# Lecture 11
# Molecular phylogenetics

*Lonicera flava*
© Kathy Melton/
Missouri Botanical Garden

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

# Lecture 11 outline

Last time: regex

This time: phylogenetics

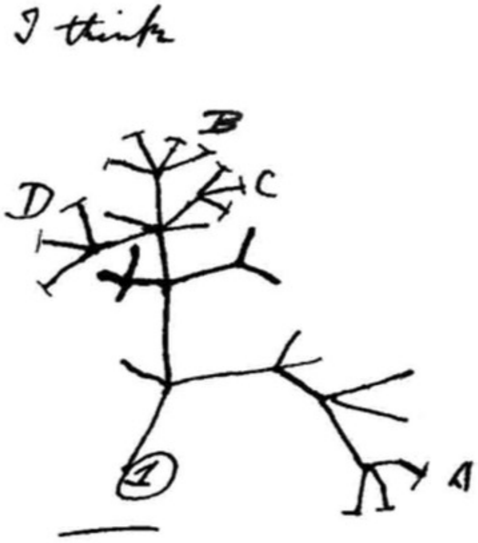
- interpreting trees
- tree-thinking
- inferring trees
- inference methods

# Phylogenetics

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

Phylogenies are useful for

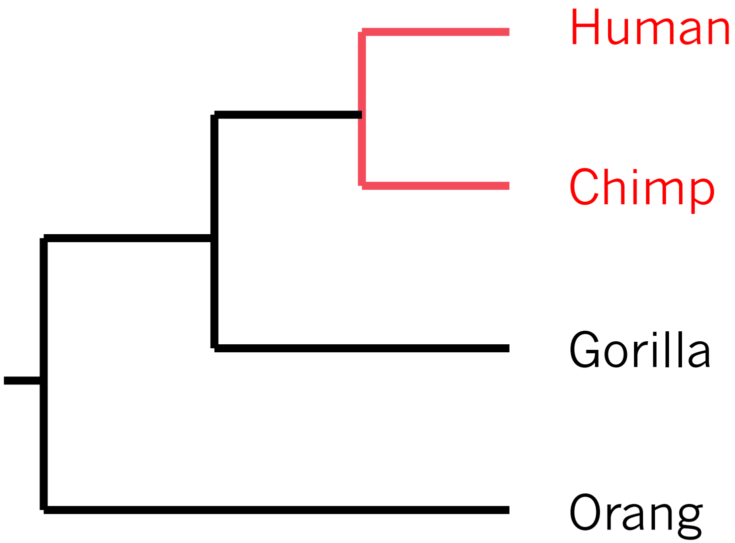
- gene annotation
- tracking viral spread
- identifying zoonosis
- reconstructing tumorogenesis
- conservation biology assays
- inferring species relationships



phylogeny sketch by Darwin

# Reading a phylogeny

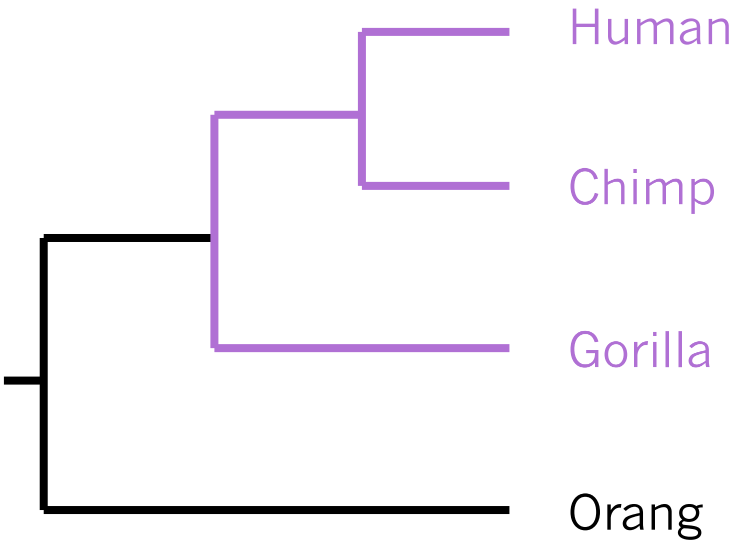
Phylogenetic relationships are hierarchical, and most often represented as bifurcating ***trees***



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***

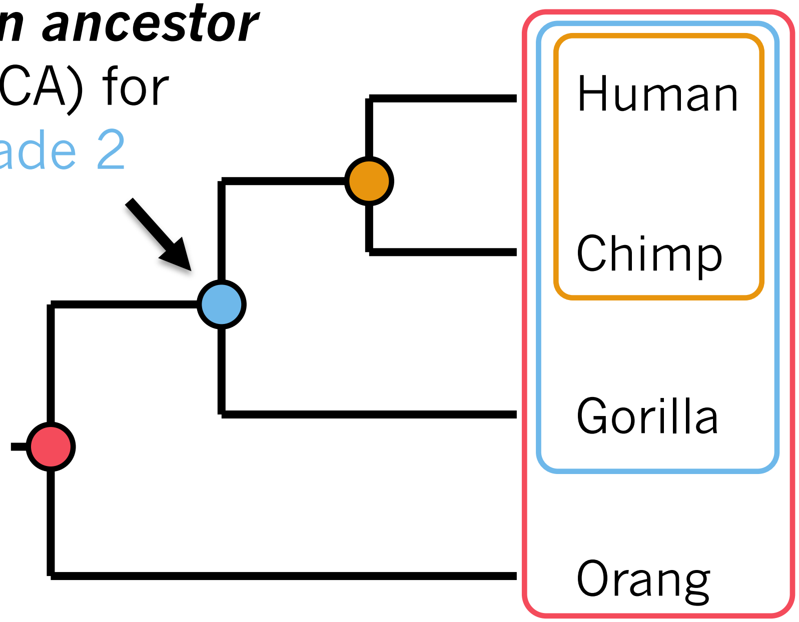


Human, Chimp and Gorilla are more closely related to each other than to Orang

# Reading a phylogeny

Taxa that are most 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

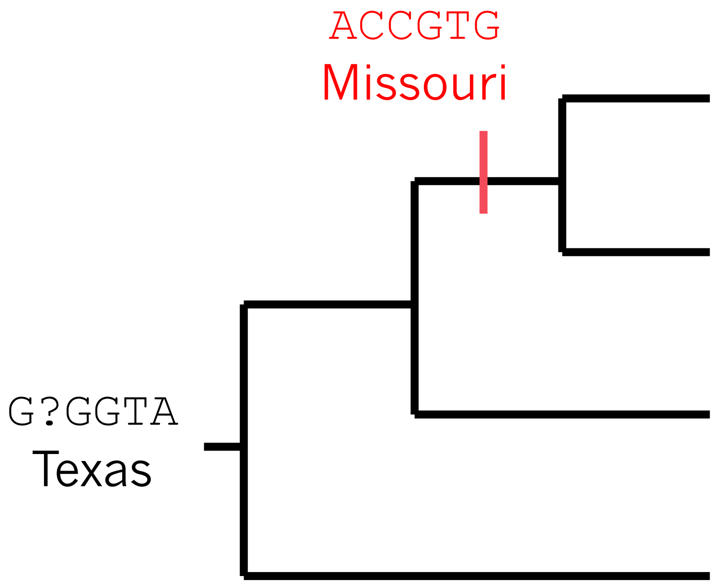
Clade 3: H+C+G+O

# “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”

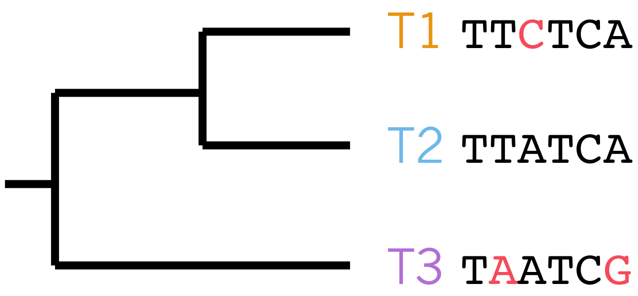


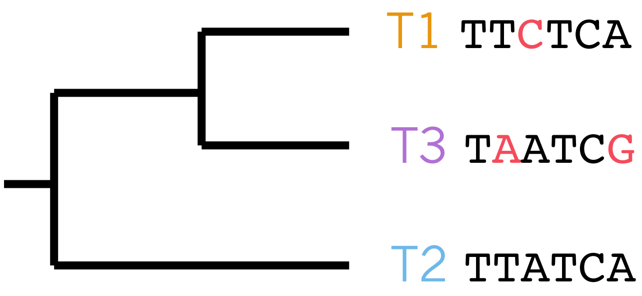
| Taxon | Gene variant | Location |
|---|---|---|
| PT1 | ACCGTG | Missouri |
| PT2 | ACCGTG | Missouri |
| PT3 | GAGGTA | Texas |
| PT4 | GCGGTA | Texas |

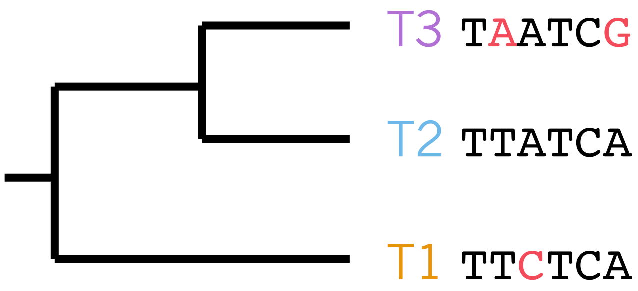
Phylogeny informs when and where variation arose, which can guide future biological research

# 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*** 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?

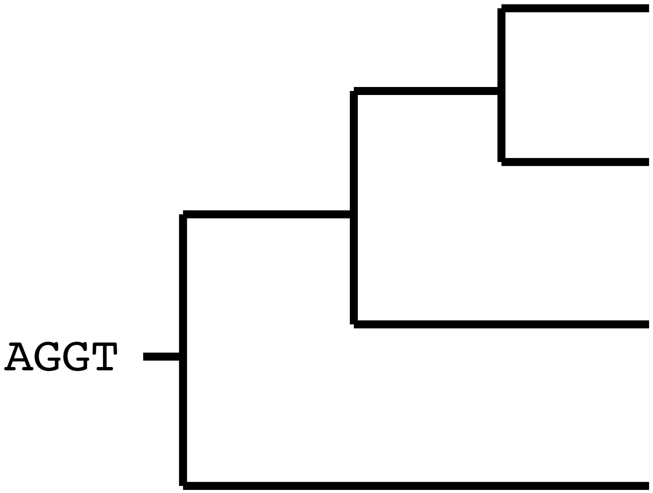
# Phylogenetic method types

Most methods used to infer phylogenies compute scores based on

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

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

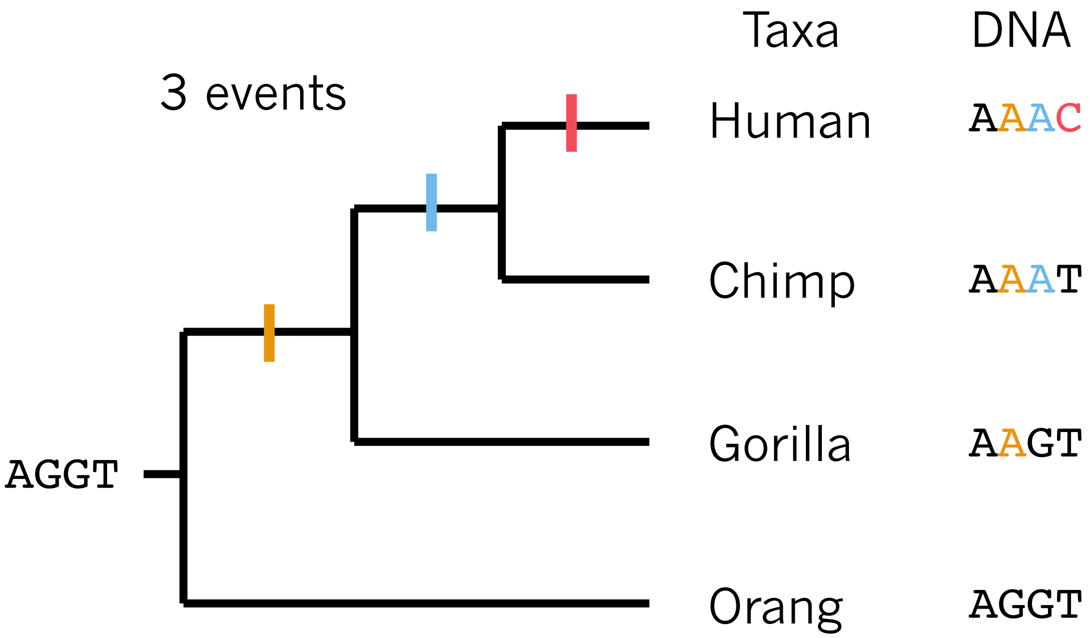
# Parsimony



| Taxa | DNA |
|---|---|
| Human | AAAC |
| Chimp | AAAT |
| Gorilla | AAGT |
| Orang | AGGT |

What phylogeny requires the fewest character change events?

# Parsimony



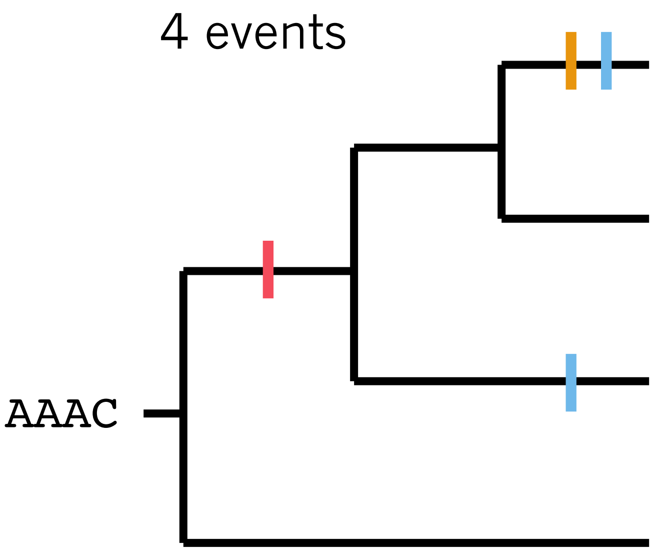
3 events

AGGT

| Taxa | DNA |
|---|---|
| Human | AAAC |
| Chimp | AAAT |
| Gorilla | AAGT |
| Orang | AGGT |

What phylogeny requires the fewest character change events?

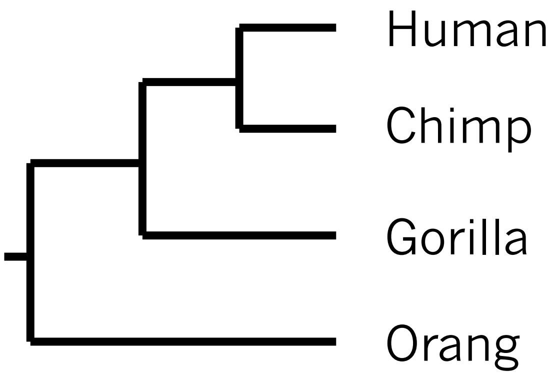
# Parsimony



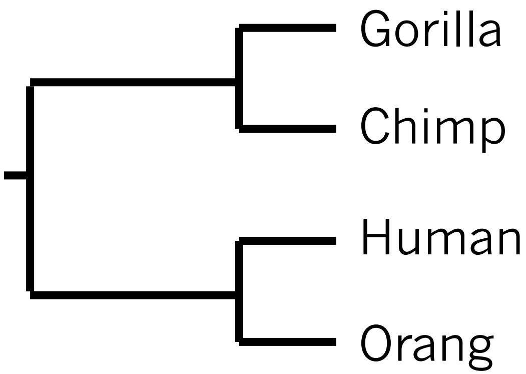
| Taxa | DNA |
|---|---|
| Orang | AGGT |
| Chimp | AAAT |
| Gorilla | AAGT |
| Human | AAAC |

What phylogeny requires the fewest character change events?

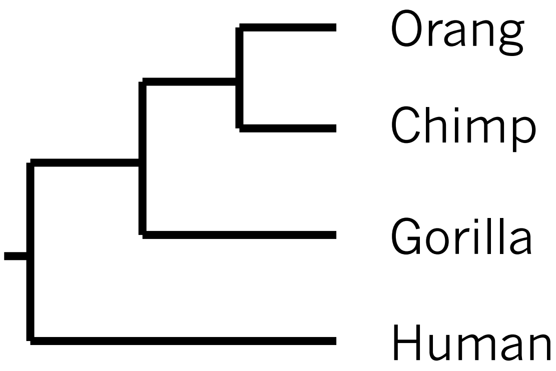
# Parsimony



3 events



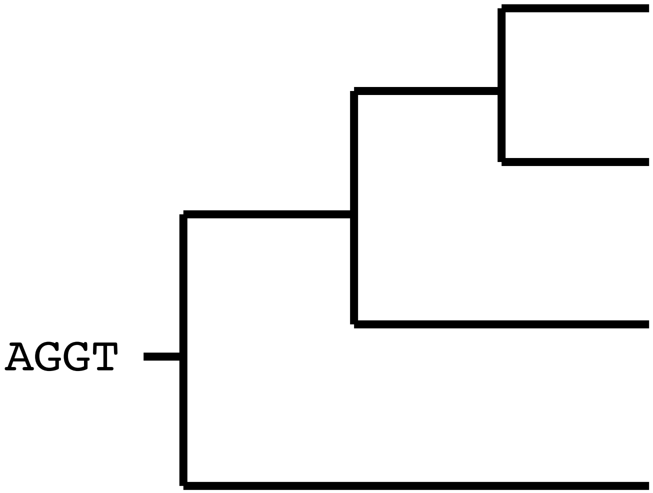
6 events



13 events

What phylogeny requires the fewest character change events?

# Likelihood

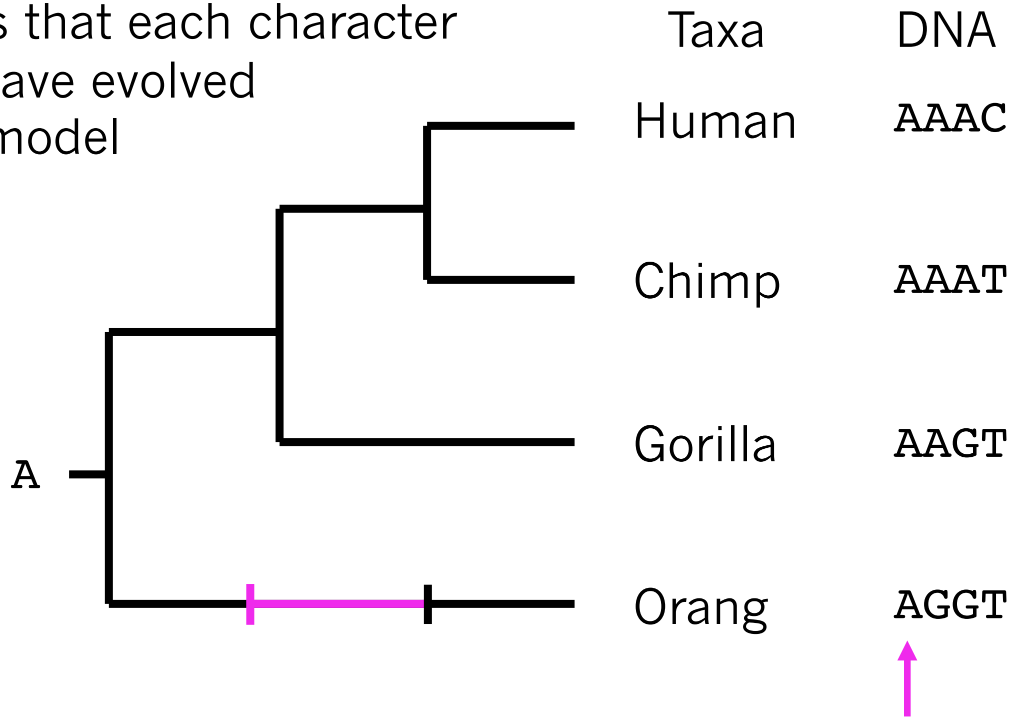


| Taxa | DNA |
|---|---|
| Human | AAAC |
| Chimp | AAAT |
| Gorilla | AAGT |
| Orang | AGGT |

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

# Likelihood

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

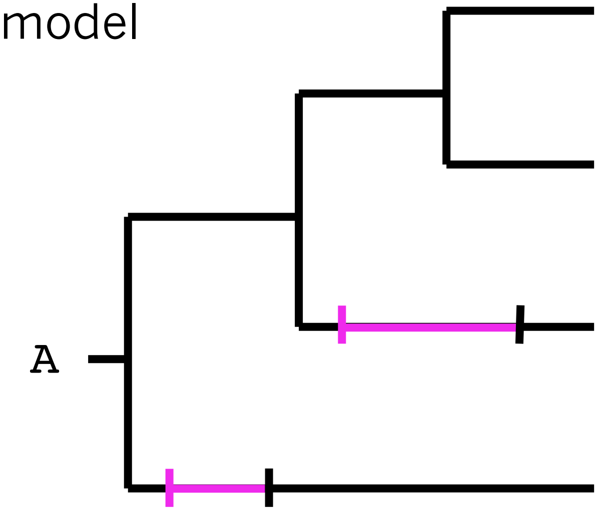


| Taxa | DNA |
|---|---|
| Human | AAAC |
| Chimp | AAAT |
| Gorilla | AAGT |
| Orang | AGGT |

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

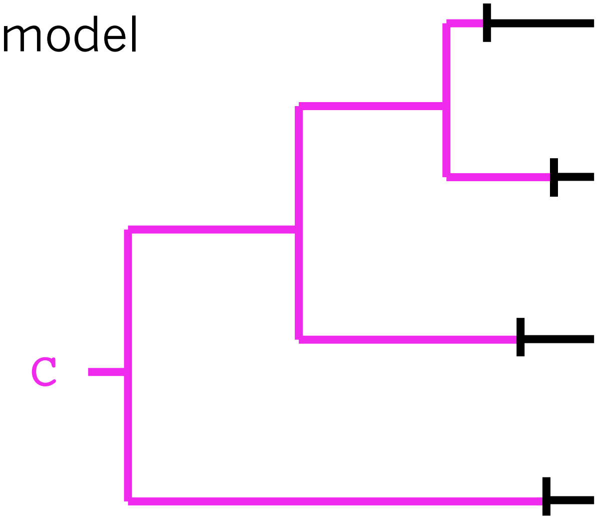


| Taxa | DNA |
|---|---|
| Human | AAAC |
| Chimp | AAAT |
| Gorilla | AAGT |
| Orang | AGGT |

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

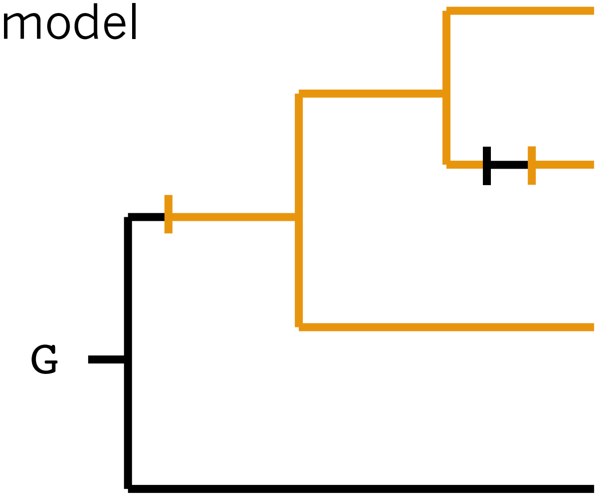


| Taxa | DNA |
|---|---|
| Human | AAAC |
| Chimp | AAAT |
| Gorilla | AAGT |
| Orang | AGGT |

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

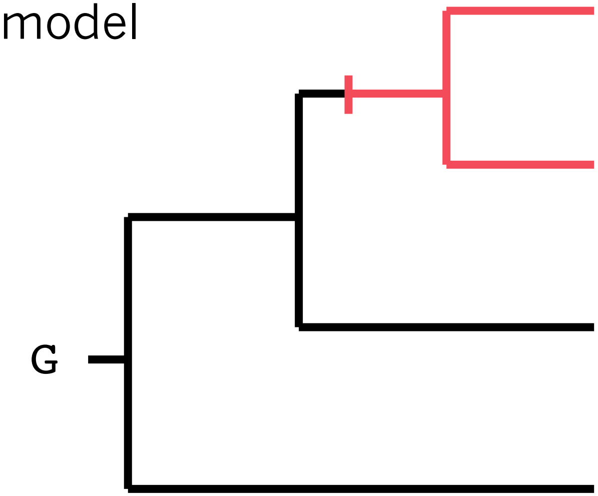


| Taxa | DNA |
|---|---|
| Human | AAAC |
| Chimp | AAAT |
| Gorilla | AAGT |
| Orang | AGGT |

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

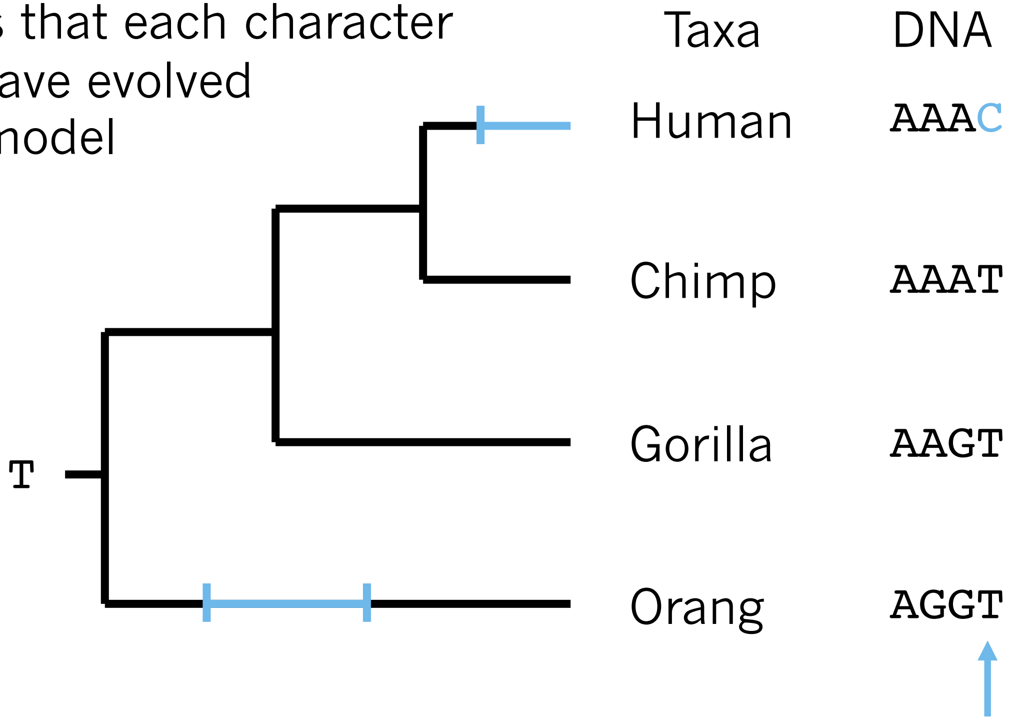


| Taxa | DNA |
|---|---|
| Human | AAAC |
| Chimp | AAAT |
| Gorilla | AAGT |
| Orang | AGGT |

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

# Likelihood

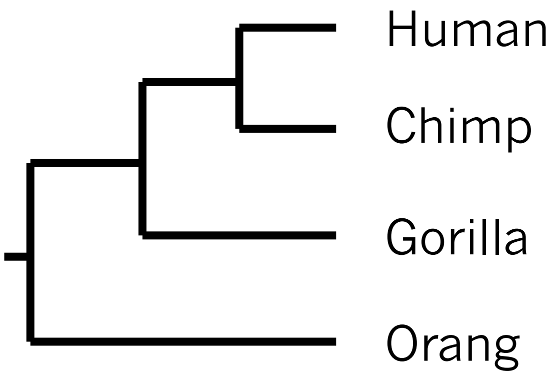
Compute probability for
all ways that each character
could have evolved
under model



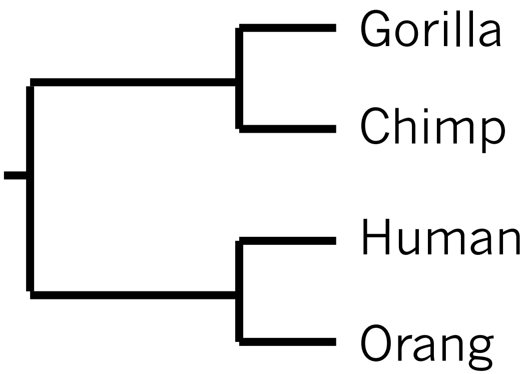
| Taxa | DNA |
|---|---|
| Human | AAAC |
| Chimp | AAAT |
| Gorilla | AAGT |
| Orang | AGGT |

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

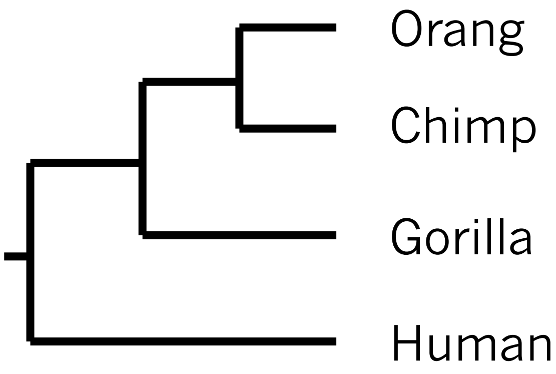
# Likelihood



log-likelihood = -32.14



log-likelihood = -42.77



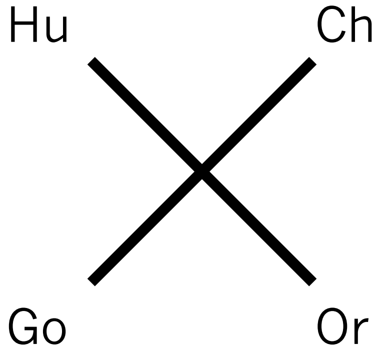
log-likelihood = -39.08

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

# Neighbor-joining

|  | 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

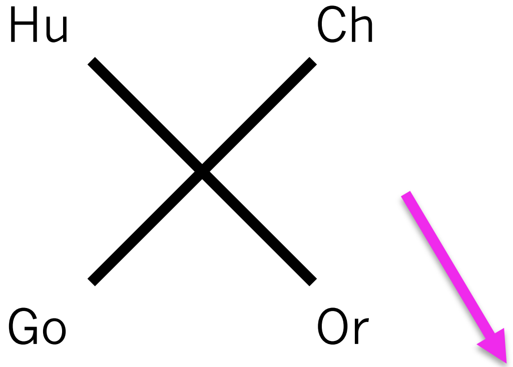


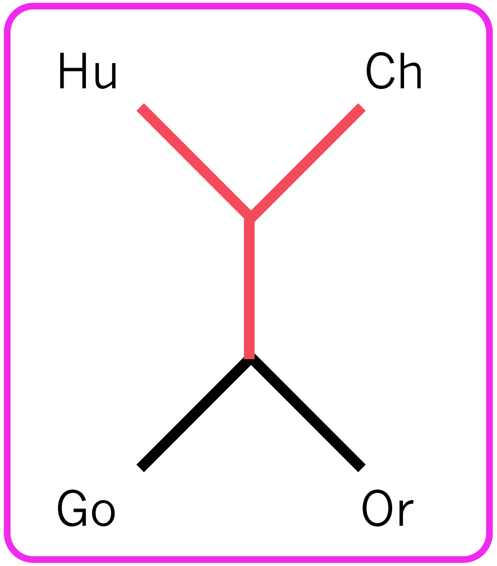
Select pairs of taxa with short sequence distances, and join them as neighbors

# 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 |

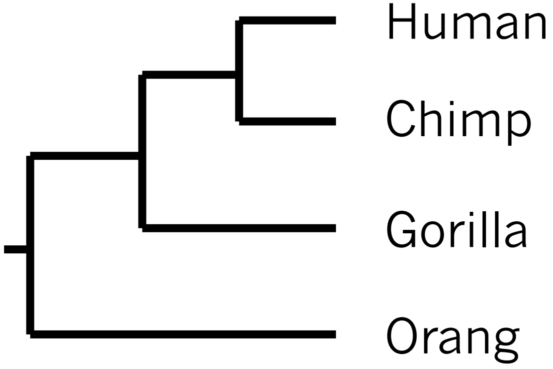
distance matrix
for sequence pairs





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

# Newick strings



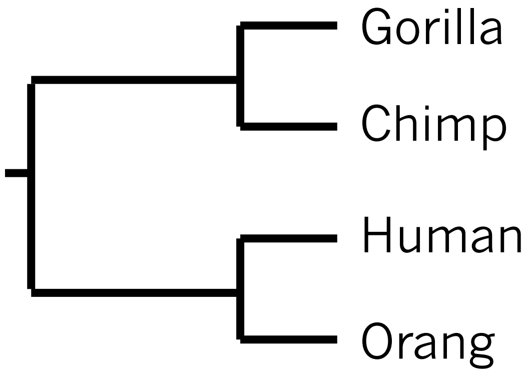
Human

Chimp

Gorilla

Orang

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



Gorilla

Chimp

Human

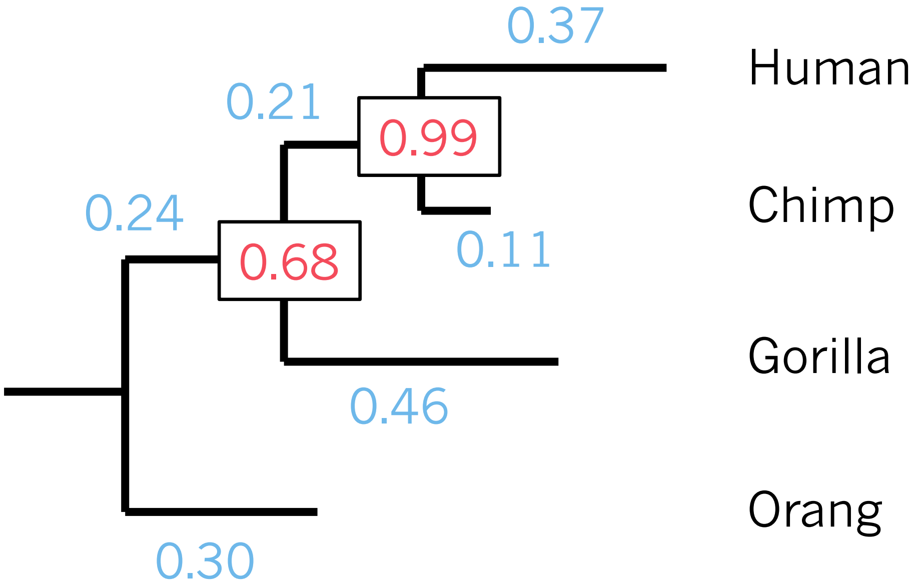
Orang

((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