

Phylogenetic reconstruction

Anne Friedrich
 Laboratory of Integrative Bioinformatics and Genomics
 IGBMC, Strasbourg, France
friedric@igbmc.fr

Phylogenetic reconstruction

- ✓ Introduction: basic concepts
- ✓ Principal methods for tree reconstruction
 - Distance based methods
 - Character based methods
- ✓ Evaluation of the reliability of a tree
- ✓ The limits of phylogeny
- ✓ Some programs

Phylogenetic reconstruction

- ✓ **Introduction: basic concepts**
- ✓ Principal methods for tree reconstruction
 - Distance based methods
 - Character based methods
- ✓ Evaluation of the reliability of a tree
- ✓ The limits of phylogeny
- ✓ Some programs

Introduction

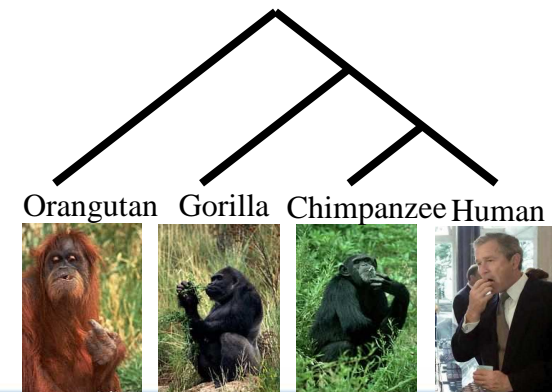
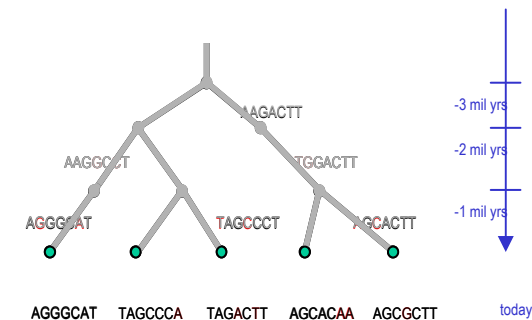
- ✓ Phylogenetics: the study of **evolutionary** relatedness among **organisms**
 - ❑ Relationships represented by a hierarchical, tree-like structure
 - ❑ Originally based on physical similarities and differences (traits or **characters**)
 - ❑ Now, mostly using gene / protein sequences (Multiple Sequence Alignment)

- ✓ What for ?

- => Reconstruction of evolutionary history of a gene family

- => Reconstruction of the evolutionary relationships between species : living (**extant**) and dead (**extinct**). eg : tree of life

- => Classification of new species. eg : viral strain



Bacteria

Archaea

Eucarya

Proteobacteria

E. coli(3), *H. infl.*, *H. pylori*(2),
C. jeju., *R. pro.*, *R. con.*, *V. chol.*,
N. men(2), *X. fasti.*, *P. aeru.*,
Buchnera sp.(2), *M. loti*, *P. multo*
C. cresc., *A. tumef.*(2), *P. aerogi.*
Brucella meli., *Rals. sola.*, *Y. pes.*
S. typhi, *Sal. typ.*, *Sino. meli.*
Xant. axo., *Xant. cam.*

Gram-positive

B. subtilis, *B. halo.*, *U. urea.*,
M. geni., *M. pneumo.*, *M. pulmo.*,
M. tuber.(2), *M. leprae*, *Stre. pyo.*(2)
L. lactis, *S. aureus*(3), *Clostr. acet.*
Clostr. perf., *Coryn. Glut.*, *Lis. inno.*,
Lis. mono., *Staph. au.*, *Stre. aga.*,
Stre. pneu. (2), *Streptom. coe.*, *T. ten.*

Cyanobacteria

Synecho., *Nostoc a.*,
Thermo. elong.,

Chlamydiae

C. tracho.(2), *C. pneumo*(3)

Flavobacteria

Bacteroides

Chloro. Tepi.

Spirochetes

T. pal., *B. burgdo.*

D. radiodurans

Green non sulfur bacteria

T. maritima

A. aeolicus

Crenarchaeota

A. pernix, *S. solfa.*,
S. toko., *Py. aero*

Euryarchaeota

A. fulgidus

Methanosarcina (2)

M. thermoauto.

M. jannaschii

Pyrococcus(3)

Halobact.

Thermoplasma (2)

Methanopyrus k.

Animals

H. sapiens
M. musculus
D. melanogaster
A. gambiae
C. elegans

Fungi

S. cerevisiae
S. pombe

Plants

A. thaliana
O. sativa (2)

Ciliates

Flagellates

Trichomonads

Microsporidia

Ence. cuni.

Diplomonads

Tree of life (based on 16S rRNA)

Homology-Orthology-Paralogy

Homology :

2 genes are homologous if they have a common ancestor

Orthology :

2 genes are orthologous if they diverged after a speciation event

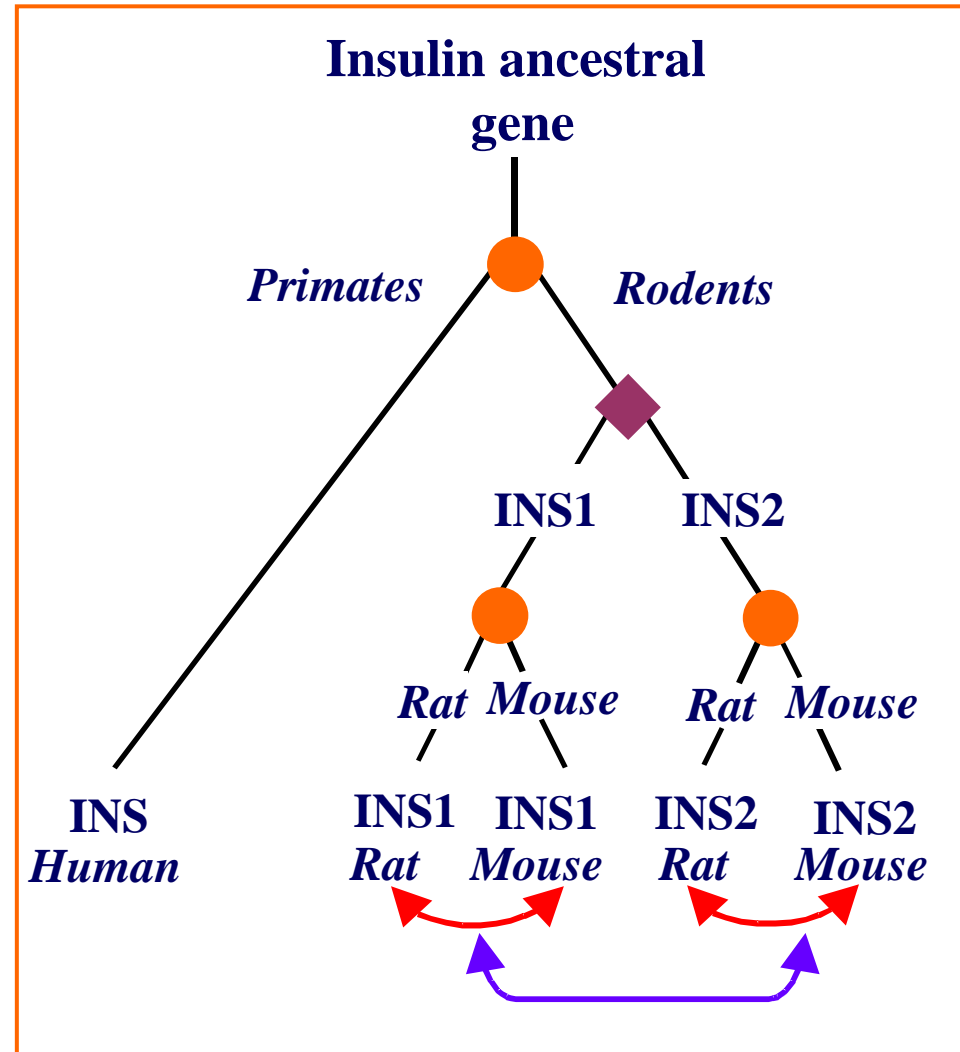


Paralogy :

2 genes are paralogous if they diverged after a duplication event



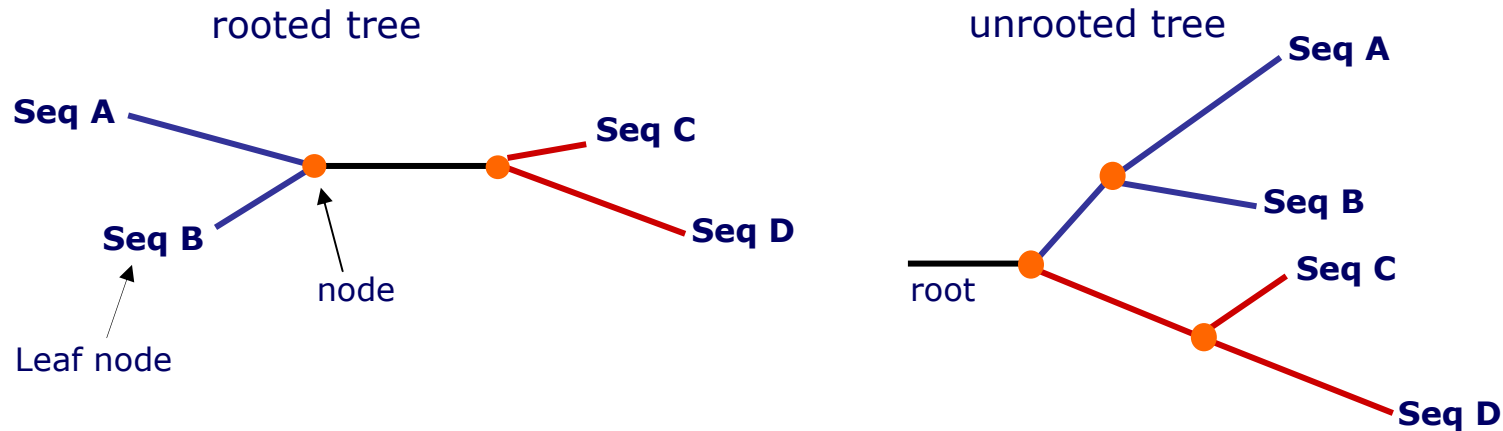
- Speciation
- ◆ Duplication



Basic concepts: phylogeny

A phylogenetic tree is characterised by :

- its topology (branching pattern)
- the lengths of the branches (possibly)



Node: represent a Taxonomic Unit TU (species, population, gene...) either existing or ancestor

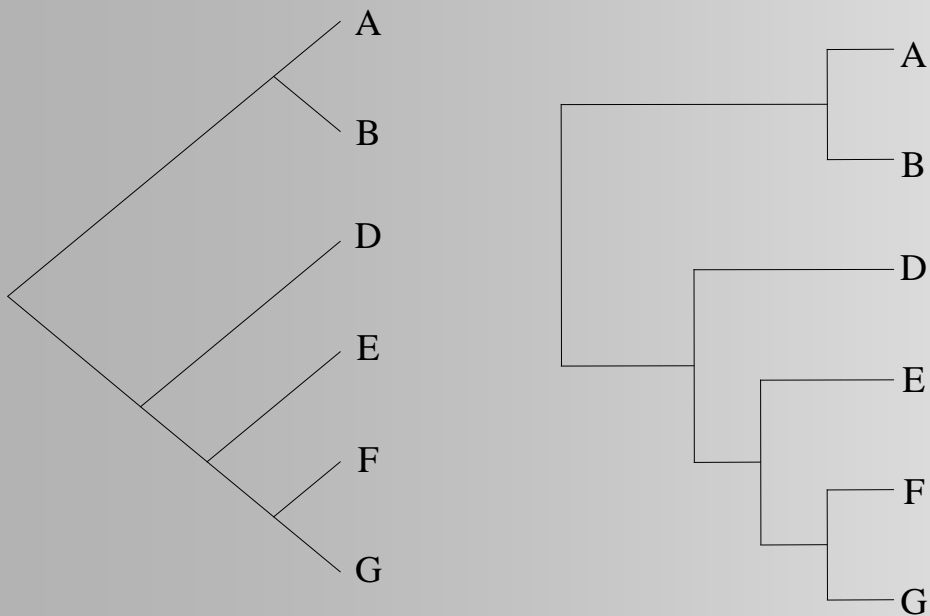
Branch: defines the relationship between the TUs (descent, ancestry)

Root: common ancestor of all taxonomic units on tree

A tree can be rooted or not

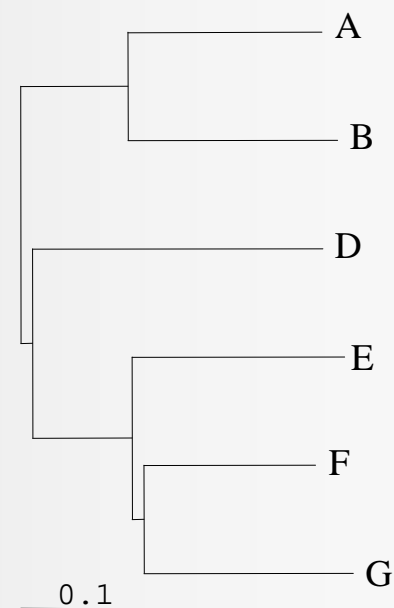
Cladograms

(branch lengths are not meaningful)



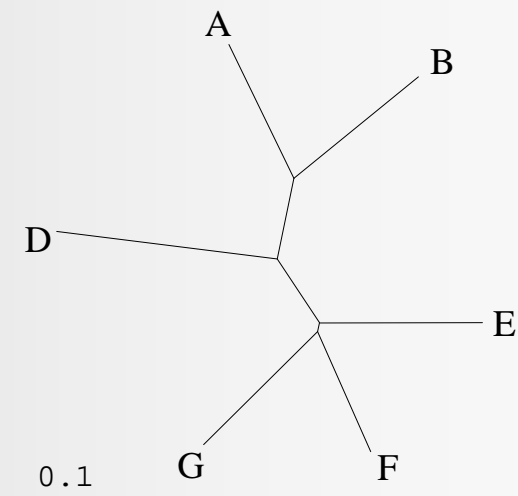
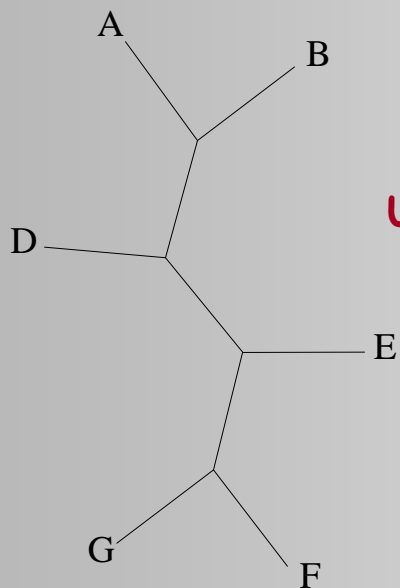
Phylogram

(branch lengths are proportional to number of change)



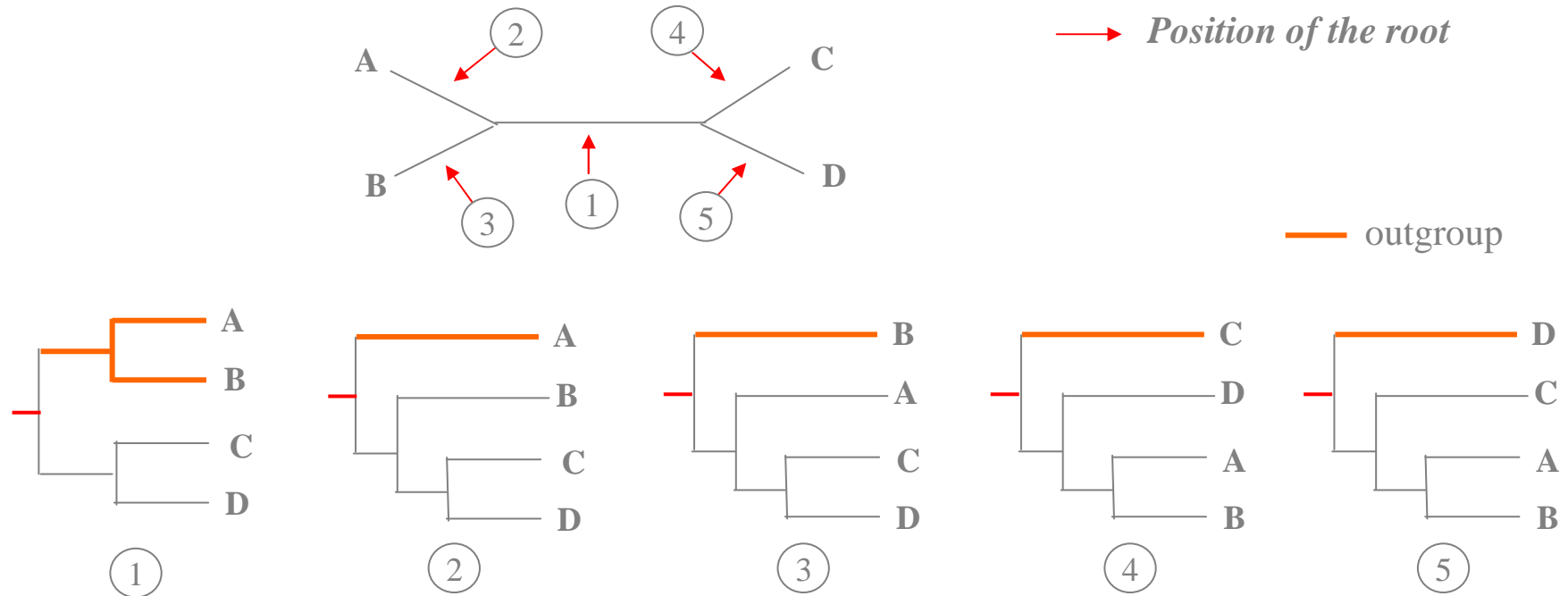
Unrooted trees (networks)

(branch lengths are not meaningful)



Basic concepts: phylogeny

For an unrooted tree, there are several possible rooted trees



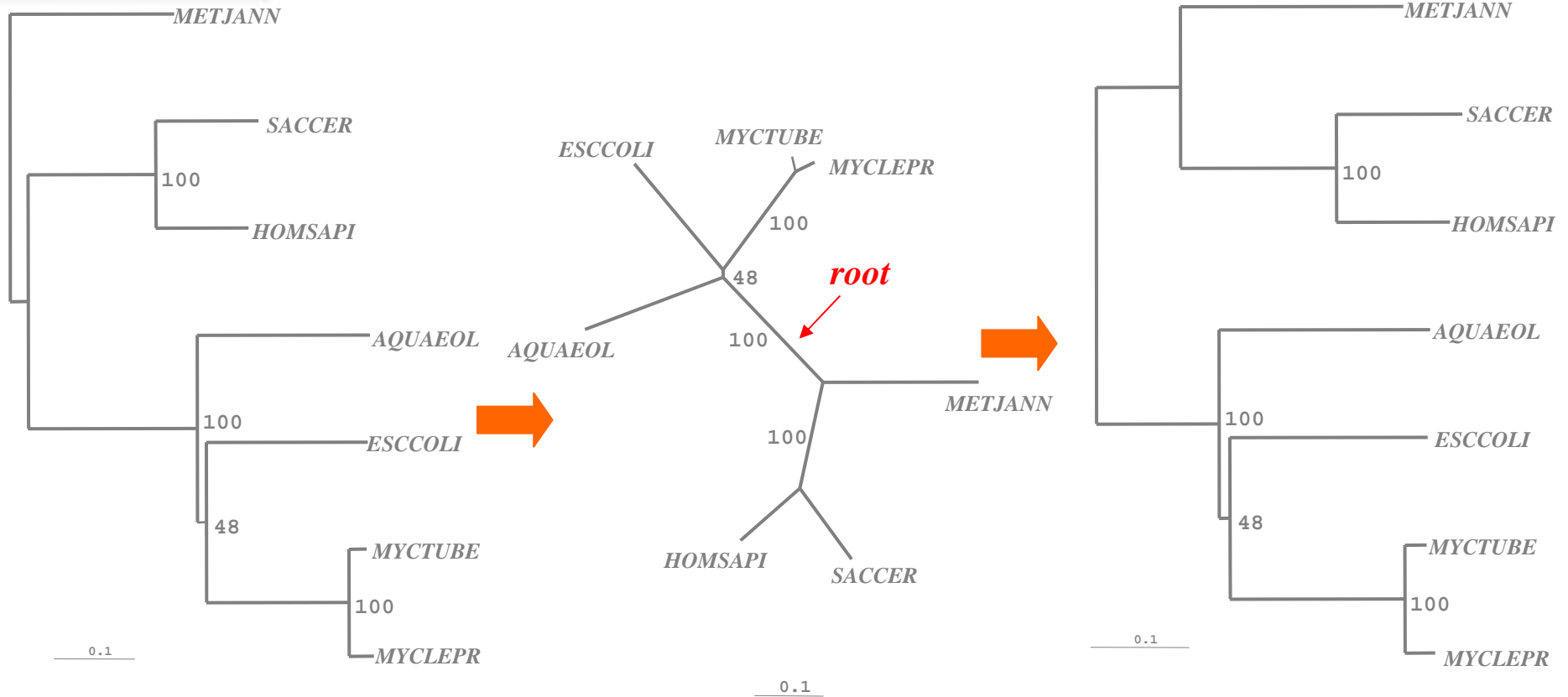
In most tree reconstruction programs, the position of the root is chosen arbitrarily :

- « midpoint rooting » (root placed in the centre of the longest branch)
- « outgroup rooting »

The user can define the sequence(s) that can be used as an **outgroup** to root the tree.

The outgroup sequence should be distantly related to all other sequences.

Basic concepts: phylogeny



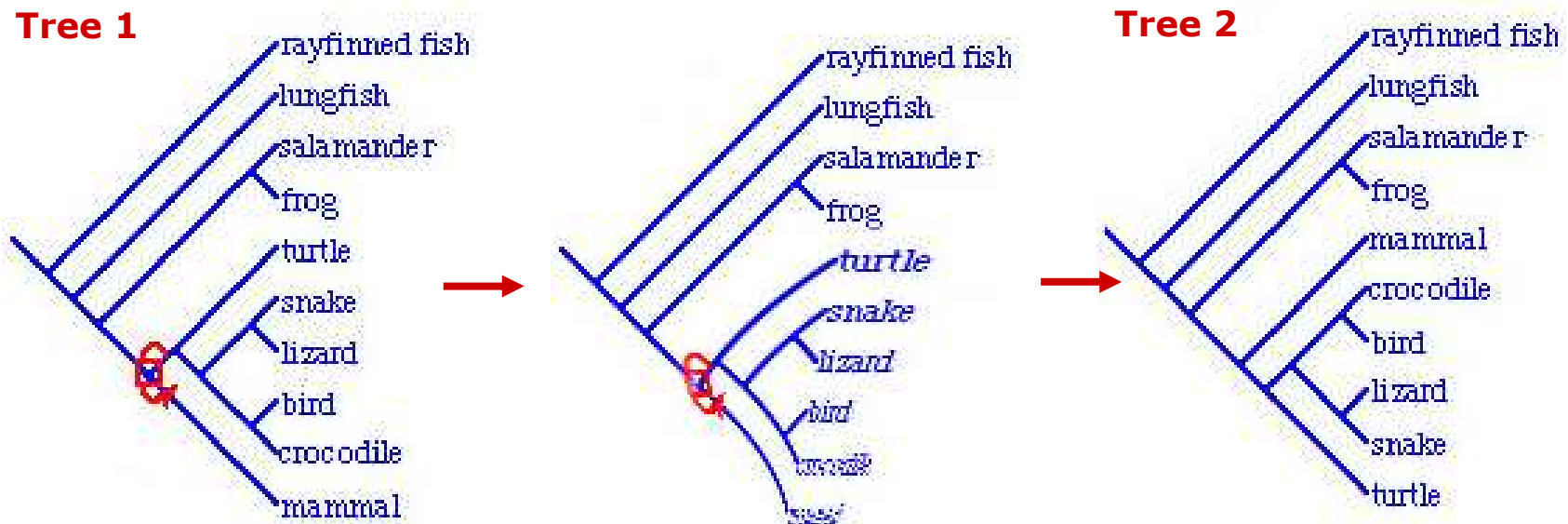
Rooted tree

Unrooted tree

Rooted tree

Basic concepts: branch order

The order of the branches belonging to the same node is not important. Rotating branches at a node does not change the topology of the tree.



Tree 1 = Tree 2

Phylogenetic reconstruction

- ✓ Introduction: basic concepts
- ✓ **Principal methods for tree reconstruction**
 - **Distance based methods**
 - **Character based methods**
- ✓ Evaluation of the reliability of a tree
- ✓ The limits of phylogeny
- ✓ Some programs

Molecular phylogenetics

- ✓ DNA, RNA, and protein sequences can be considered as phenotypic traits
- ✓ Molecular phylogenetics attempts to determine the rates and patterns of change occurring in the sequences and to reconstruct the evolutionary history of genes and organisms
- ✓ How? Several steps
 - ❑ build a multiple alignment
 - ❑ validate, edit or mask the resulting alignment
 - ❑ reconstruct the tree
 - ❑ statistically evaluate the reliability of the tree

Multiple sequence alignment

Errors in the initial alignment will lead to inaccurate trees

Alignment of 18s rRNA sequences (Morrison and Ellis, J Mol Evol, 1997)

Alignment algorithms: Pileup, ClustalW, TreeAlign, MALIGN, SAM.

Trees construction: neighbor-joining, weighted-parsimony and maximum-likelihood.

→ different alignments produced trees that were more dissimilar than did the different tree-building methods

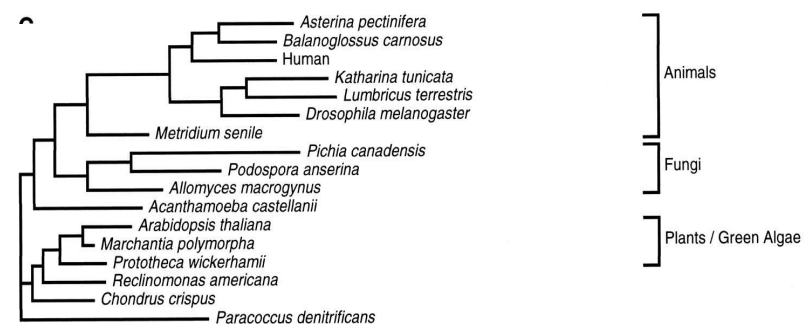
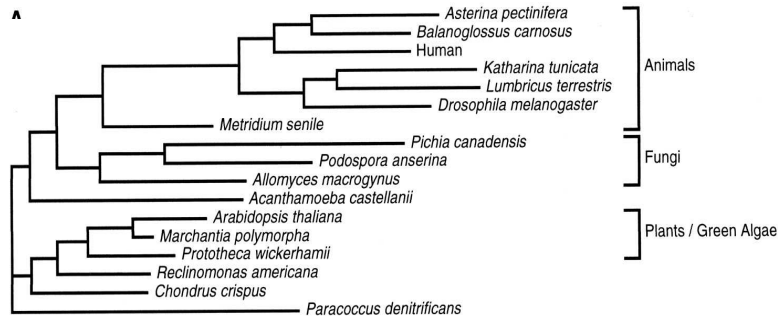
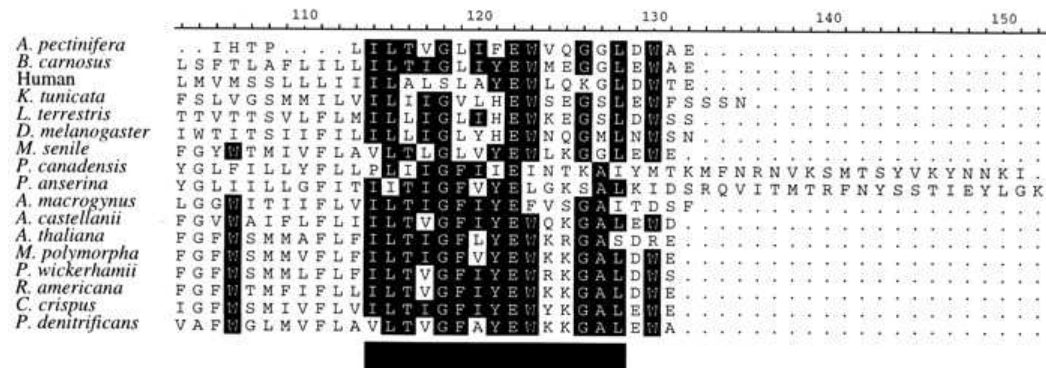
Using genomic data from seven yeast species (Wong et al, Science 2008)

build alignments using 7 different alignment programs and then estimate phylogeny (using maximum parsimony)

→ 46.2% of the 1502 genes had one or more differing trees depending on the alignment procedure used

Alignment masking

Remove or mask columns with gaps and unreliable regions
E.g. Gblocks (Castresana, 2000)



Multiple sequence alignment

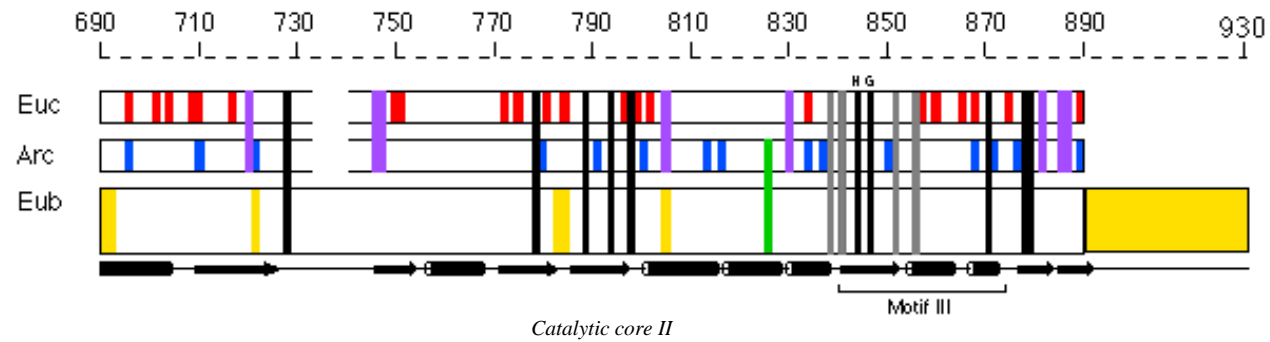
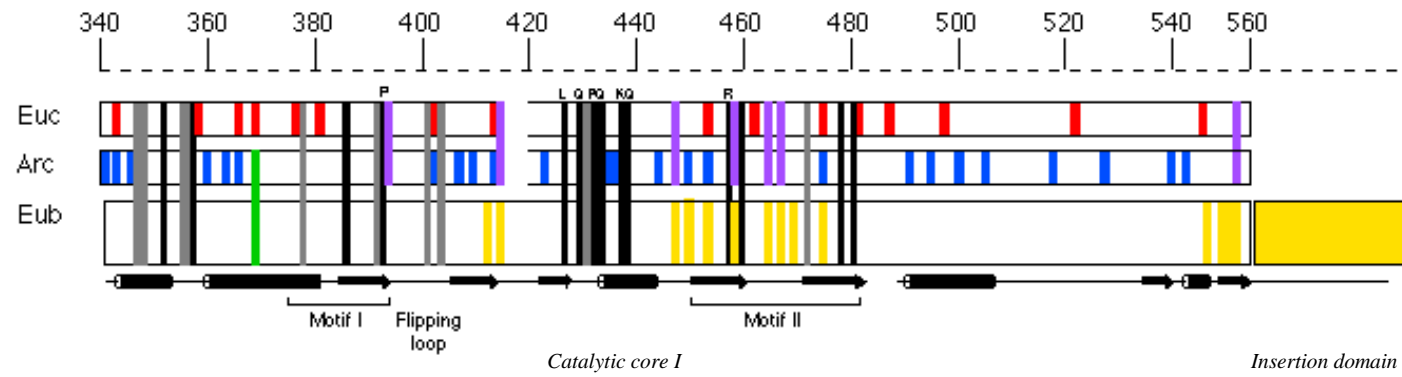
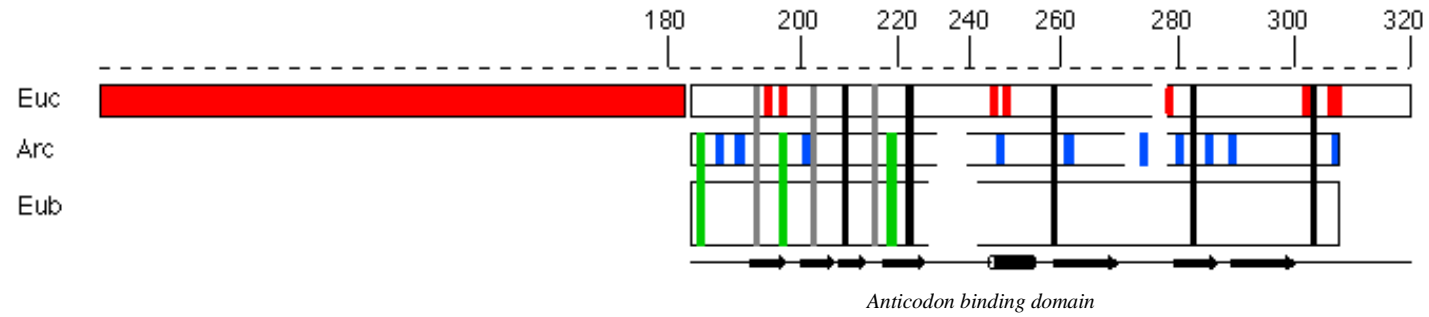


The quality of the multiple alignment is crucial for the quality of the tree !!

The trees can vary depending on the region of the alignment selected.



AspRS

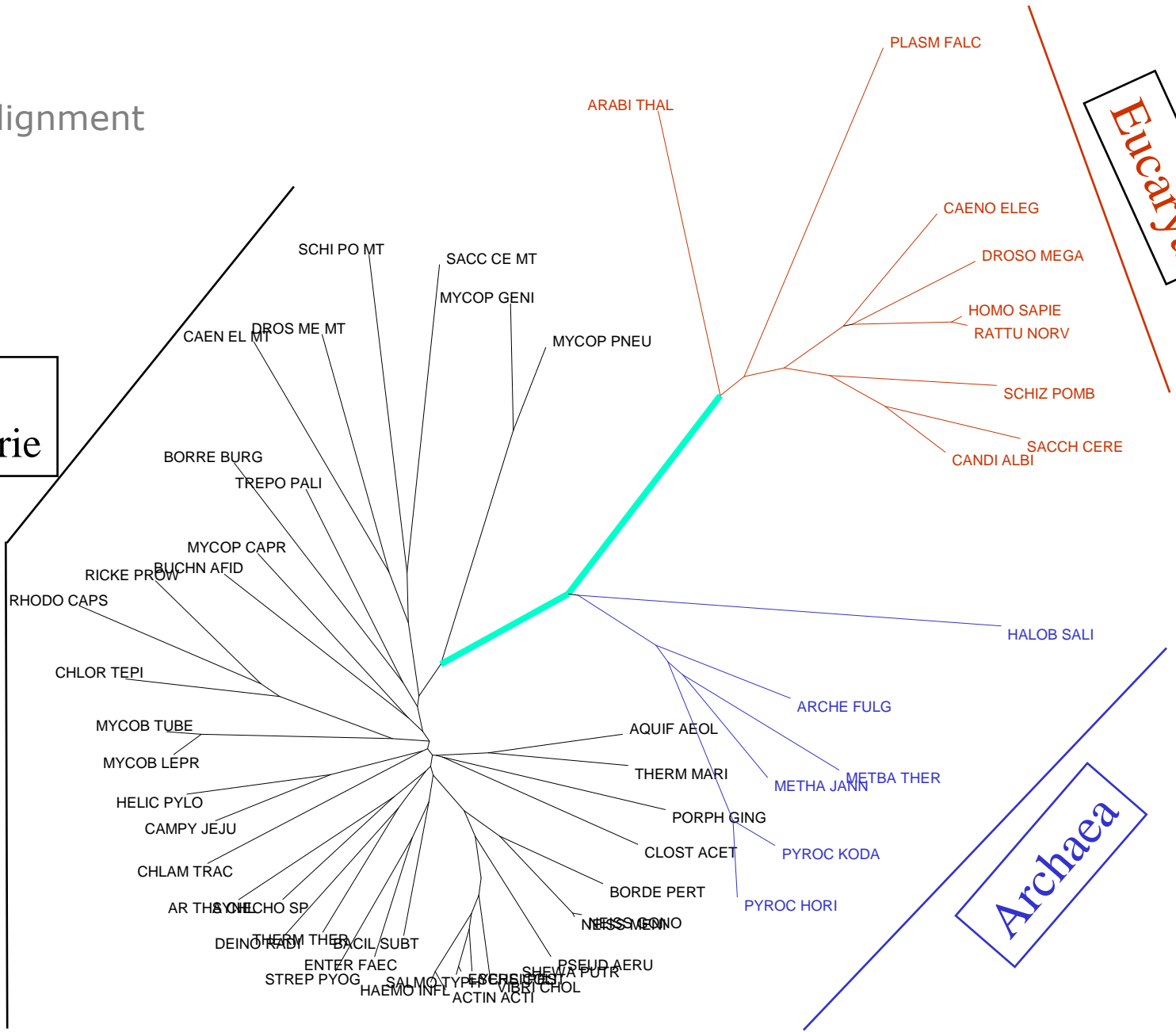


Whole alignment

Bacteria +
Mitochondrie

Eucarya

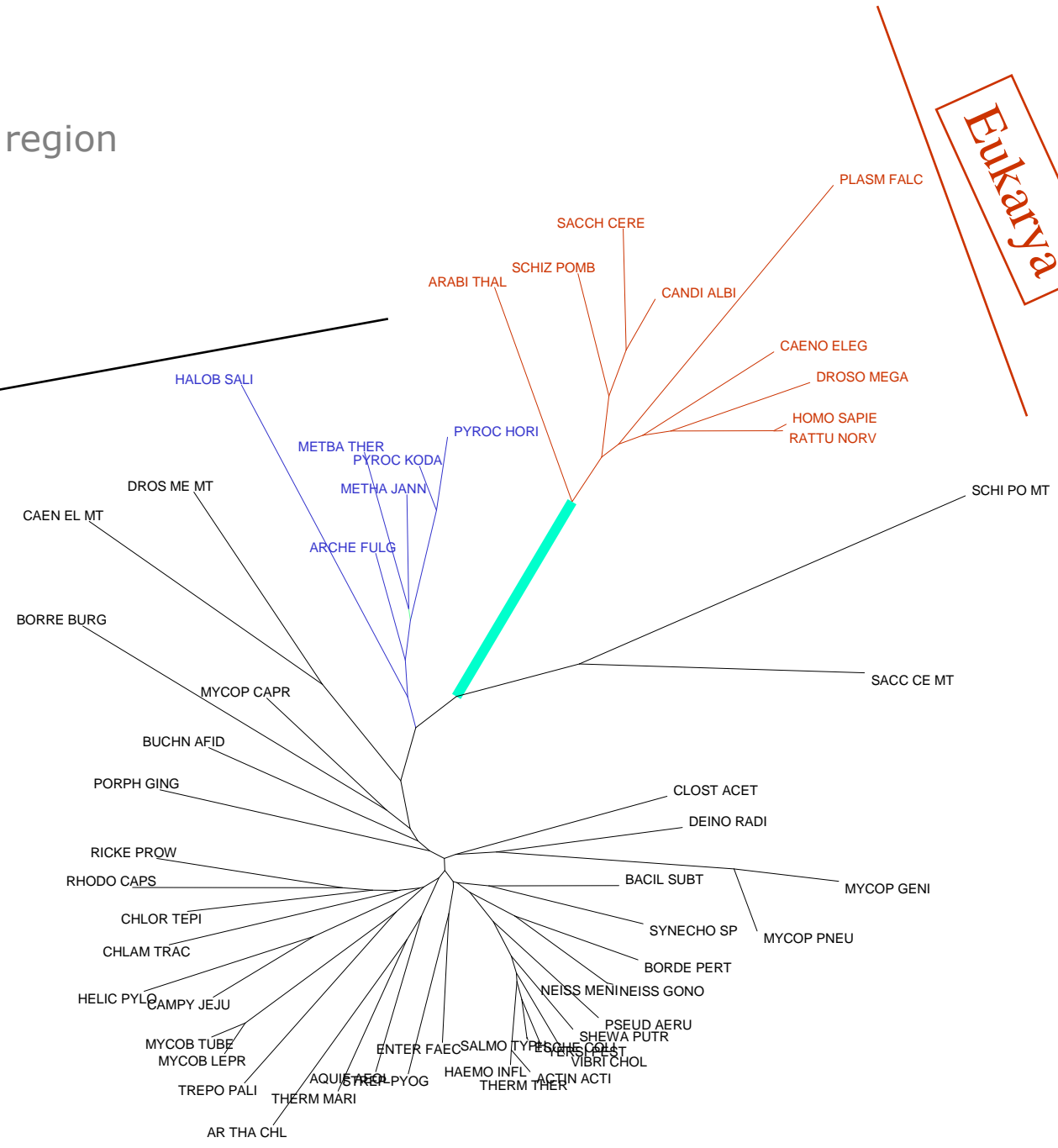
Archaea



N terminal region

Bacteria
Archaea
Mitoch.

Eukarya



0.1

Tree reconstruction method

- ✓ Distance based methods

- UPGMA
- Neighbor-Joining

- ✓ Character based methods

- Maximum parsimony
- Maximum likelihood
- Bayesian inference

Tree reconstruction method

Distance based methods

Calculate the distances between **pairs of sequences**

=> distance matrix

Group the sequences based on the distances

- UPGMA
- Neighbor-Joining

→ relatively simple and fast

→ but, the sequences themselves are not taken into account

Distance calculation

observed distance = mean number of substitution per site

$$\text{Observed distance} = \frac{\text{Nb substitutions}}{\text{Nb sites considered}}$$

Site considered

- ❑ for DNA, the 3rd base of each codon can be excluded from analysis
- ❑ alignment positions with gaps are generally eliminated
 - removal of positions with at least 1 gap in the alignment
→ **global gap removal**
 - removal of positions with 1 gap in 1 of the 2 considered sequences → **pairwise gap removal**

```
Seq1 MAIKKIISRSNSGIHNA TVI
Seq2 MPIKK. ISRSNSGIHHSTVI
Seq3 MPIKKIISRSNTGI. HSTVI
```

Total nb of sites = 20

Nb substitutions (Seq1,seq2) = 3
Nb substitutions (Seq1,seq3) = 4
Nb substitutions (Seq2,seq3) = 1

Global gap removal (18 sites considered)

Dist(Seq1,Seq2) = $3/18 = 0,1667$
Dist(Seq1,Seq3) = $4/18 = 0,2222$
Dist(Seq2,Seq3) = $1/18 = 0,0556$

Pairwise gap removal

Dist(Seq1,Seq2) = $3/19 = 0,1579$
Dist(Seq1,Seq3) = $4/19 = 0,2105$
Dist(Seq2,Seq3) = $1/18 = 0,0556$

Distance correction

For more distantly related sequences, the probability of several substitutions occurring at the same site increases.

=> The number of **observed** substitutions **under-estimates** the true number of substitutions between distantly related sequences.

	Sequence1	Sequence2	Observed no. substitutions	True no. substitutions
Single substitution	C	C=>A	1	1
Multiple substitutions	C	C=>A=>T	1	2
Coincident substitutions	C=>G	C=>A	1	2
Parallel substitutions	C=>A	C=>A	0	2
Convergent substitutions	C=>A	C=>T=>A	0	3
Reverse substitutions	C	C=>T=>C	0	2

=> Numerous methods available to estimate the true distance between sequences

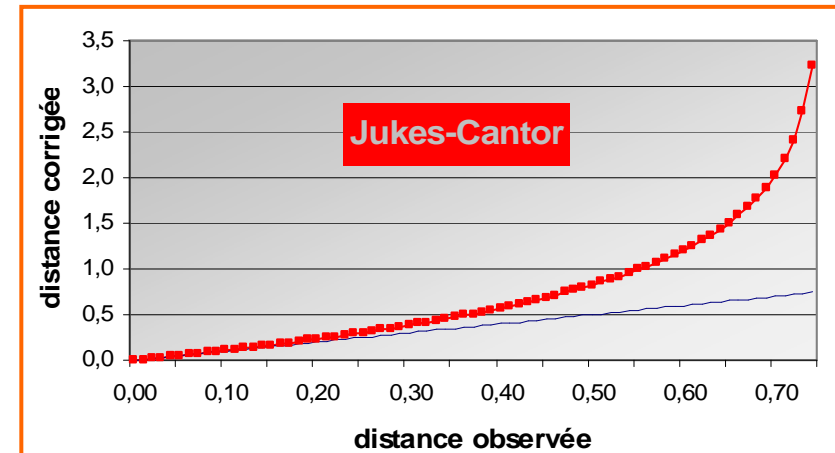
Distance correction: DNA substitution models

Jukes-Cantor (JC) model

- waiting time before mutation is exponential
- 4 bases have the same frequency
- all substitutions occur with equal probabilities

$$d = -\frac{3}{4} \ln(1 - \frac{4}{3} D)$$

with D observed distance



Transition : purine(A,G)-purine or pyrimidine(C,T)-pyrimidine substitution

Transversion : purine-pyrimidine or pyrimidine-purine substitution

Kimura 2 parameters (K2P) model

- 4 bases have the same frequency
- transitions (substitutions A-G or C-T) are more common than transversions

$$d = -\frac{1}{2} \ln(1 - 2P - Q) - \frac{1}{4} \ln(1 - 2Q)$$

where P is mean no. of transitions
 Q is mean no. of transversions

Tamura-Nei model

- 4 bases have different frequencies
- different transition frequencies (α_1 between purines; α_2 between pyrimidines), equal transversion frequencies (β)

Distance calculation: amino acid substitution models

✓ PAM matrices (Dayhoff, 1978)

- ❑ A 1-PAM mutation matrix describes an amount of evolution which will change, on the average, 1% of the amino acids. In mathematical terms:
- ❑ estimated from the observation of accepted mutations between 34 superfamilies of closely related sequences
- ❑ extrapolated to other evolutionary distances (eg. PAM250)

✓ Blosum matrices (1992)

- ❑ local, ungapped alignments of distantly related sequences to derive the BLOSUM series of matrices. Matrices of this series are identified by a number after the matrix (e.g. BLOSUM50), which refers to the minimum percentage identity of the blocks of multiple aligned amino acids used to construct the matrix.

✓ Gonnet matrices (1992)

- ❑ similar to Dayhoff, but with more modern sequence databases

UPGMA

UPGMA = Unweighted Pair Group Method with Arithmetic Mean
UPGMA is the simplest clustering algorithm

Assumption

mutation rate is the same in all lineages (molecular clock)

Method

Group the 2 closest OTUs

The node is placed at distance d from each sequence

$$d = (\text{dist OTU1,OTU2})/2$$

Calculate distance between this node and all other sequences :

$$\text{dist (OTU1,OTU2),OTUx} = (\text{dist OTU1,OTUx} + \text{dist OTU2,OTUx}) / 2$$

etc...

OTU (Operational Taxonomic Unit) : one or several grouped sequences

UPGMA

Distance matrix

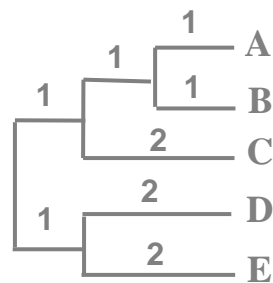
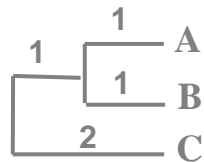
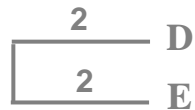
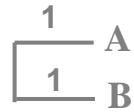
	A	B	C	D	E
B	2				
C	4	4			
D	6	6	6		
E	6	6	6	4	
F	8	8	8	8	8

	(A,B)	C	D	E
C	4			
D	6	6		
E	6	6	4	
F	8	8	8	8

	(A,B)	C	E
C	4		
(D,E)	6	6	
F	8	8	8

	(A,B,C)	(D,E)
(D,E)	6	
F	8	8

Clustering



Calculation of new distance matrix

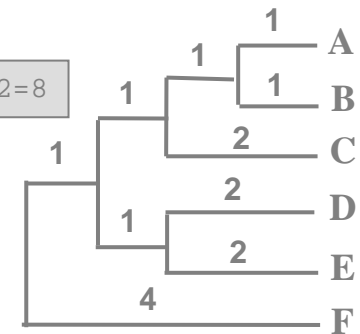
$$\begin{aligned} \text{dist}(A,B),C &= (\text{dist } AC + \text{dist } BC) / 2 = 4 \\ \text{dist}(A,B),D &= (\text{dist } AD + \text{dist } BD) / 2 = 6 \\ \text{dist}(A,B),E &= (\text{dist } AE + \text{dist } BE) / 2 = 6 \\ \text{dist}(A,B),F &= (\text{dist } AF + \text{dist } BF) / 2 = 8 \end{aligned}$$

$$\begin{aligned} \text{dist}(D,E), (A,B) &= (\text{dist } D(A,B) + \text{dist } E(A,B)) / 2 = 6 \\ \text{dist}(D,E),C &= (\text{dist } DC + \text{dist } EC) / 2 = 6 \\ \text{dist}(D,E),F &= (\text{dist } DF + \text{dist } EF) / 2 = 8 \end{aligned}$$

$$\begin{aligned} \text{dist}(AB,C), (D,E) &= (\text{dist } (A,B)(D,E) + \text{dist } C(D,E)) / 2 = 6 \\ \text{dist}(AB,C),F &= (\text{dist } (A,B)F + \text{dist } CF) / 2 = 8 \end{aligned}$$

$$\text{dist}(ABC,DE),F = (\text{dist } (ABC)F + \text{dist } (D,E)F) / 2 = 8$$

(ABC,DE)
F 8

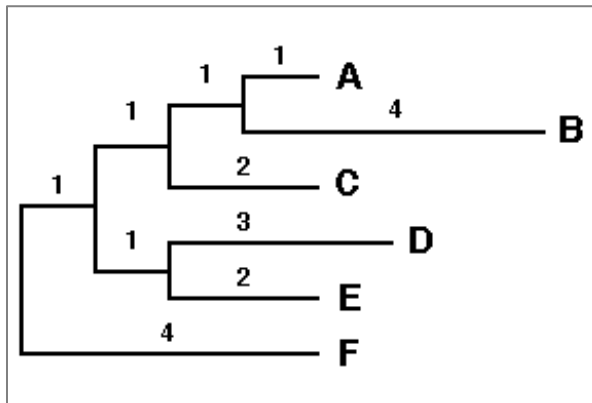


UPGMA

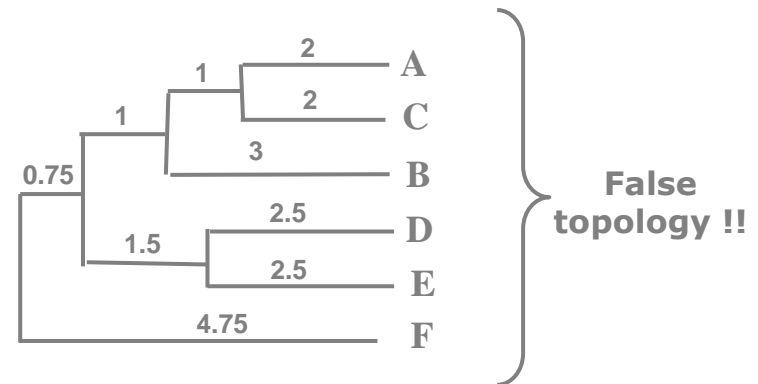
Major problem

If the mutation rate is different between branches, UPGMA can give a completely false topology

Ex: Mutation rate of B is much faster than A



Tree obtained by UPGMA



Initial matrix

	A	B	C	D	E
B	5				
C	4	7			
D	7	10	7		
E	6	9	6	5	
F	8	11	8	9	8

Clustering by UPGMA

	A,C	B	D	E
B	6			
D	7	10		
E	6	9	5	
F	8	11	9	8

	A,C	B	D,E
B	6		
D,E	6.5	9.5	
F	8	11	8.5

	AC,B	D,E
D,E	8	
F	9.5	9.5

	ACB,DE
F	9.5

Neighbor-Joining (NJ)

✓ Reference

Saitou & Nei Mol. Biol. Evol. 4(4):406-25 1987

✓ Method

- ❑ the 2 OTUs to be clustered are chosen such that they minimise the sum of lengths of all branches
- ❑ the calculation of branch lengths takes into account the mean distance between each sequence and all other sequences

=> allows different mutation rates between branches

Neighbor-Joining (NJ)

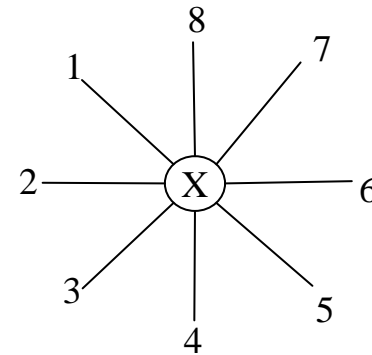
- ✓ Estimate sum of branch lengths for all possible clusters (S_{12}, S_{13}, \dots)
- ⇒ Choose cluster that minimises sum of branch lengths
- ✓ How to calculate the sum of branch lengths ?

For a star-shaped tree, the sum of N branch lengths :

$$S_0 = \sum_{i=1}^N L_{iX} = \left(\sum_{i < j} D_{ij} \right) / (N-1)$$

Where D_{ij} is distance between TU_i and TU_j

L_{iX} is length between node i and node X

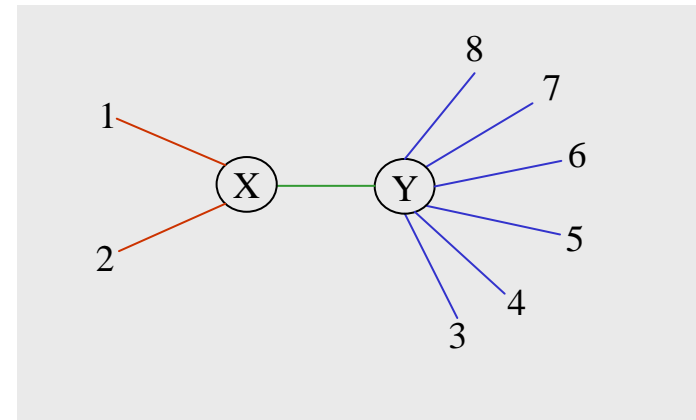


Neighbor-Joining (NJ)

Assume TUs 1 and 2 have been joined,
then sum of branch lengths :

$$S_{12} = (L_{1X} + L_{2X}) + L_{XY} + \sum_{i=3}^N L_{iY}$$

$$S_{12} = D_{12} + L_{XY} + \left(\sum_{3 \leq i < j}^N D_{ij} \right) / (N-3)$$



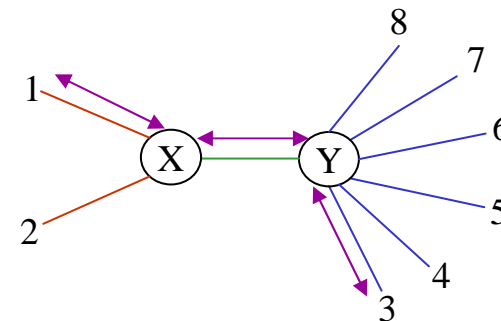
$$L_{XY} = \frac{1}{2(N-2)} \left[\sum_{k=3}^N (D_{1k} + D_{2k}) - (N-2)(L_{1X} + L_{2X}) - 2 \sum_{i=3}^N L_{iY} \right]$$

In the example :

L_{1X} and L_{2X} have been counted 6 times each.

Lengths L_{3Y}, \dots, L_{8Y} have been counted twice each.

Length L_{XY} has been counted 12 times.



$$S_{12} = \frac{1}{2(N-2)} \sum_{k=3}^N (D_{1k} + D_{2k}) + \frac{1}{2} D_{12} + \frac{1}{N-2} \sum_{3 \leq i < j} D_{ij}$$

Neighbor-Joining (NJ)

estimate sum of branch lengths for all possible clusters (S_{12}, S_{13}, \dots)

	A	B	C	D	E
A					
B	5				
C	4	7			
D	7	10	7		
E	6	9	6	5	
F	8	11	8	9	8

$$S_{12} = \frac{1}{2(N-2)} \sum_{k=3}^N (D_{1k} + D_{2k}) + \frac{1}{2} D_{12} + \frac{1}{N-2} \sum_{3 \leq i < j} D_{ij}$$

Application :

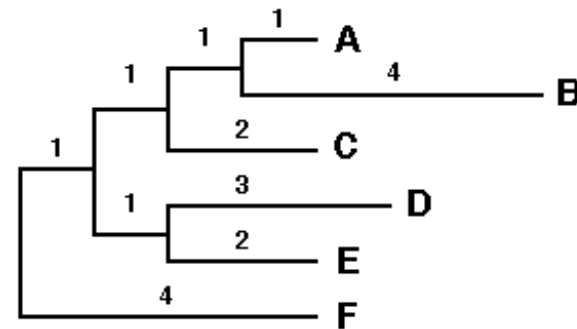
$$S_{AB} = 1/8 * 62 + 5/2 + 1/4 * 43 = 21,00$$

$$S_{AC} = 1/8 * 54 + 4/2 + 1/4 * 52 = 21,75$$

...



Cluster AB



Neighbor-Joining (NJ)

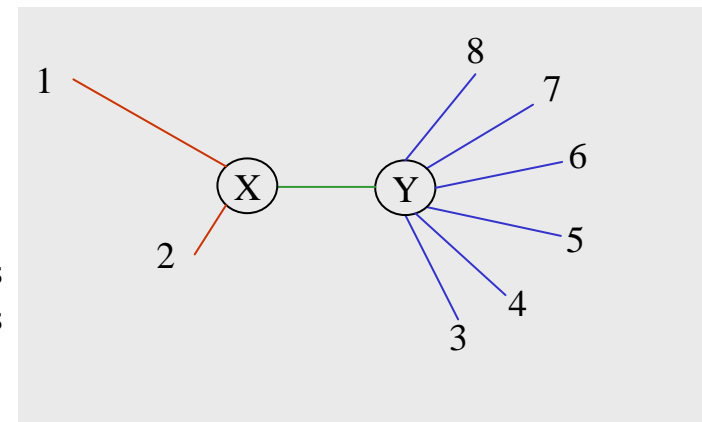
Branch lengths are estimated by Fitch, Margoliash method

$$L_{1X} = (D_{12} + D_{1Z} - D_{2Z}) / 2$$

$$L_{2X} = (D_{12} + D_{2Z} - D_{1Z}) / 2$$

D1Z represents mean distance between 1 and all other sequences

D2Z represents mean distance between 2 and all other sequences



Application :

	A	B	C	D	E
B	5				
C	4	7			
D	7	10	7		
E	6	9	6	5	
F	8	11	8	9	8

$$L_{AX} = (5+6.25-9.25)/2 = 1$$

$$L_{BX} = (5+9.25-6.25)/2 = 4$$



Allows different mutation rates between branches

Tree reconstruction method

✓ Distance based methods

- ❑ UPGMA

- ❑ Neighbor-Joining

=> *relatively fast and simple*

=> *but, the sequences themselves are not taken into account*

✓ Character based methods

- ❑ Maximum parsimony

- ❑ Maximum likelihood

- ❑ Bayesian inference

each position in the sequence is considered as a trait or character

=> *calculation time is very long*

Maximum Parsimony

Used historically to study morphological characters

→ prefer evolutionary scenario that involves the smallest number of events

if 2 species share the same character, they may have inherited from a common ancestor in which this character has appeared

Application to sequences:

one column in the alignment = one character

☞ Search for most parsimonious trees, i.e. those for which the topology involves minimum number of substitutions

☞ restriction: sequences must be fairly close

Steps:

1) Search for all possible topologies

2) Select the most parsimonious tree

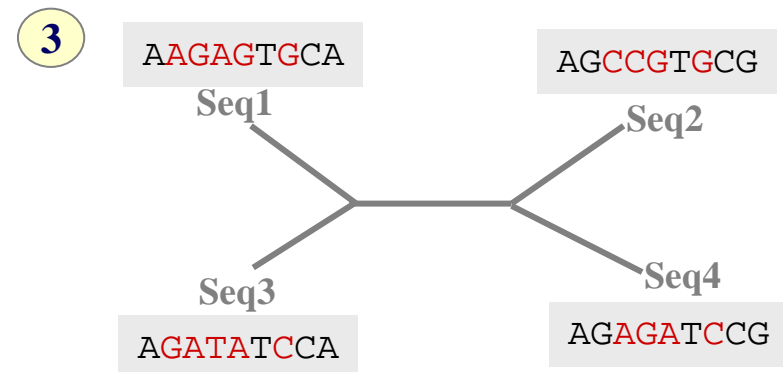
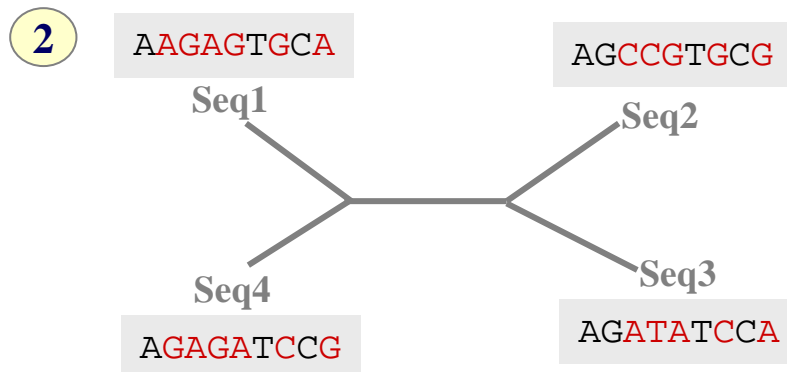
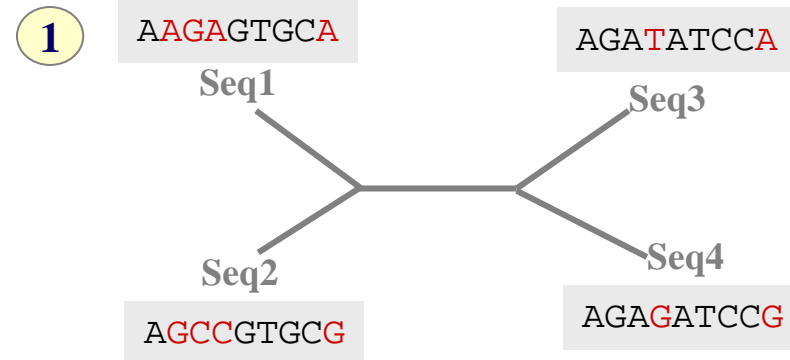
- find smallest no. of substitutions for each topology

- find informative sites => sites that support one topology

Maximum Parsimony

Seq1 AAGAGTGCA
Seq2 AGCCGTGCG
Seq3 AGATATCCA
Seq4 AGAGATCCG

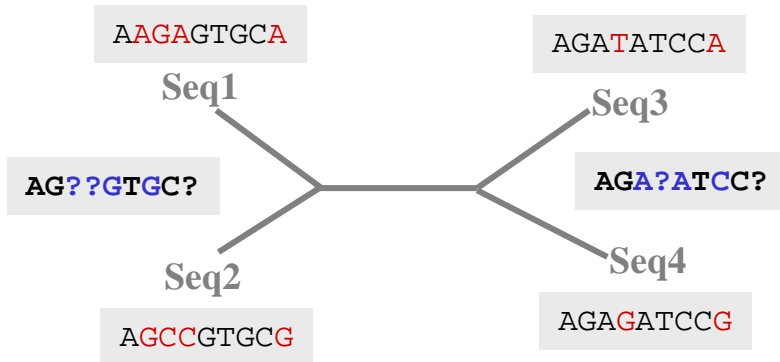
No. of possible topologies (unrooted trees) : 3



Maximum Parsimony

Second step : find smallest no. of substitutions for a given topology (Fitch method)

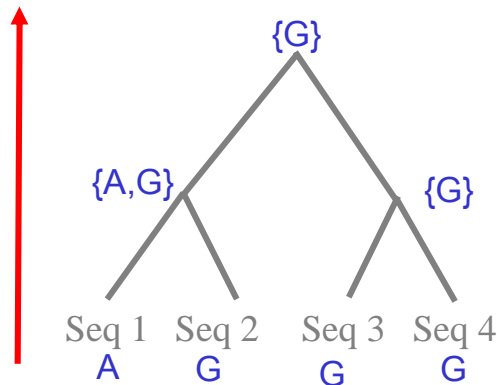
1) Root the tree (anywhere)



3) From root to leaves:

- Choose one nucleotide x from set N of root n
- For child node u , choose one nucleotide :
 x if $x \in U$
 any nucleotide from set U otherwise

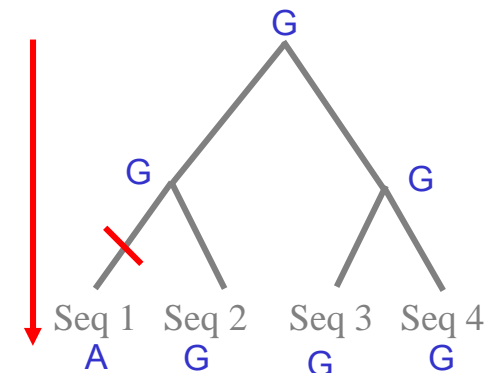
2) From leaves to root



Nodes u and v are children of n
 U, V and N are sets of nucleotides associated with these nodes

$$N = U \cap V \text{ si } U \cap V \neq \emptyset$$

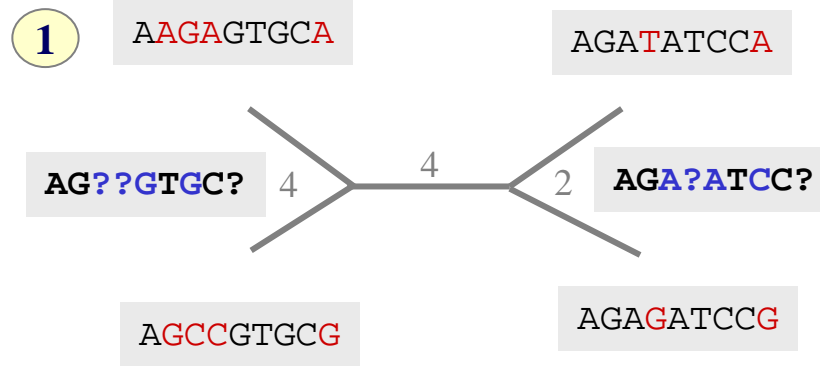
$$N = U \cup V \text{ otherwise}$$



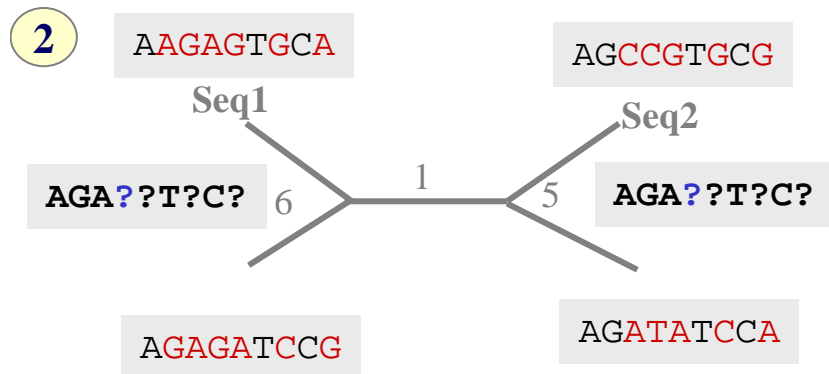
1 substitution

Maximum Parsimony

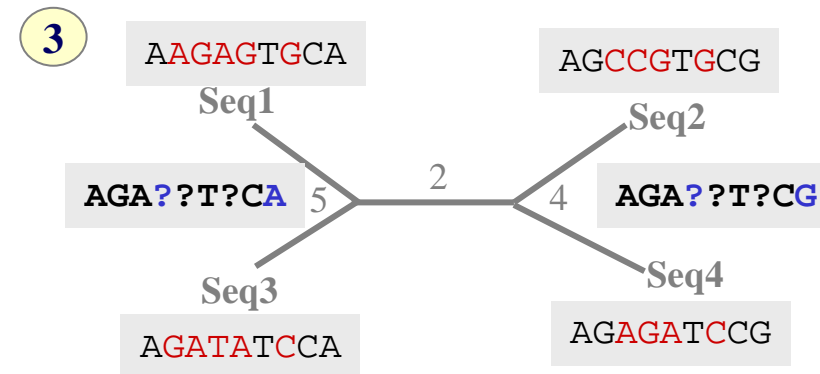
Comparison of « length » of tree for the different topologies



10 substitutions



12 substitutions



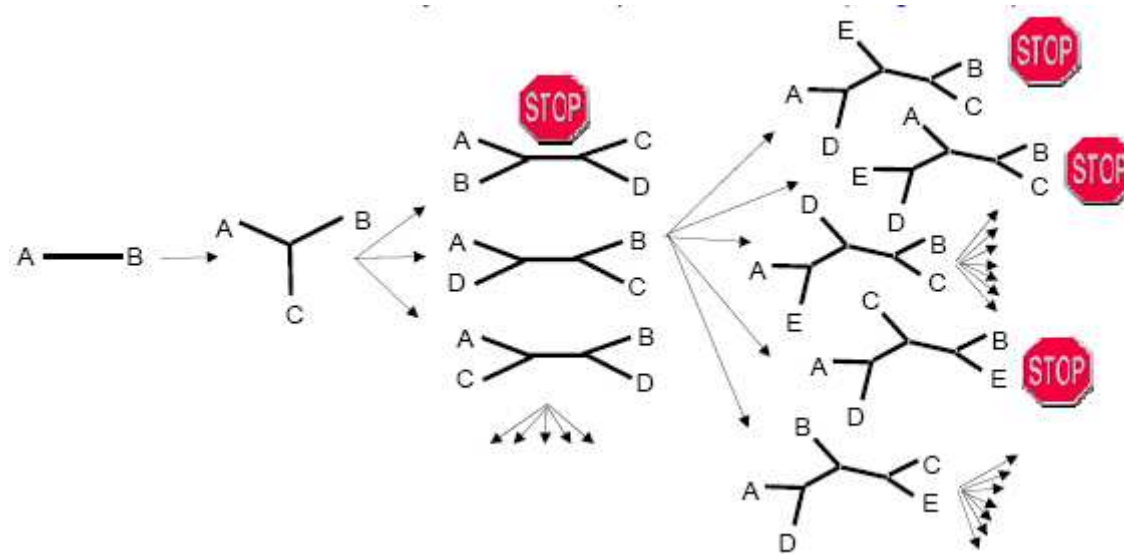
11 substitutions

Search for shortest tree

Branch-and-bound Method (Hendy & Penny, 1982) :
Exact algorithm, guaranteeing the optimal solution without
exhaustive searching

Steps:

1. add sequences one by one (following order of alignment)
→ calculate no. of substitutions for each topology
2. explore different topologies
→ if no. of substitutions during addition > no. found for best topology
→ stop



Search for shortest tree

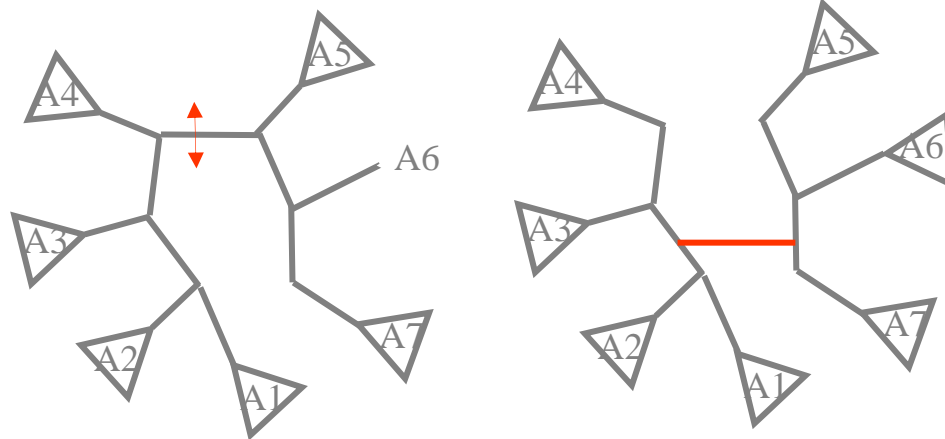
Heuristic methods

Can handle more sequences and longer sequences, but do not guarantee optimal solution

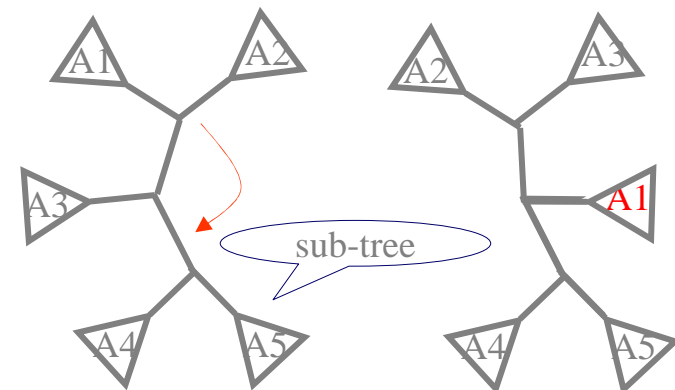
- 1) construct initial tree by progressively adding TUs
- 2) then, rearrange initial tree to reduce length (branch swapping)



Rearrangement by bisection and reconnexion
(tree bisection and reconnection)



Global rearrangement
(subtree pruning and regrafting)



result depends on initial order of sequences

☞ perform several tests with different initial ordering : option « jumble »

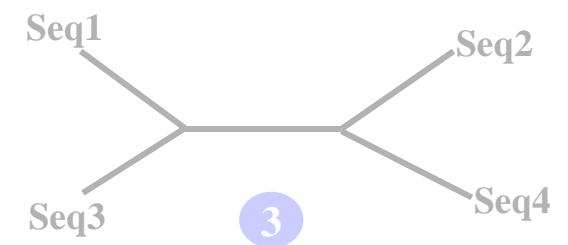
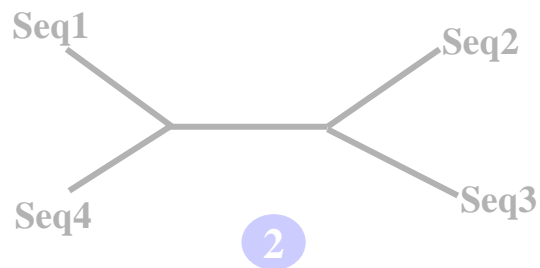
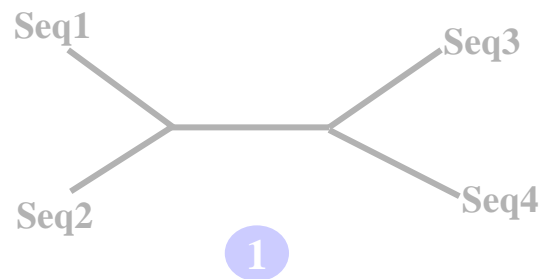
Maximum Parsimony

Variant :

- find all possible topologies
- only include *informative sites*
- => sites that support one topology

Seq1 A A G A G T G C A
Seq2 A G C C G T G C G
Seq3 A G A T A T C C A
Seq4 A G A G A T C C G

1 1 1 3



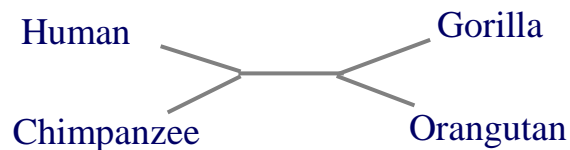
Maximum Parsimony

Table 1. The observed site-pattern frequencies (n_i) in the 895-bp mtDNA sequences of human (H), chimpanzee (C), gorilla (G), and orangutan (O) (Brown et al. 1982)

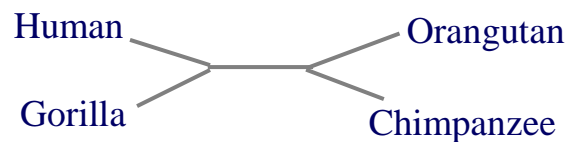
n_i	2 2 1	2 1 1	1 1	1 1	6 7 7 1 9	6 3 2 4 3	1 1 1 2 4	1 1 1 1 2	1 1 3 1 1	1	
Pattern i	AGCTG	TCATC	ACCCA	CCTTT	TGCAG	AGAAT	TATAC	GGTGC	ACTGC	C	Human
	AGCTA	TTACC	ACCCA	TCCTC	TGCCA	GGTAC	CGTAT	AAAGC	GTTAA	C	Chimpanzee
	AGCTG	TTGTT	ATCAA	CACCT	CGCAA	AAAAC	TGCCC	AGAGT	ATTAA	T	Gorilla
	AGCTA	CCACC	GTTCC	CACCT	TAATA	AAATT	AAAAT	GGGCG	CTACA	A	Orangutan
Supported tree	2	3 2	1	1 1		1 3	3 2 3				

^a The 46 observed patterns are arranged columnwise in the order of occurrence in the data, for example, the first two patterns, AAAA and GGGG, are observed at 222 and 71 sites, respectively. The three tree topologies supported by the “informative” patterns are $T_1 = ((H,C),G,O)$, $T_2 = ((H,G),C,O)$, $T_3 = ((C,G),H,O)$; other site patterns are “noninformative” by the parsimony analysis. So T_1 , T_2 , and T_3 are supported by 17 (= 5 + 3 + 6 + 3), 9 (= 2 + 3 + 4), and 13 (= 8 + 3 + 1 + 1) sites, respectively, and T_1 is the most parsimonious tree

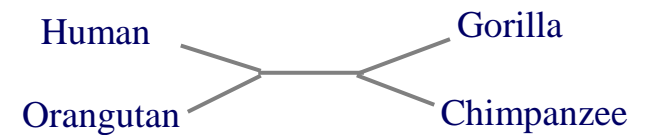
Yang *J. Mol. Evol.* 42:294-307 (1996)



Topology 1
(17 sites)



Topology 2
(9 sites)

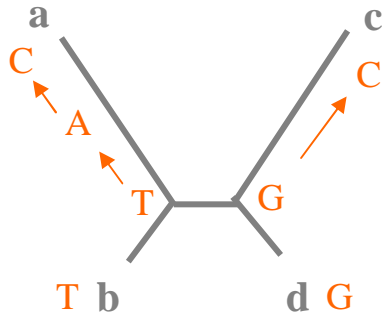


Topology 3
(13 sites)

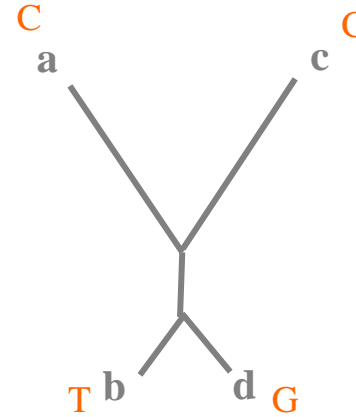
Maximum Parsimony

Problem: long branch attraction

True tree



Parsimonious tree



If the sequences are evolving at very different rates, the probability of convergent substitutions is significant in long branches
=> The most parsimonious clustering can lead to false topology

- ✓ All changes occur at equal rates
- ✓ no correction for multiple substitutions
- ✓ Can lead to several equally parsimonious trees
- ✓ Relatively slow, not suitable for a large number of sequences
- ✓ No information about branch lengths (generally)

Maximum likelihood

maximizes the probability that a given tree could have produced the observed data (that is, the likelihood)

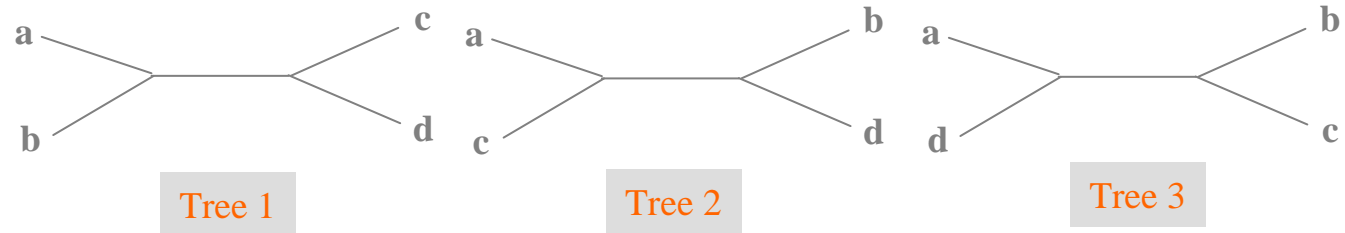
- ✓ As for parsimony method :
 - ❑ each column is considered to be a character
 - ❑ all possible trees are considered
 - ❑ for trees requiring many mutations, probability is low
=> trees requiring few mutations are preferred
- ✓ Differences :
 - ❑ use of an **explicit evolutionary model**
 - ❑ allows **variable substitution rates** for each branch
- ✓ Can be used to estimate reliability of tree

Maximum likelihood

Alignment of 4 sequences

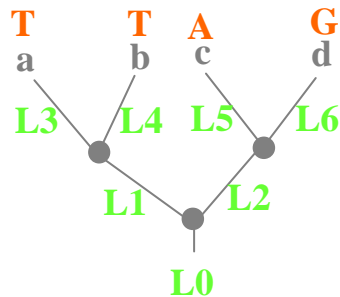
	1	2	3	4
Seq a	T	T	G	C...
Seq b	T	T	G	C...
Seq c	A	T	A	C...
Seq d	G	T	A	C...

3 possible unrooted trees



The root position do not influence the tree likelihood

For Tree 1, one of the 5 rooted tree is:



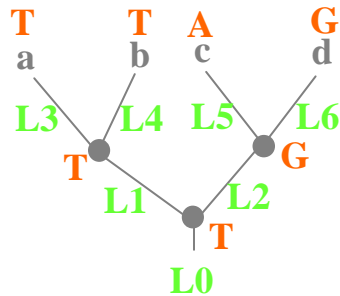
For position 1, possible combinations of bases at each node :

- 3 internal nodes
- 4 possible bases at each node

No. of possible combinations: $4 \times 4 \times 4 = 64$

Maximum likelihood

Example 1:



Estimation of likelihood of tree 1 for position 1 :

sum of probabilities for each of 64 combinations

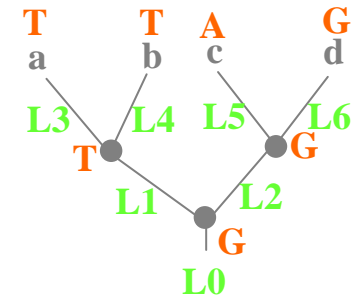
Estimation of likelihood of tree 1 :

Sum of probabilities obtained for each position

The calculation is performed for all possible tree (3 here).

The tree with the maximum likelihood is selected.

Example 2:



Evolutionary model example:

L0 = frequency of T ~ 0.25

L2 = probability of transversion of T=>G

L5 = probability of transition of G=>A

L1, L3, L4, L6 = ~1

Probability of this combination for position 1:

$L = L0 \times L1 \times L2 \times L3 \times L4 \times L5 \times L6$

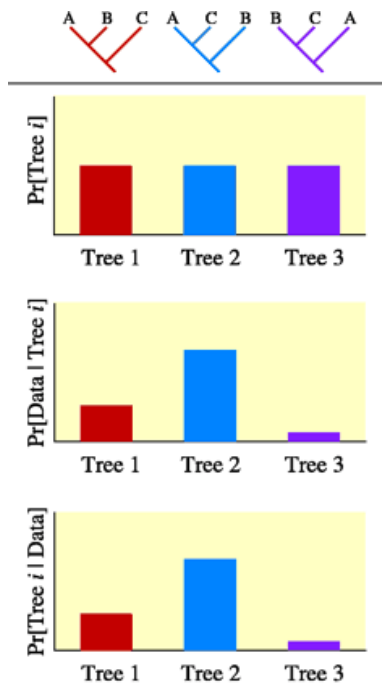
No of sequences	No of possible unrooted trees
3	1
4	3
5	15
6	105
7	945
8	10 395
9	135 135

The number of possible trees is growing very quickly! The method is only applicable for a small number of sequences.

Bayesian inference

- ✓ Use Bayes's theorem to combine the prior probability of a phylogeny with the likelihood to produce a posterior probability distribution on trees

$$\Pr[Tree | Data] = \frac{\Pr[Data | Tree] \times \Pr[Tree]}{\Pr[Data]}$$



- Prior probability of a tree before observations are made (generally, all trees are equally probable)
- Likelihood is proportional to probability of the observations, conditional on the tree (makes assumptions about processes generating observations)
- Posterior probability of a tree is the probability of the tree conditional on the observations, obtained by combining prior and likelihood for each tree

Bayesian inference

- ✓ choose tree with the highest posterior probability as the best estimate of phylogeny
- ✓ some numerical methods are available that allow the posterior probability of a tree to be approximated
- ✓ e.g. Markov chain Monte Carlo (MCMC) in **MrBayes** (Huelsenbeck et al, 2001)



Phylogenetic reconstruction

- ✓ Introduction: basic concepts
- ✓ Principal methods for tree reconstruction
 - Distance based methods
 - Character based methods
- ✓ **Evaluation of the reliability of a tree**
- ✓ The limits of phylogeny
- ✓ Some programs

Evaluation of the reliability of a tree

Goal

Statistically estimate the reliability of a given tree topology

Example: bootstrapping

Build n pseudo-alignments by random sampling of the columns in the initial alignment

- => each column can be used 0,1 or more times
- => the pseudo-alignments have the same length as the initial alignment
- => the number of pseudo-alignments should allow for significant statistical testing ($n \geq$ number of columns)

For each pseudo-alignment, build a tree

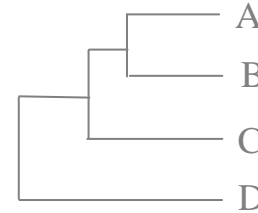
For each branch in the initial tree, count the number of times this branch is found in the n trees

Bootstrap method

Initial alignment

Seq A	A	G	G	C	T	C	C	A	A	A
Seq B	A	G	G	T	T	C	G	A	A	A
Seq C	A	G	C	C	C	C	G	A	A	A
Seq D	A	T	T	T	C	C	G	A	A	C

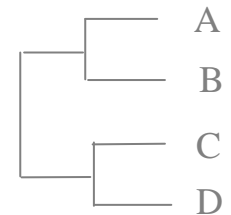
	A	B	C
B	2		
C	3	3	
D	6	4	4



1

Seq A	G	G	G	T	T	T	C	A	A	A
Seq B	G	G	G	T	T	T	G	A	A	A
Seq C	G	C	C	C	C	C	G	A	A	A
Seq D	T	T	T	C	C	C	G	A	A	C

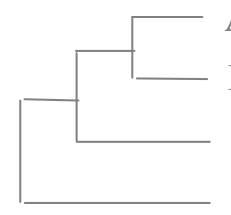
	A	B	C
B	1		
C	6	5	
D	8	7	4



2

Seq A	A	T	T	C	C	C	C	A	A	A
Seq B	A	T	T	C	C	G	G	A	A	A
Seq C	A	C	C	C	C	G	G	A	A	A
Seq D	A	C	C	C	C	G	G	C	C	C

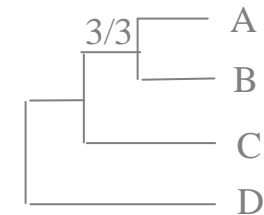
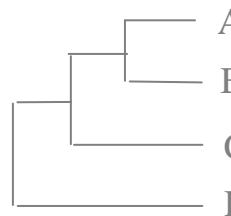
	A	B	C
B	2		
C	4	2	
D	7	5	3



3

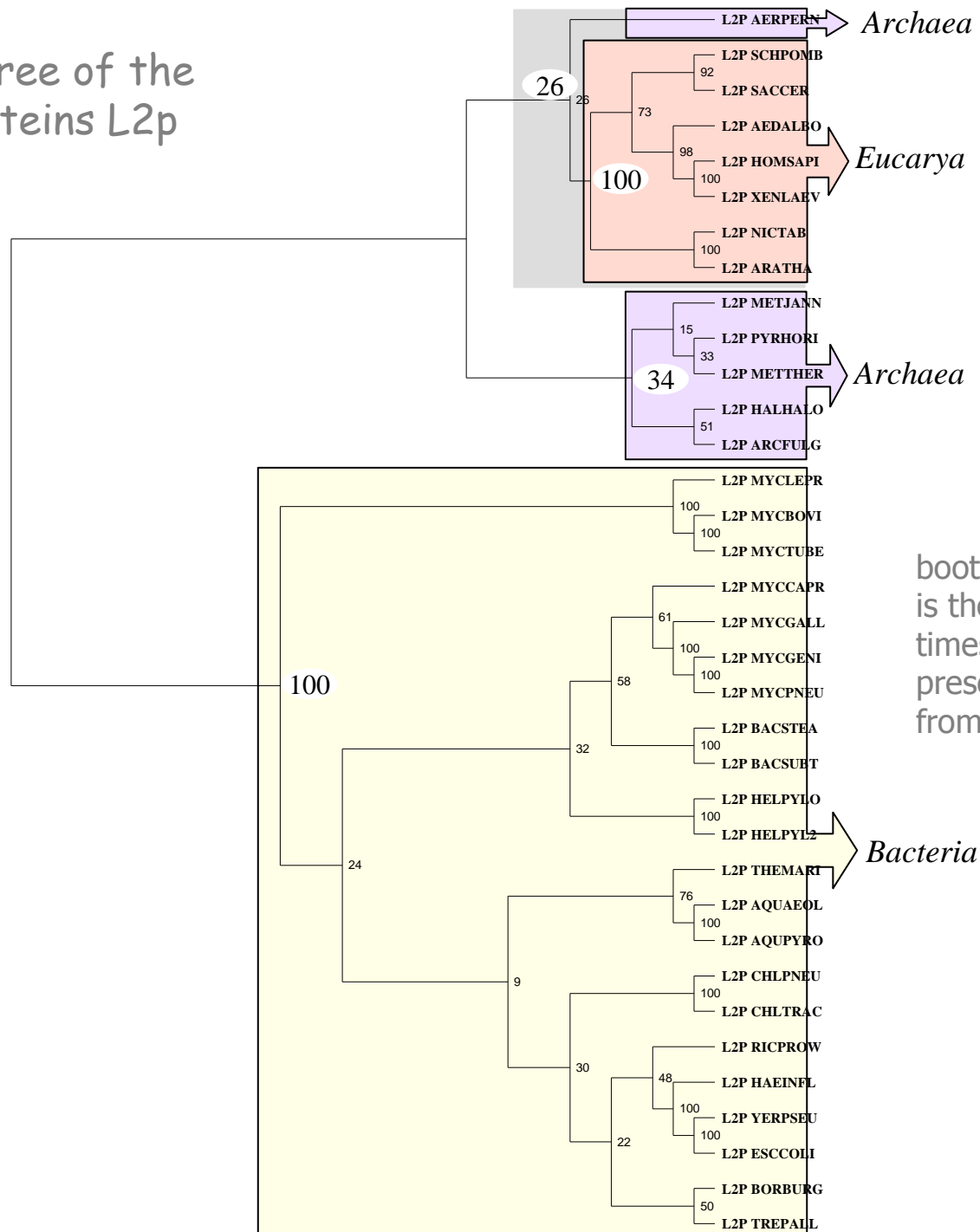
Seq A	A	G	T	T	C	C	C	A	A	A
Seq B	A	G	T	T	C	C	G	A	A	A
Seq C	A	G	C	C	C	C	G	A	A	A
Seq D	A	T	C	C	C	C	G	A	C	C

	A	B	C
B	1		
C	3	2	
D	6	5	3



Random sampling

Phylogenetic tree of the ribosomal proteins L2p



bootstrap value of a node is the percentage of times that node is present in the trees built from the random samples

Phylogenetic reconstruction

- ✓ Introduction: basic concepts
- ✓ Principal methods for tree reconstruction
 - Distance based methods
 - Character based methods
- ✓ Evaluation of the reliability of a tree
- ✓ **The limits of phylogeny**
- ✓ Some programs

Limits of phylogeny

- ✓ Phylogenetic tree-building models make certain assumptions:
 - ❑ sequences are homologous (descended from a shared ancestral sequence)
 - ❑ each of the sequences has a common phylogenetic history with the other sequences
 - ❑ at each position in the alignment, the characters are homologous with each other
 - ❑ the sequence variability in the sample contains phylogenetic signal adequate to resolve the problem under study.

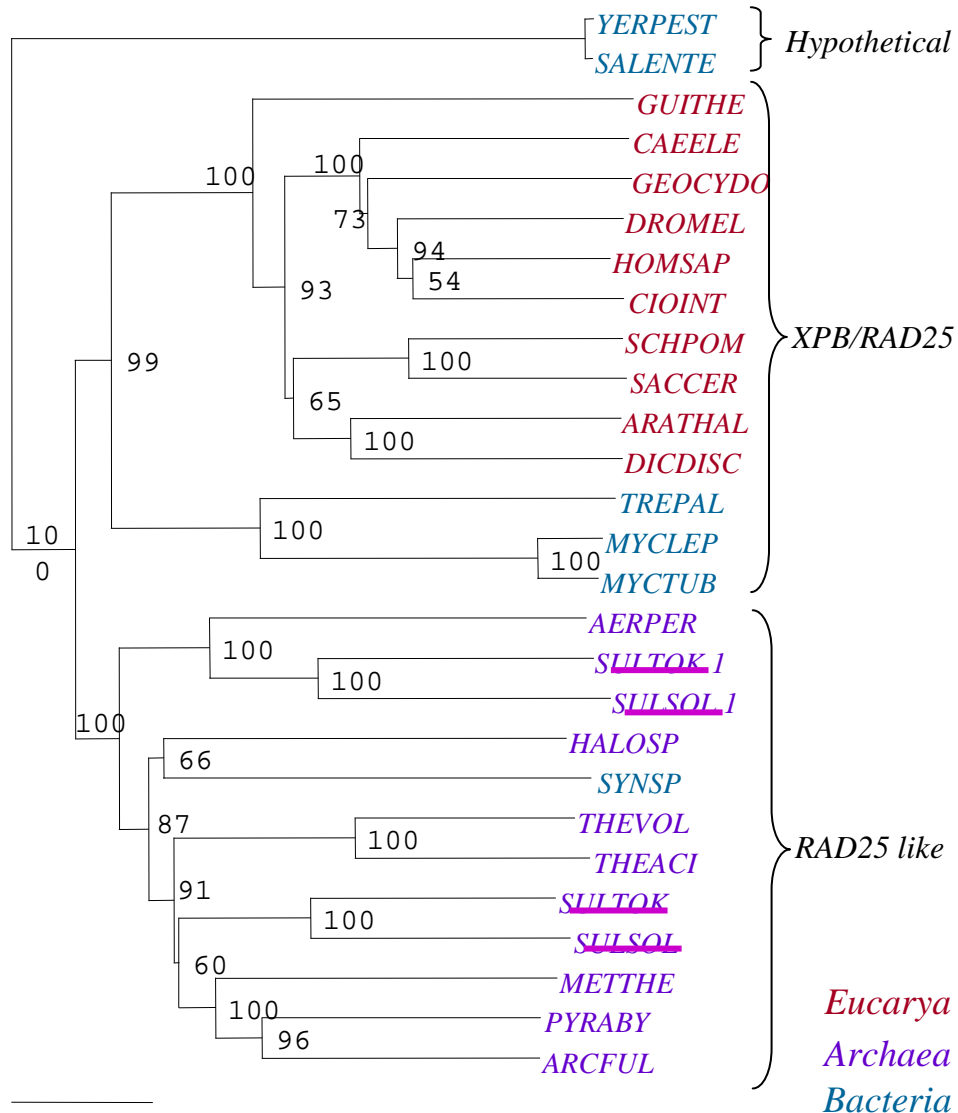
Careful selection of sequences and evolutionary model

Limits of phylogeny

- ✓ No algorithm is perfect
 - ❑ it is never certain that the reconstructed tree is the real one!
 - ❑ the same data can result in different trees depending on the algorithm used
- ✓ The evolutionary history of a gene is not always transposable to the species
 - ❑ not all genes evolve at the same rate (different selection pressure)
 - ❑ horizontal transfers
 - ❑ paralogy
- ✓ The tree topology must be confronted to:
 - ❑ taxonomic data (species phylogeny)
 - ❑ bibliographic data related to the studied gene family

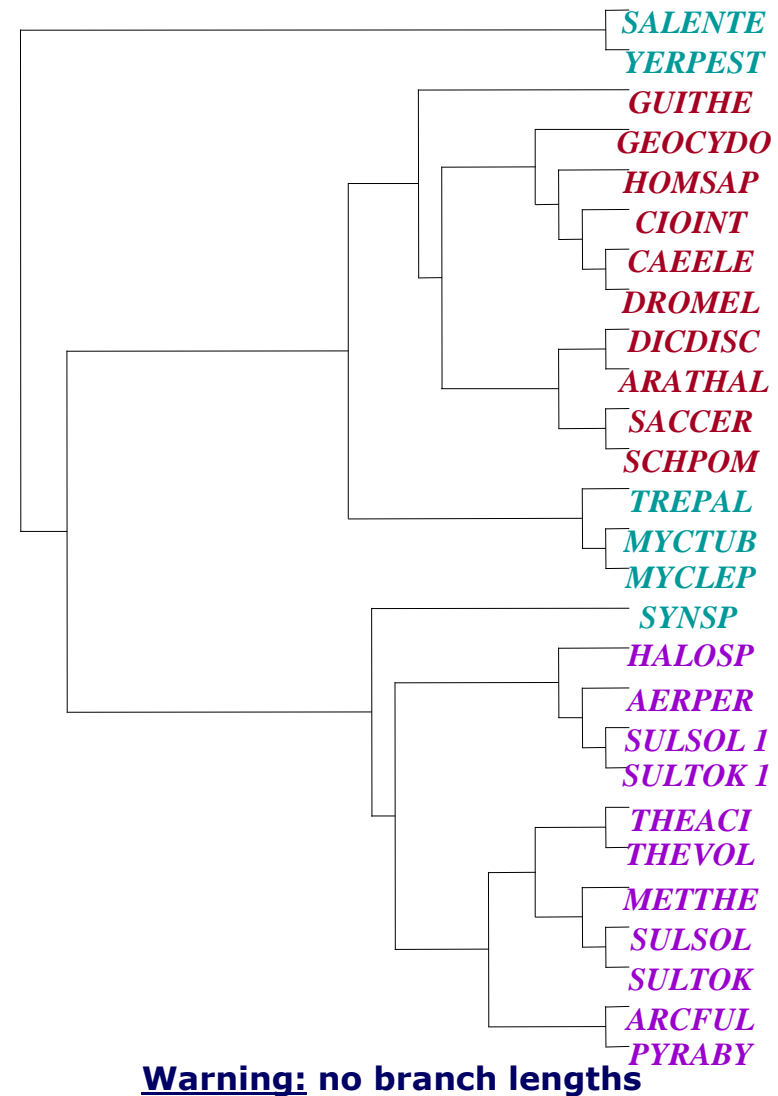
Phlogenetic tree of RAD25 and RAD25-like proteins

- Neighbor-Joining -

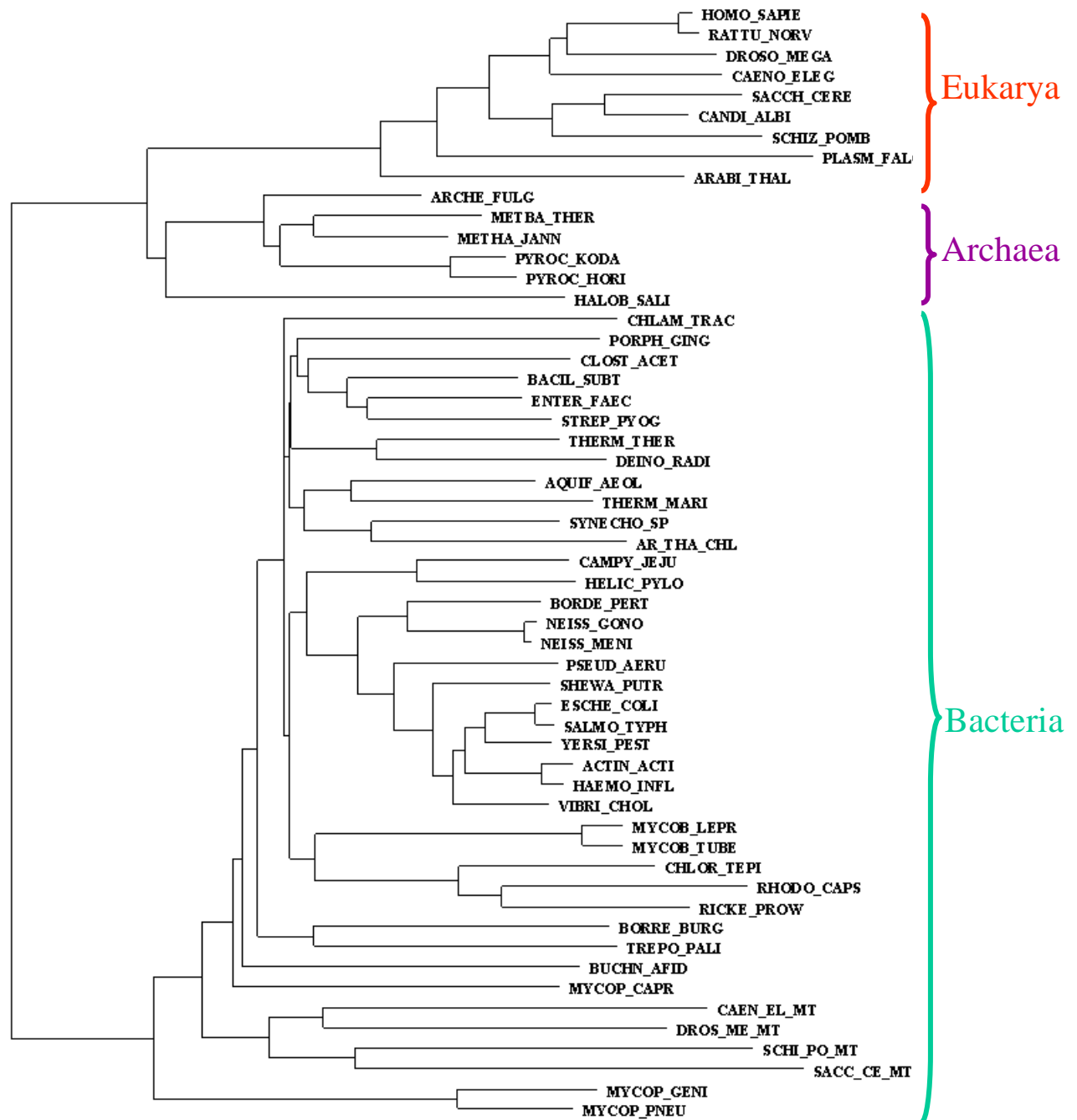


Phlogenetic tree of RAD25 and RAD25-like proteins

- Maximum parsimony -

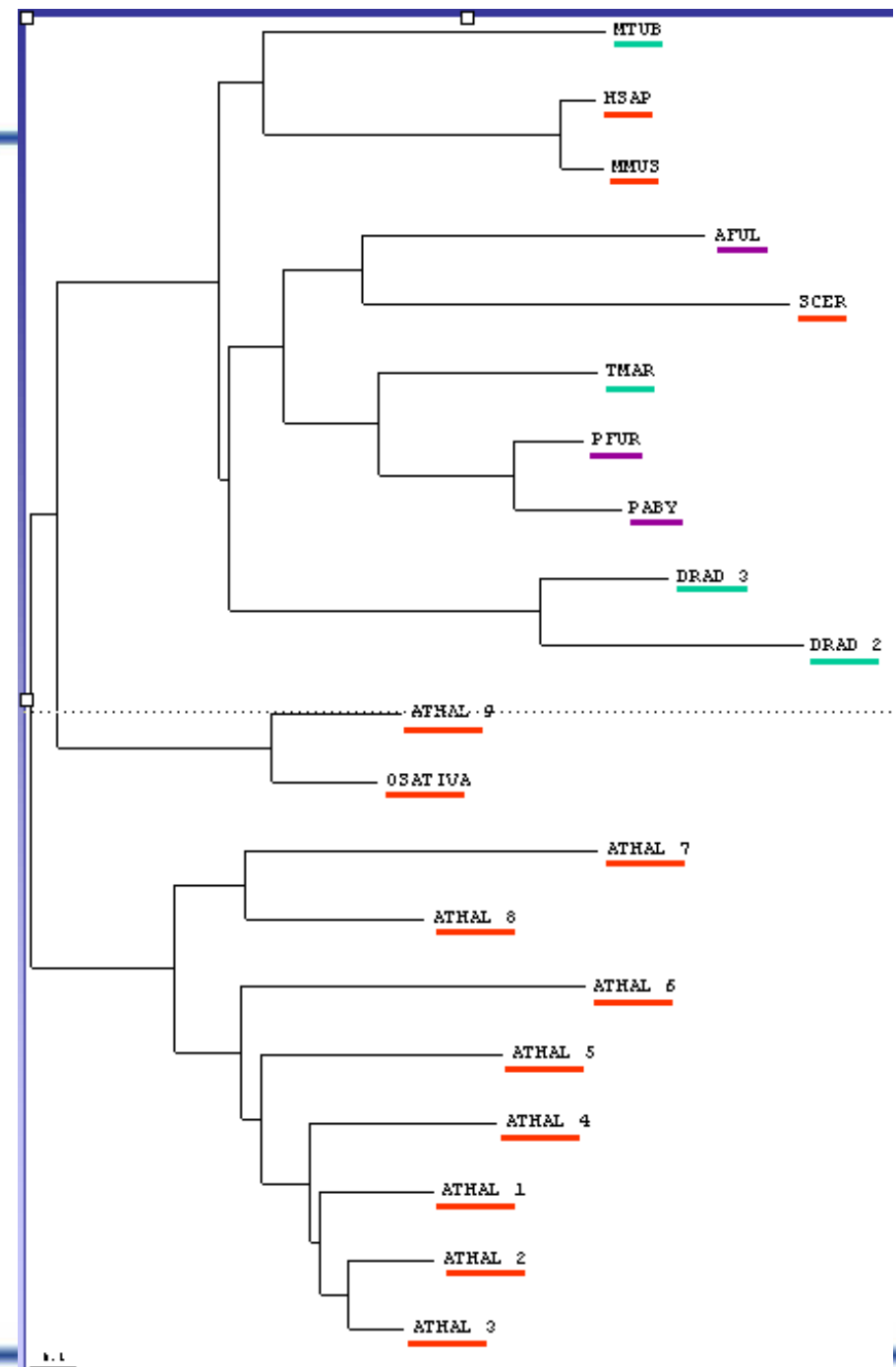
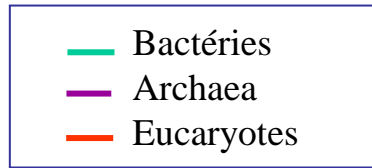


An informational protein - AspRS -



An operational protein : lysophospholipase

- duplications
- losses
- transfers



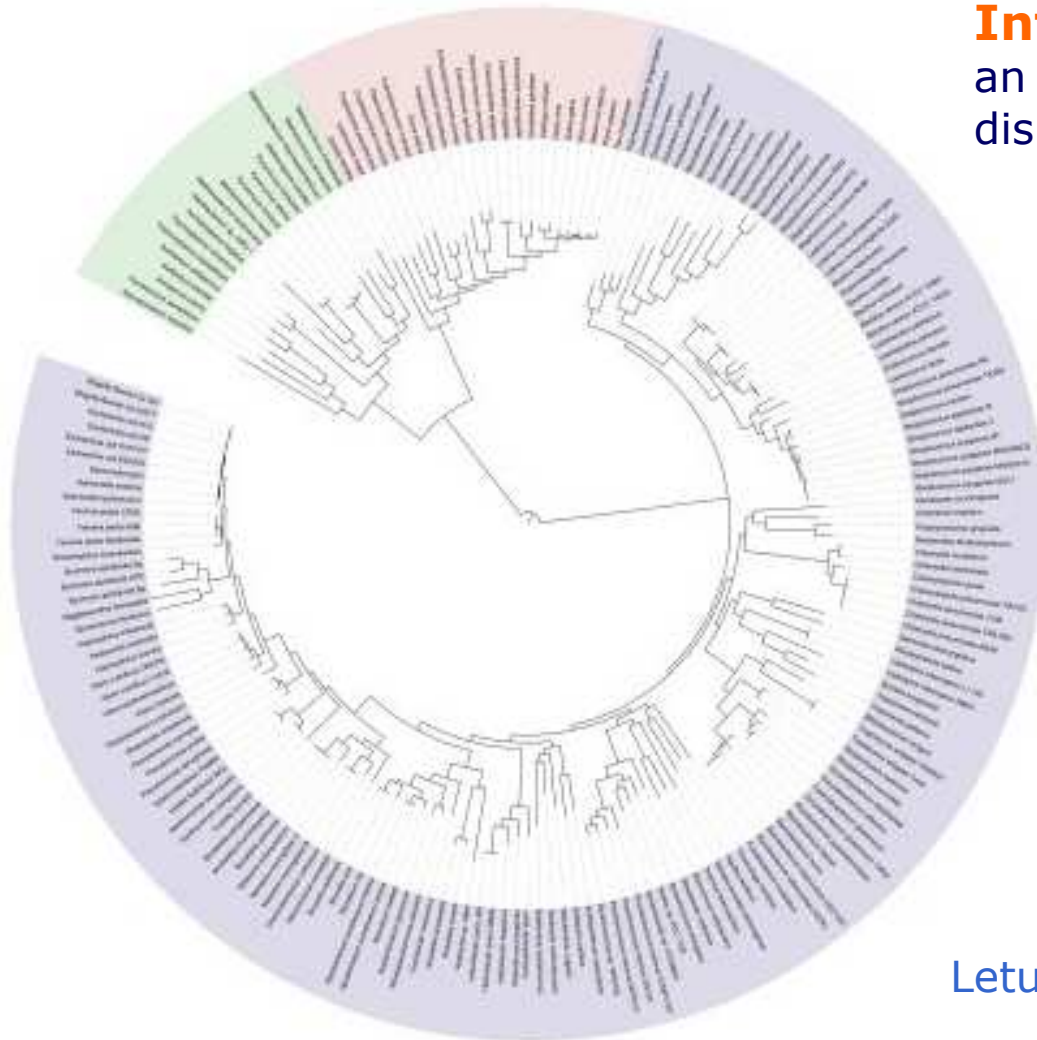
Phylogenetic reconstruction

- ✓ Introduction: basic concepts
- ✓ Principal methods for tree reconstruction
 - Distance based methods
 - Character based methods
- ✓ Evaluation of the reliability of a tree
- ✓ The limits of phylogeny
- ✓ **Some programs**

Some programs

- ✓ Software suites :
 - ❑ **Phylip** (very comprehensive)
<http://evolution.genetics.washington.edu/phylip.html>
 - ❑ **PAUP** (Phylogenetic Analysis Using Parsimony)
<http://www.lms.si.edu/PAUP/about.html>
 - ❑ **TREE-PUZZLE**
<http://www.tree-puzzle.de/>
 - ❑ **phylowin** (graphical interface)
<http://pbil.univ-lyon1.fr/software/phylowin.html>
- ✓ Visualisation, manipulation
 - ❑ **Njplot, baobab, treeedit, phylodendron...**
 - ❑ **Treeview**
<http://taxonomy.zoology.gla.ac.uk/rod/treeview.html>

Some programs



Interactive Tree Of Life (iTOL)
an online tool for phylogenetic tree
display and annotation.

<http://itol.embl.de/>

Letunic I, Bork P. Bioinformatics 2007