Phylogenetic Tree Construction

Distance Metric Behavior

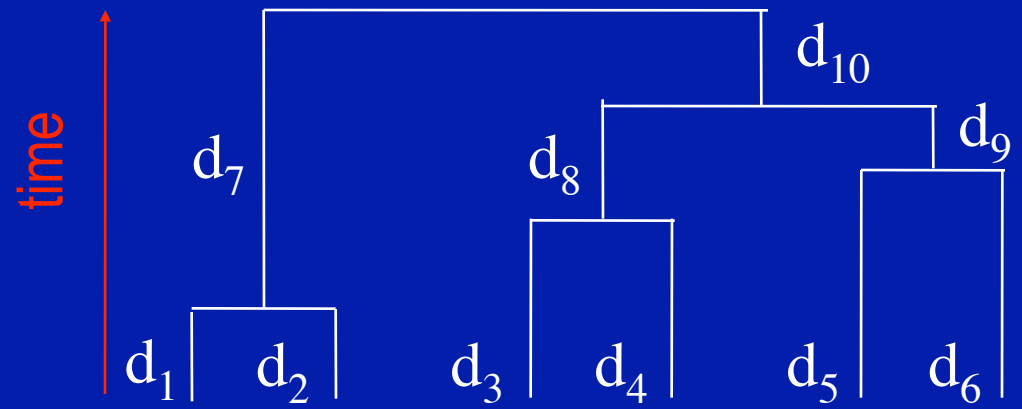
Under the Jukes-Cantor model (all nucleotides equally probable; all transitions between them equally probable), how do you expect the following distance metrics to behave as a function of evolutionary time t? (sketch a basic graph with t as the horizontal axis and distance as the vertical axis):

- *f: observed letter differences per site*
- δ: *inferred mutation events per site*

Distance Metric Behavior Answer

- f grows linearly from zero for small t.
- f follows exponential decay to $1 - \pi = 3/4$ as $t \to \infty$.
- $\delta \approx f$ for small t.
- δ grows linearly for all times t.

A Source of Systematic Error: Molecular Clock Assumption



$d_1 = d_2$

$d_3 = d_4$

$d_5 = d_6$

$d_3 + d_8 = d_5 + d_9$

$d_1 + d_7 = d_3 + d_8 + d_{10}$

Assumes that rates of mutation are equal on all branches, so mutation distances should be equal between any pair of modern sequences and their most recent common ancestor (MRCA). Under this assumption, mutation distance is directly proportional to *time of divergence*.

Additive Distances

- In general the symmetric distance matrix D_{ij} has (n choose 2) degrees of freedom.
- But if we assume the sequences evolved on the edges of a tree, we assert the D_{ij} are just equal to the sum of the edges from *i* to *j*.
- Far fewer degrees of freedom (*2n-3* edge lengths).

Ultrametric vs. Additive Distances

- *reversibility*: if mutation rate matrix obeys the detailed balance equations, then its equilibrium transition rates are equal in the time-forward direction vs. the time-reverse direction. Use the same matrix regardless of whether you are traversing an edge in time-forward vs. time-reverse direction.
- "additive distances": if the transition matrices on different edges differ only by a scalar factor (i.e. $T_1 = \alpha T_2$), then the total distance for traversing multiple edges is just equal to the sum of the distances associated with each edge.
- "Molecular clock" assumption: if mutation rates constant at all times, then branch length (distance) on any edge is just proportional to time.
- "ultrametric" distances: the distance between any pair of (modern) sequences is just proportional to the time since they diverged (i.e. since their MRCA).

Ultrametric vs. Additive

- Can the Jukes-Cantor model fit the *additive distance* assumption?
- Can the Jukes-Cantor model fit the *ultrametric distance* assumption?

If so, how? If not, why not?

Ultrametric vs. Additive Answer

- Yes, it produces an additive distance metric. Let $\delta = \lambda t$, and Π be the JK unit-distance rate matrix whose off-diagonal elements are π, and whose diagonal elements are just $-(1-\pi)$. Then the transition matrix for two edges δ_1, δ_2 are $e^{\delta_1 \Pi}, e^{\delta_2 \Pi}$ respectively. The transition matrix for traversing both edges is

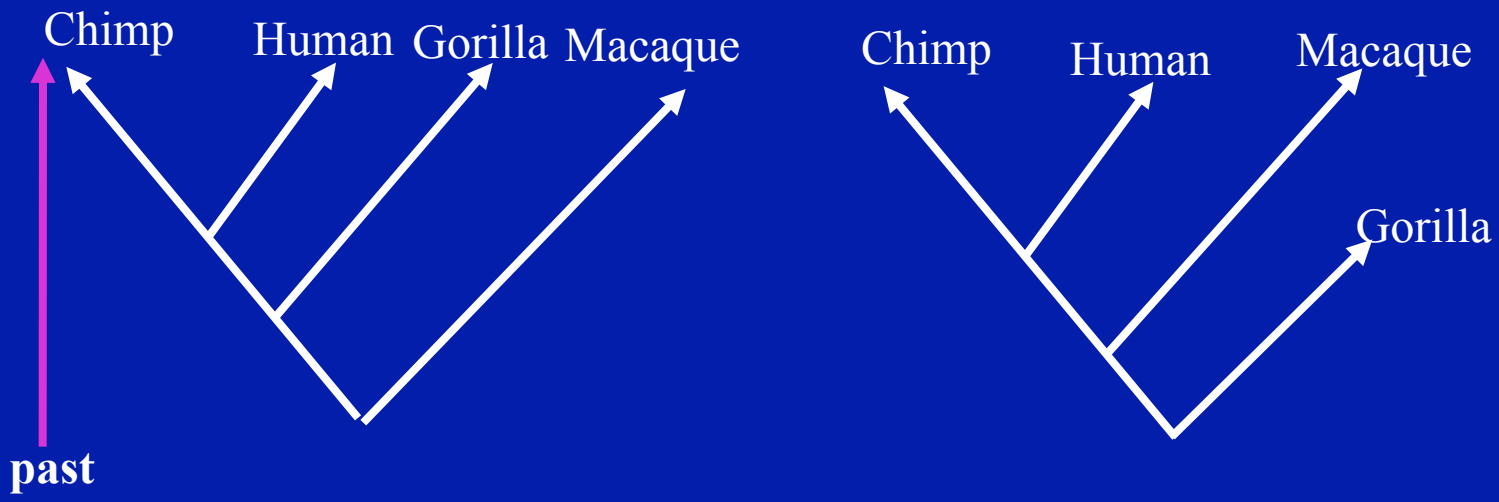
$$e^{\delta_3 \Pi} = e^{\delta_1 \Pi} e^{\delta_2 \Pi} = e^{(\delta_1 + \delta_2)\Pi}$$

  So the effective total distance is just $\delta_3 = \delta_1 + \delta_2$. The key is that JK assumes that same rate matrix on all edges; if that weren't the case, it wouldn't give this simple, additive form.

- Yes, it can also produce an ultrametric distance, in the case where the value of λ is the same on all edges. Then the distance is always the total time multiplied by a constant, as required by the ultrametric assumption.

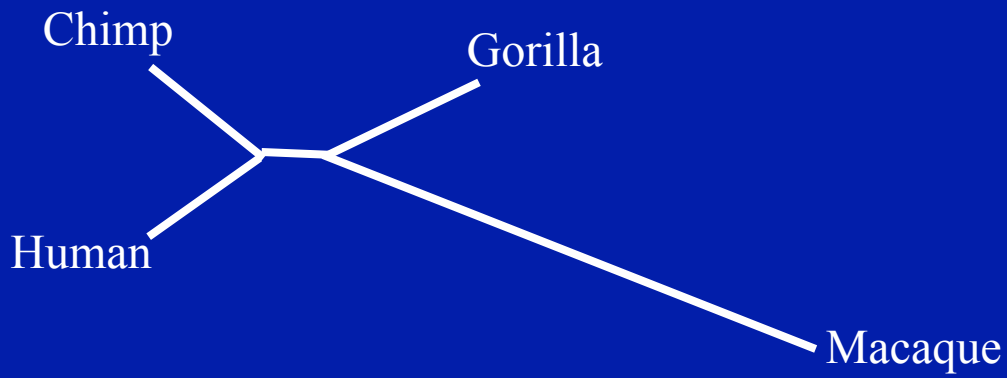
Rooted Versus Unrooted Trees

Rooted Trees



These two rooted trees are different
But both have the same Unrooted tree:

Unrooted Tree

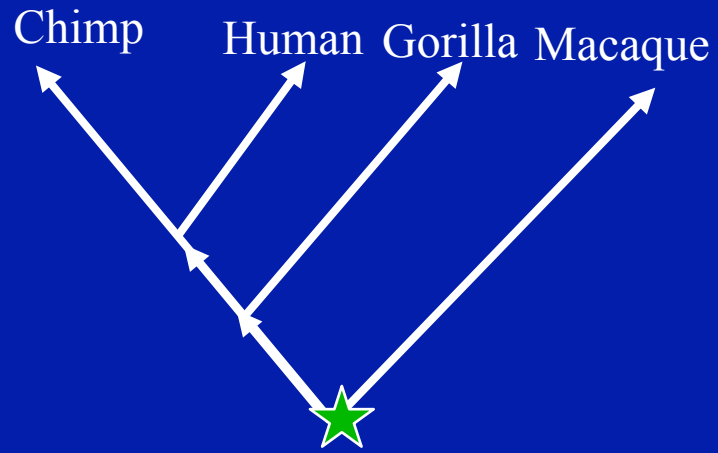


The direction of evolution is specified in a rooted tree but not in an unrooted tree

Rooted Trees Are Directed Graphs

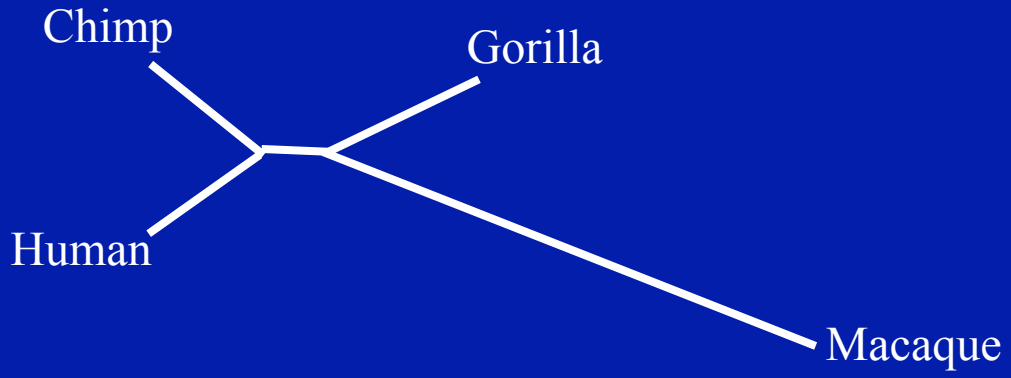
In a rooted tree each edge is directed, pointing from root to "leaves".

Rooted Tree



In an unrooted tree, the edges are "undirected".

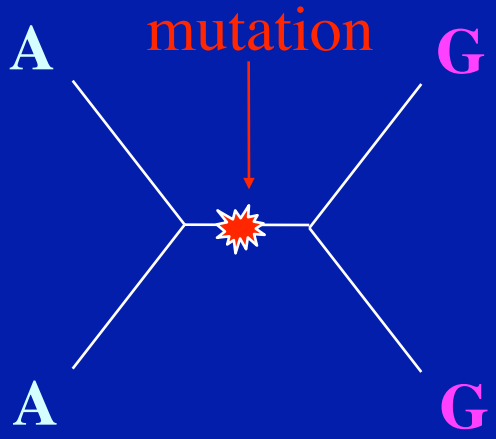
Unrooted Tree



The Principle of Parsimony

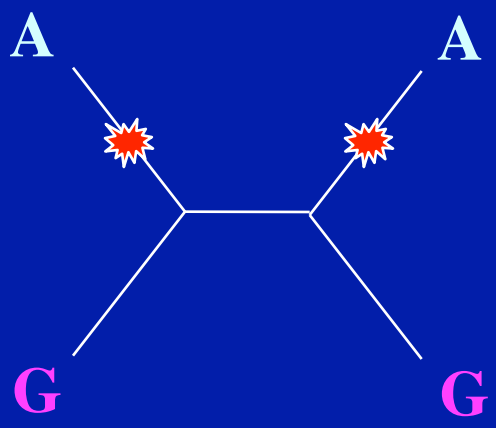
- Out of the possible trees, the one with the fewest required “mutation events” is best.
- A mutation event is a letter substitution at a particular point on the tree; sequences on one side of the event have letter X, and on the other side have letter Y.

Parsimony: Prefer Tree with Least Mutation Events



✓

Requires only one mutation event



Requires at least two mutation events

#mutations depends on tree

Tree Topology Information

You are given an alignment of four sequences, and asked to construct an unrooted tree that minimizes the number of total mutations on the tree, using the alignment columns where all four are aligned. Select all of the following classes of alignment columns that contain information about the tree topology (i.e. *excluding* those alignment columns from the analysis could change *which* tree has the minimal number of total mutations):

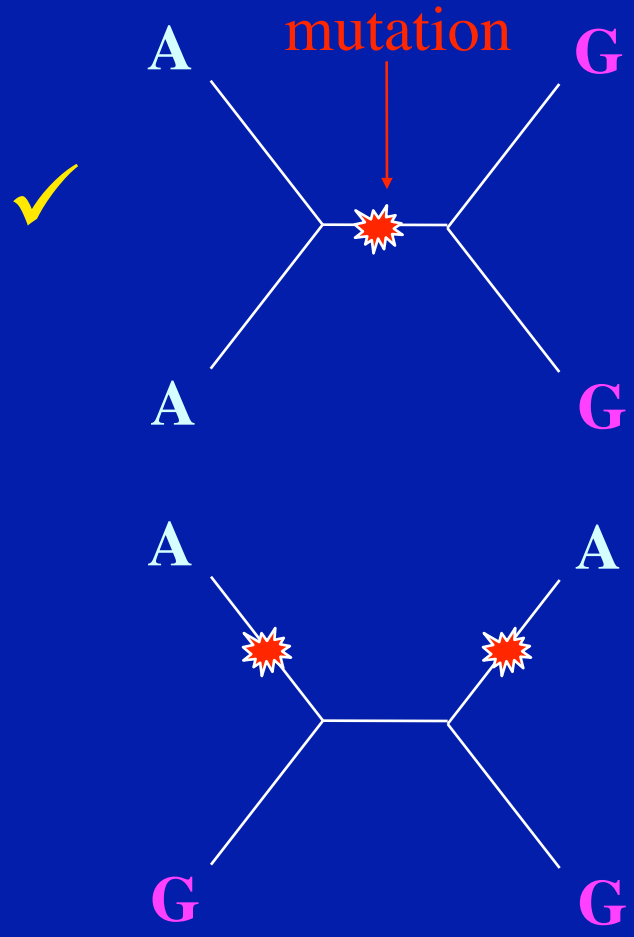
- A. alignment columns where all four sequences share the same letter.
- B. alignment columns where three sequences share the same letter, and the remaining sequence has a different letter.
- C. alignment columns where two sequences share a letter, and the other two sequences share another letter.
- D. alignment columns where all four sequences disagree with each other.

Tree Topology Information Answer

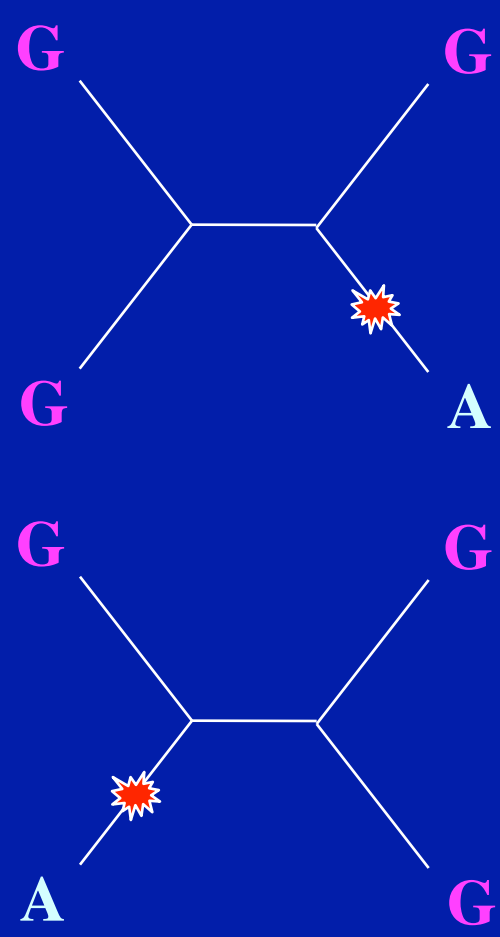
This is the classic parsimony criterion.

- sites where two sequences diverge from the other two sequences are informative, i.e. the correct tree has a lower mutation count from the other possible trees.
- sites where one sequence diverges from the other three are not informative (all possible trees have exactly one mutation).
- (obviously, sites where are four sequences agree are not informative).

Parsimony on a simple Unrooted Tree



✓

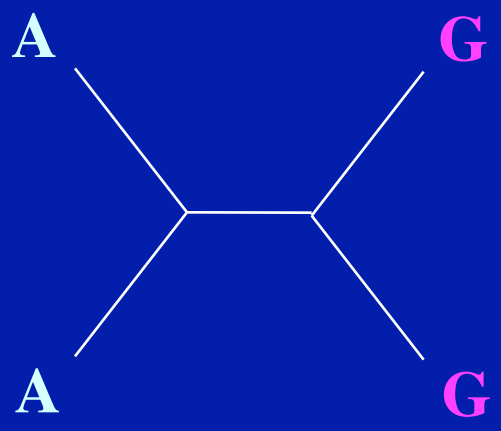


#mutations depends on tree

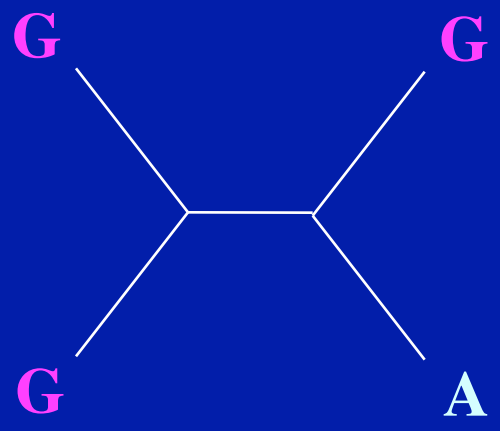
Same #mutations for all trees

What Characters Informative for an Unrooted Tree?

```
Chimp       ..AGCTAAAGGGTCAGGGGAAGGGCA..
Human       ..AGCAAAAGGGTCAGGGGAAGGGGA..
Gorilla     ..AGCATAGGGGTCAGGGGAAAGGCT..
Squirrel    ..AGCGGACCGGTAAGGAGAAAGGAC..
informative     ** *               *  **
```



2:2 is Informative



3:1 is not informative

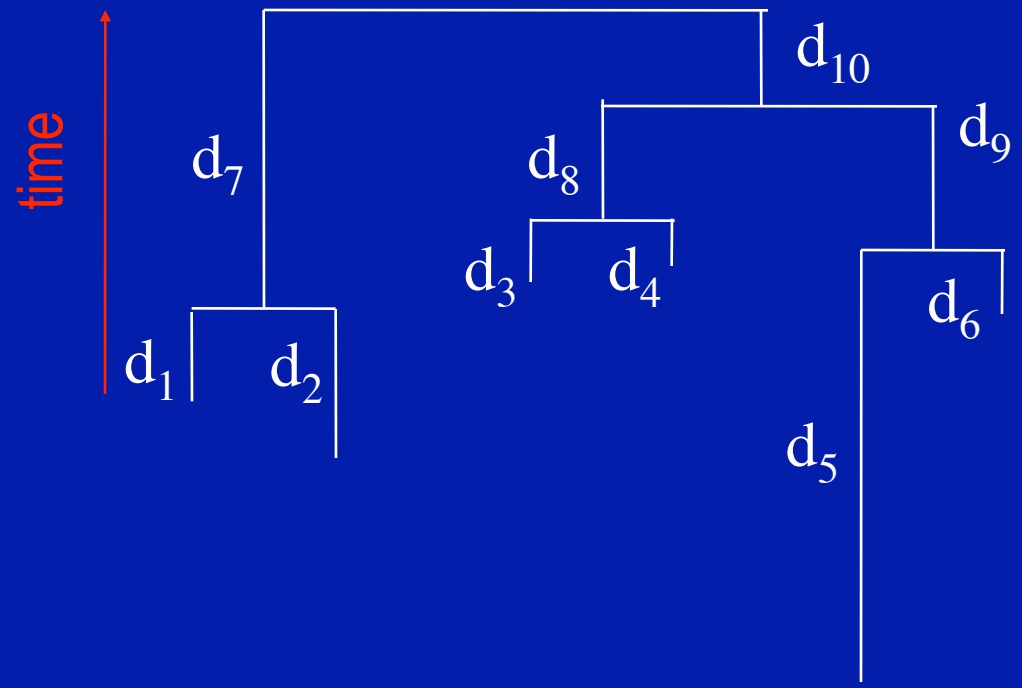
UPGMA Applications

UPGMA is a general-purpose clustering algorithm that can be applied to inferring the family tree structure of a set of related sequences. How general is this algorithm? I.e. do you expect it to work on all kinds of evolutionarily related sequences? If yes, briefly explain why; if not, provide an example scenario where you would expect it to fail.

UPGMA Applications Answer

- UPGMA assumes that the cluster pair with minimum distance must be neighbors in the tree. This is equivalent to assuming the molecular clock assumption, i.e. that all distances are just propotional to time of divergence.
- Any case where one branch has a much faster or slower mutation rate than the rest of the tree can cause UPGMA to infer the wrong tree. For example, a branch with a much higher mutation rate (yielding much larger distances) will be treated as an “outgroup” insted of being clustered with its actual neighbor.

Violations of the Molecular Clock Assumption



$d_1 \neq d_2$

$d_3 \neq d_4$

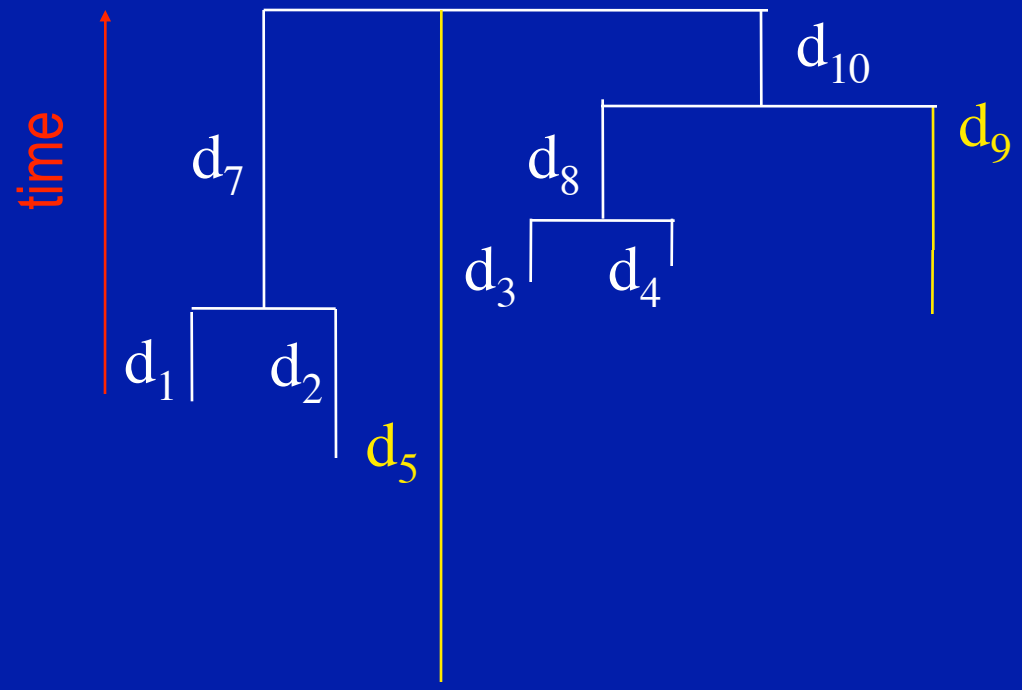
$d_5 \neq d_6$

$d_3 + d_8 \neq d_5 + d_9$

$d_1 + d_7 \neq d_3 + d_8 + d_{10}$

Mutation rates can be different on each branch, so distances to the most recent common ancestor (MRCA) will not be equal.

Molecular Clock Errors: Incorrect Tree Inference

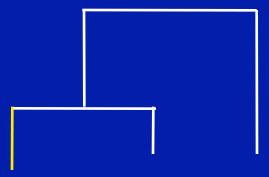


These branches should be placed together, but get put in different parts of the tree due to faster evolution on one of the branches.

Mutation rates can be different on each branch, so distances to the most recent common ancestor (MRCA) will not be equal.

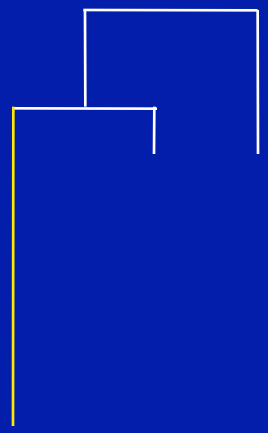
Molecular Clock Errors: Paralog Divergence

Actual history

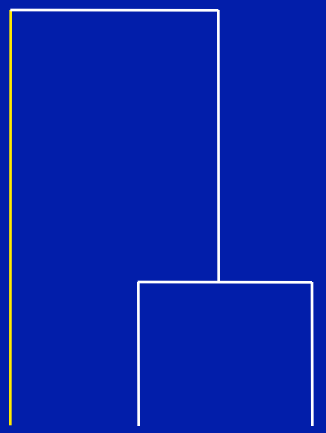


Paralog evolves more rapidly (more mutation per unit time)

Measured distances



Parsimony Tree



A higher mutation rates on one branch (e.g. paralog evolution) can cause parsimony to infer the wrong branching structure.

Violations of the Molecular Clock Assumption

- Variations in mutation rate (in different organisms, or different genome regions).
- Variations in selection pressure, e.g.
  - Housekeeping genes are under stronger negative selection pressure than tissue-specific genes.
  - Gene duplications typically mutate rapidly and asymmetrically (paralogs).