

Lecture 11

Molecular phylogenetics



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



Lecture 11 outline

Last time: regex

This time: phylogenetics

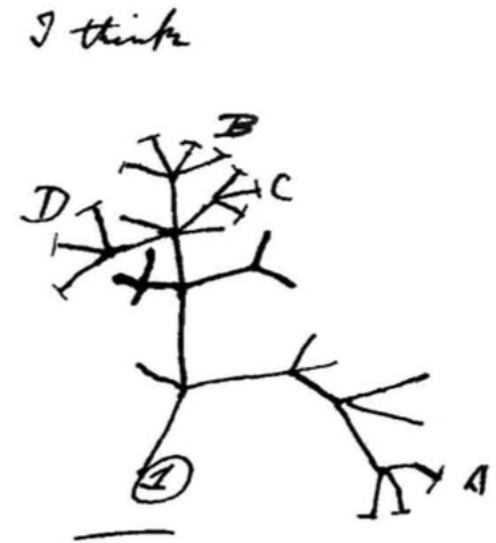
- interpreting trees
- tree-thinking
- inferring trees
- inference methods

Phylogenetics

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

Phylogenies are useful for

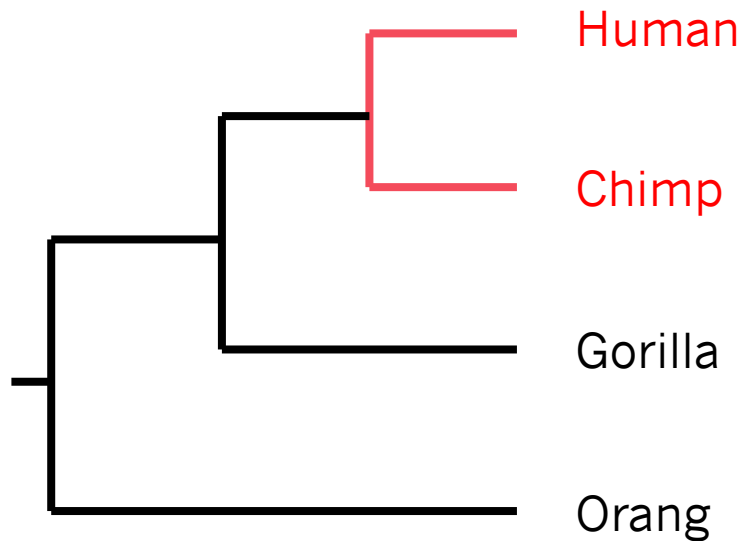
- gene annotation
- tracking viral spread
- identifying zoonosis
- reconstructing tumorigenesis
- conservation biology assays
- inferring species relationships



phylogeny sketch
by Darwin

Reading a phylogeny

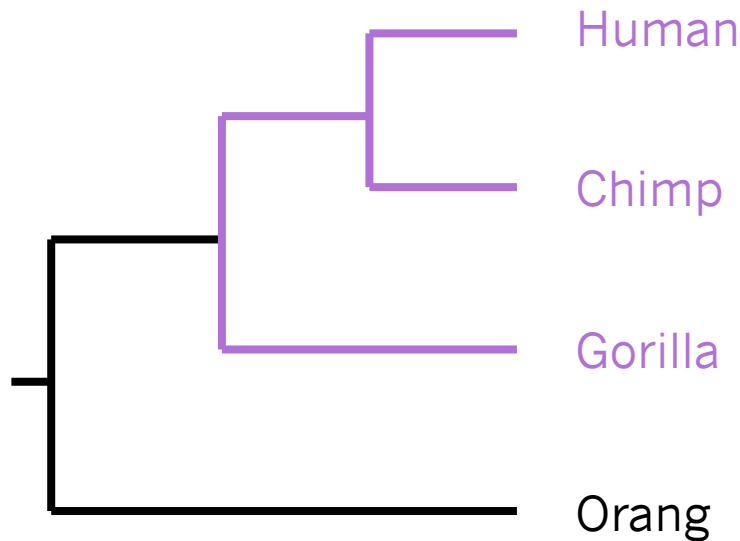
Phylogenetic relationships are hierarchical,
and most often represented as bifurcating **trees**



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**

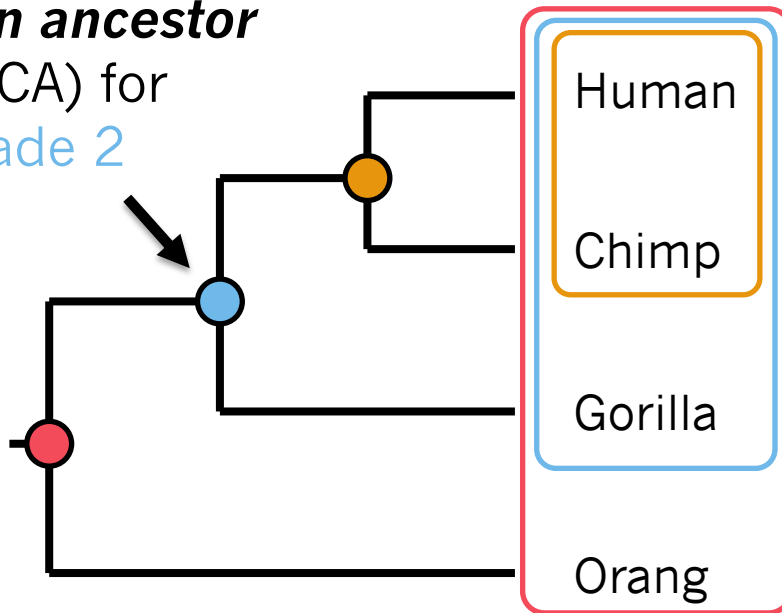


Human, Chimp and Gorilla are more closely related to each other than to Orang

Reading a phylogeny

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

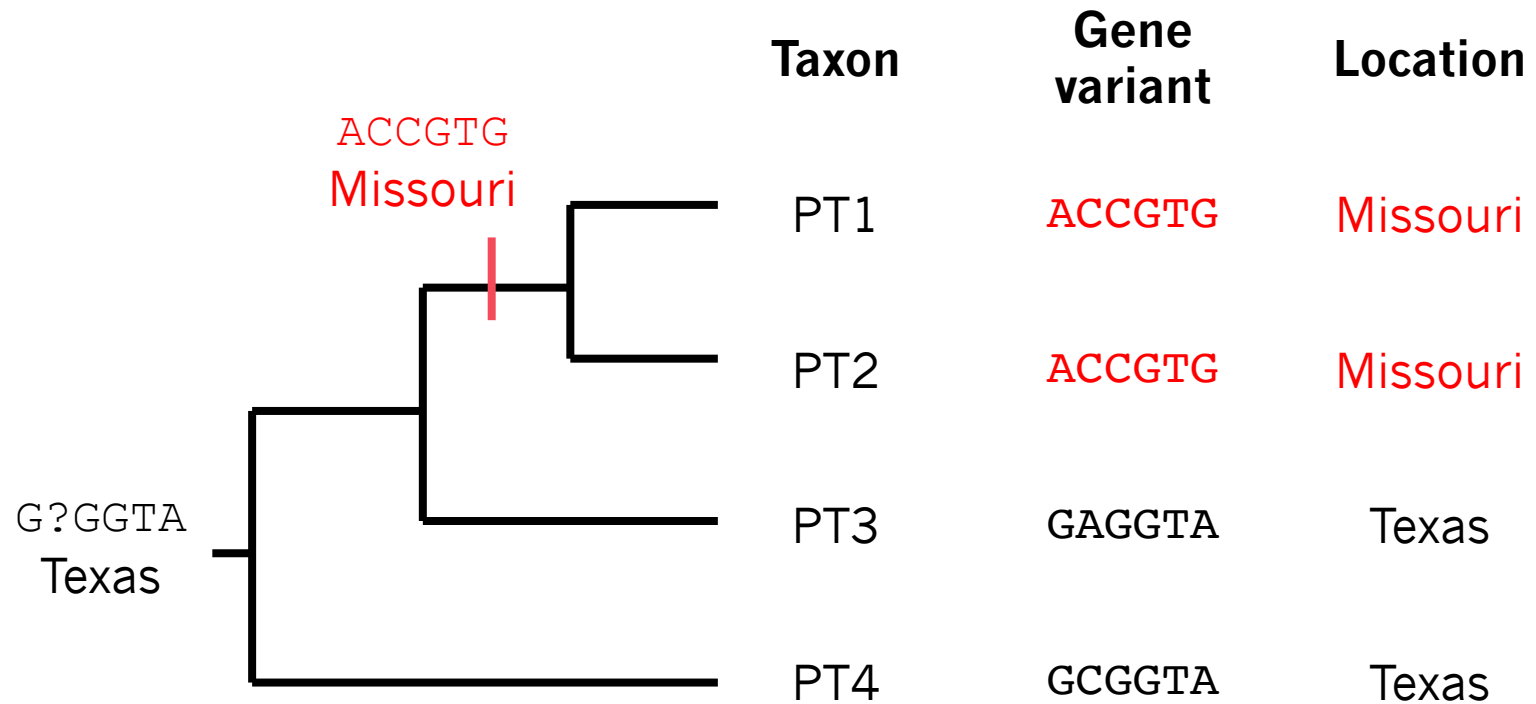
Clade 3: H+C+G+O

“Tree-thinking”

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

Four sequences,
but no historical context

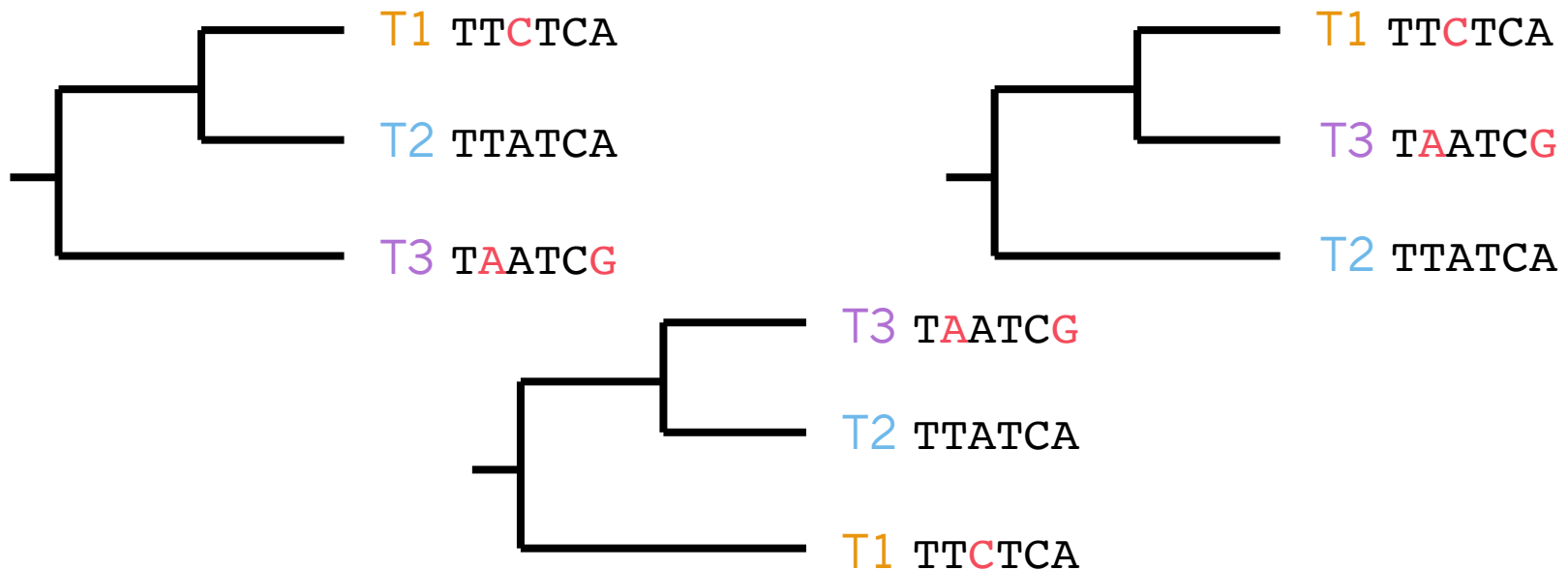
“Tree-thinking”



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

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?

Phylogenetic method types

Most methods used to infer phylogenies compute scores based on

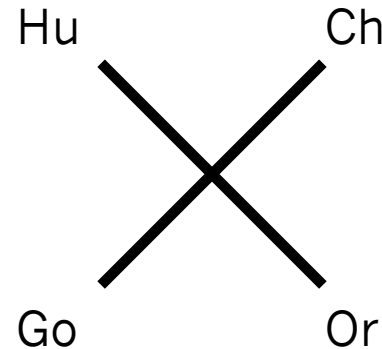
1. pattern distances (e.g. ***neighbor joining***)
2. event counting (***parsimony***)
3. event probabilities (***likelihood***)

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

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

distance matrix
for sequence pairs



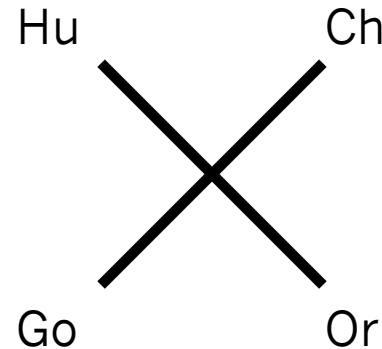
Select pairs of taxa with short
sequence distances, and join
them as neighbors

Neighbor-joining

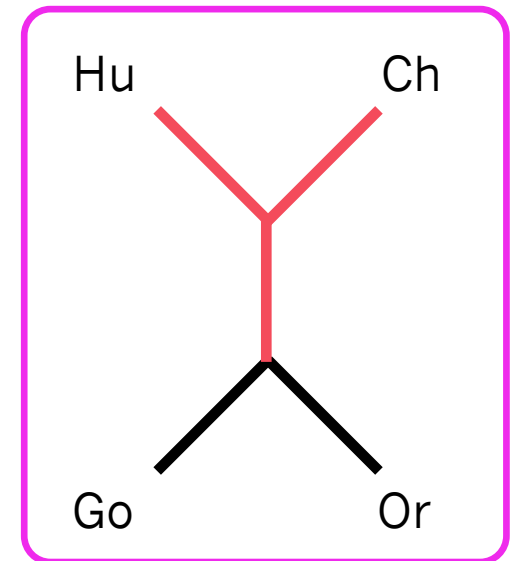
	Hu	Ch	Go	Or
Hu	0	1	3	5
Ch	1	0	3	5
Go	3	3	0	2
Or	5	5	2	0

distance matrix
for sequence pairs

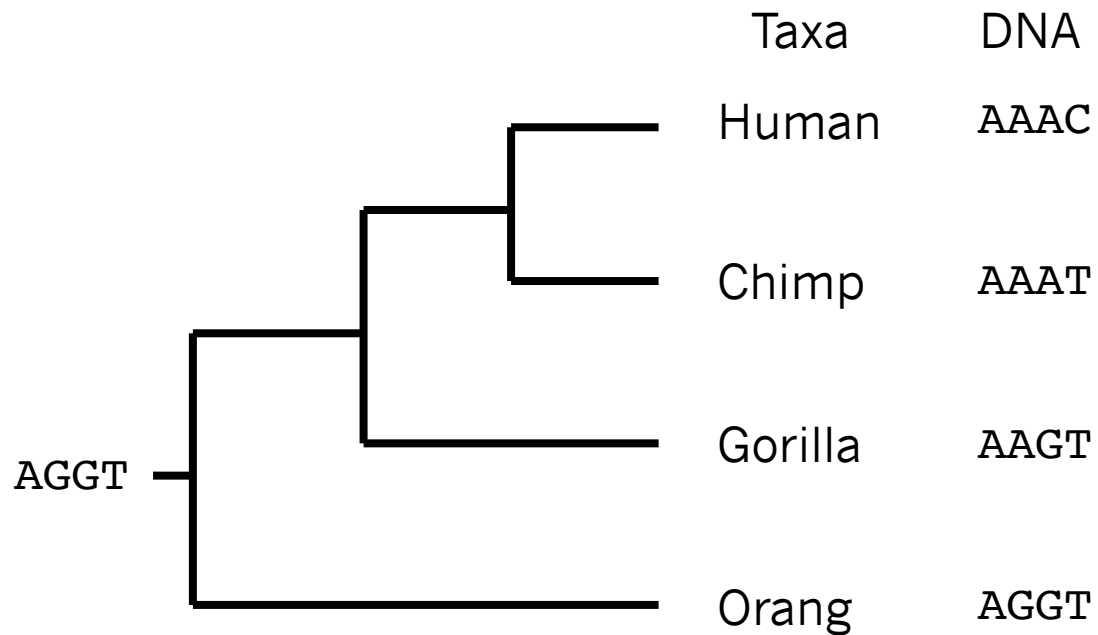
Select pairs of taxa with short
sequence distances, and join
them as neighbors



Hu and Ch
form a cluster



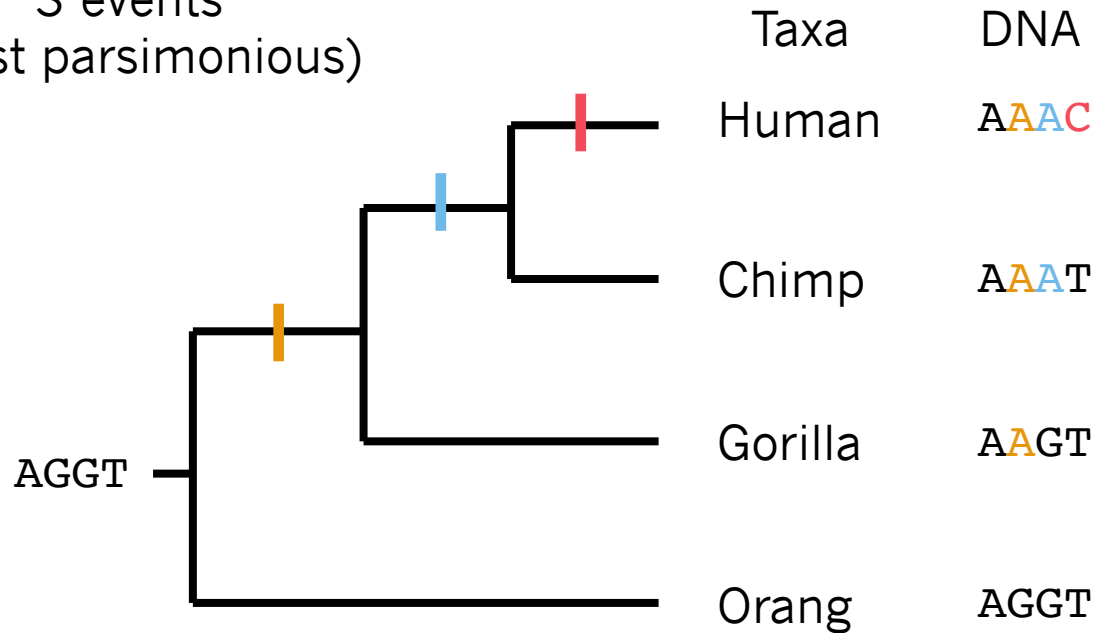
Parsimony



What phylogeny requires the fewest character change events?

Parsimony

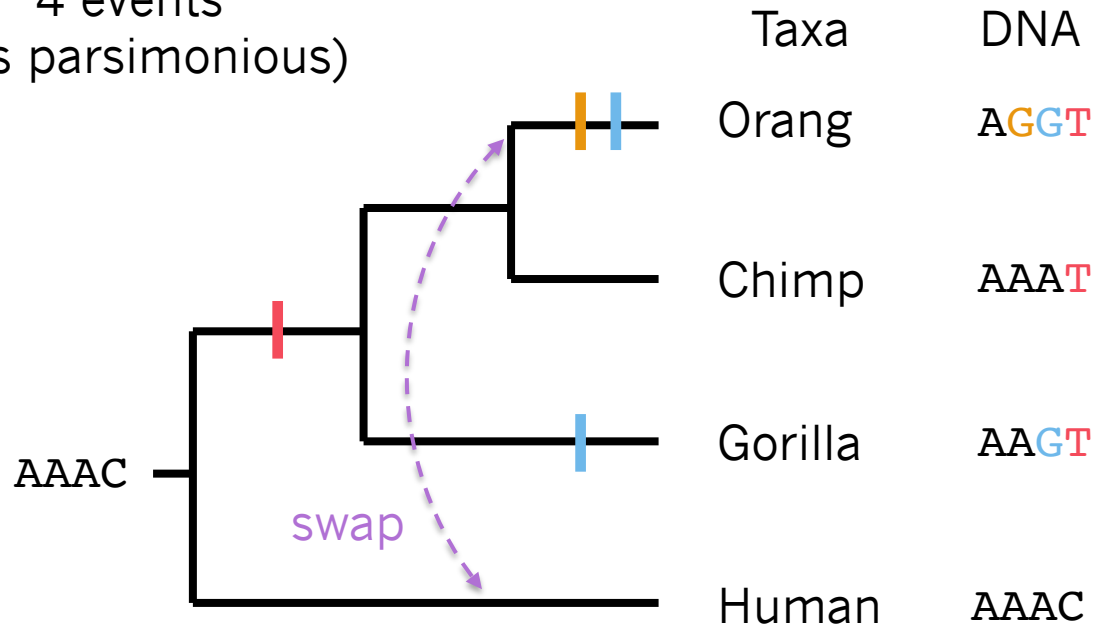
3 events
(most parsimonious)



What phylogeny requires the fewest
character change events?

Parsimony

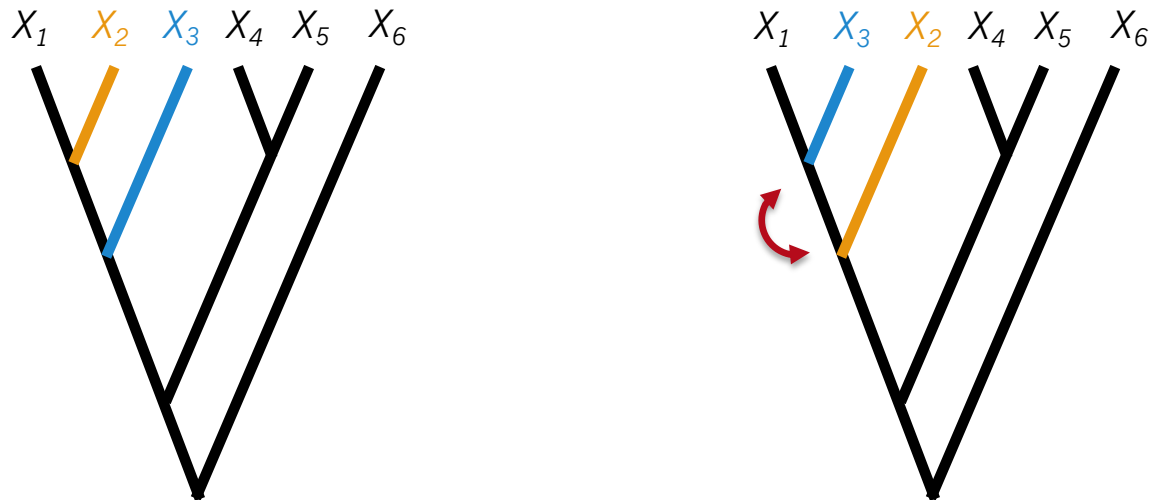
4 events
(less parsimonious)



What phylogeny requires the fewest
character change events?

Exploring tree space

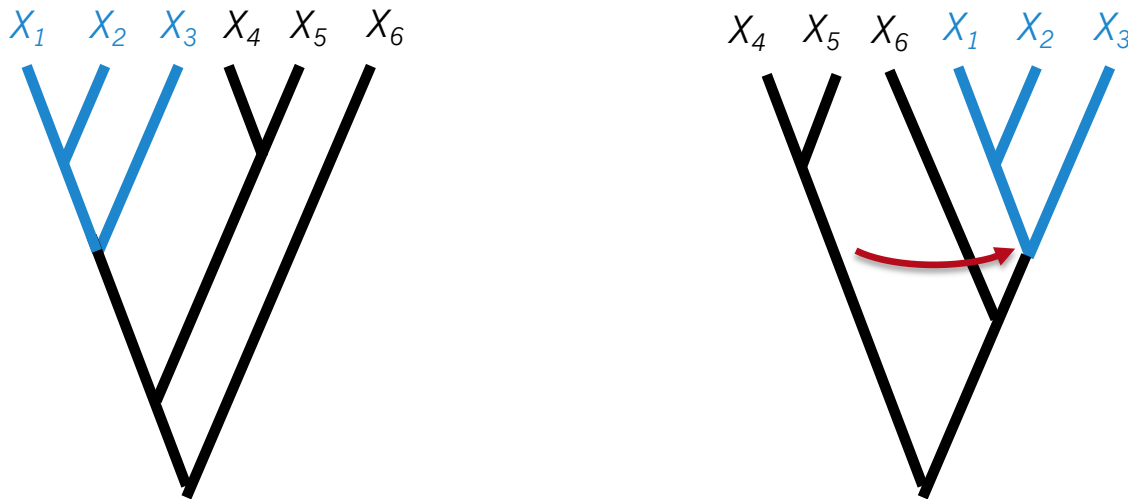
Define stochastic “moves” that modify topology,
prefer moves that improve tree score



Nearest neighbor interchange (NNI)

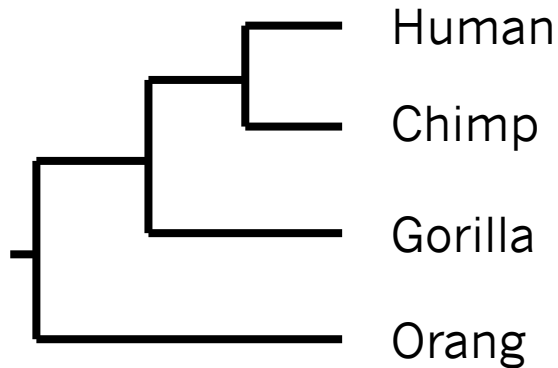
Exploring tree space

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

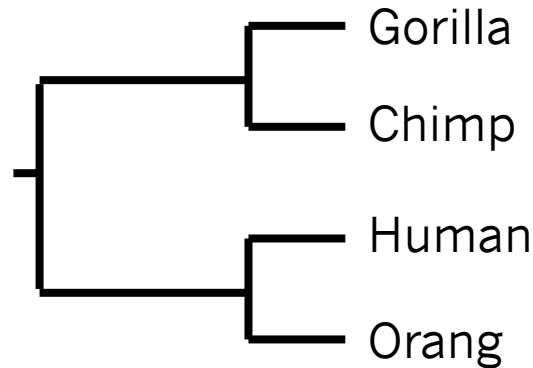


Subtree-prune-regraft (SPR)

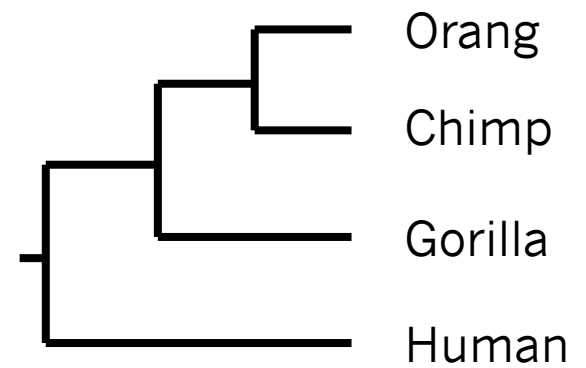
Parsimony



3 events



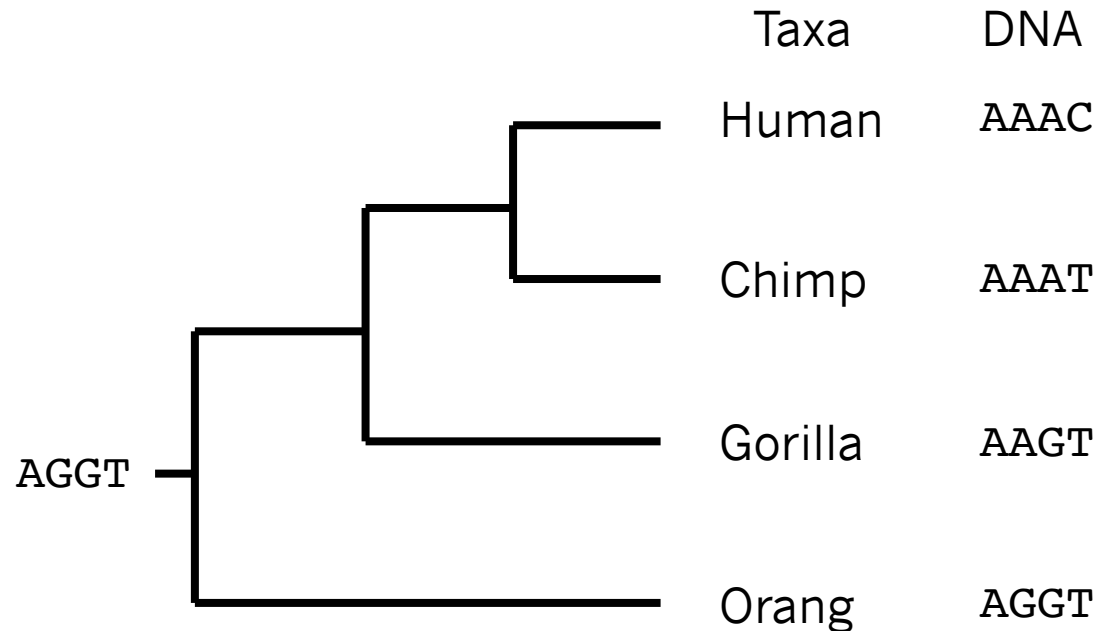
7 events



4 events

What phylogeny requires the fewest character change events?

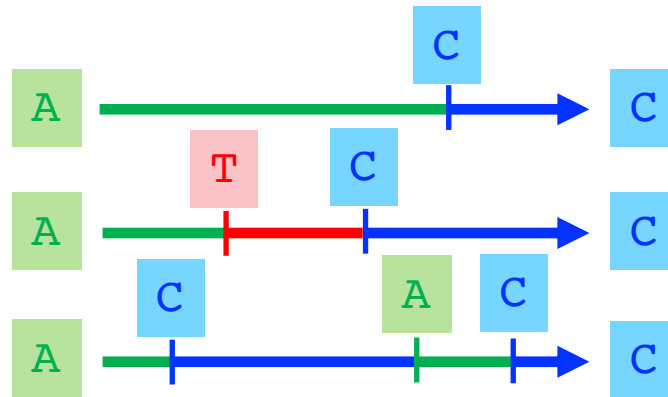
Likelihood



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

DNA evolution on branch

A single DNA site for branch k at time t can be in one four discrete states: A, C, G, T

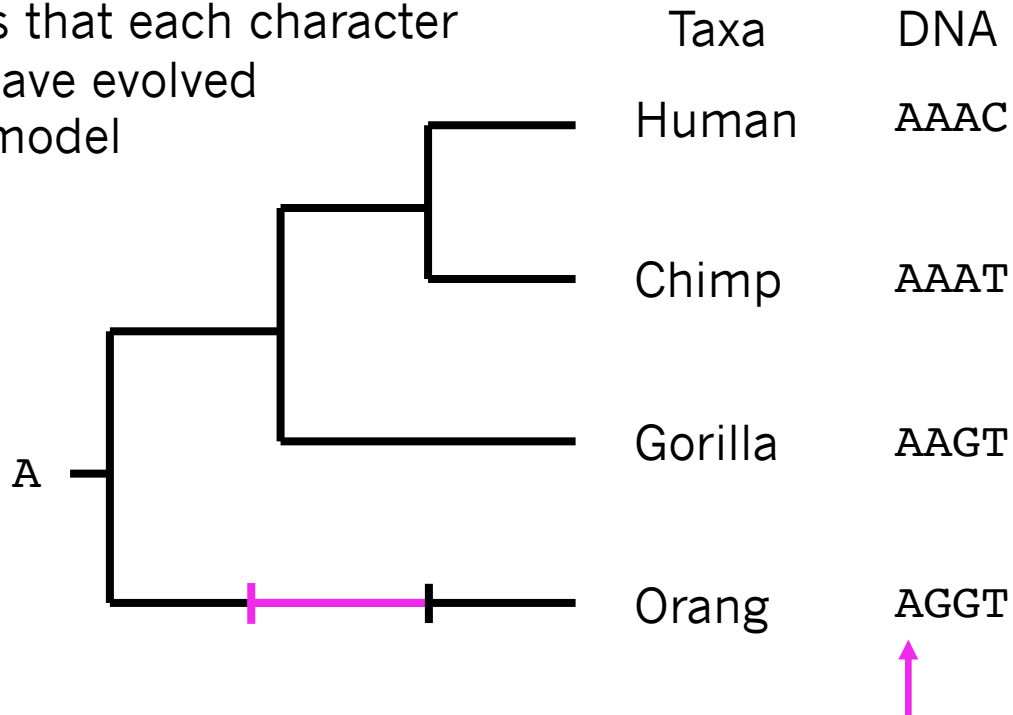


(possible evolutionary histories)

What is the probability that a DNA site in state i will end in state j after time t for branch k ?

Likelihood

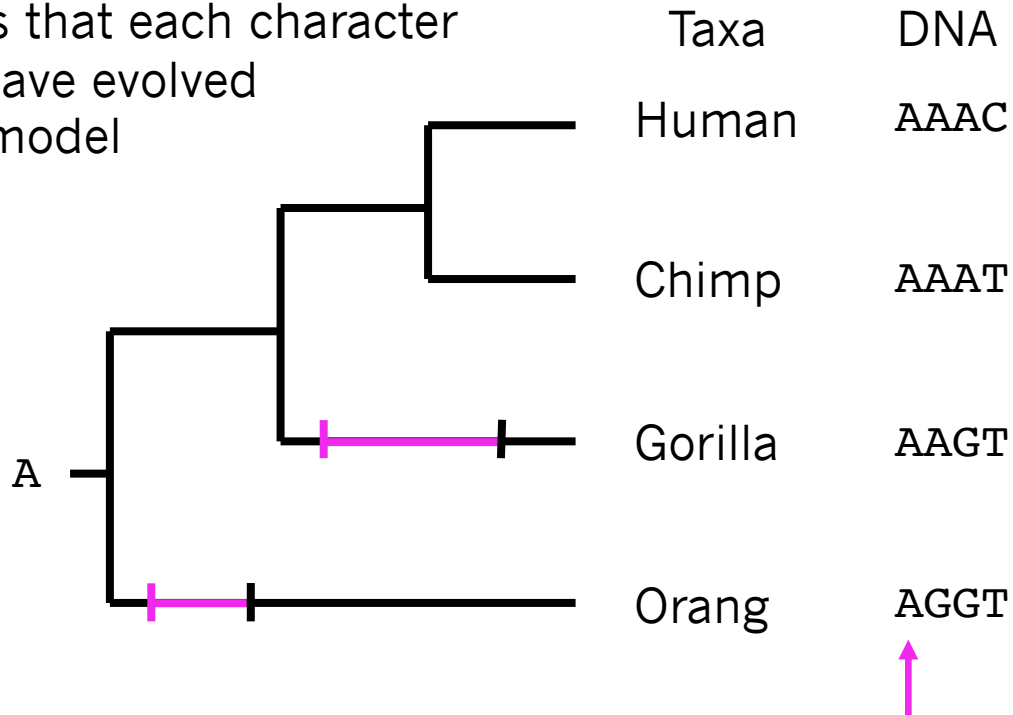
Compute probability for
all ways that each character
could have evolved
under model



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

Likelihood

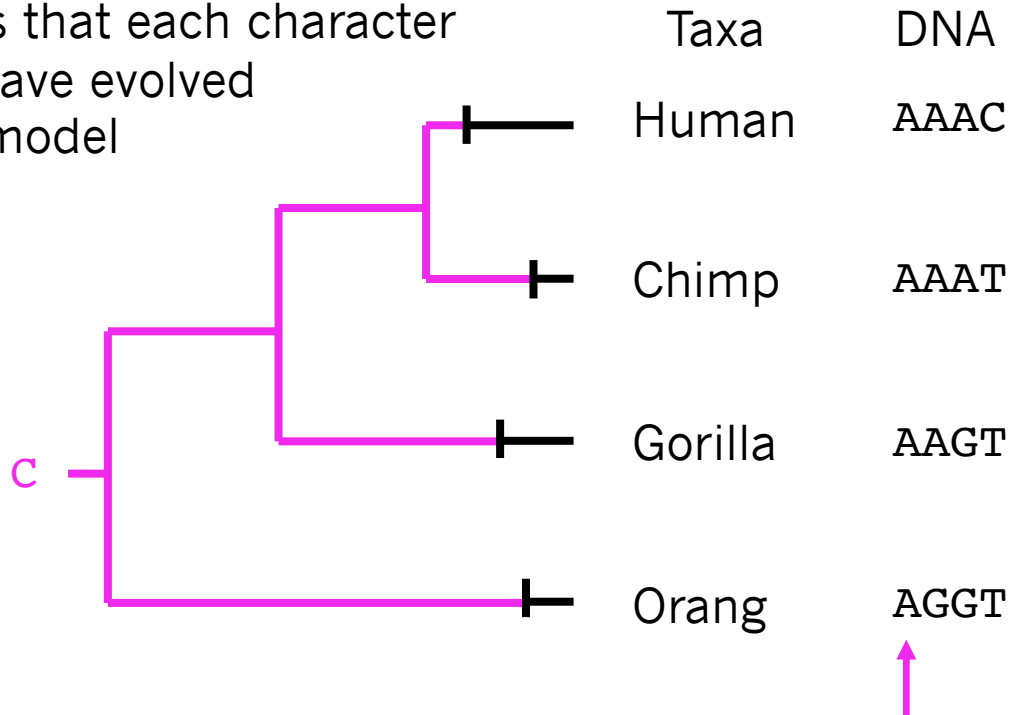
Compute probability for
all ways that each character
could have evolved
under model



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

Likelihood

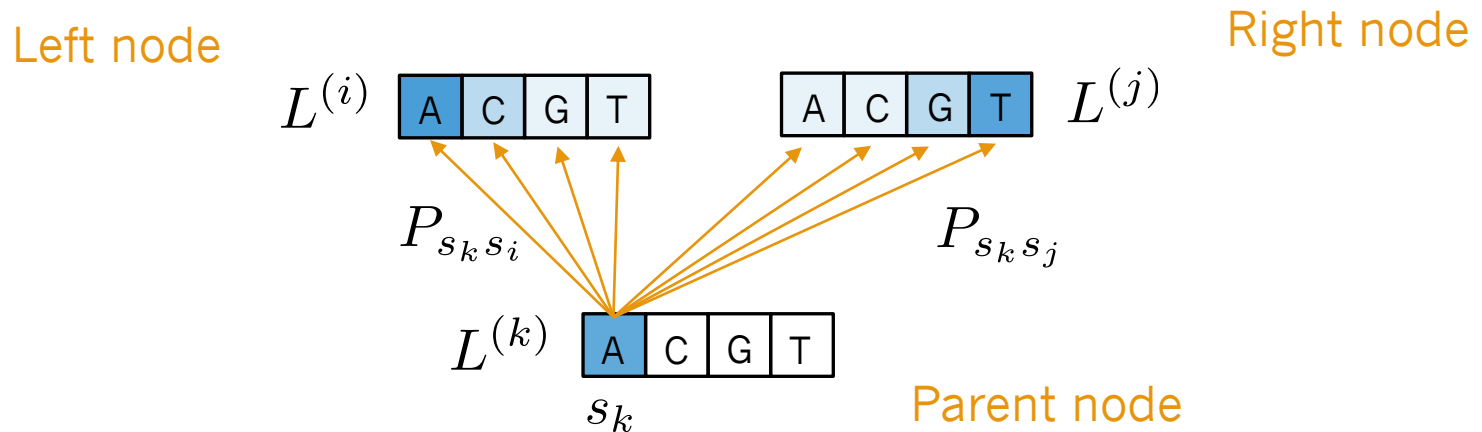
Compute probability for
all ways that each character
could have evolved
under model



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

Phylogenetic marginalization

Compute partial likelihood ($L^{(k)}$) for each start state (s_k) against all end states (s_i, s_j)

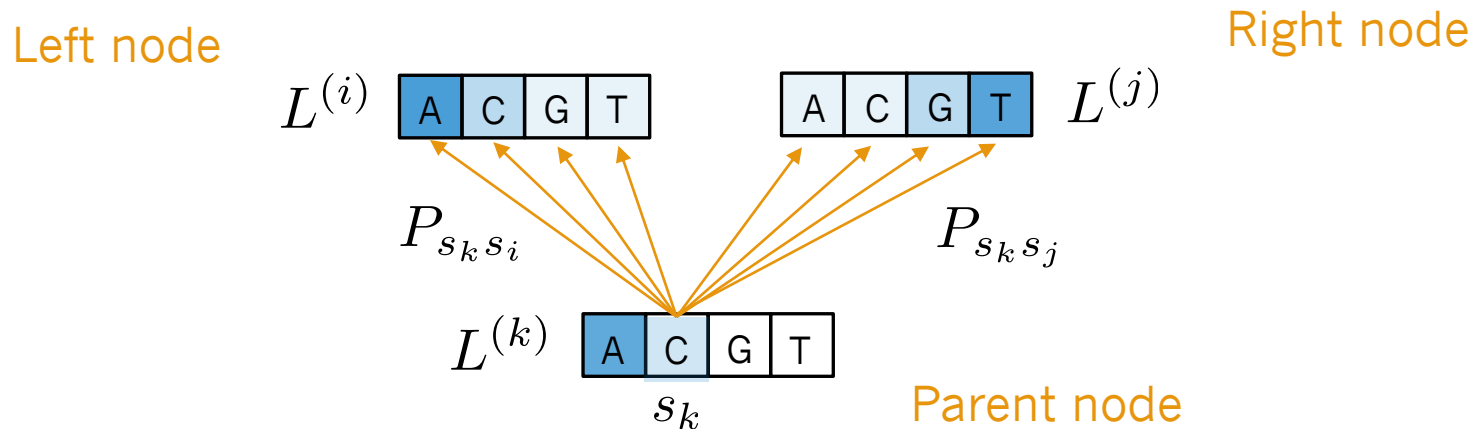


Example for partial likelihood $L^{(k)}$ at node X_k for state $s_k =$ A

$$L_A^{(k)} = \left(P_{AA}(t_i) L_A^{(i)} + P_{AC}(t_i) L_C^{(i)} + P_{AG}(t_i) L_G^{(i)} + P_{AT}(t_i) L_T^{(i)} \right) \\ \times \left(P_{AA}(t_j) L_A^{(j)} + P_{AC}(t_j) L_C^{(j)} + P_{AG}(t_j) L_G^{(j)} + P_{AT}(t_j) L_T^{(j)} \right)$$

Phylogenetic marginalization

Compute partial likelihood ($L^{(k)}$) for each start state (s_k) against all end states (s_i, s_j)

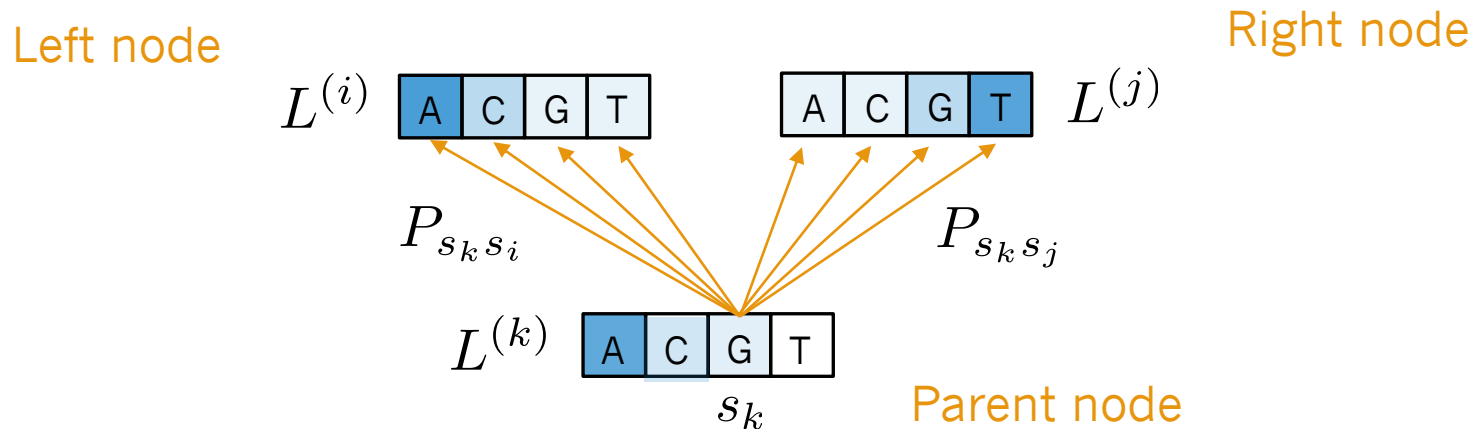


Example for partial likelihood $L^{(k)}$ at node X_k for state $s_k =$ C

$$L_C^{(k)} = \left(P_{CA}(t_i)L_A^{(i)} + P_{CC}(t_i)L_C^{(i)} + P_{CG}(t_i)L_G^{(i)} + P_{CT}(t_i)L_T^{(i)} \right) \\ \times \left(P_{CA}(t_j)L_A^{(j)} + P_{CC}(t_j)L_C^{(j)} + P_{CG}(t_j)L_G^{(j)} + P_{CT}(t_j)L_T^{(j)} \right)$$

Phylogenetic marginalization

Compute partial likelihood ($L^{(k)}$) for each start state (s_k) against all end states (s_i, s_j)

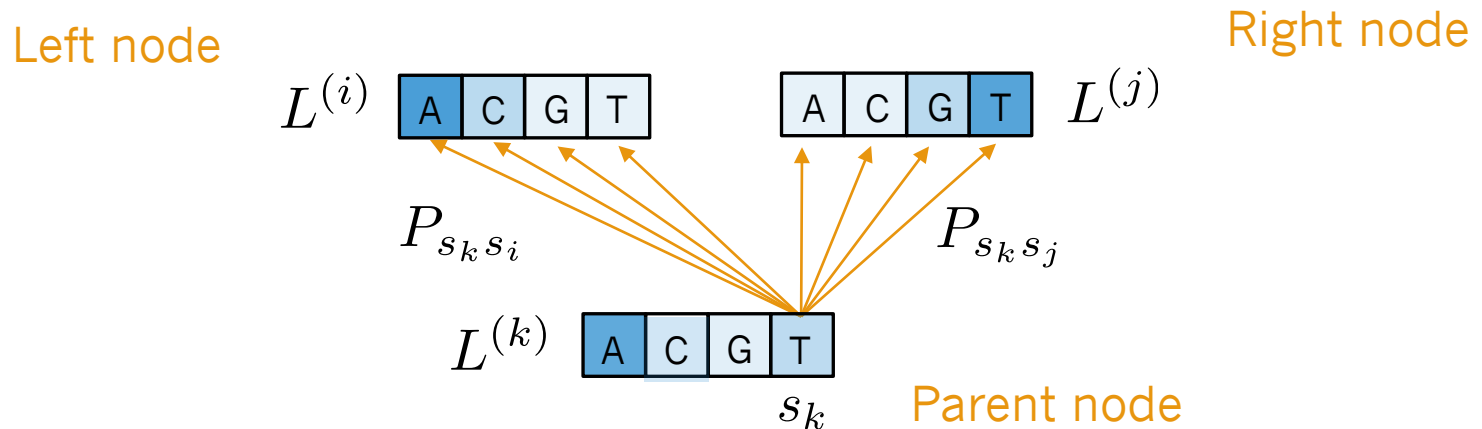


Example for partial likelihood $L^{(k)}$ at node X_k for state $s_k = \text{G}$

$$L_G^{(k)} = \left(P_{GA}(t_i) L_A^{(i)} + P_{GC}(t_i) L_C^{(i)} + P_{GG}(t_i) L_G^{(i)} + P_{GT}(t_i) L_T^{(i)} \right) \\ \times \left(P_{GA}(t_j) L_A^{(j)} + P_{GC}(t_j) L_C^{(j)} + P_{GG}(t_j) L_G^{(j)} + P_{GT}(t_j) L_T^{(j)} \right)$$

Phylogenetic marginalization

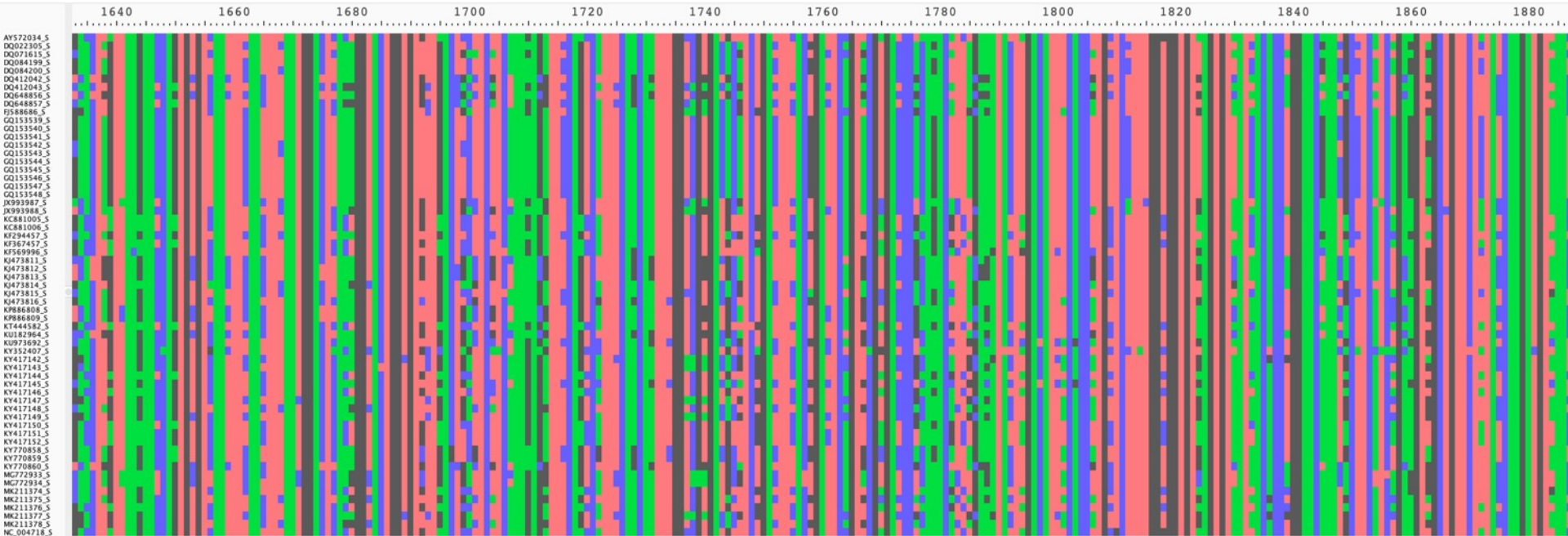
Compute partial likelihood ($L^{(k)}$) for each start state (s_k) against all end states (s_i, s_j)



Example for partial likelihood $L^{(k)}$ at node X_k for state $s_k =$ T

$$L_T^{(k)} = \left(P_{TA}(t_i)L_A^{(i)} + P_{TC}(t_i)L_C^{(i)} + P_{TG}(t_i)L_G^{(i)} + P_{TT}(t_i)L_T^{(i)} \right) \\ \times \left(P_{TA}(t_j)L_A^{(j)} + P_{TC}(t_j)L_C^{(j)} + P_{TG}(t_j)L_G^{(j)} + P_{TT}(t_j)L_T^{(j)} \right)$$

Phylogenetic likelihood

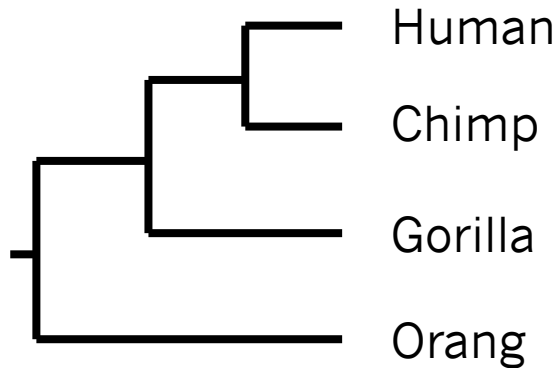


↑
single
site

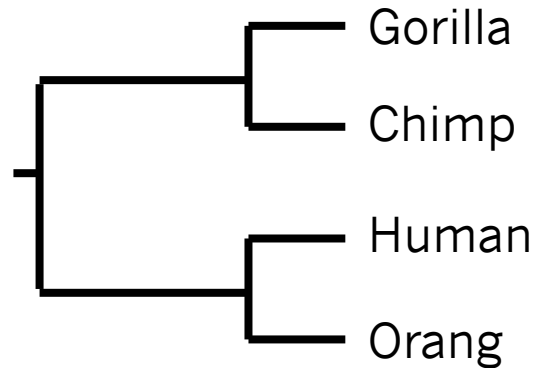
What is the probability of observing a set of genetic variants
for a given underlying tree?

$$\text{Prob}(\text{all sites} \mid \text{tree}) = \text{Prob}(\text{site 1} \mid \text{tree}) \times \dots \times \text{Prob}(\text{site N} \mid \text{tree})$$

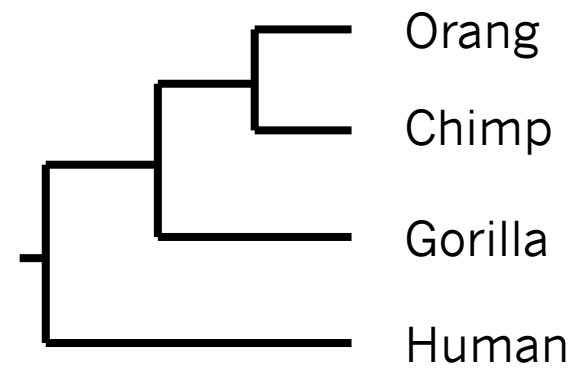
Likelihood



log-likelihood = -32.14



log-likelihood = -42.77



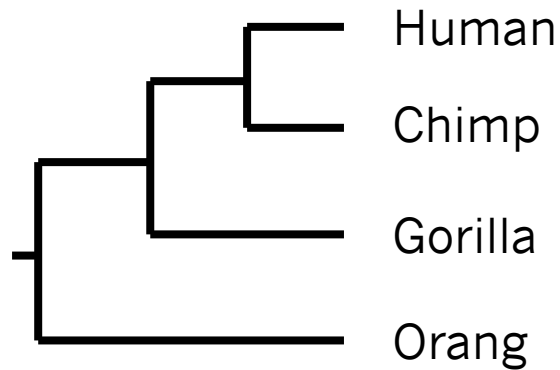
log-likelihood = -39.08

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

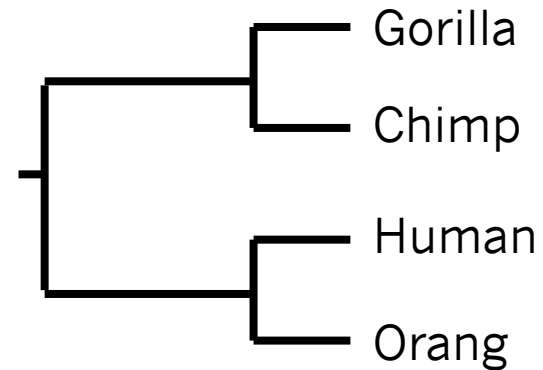
Method comparison

Method	Pros	Cons
Neighbor-joining	Extremely fast Scalable	Does not use evolutionary events to infer tree
Parsimony	Intuitive Fairly fast	Assumes change is rare; Unclear how to choose "costs" for events
Likelihood	Most accurate Most realistic Can simulate data	Slower More complex theory + algorithms

Newick strings



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



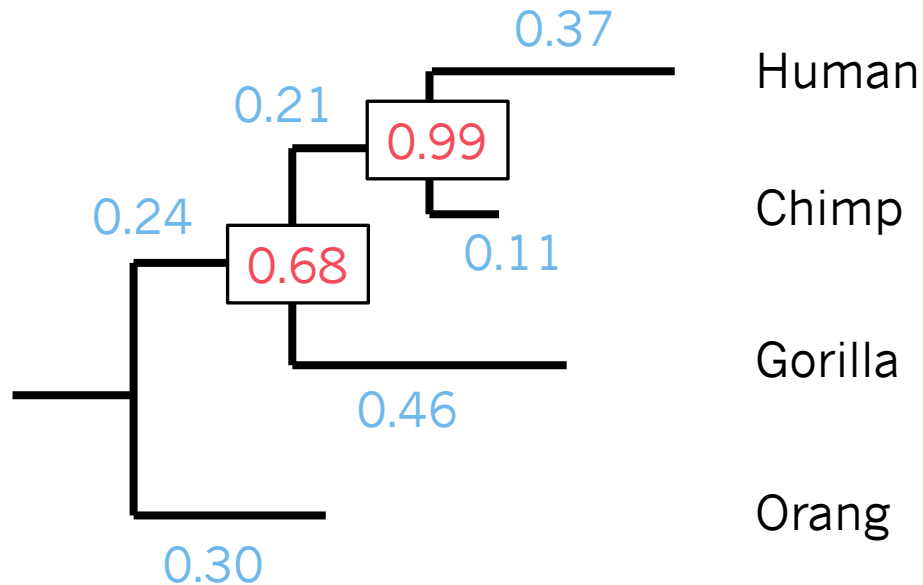
`((Gorilla,Chimp),(Human,Orang));`

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 11