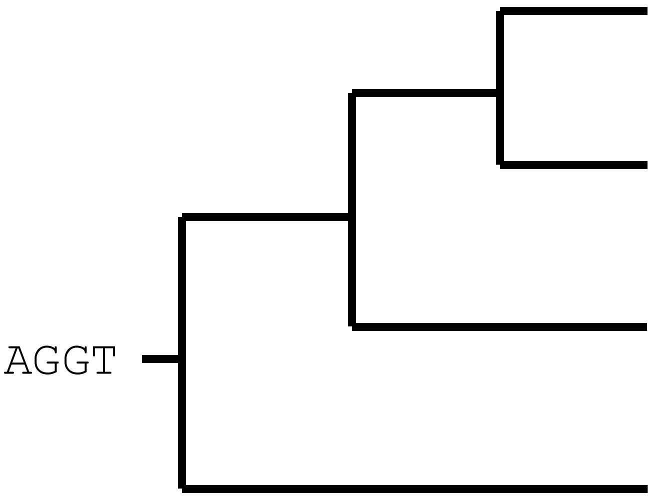
7 events



4 events

What phylogeny requires the fewest character change events?

Likelihood

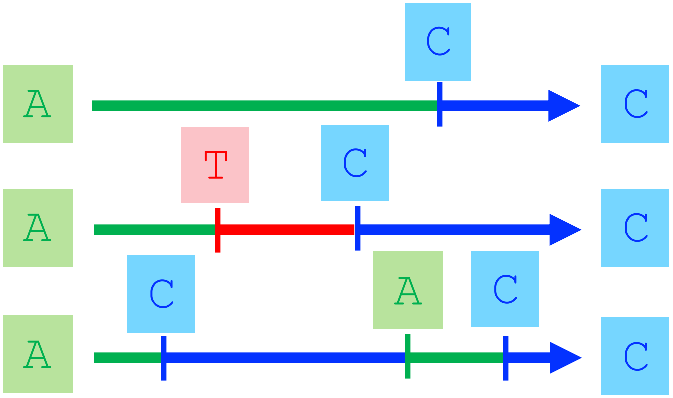


Taxa	DNA
Human	AAAC
Chimp	AAAT
Gorilla	AAGT
Orang	AGGT

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

Likelihood for single site and single branch

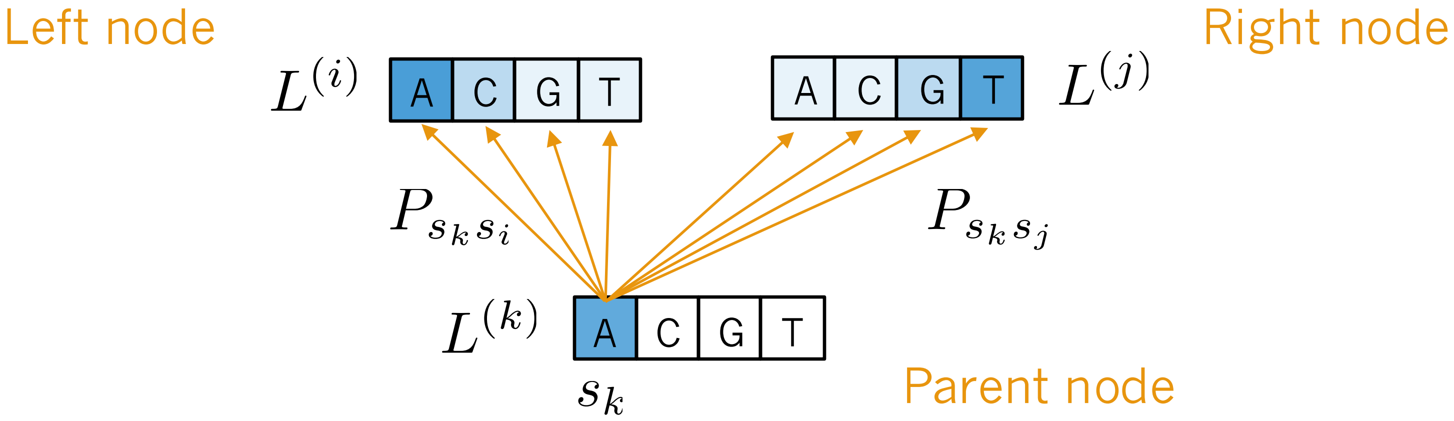
A single site is in one four discrete states: A, C, G, T



(possible evolutionary histories)

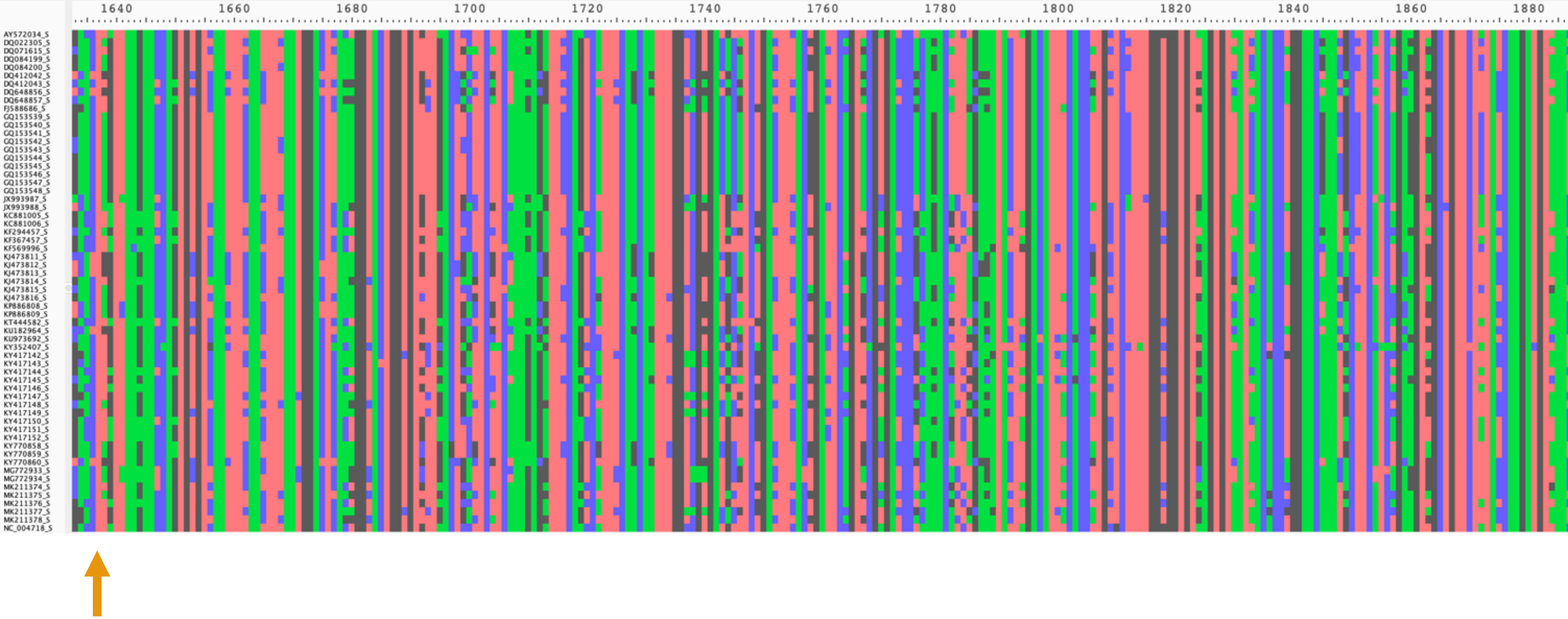
What is the probability to start in state A and end in state C along branch of length *t*?

Phylogenetic marginalization



Using the ***pruning algorithm*** to move “backwards in time”; compute the ***partial likelihood*** for each ancestral node state based on its descendants’ states

Phylogenetic likelihood

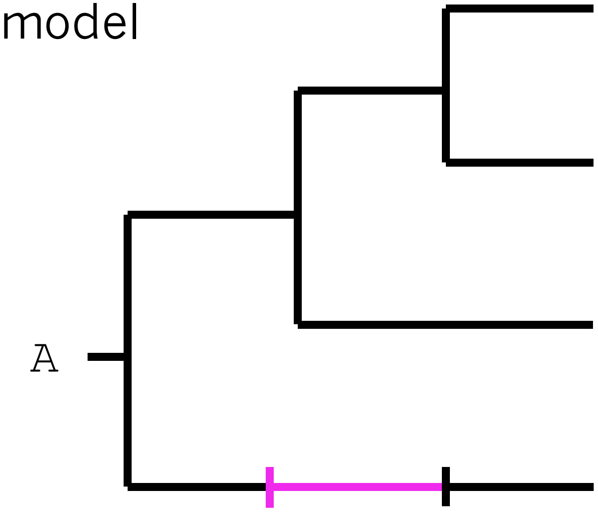


single
site

Compute the total phylogenetic likelihood as the product of likelihoods across all sites

Likelihood

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

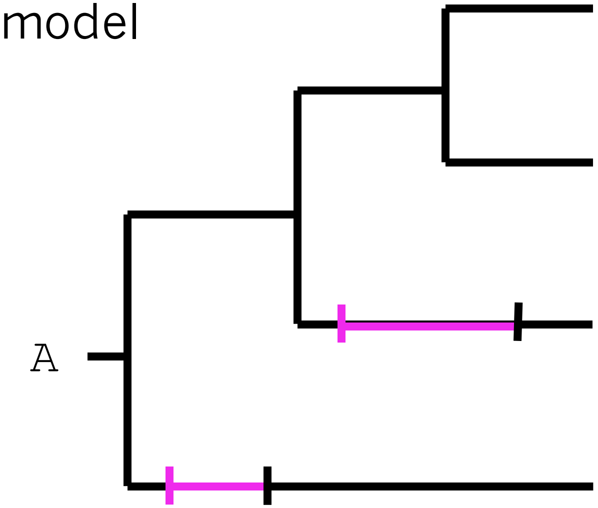


Taxa	DNA
Human	AAAC
Chimp	AAAT
Gorilla	AAGT
Orang	AGGT

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

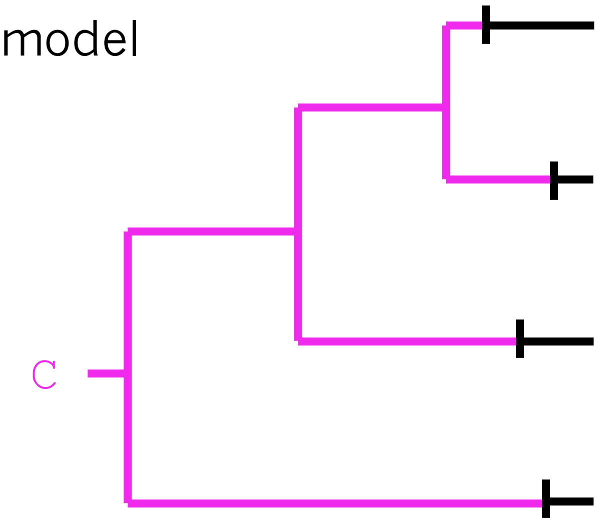


Taxa	DNA
Human	AAAC
Chimp	AAAT
Gorilla	AAGT
Orang	AGGT

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

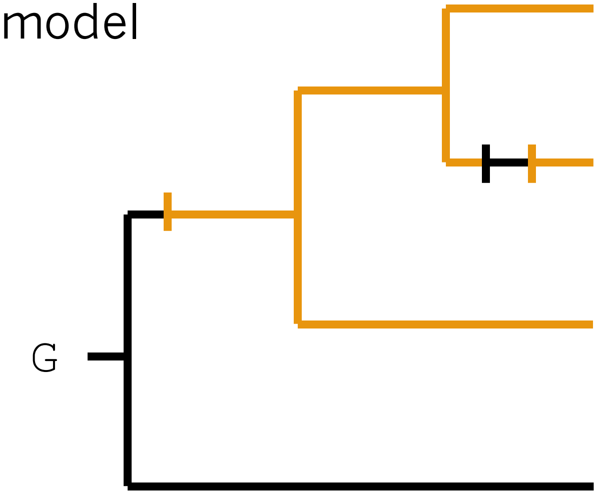


Taxa	DNA
Human	AAAC
Chimp	AAAT
Gorilla	AAGT
Orang	AGGT

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

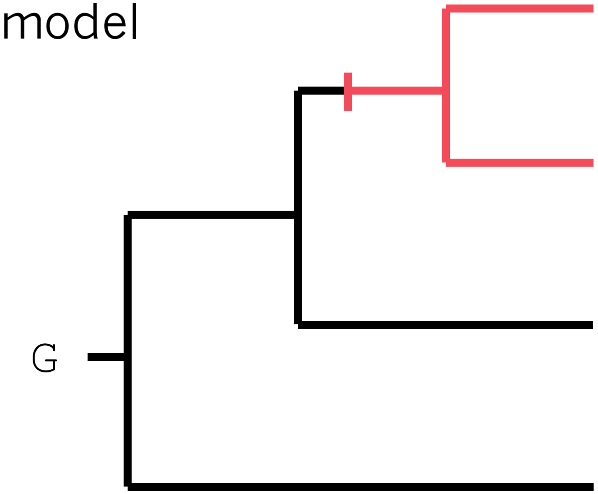


Taxa	DNA
Human	AAAC
Chimp	AAAT
Gorilla	AAGT
Orang	AGGT

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

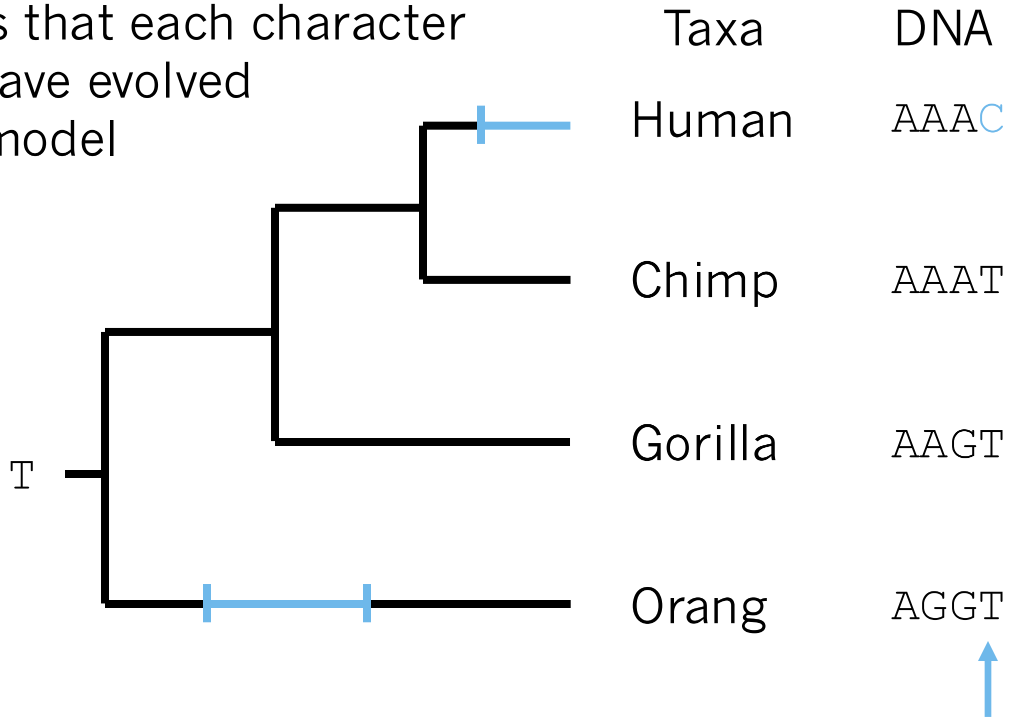


Taxa	DNA
Human	AAAC
Chimp	AAAT
Gorilla	AAGT
Orang	AGGT

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

Likelihood

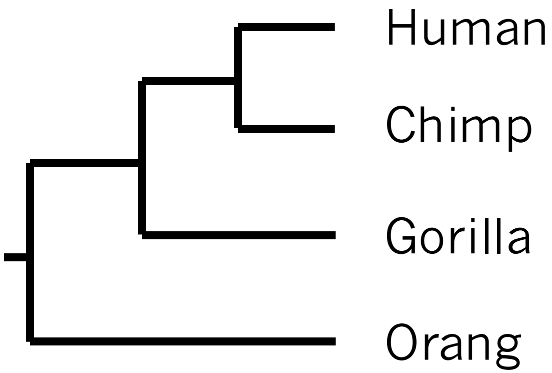
Compute probability for all ways that each character could have evolved under model



Taxa	DNA
Human	AAAC
Chimp	AAAT
Gorilla	AAGT
Orang	AGGT

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

Likelihood



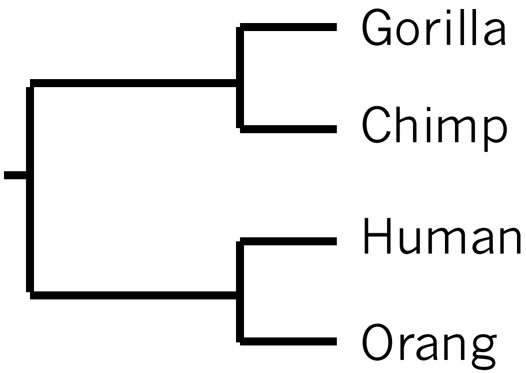
Human

Chimp

Gorilla

Orang

log-likelihood = -32.14



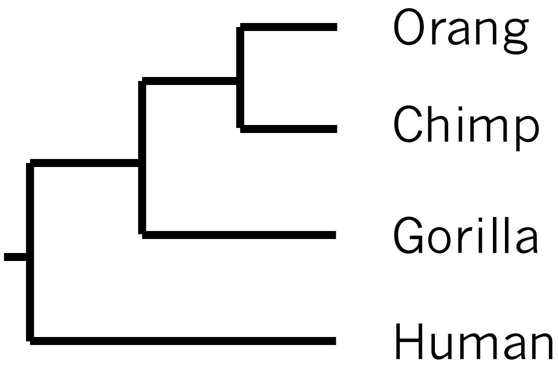
Gorilla

Chimp

Human

Orang

log-likelihood = -42.77



Orang

Chimp

Gorilla

Human

log-likelihood = -39.08

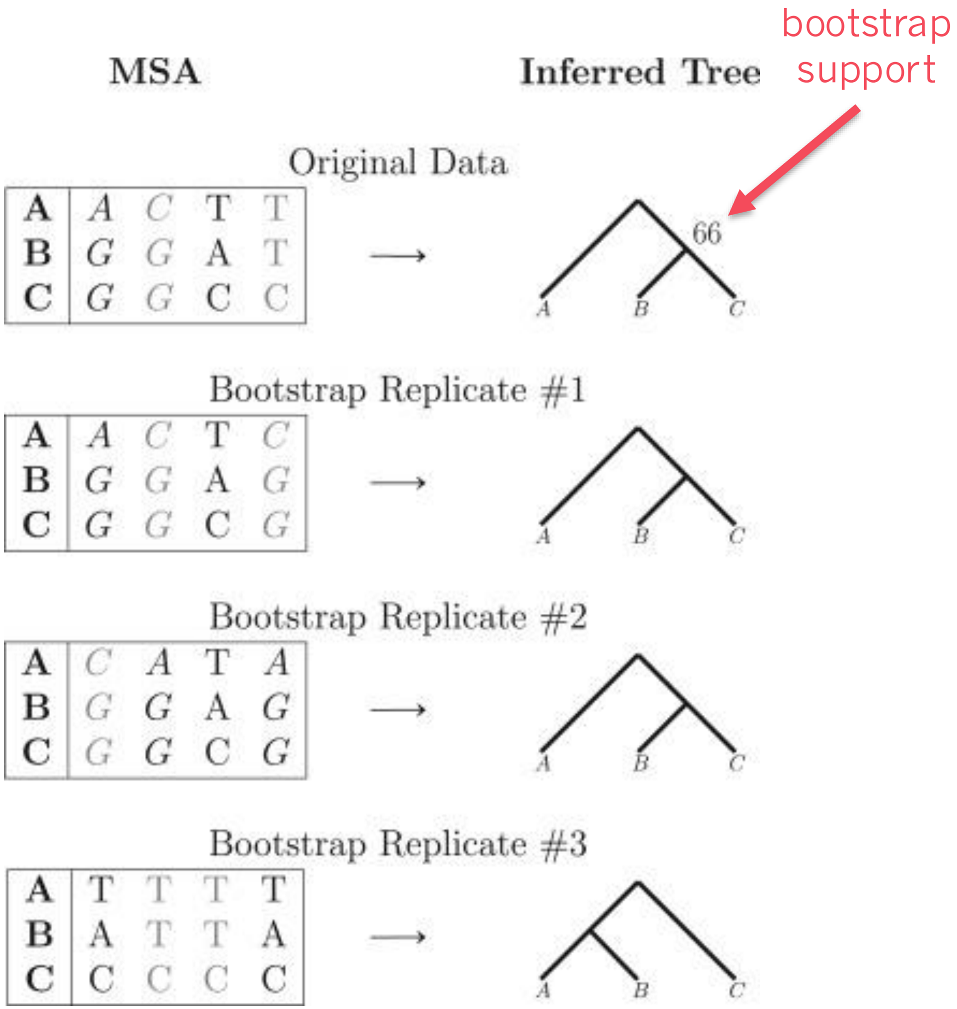
What phylogeny and model of evolution is *most likely* to generate the character data?

Clade support

Clade support measures our statistical confidence for each clade

Bootstrap support:

1. estimate a tree from the original dataset
2. simulate *K* replicate datasets by resampling sites *with replacement* from the original dataset
3. estimate a new tree for each of the *K* replicates
4. report the frequency (*k*/*K* %) for each clade in the original tree



bootstrap support

Original Data

A	A	C	T	T
B	G	G	A	T
C	G	G	C	C

Bootstrap Replicate #1

A	A	C	T	C
B	G	G	A	G
C	G	G	C	G

Bootstrap Replicate #2

A	C	A	T	A
B	G	G	A	G
C	G	G	C	G

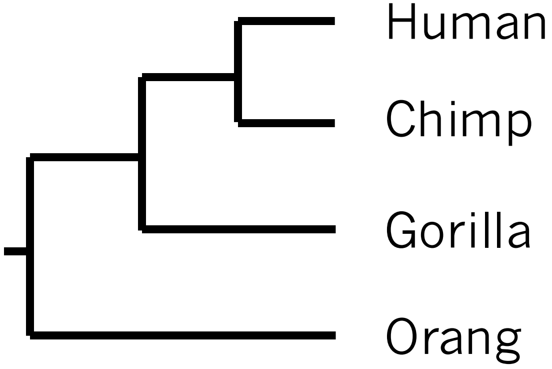
Bootstrap Replicate #3

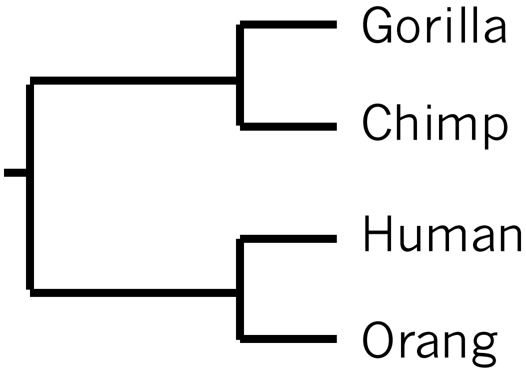
A	T	T	T	T
B	A	T	T	A
C	C	C	C	C

Method comparison

Method	Pros	Cons
Neighbor-joining	Extremely fast Scalable	Does not use evolutionary events to infer tree
Parsimony	Intuitive Fairly fast	Assumes change is rare; Event costs are arbitrary
Likelihood	Most accurate Most realistic Can simulate data	Slower Complex theory + algorithms

Newick strings





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

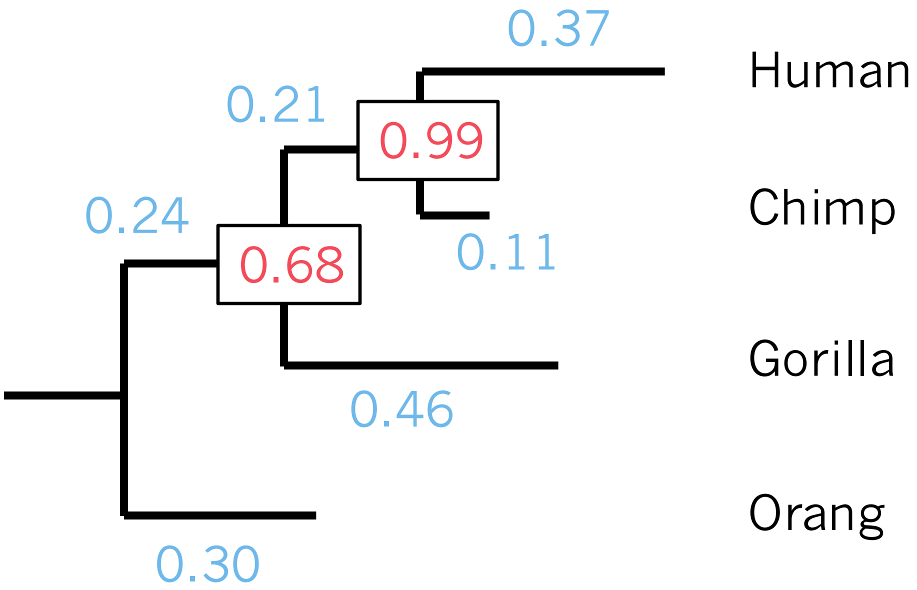
((Gorilla,Chimp),(Human,Orang));

Taxa in parentheses define clades;
commas define divergence events

Newick strings

Branch lengths measure molecular distances in expected # substitutions per site

Clade support measures reliability of clade in a tree estimate



(((Human:0.37,Chimp:0.11)0.99:0.21
,
Gorilla:0.46)0.68:0.24,Orang:0.30);

Overview for Lab 12