Computational Molecular Evolution

by Mátyás Pajkos

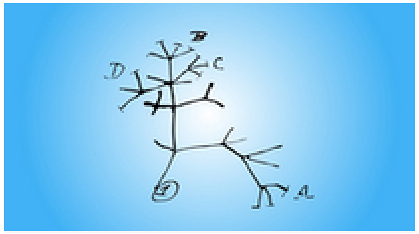
Dosztányi Lab

Eötvös Loránd University

2019 May

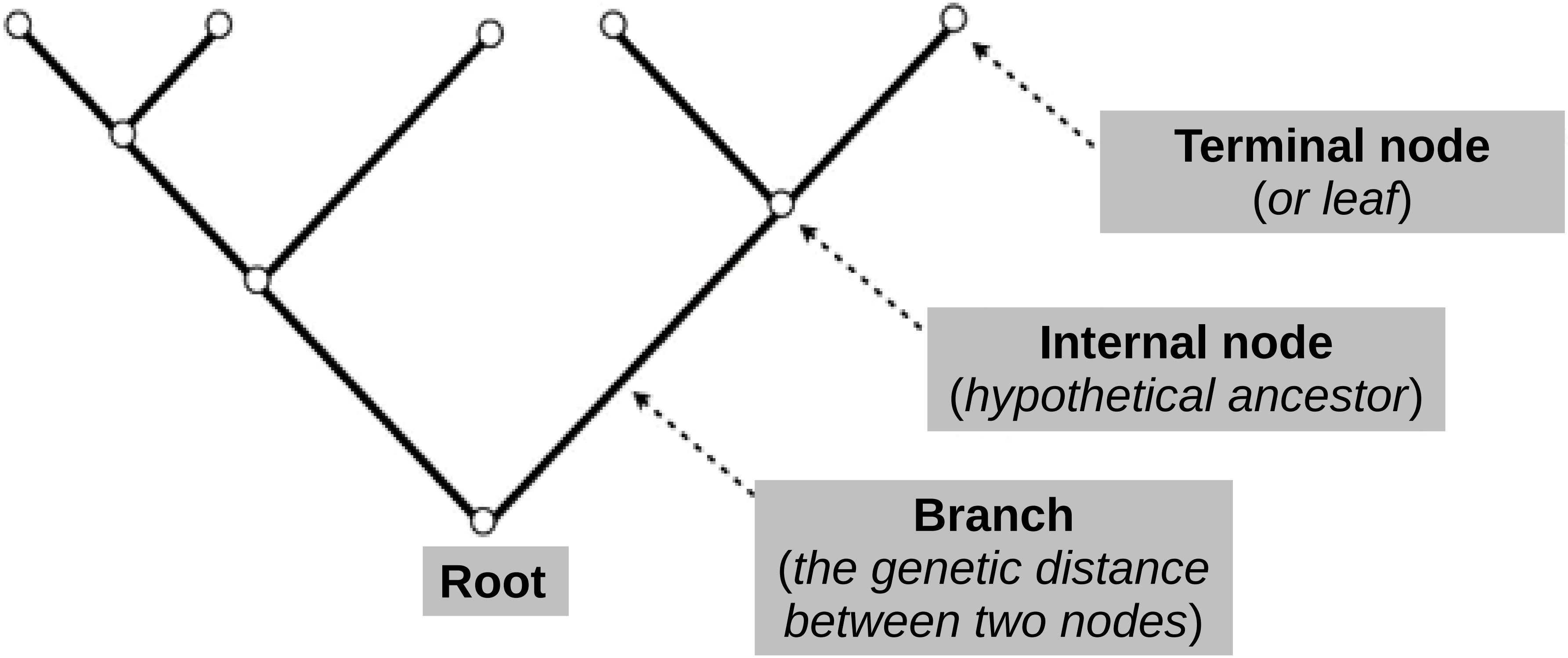
- *Phylogenetic trees: Terminology and representation*
- *Reconstructing trees using present-day data*
- *Orthologous groups*
- *Detection of molecular selection*
- *Tutorial*

Phylogenetic trees: Terminology and representation



- Terminology
- Representation
- The newick format

Phylogenetic trees: Terminology



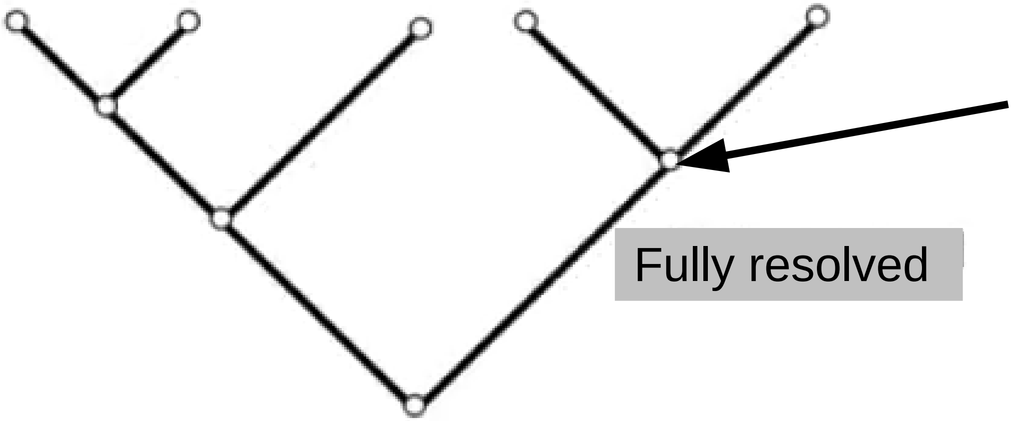
Terminal node
(*or leaf*)

Internal node
(*hypothetical ancestor*)

Branch
(*the genetic distance between two nodes*)

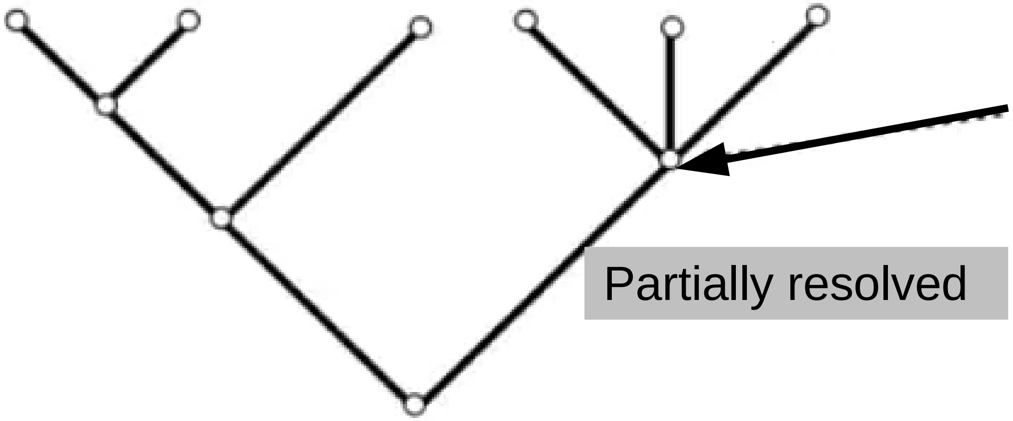
Root

Phylogenetic trees: Representation



Fully resolved

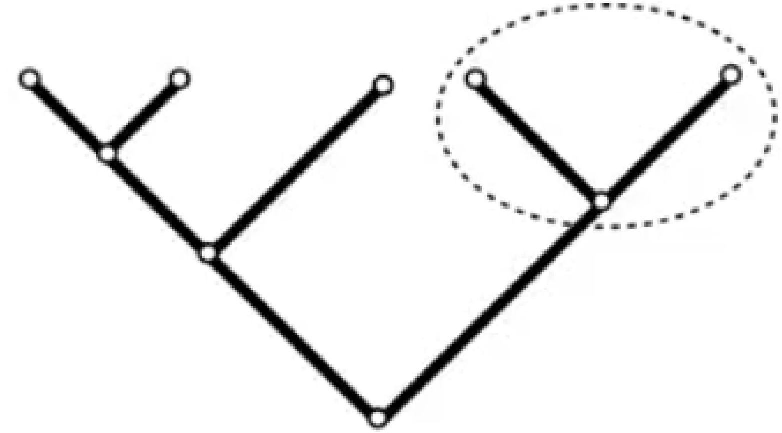
- An internal node has exactly two branches going out
- Reason: A population split into two



Partially resolved

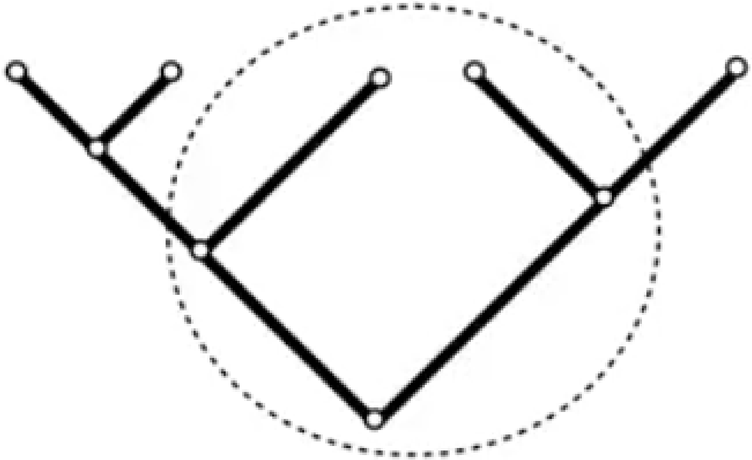
- An internal node has more than two branches going out
- Reason: We do not have enought data

Phylogenetic trees: Representation



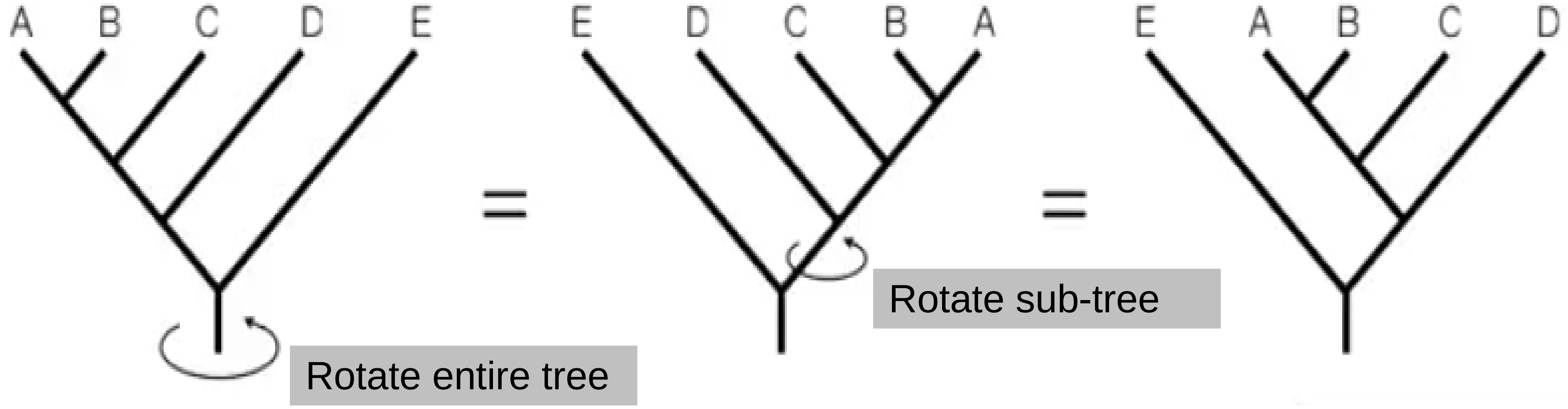
Monophyletic (clade)

- A clade is a group of organisms that group includes all descendants of their common ancester
- All the member of such a group, they have several shared features



Non-monophyletic (paraphyletic)

Phylogenetic trees: Representation of topology

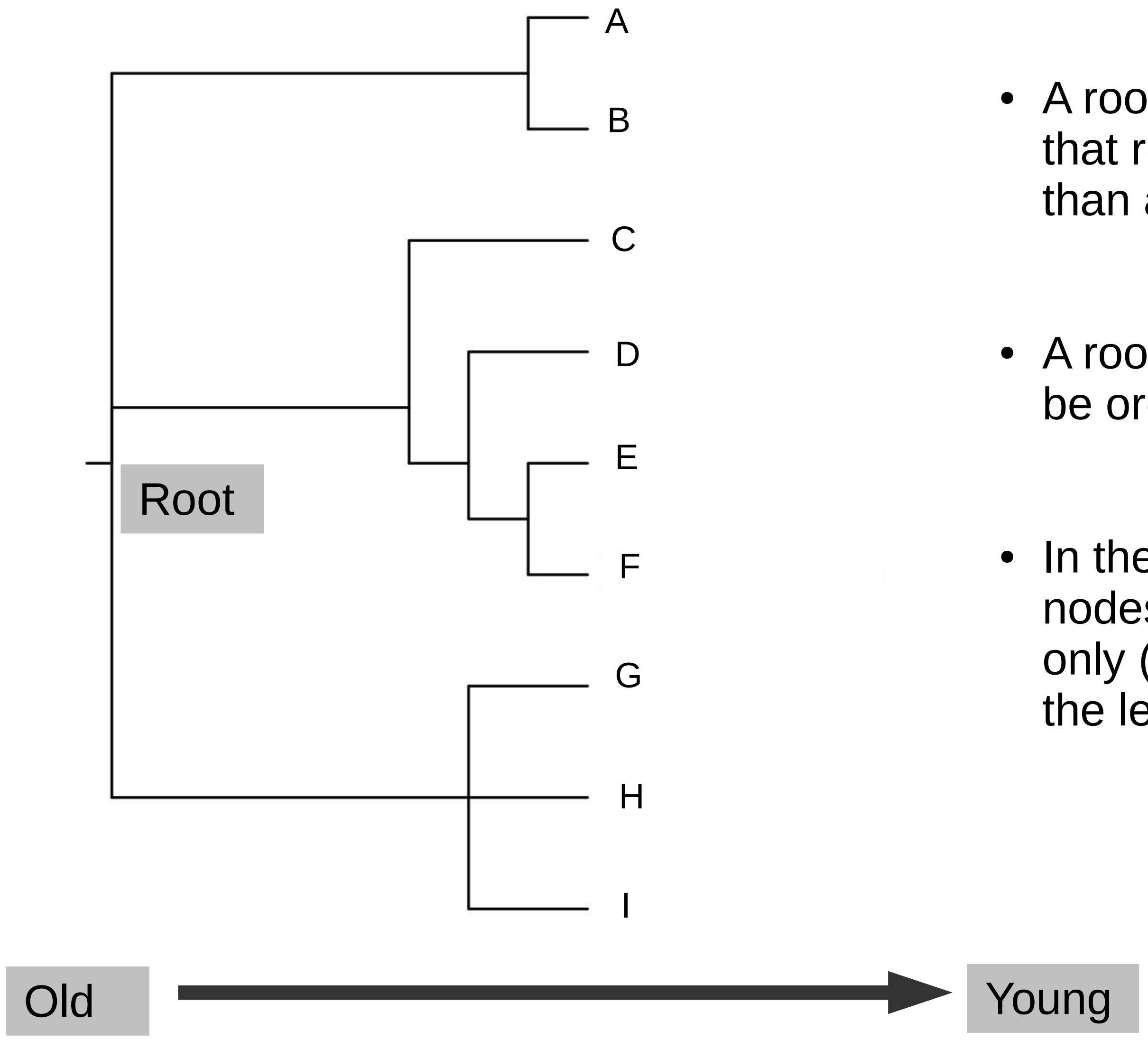


Rotate entire tree

Rotate sub-tree

- Three different representation of the same tree-topology
- It is not always true that the neighbours are closely related to each other

Phylogenetic trees: rootedness

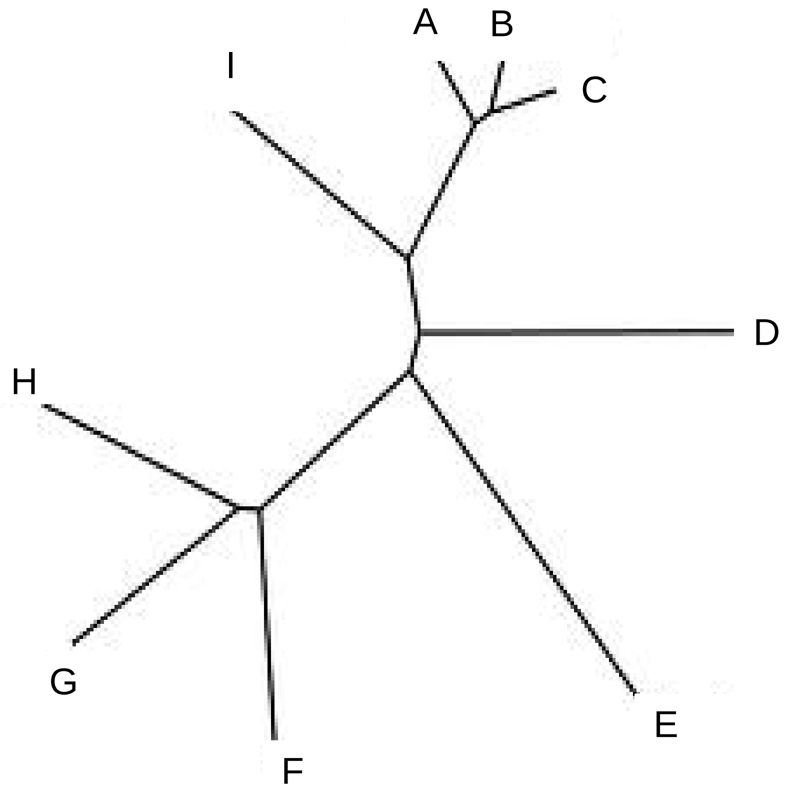


- A rooted tree has a single node (**The root)** that represents a point in time that is earlier than any other node in the tree

- A rooted tree has directionality (nodes can be ordered in terms of "earlier" or "later")

- In the rooted tree, distance between two nodes is represented along the time-axis only (the second axis just helps spread out the leaves)

Phylogenetic trees: Unrootedness

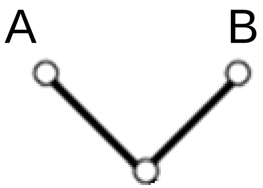


- In unrooted trees there is no directionality: we do not know if a node is younger or older than another node

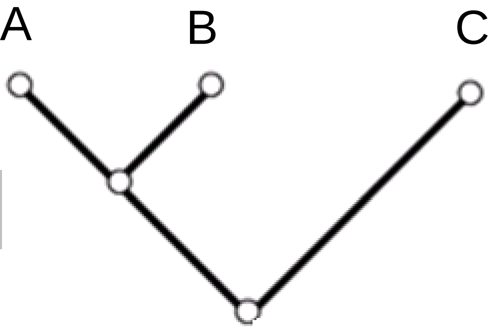
- Distance along branches directly represents node distance

Phylogenetic trees: The Newick format

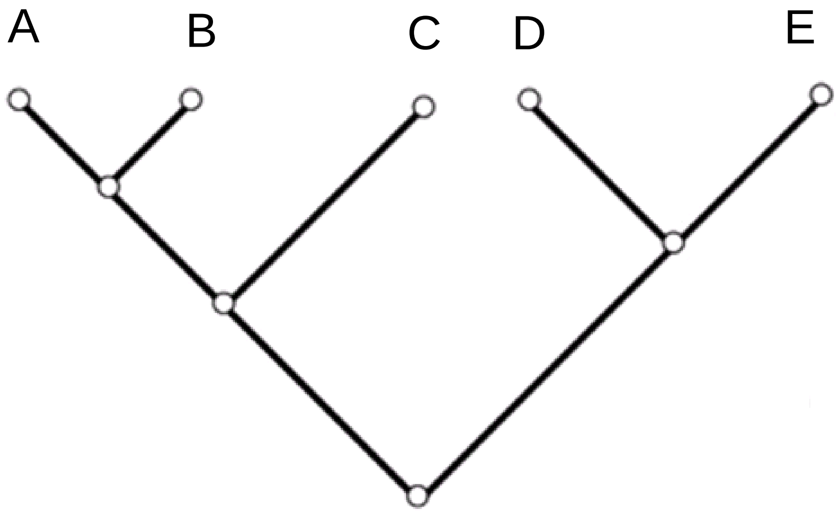
(A , B);



((A , B) , C);



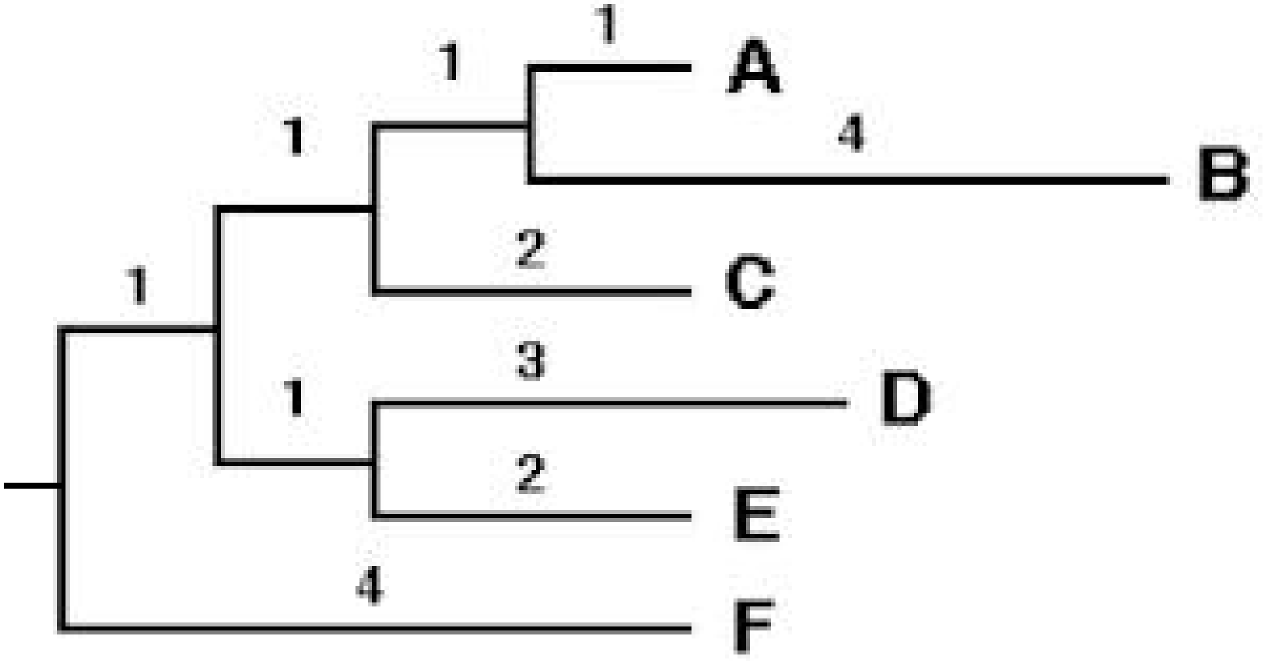
(((A , B) , C) , (D , E));



- Standard computer-readable format
- It is based on nested brackets, commas and a terminal semicolon

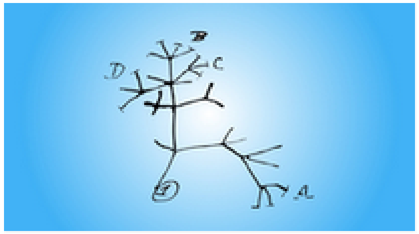
Phylogenetic trees: The Newick format

(((((A:1 , B:4):1 , C:2):1) , (D:3 , E:2):1):1 , F:4);



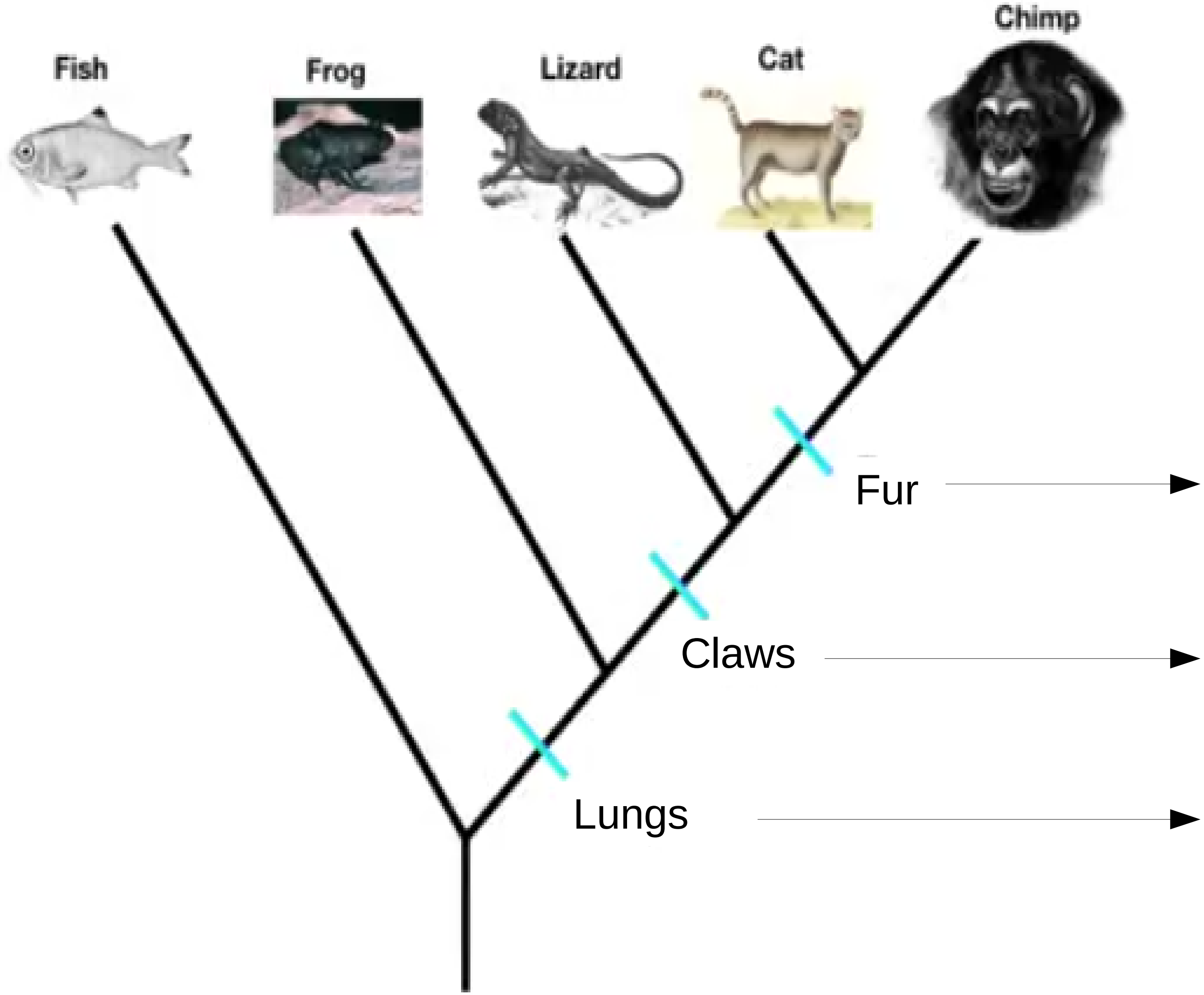
- In the Newick format the branch lengths can be indicated

Reconstructing a tree using present-day data



- Homology
- Homologous alignment characters

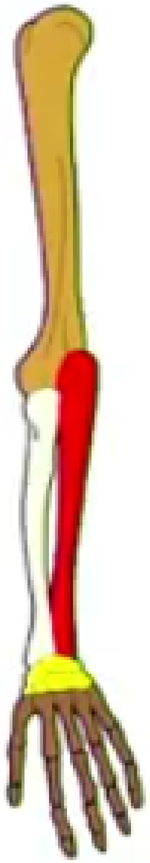
Reconstructing a tree using present-day data: The basic idea



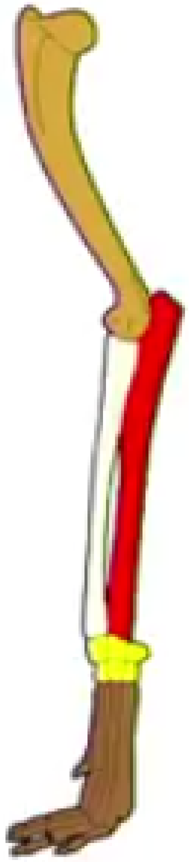
- Trying to group animals that share the largest number of derived traits

- From this point the organisms have lungs, claws and fur
- From this point the organisms have lungs, claws
- From this point the organisms have lungs

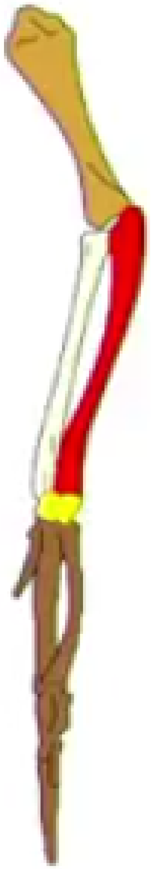
Reconstructing a tree using present-day data: Homology



Human



Dog

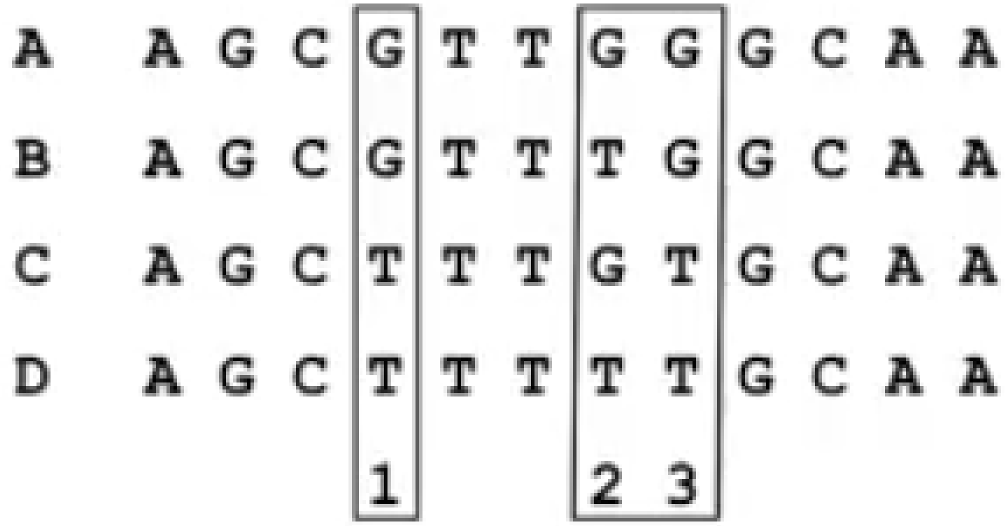


Bird

- Homologous trait means the organisms have derived from a common ancestor

- Human, dog, bird have derived from a common ancester and they all have the same bone structure that is a homologous feature

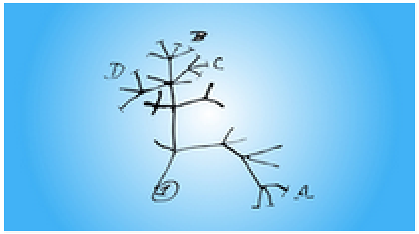
Reconstructing a tree using present-day data: Molecular phylogenetic



A	A	G	C	G	T	T	G	G	G	C	A	A
B	A	G	C	G	T	T	T	G	G	C	A	A
C	A	G	C	T	T	T	G	T	G	C	A	A
D	A	G	C	T	T	T	T	T	G	C	A	A
				1			2	3				

- Homology means the homologous characters
- Homologous characters mean columns in alignment

Evolutionary implications of gene orthology



- Orthologs and paralogs
- Prediction of orthologous sequences

Orthologs and paralogs

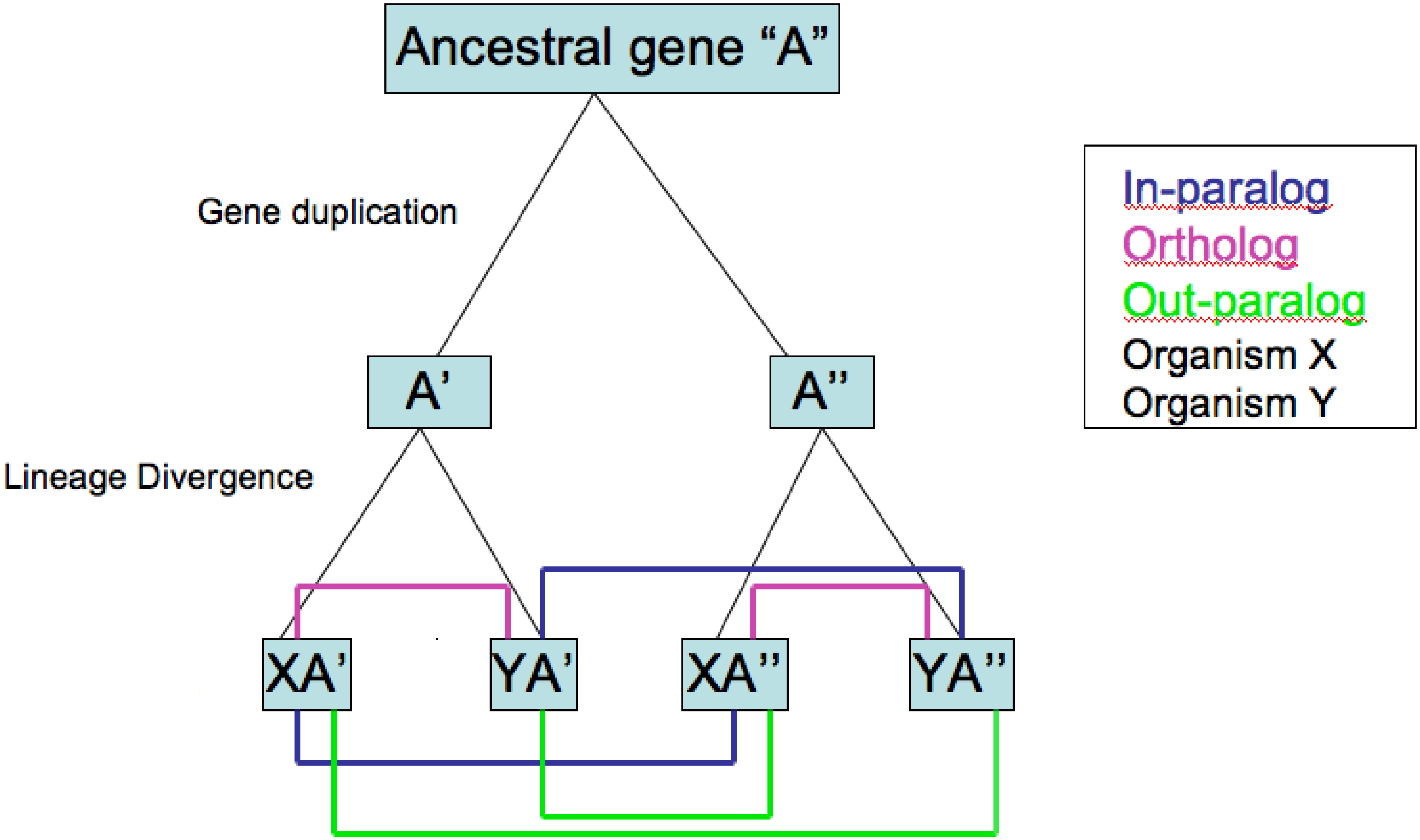
- **(In-)Paralogs** are genes related by duplication within a genom

- **Out-Paralogs** are in different species, and derived from a more ancient shared duplication event

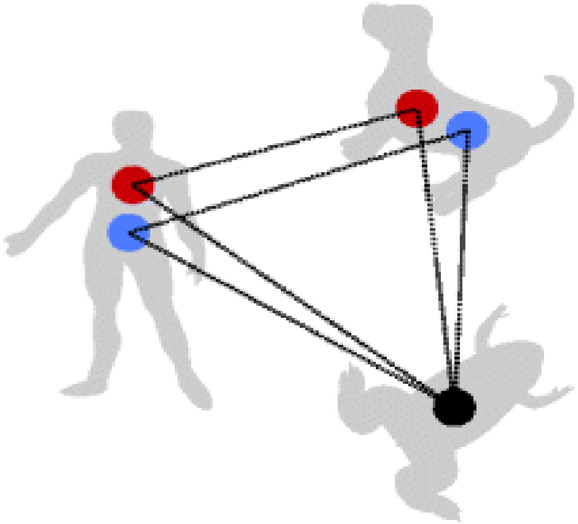
- **Orthologs** are genes in different species that evolved from a common ancestral gene.

- Paralogs evolve new functions, even if these are related to the original one
- Normally, orthologs retain the same function in the course of evolution

Diagram depicting evolutionary relationship between orthologs, out-paralogs and in-paralogs

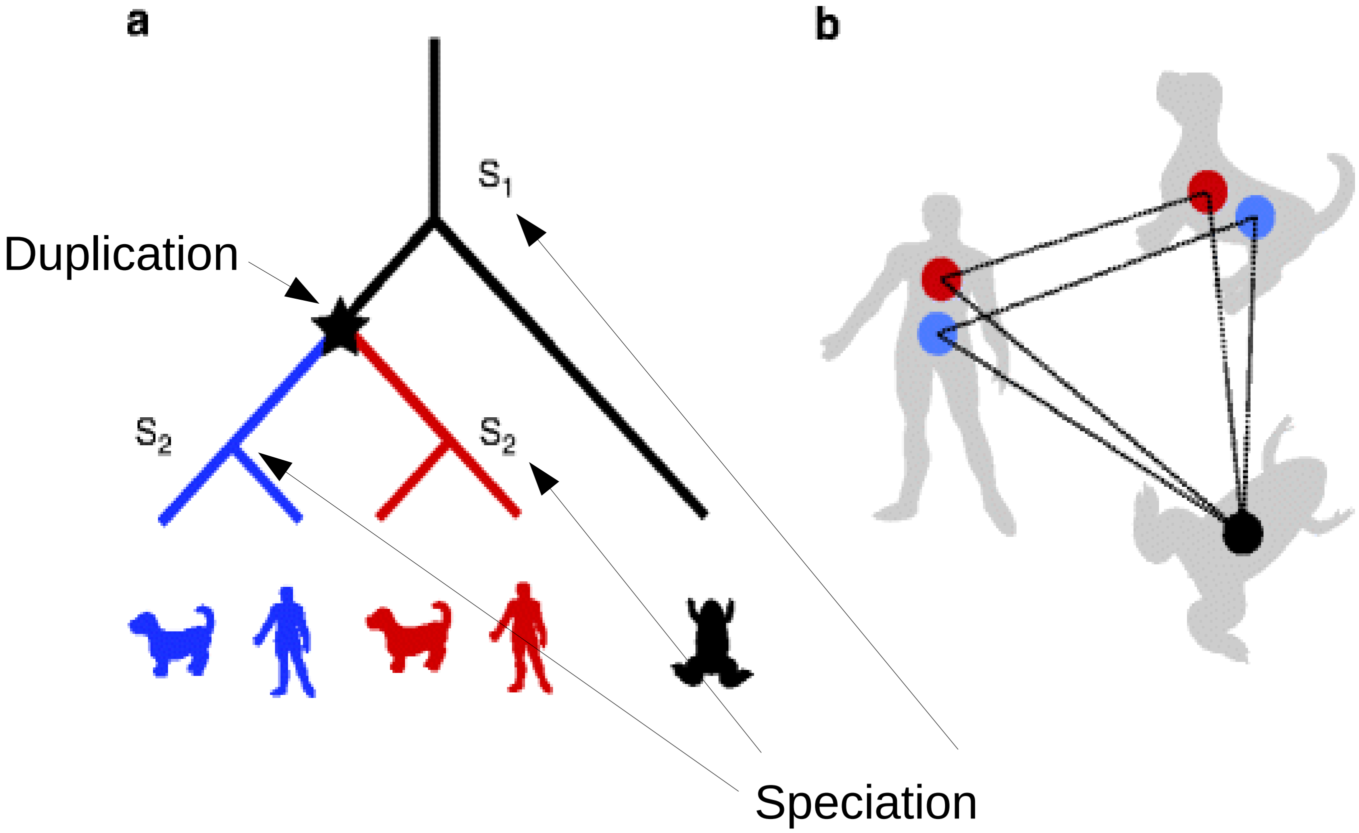


- A gene in human, frog and dog
- In frog there is one copy
- But in dog and human there are two copies



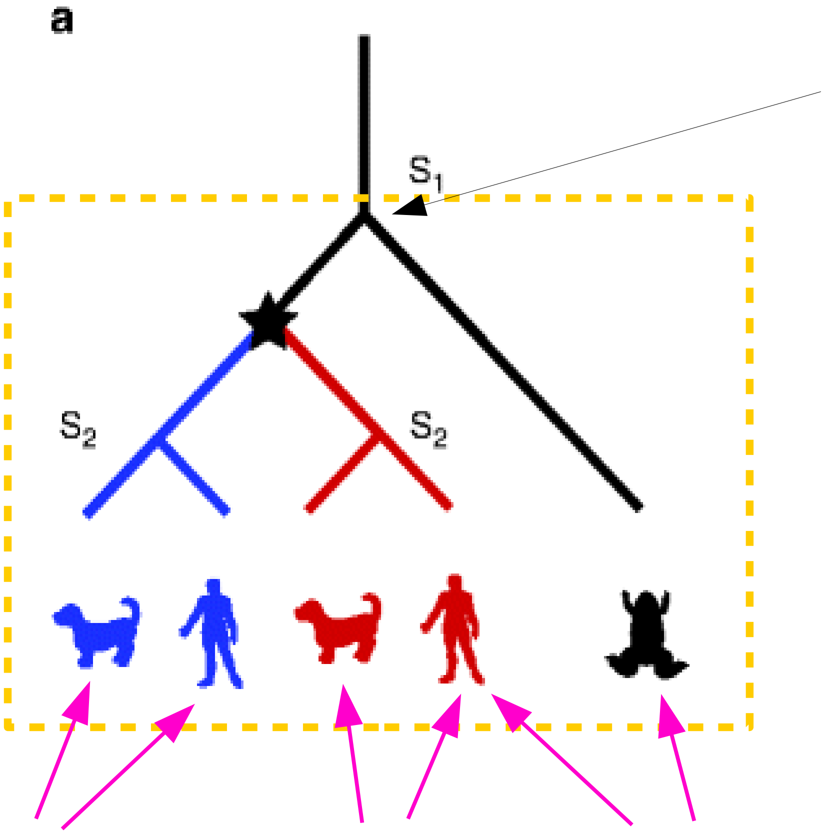
What's going on here?

Look at the history of these five genes which is depicted in a phylogenetic tree



Duplication

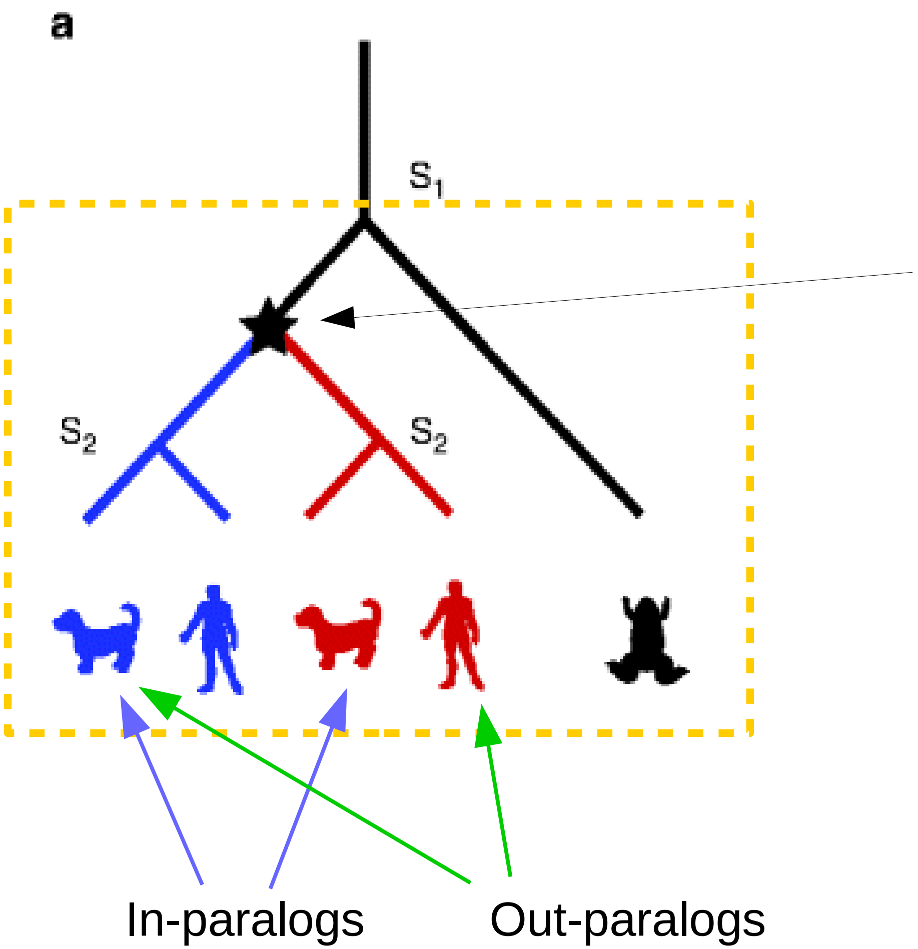
Speciation



Orthologous Orthologous Orthologous

- In ancestral vertebrata: one gene
- All genes are derived from it
- This clade includes the five genes

Orthologous: Genes related by speciation



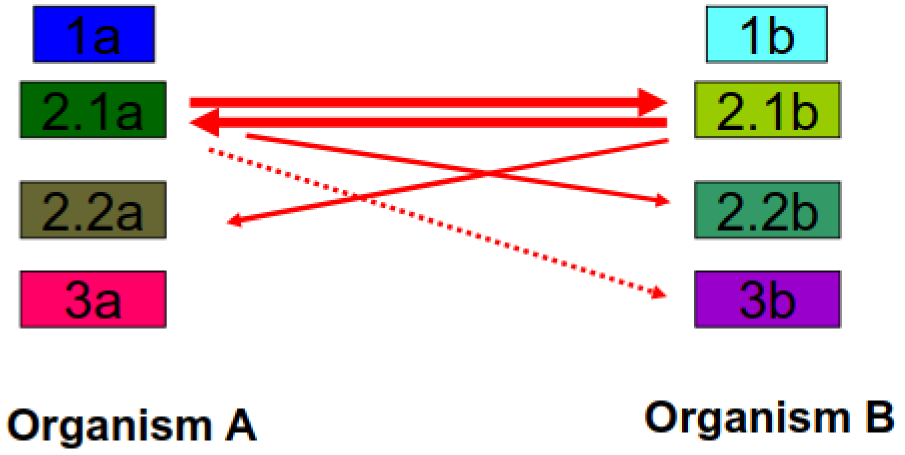
The duplication must have happened within the clade in question

In-paralogs

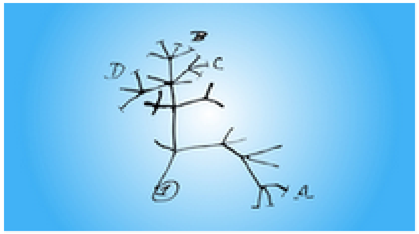
Out-paralogs

paralogs: Genes related by duplications

- How to detect orthologous genes?
  - Easy way: **R**eciprocal **B**est **H**it (RBH)

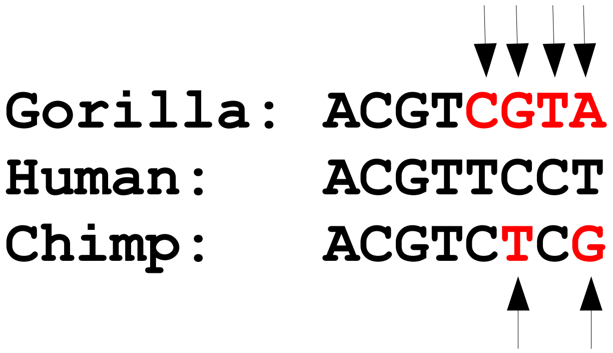


Phylogenetic tree building methods



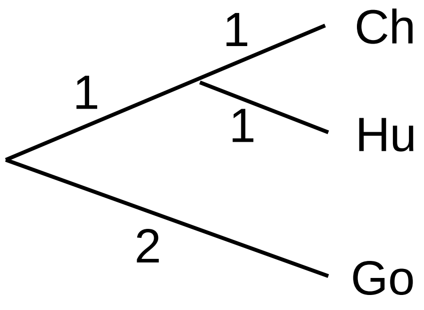
- Distance-based method
- Maximum parsimony method
- Maximum likelihood method

Phylogenetic tree building methods: Distance matrix



```
Gorilla: ACGTCGTA
Human:   ACGTTCCT
Chimp:   ACGTCTCG
```

	Go	Hu	Ch
Go	-	4	4
Hu		-	2
Ch			-



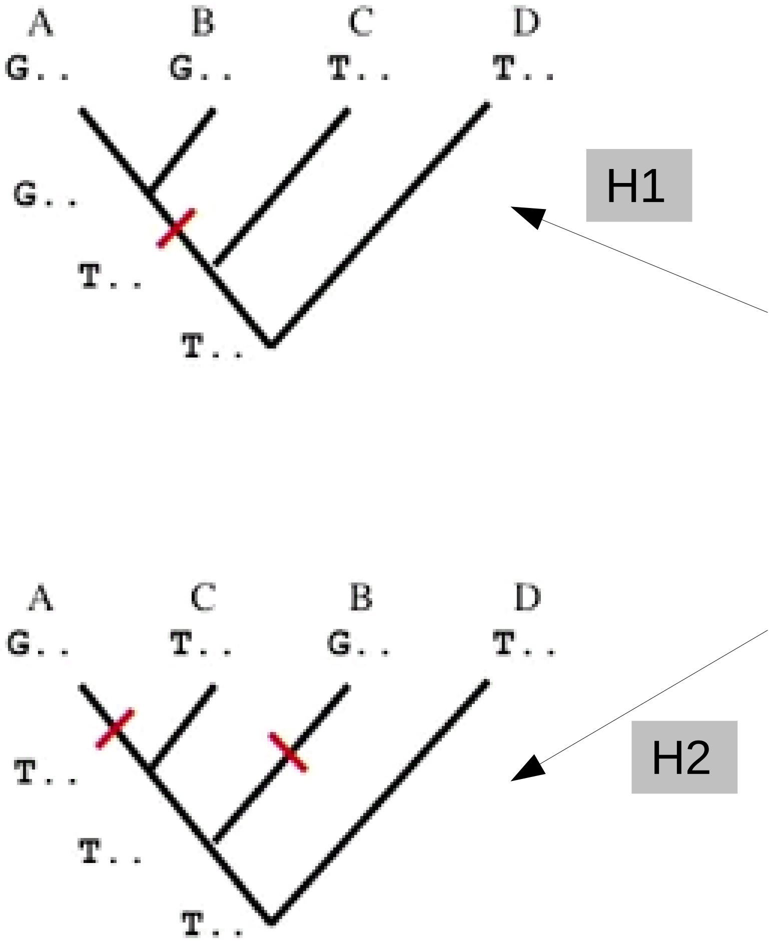
- Count the number of substitutions among the sequences

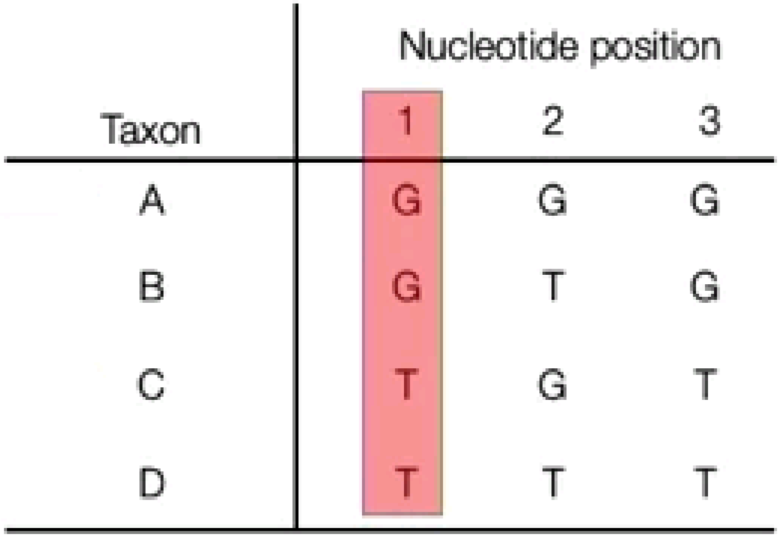
- Write these number in a matrix to get the distance matrix

- According to the matrix the phylogenetic tree can be built

Phylogenetic tree building methods: Maximum Parsimony

- Maximum parsimony: choose the simplest possible hypothesis

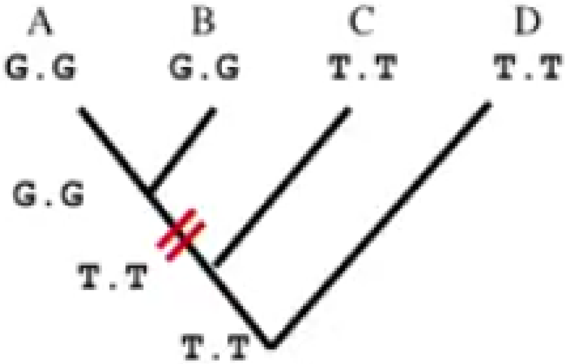


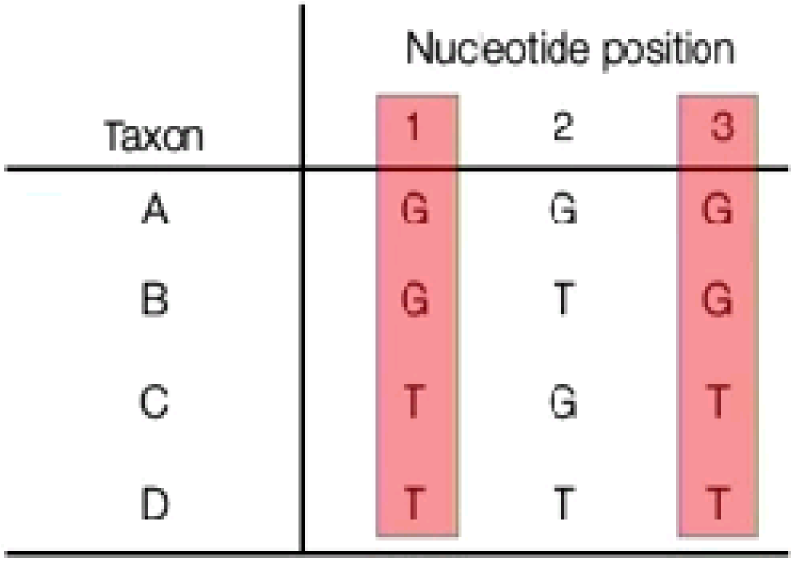


Taxon	Nucleotide position		
	1	2	3
A	G	G	G
B	G	T	G
C	T	G	T
D	T	T	T

- H1 is the simplest possible hypothesis
- The tree has 1 mutation

Phylogenetic tree building methods: Maximum Parsimony

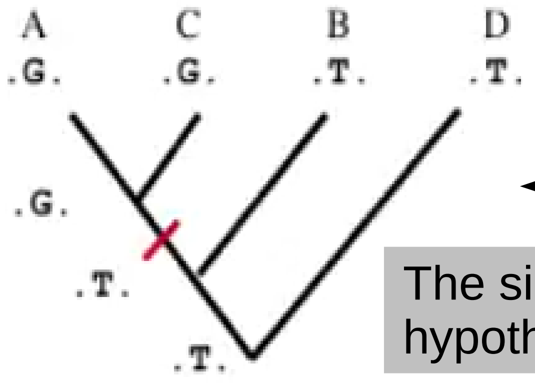




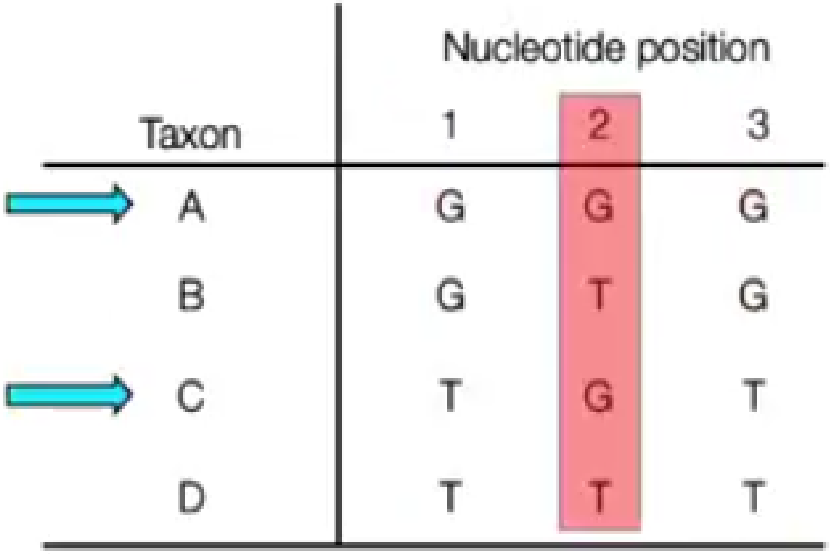
Taxon	Nuceotide position 1	2	3
A	G	G	G
B	G	T	G
C	T	G	T
D	T	T	T

- This is the same as the first column
- The tree has 2 mutations

Phylogenetic tree building methods: Maximum Parsimony



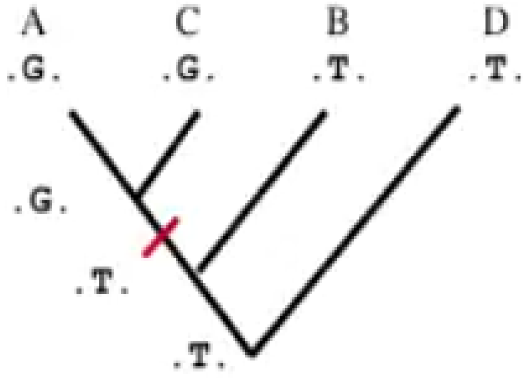
The simplest possible hypothesis



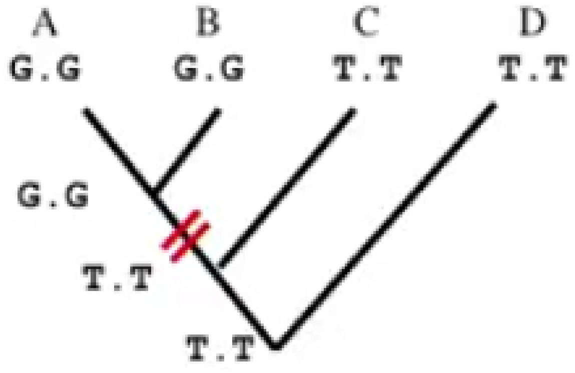
Taxon	Nucleotide position 1	2	3
A	G	G	G
B	G	T	G
C	T	G	T
D	T	T	T

The best hypothesis differs from the others

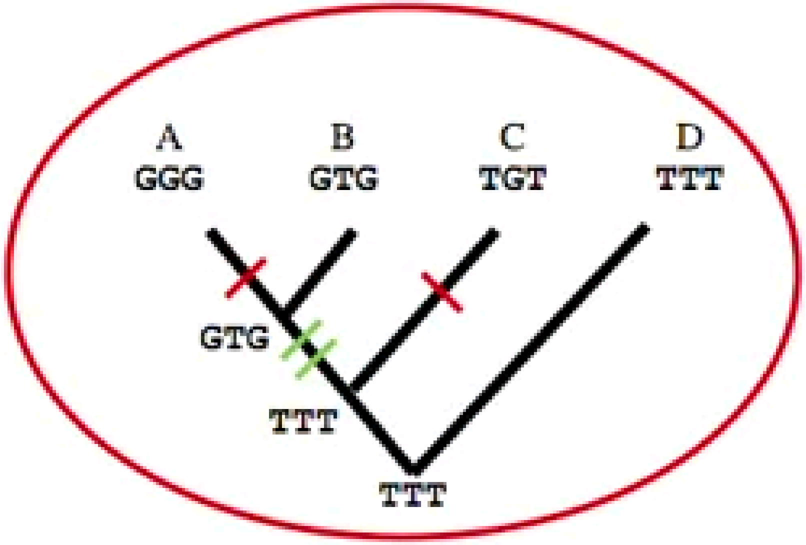
Conflict:



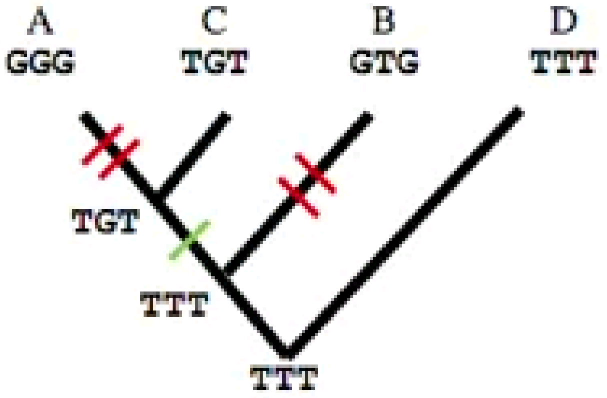
?



Phylogenetic tree building methods: Maximum Parsimony



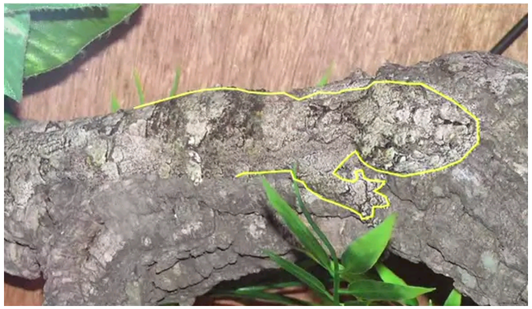
Total length of tree: 4



Total length of tree: 5

- The best tree: the smallest number of the mutations
- Count the total number of the mutations for the two versions
- Compare them and choose the smaller
- In this case we have to reject the best hypothesis at position 2 in order to get the best tree

Detection of positive/negative molecular selection



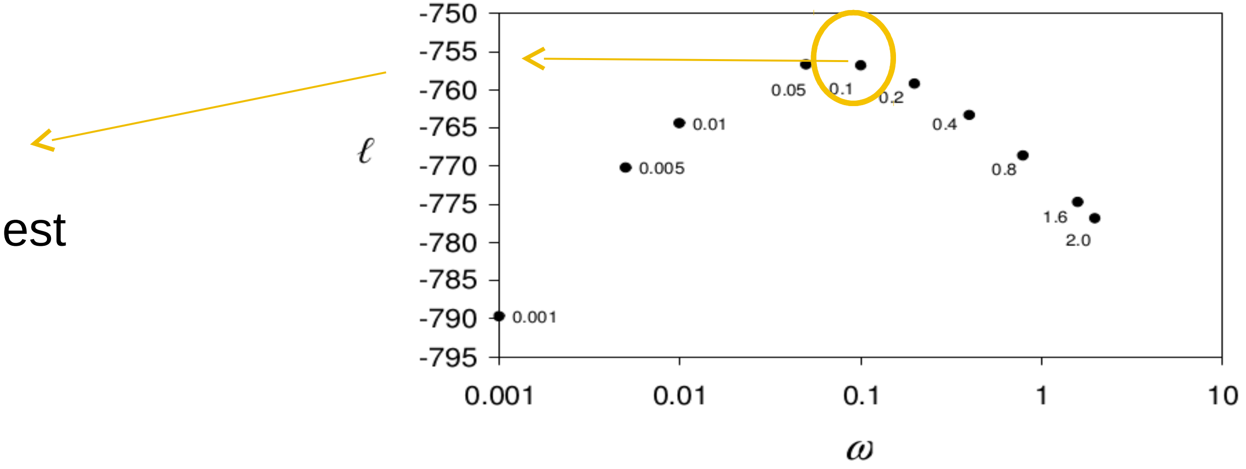
- Substitutions
- Detection of molecular selection
- The levels of detection

Detection of positive/negative molecular selection: substitutions

- **The main parameter: ω (omega)**

ω = dN/dS

- **dN:** rate of non-synonymous substitutions
- **dS:** rate of synonymous substitutions



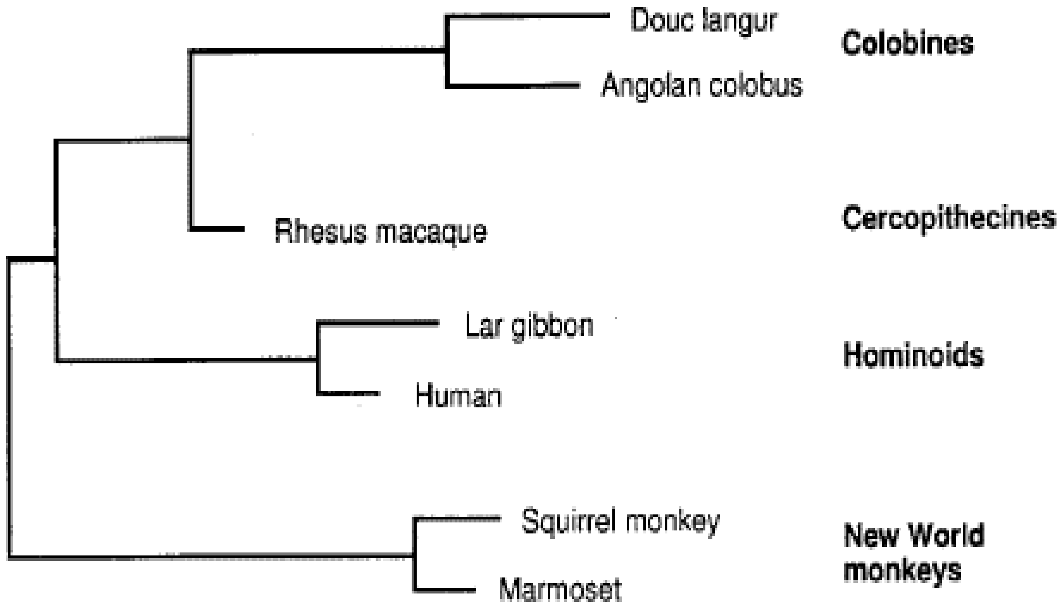
Select the smallest likelihood value

- **ω = 1 → neutral selection**
- **ω < 1 → negative selection**
- **ω > 1 → positive selection**

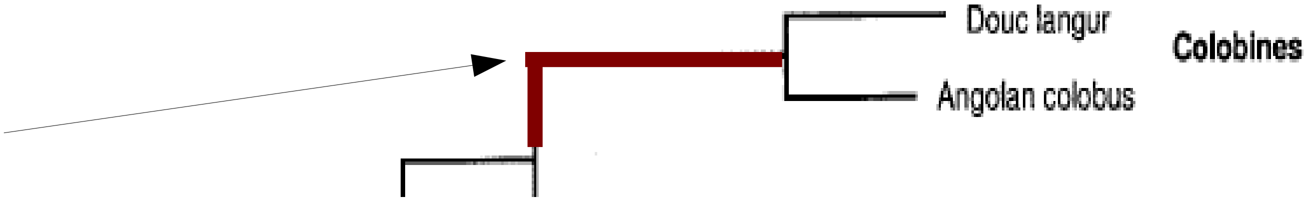
Detection of positive molecular selection: Models

- **Null-model:**
  - Global average omega value
  - It describes the evolution of the entire tree

  ω = 0.34 → The given gene spent the overwhelmed majority of time under negative selection



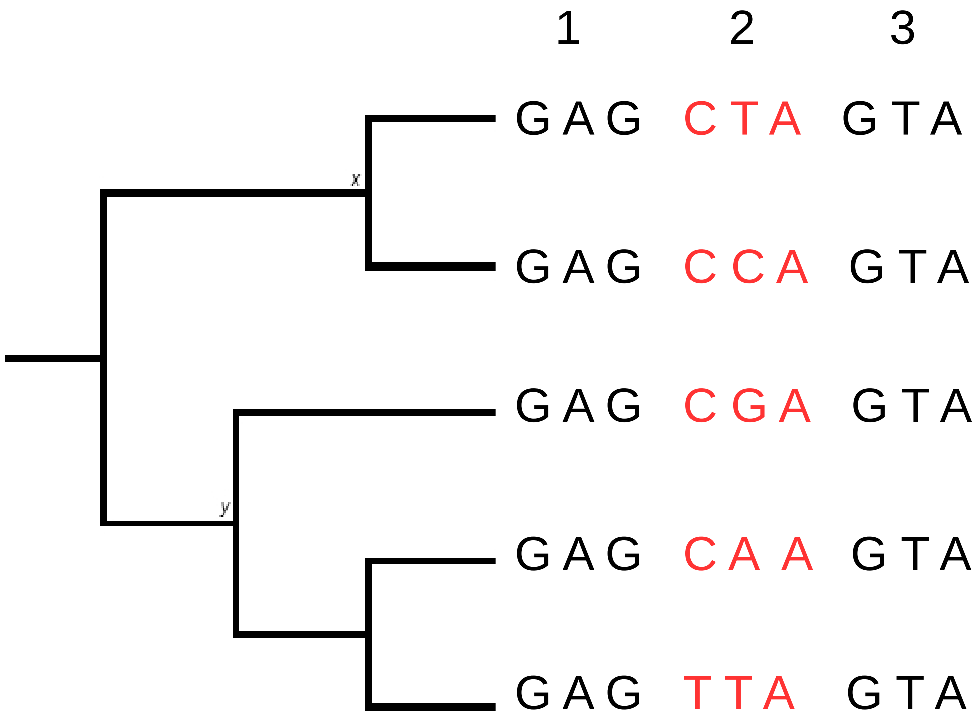
- **Branch-model:**
  - Partial omega value
  - Global omega value
  - It describes the evolution of a given branch



ω = 2.21 → Positive selection can be observed on this branch

Detection of positive molecular selection: Models

- **Site-model:**
  - Global average omega value for each codon independently
  - It describes the evolution of each codon



1 2 3

GAG CTA GTA

GAG CCA GTA

GAG CGA GTA

GAG CAA GTA

GAG TTA GTA

ω_1 = 1 → There is not any selection

ω_2 = 2.81 → There is positive selection

ω_3 = 1 → There is not any selection

Detection of positive molecular selection: Likelihood Ratio Test

- We have to declare hypotheses to calculate some kind of statistics

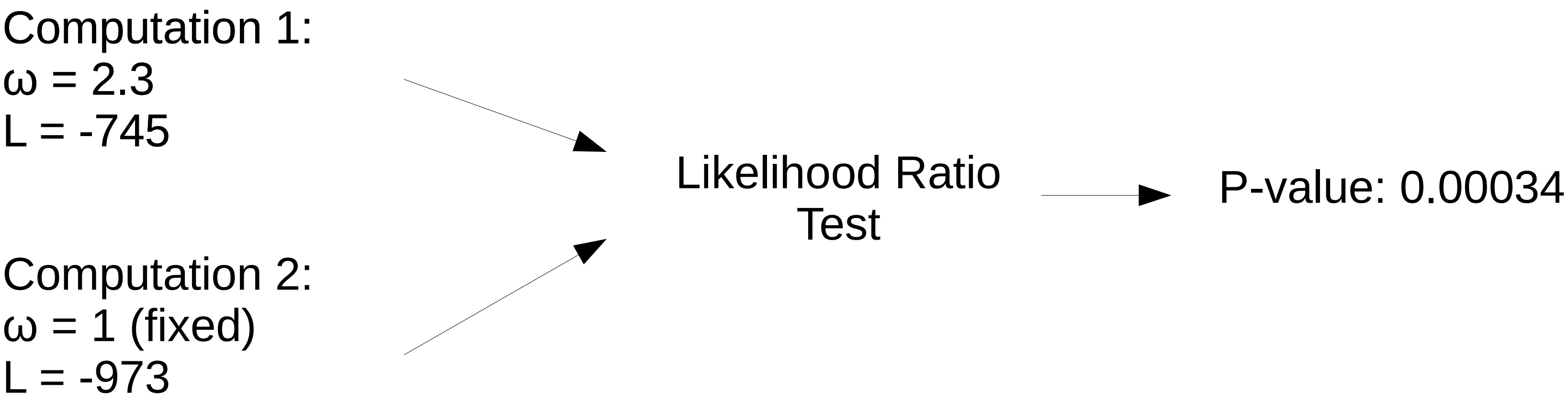
Hypothesis 1:

- ω < 1 or ω > 1
- Likelihood value 1

Hypothesis 0:

- ω = 1
- Likelihood value 2

- Using the two likelihood values we can decide whether the selection is statistically significant or not



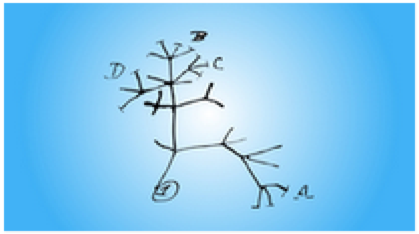
Computation 1:
ω = 2.3
L = -745

Computation 2:
ω = 1 (fixed)
L = -973

Likelihood Ratio Test

P-value: 0.00034

Tutorial



- Prepare and view trees in FigTree viewer
- Prepare distance matrix
- Computational molecular evolution

Please download the files below:

- http://matyaspajkos.web.elte.hu/Evolutionary_analyses
  - Windows: codeml.zip
  - Linux: codeml_linux.zip