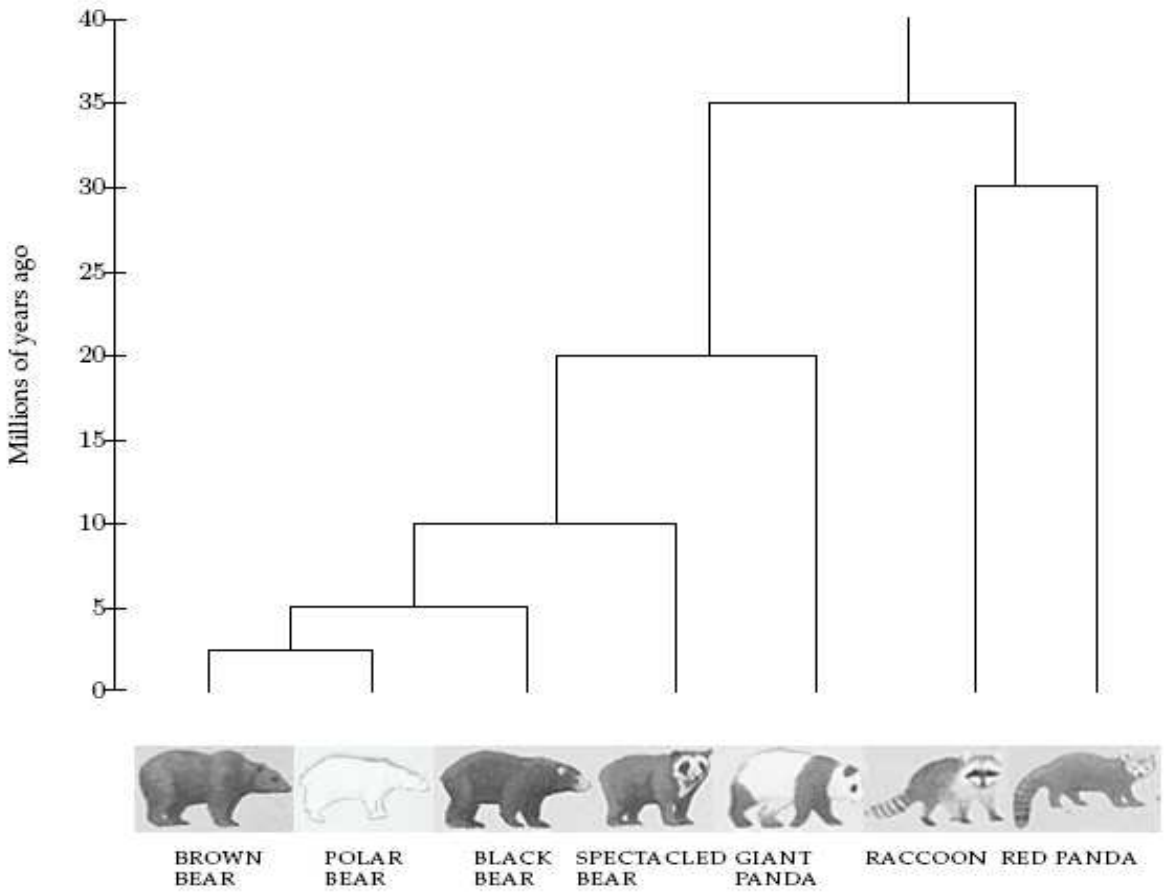
MOLECULAR PHYLOGENETICS



Evolutionary Tree for Bears and Raccoons

Outline

- Theoretical aspects of molecular phylogenetics
  - Definition and concepts
  - Distance matrices versus character-state approaches
  - Phylogenetic trees
    - Optimality criterion
    - Search strategy (heuristics)
  - Probabilistic model of sequence evolution
- Phylogenetic methods
  - Distance-based methods
    - UPGMA, Fish-Margoliash, Neigbour-Joining
  - The parsimony method (character-state approach)
  - The maximum likelihood method (character data)
- Bootstrapping

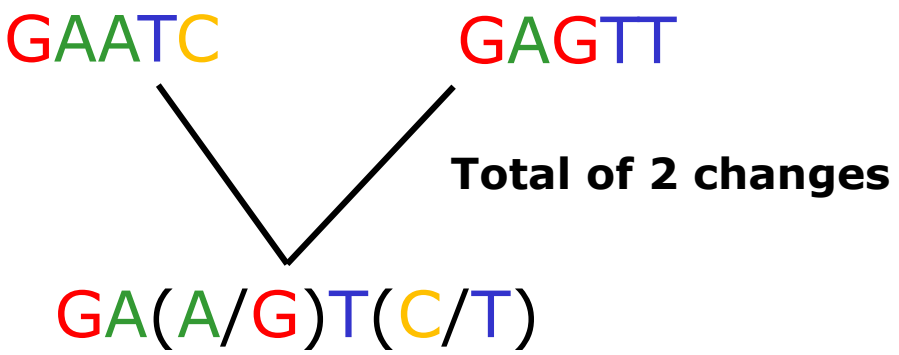
1. Molecular phylogenetics : definition and concepts

- Phylogeny is the inference of evolutionary relationships among biological species
  - To determine the closest homologs
  - To estimate time of species divergence
- Molecular phylogenetics is the use of molecular sequences to construct evolutionary trees
  - Evolution of proteins and nucleic acids is much more regular
  - Sequence analysis is more amenable to quantitative treatments and statistical evaluation
- A phylogenetic tree is a branching diagram that shows:
  - changes in genes and proteins,
  - the order these sequences diverged from one another
  - the orthologs and paralogs within a gene family

Phylogenetic tree : A two-dimensional graph showing ancestral relationships among homologous genes

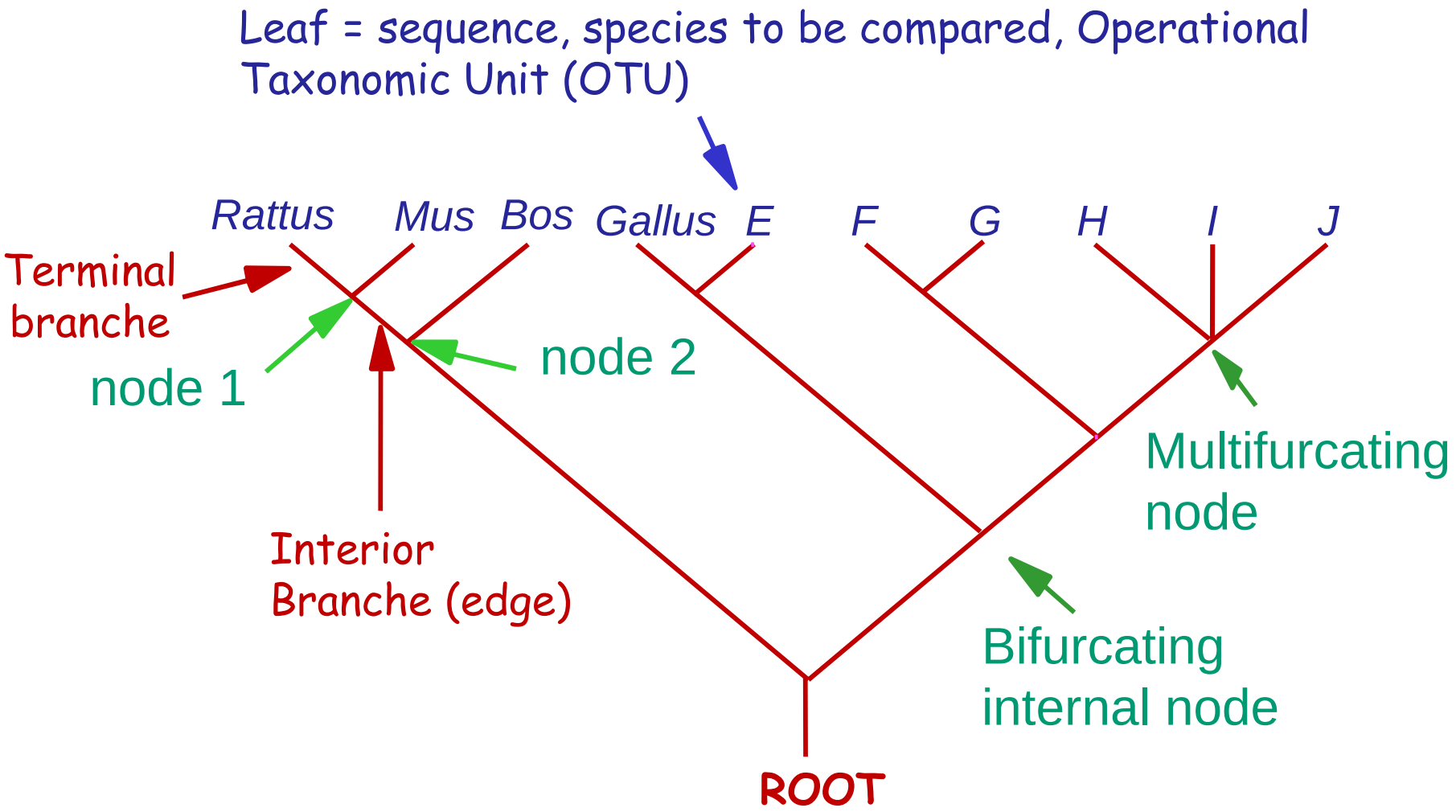
Sequences 1 and 2 are assumed to evolved from a common ancestor

GAATC sequence 1
GAGTT sequence 2



Some residues of the ancestor sequence can be inferred from conserved positions

Tree components: nodes, leaves, branches, (root)



Node = common ancestor taxonomic unit (taxon)

Molecular phylogenetics consists of four steps:

1. Selection of sequences to be analysed
2. Multiple sequence alignment
   - Overall similarity versus evolved character
3. Tree building
   - Distance matrix
   - Character-state
4. Tree evaluation
   - Bootstrapping

Enumerating trees

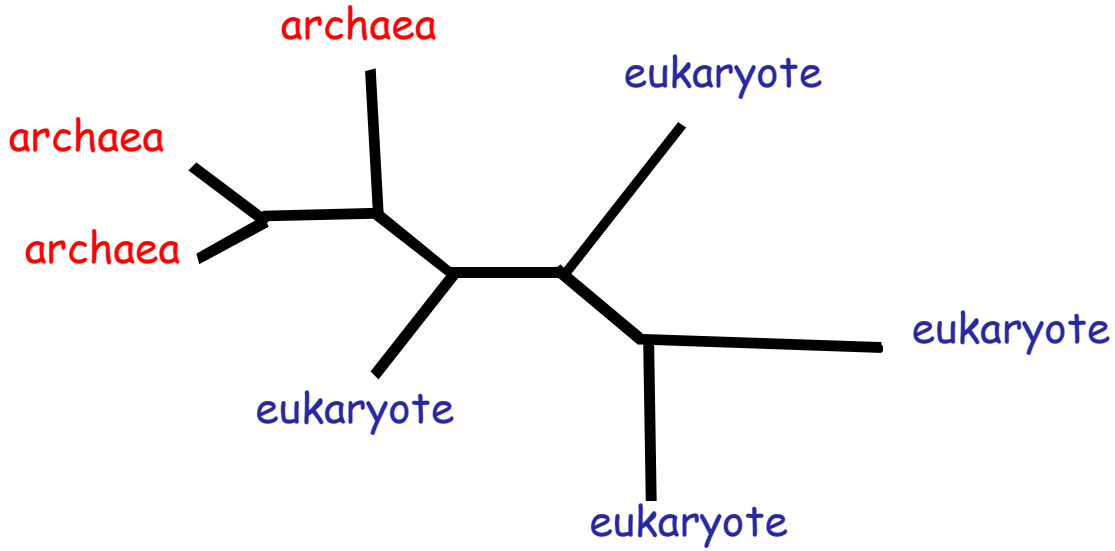
Number of OTUs	Number of rooted trees	Number of unrooted trees
2	1	1
3	3	1
4	15	3
5	105	15
10	34,459,425	2×10^6

$$N_r = (2n-3)!/2^{n-2}(n-2)! \quad \text{for } n \geq 2$$

$$N_u = (2n-5)!/2^{n-3}(n-3)! \quad \text{for } n \geq 3$$

Most phylogenetic methods produce unrooted trees

Heuristic search strategies for identifying the best tree

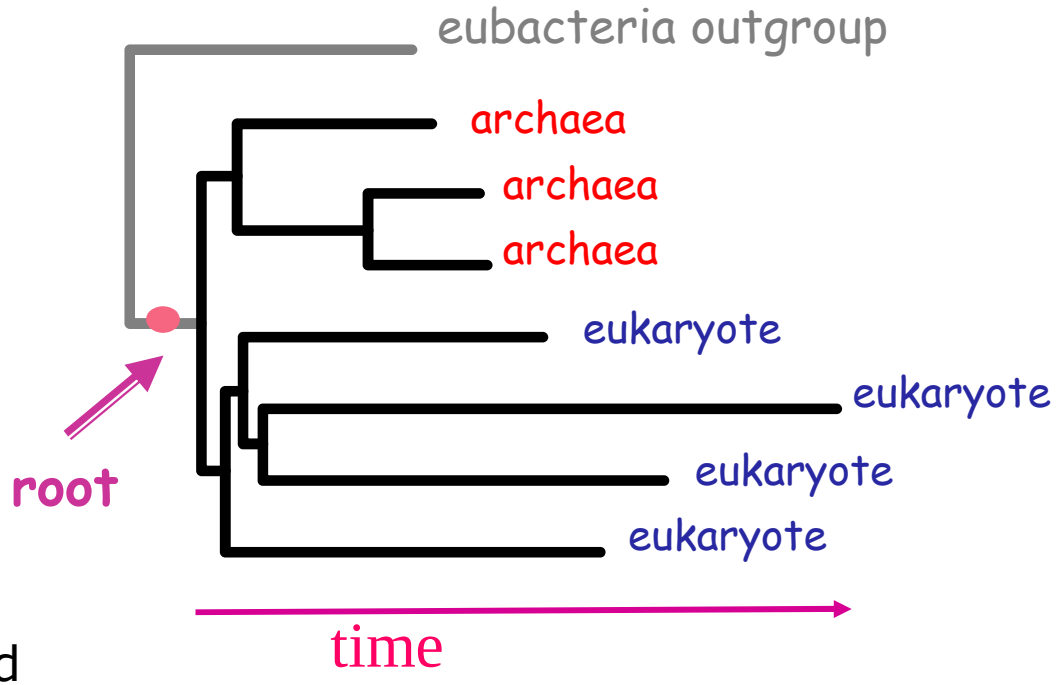


Unrooted tree

We cannot tell in which direction the change was happening

Rooting with an outgroup

- An OTU known to fall outside of the groups of interest
- An evolutionary path is specified



A midpoint method is illustrated with the UPGMA method (slide 10)

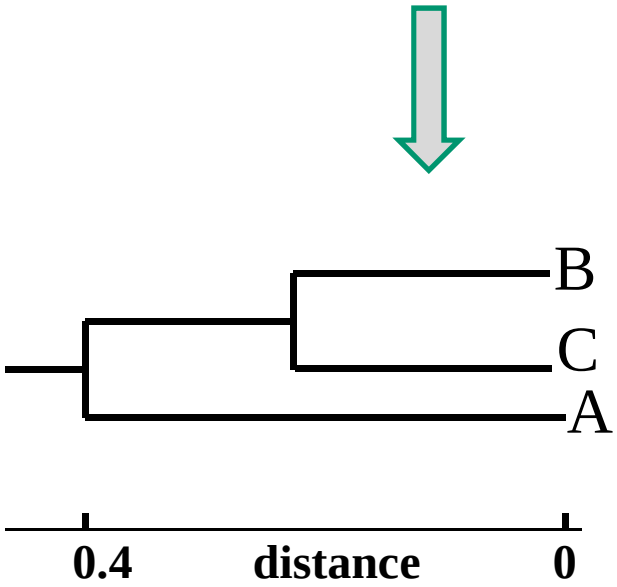
Phenetics

- Based on quantitative global dissimilarities
- Does not directly consider evolutionary history
- The inferred tree is named a phenogram
- UPGMA is a typical phenetic method

The UPGMA method

	A	B	C
A	0		
B	0.7	0	
C	0.9	0.4	0

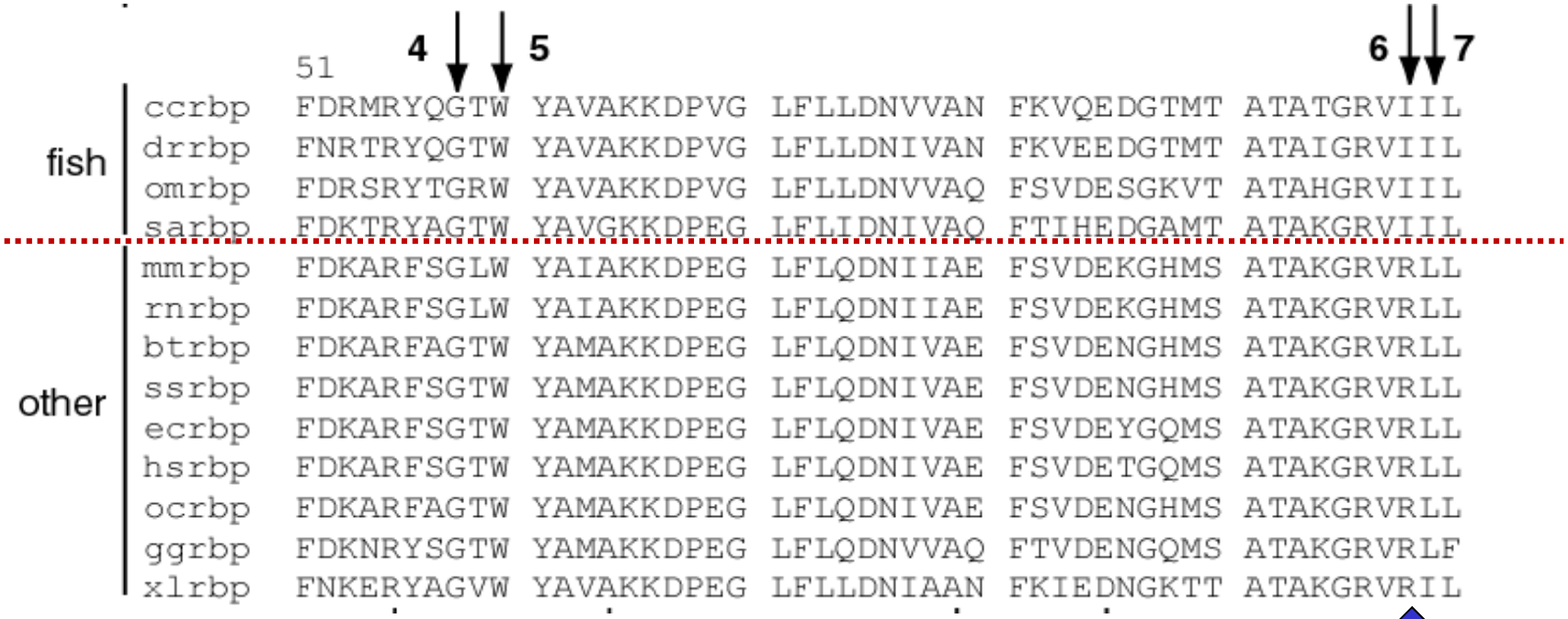
Distance matrix



Rooted Tree (the mean distance measured along the tree to the sequences on one side of the root is equal to the mean distance to the sequences on the other side = midpoint method; hypothesis of a molecular clock or similar rate of evolution

Cladistics methods

- Emphasize genealogical relationship and evolutionary history
- Use subsets of characters /sites
  - Make a distinction between evolved (apomorphic) versus ancestral (plesiomorphic)
- Consider shared characters that come from last common ancestor and are absent in more distant ancestors
- The cladogram focuses more on the tree topology (branch lengths are meaningless)
- Parsimony and maximum likelihood methods



Character-based tree: identify positions that best describe how characters (amino acids) are derived from common ancestors

- Only certain sequence variations at a given site are useful!

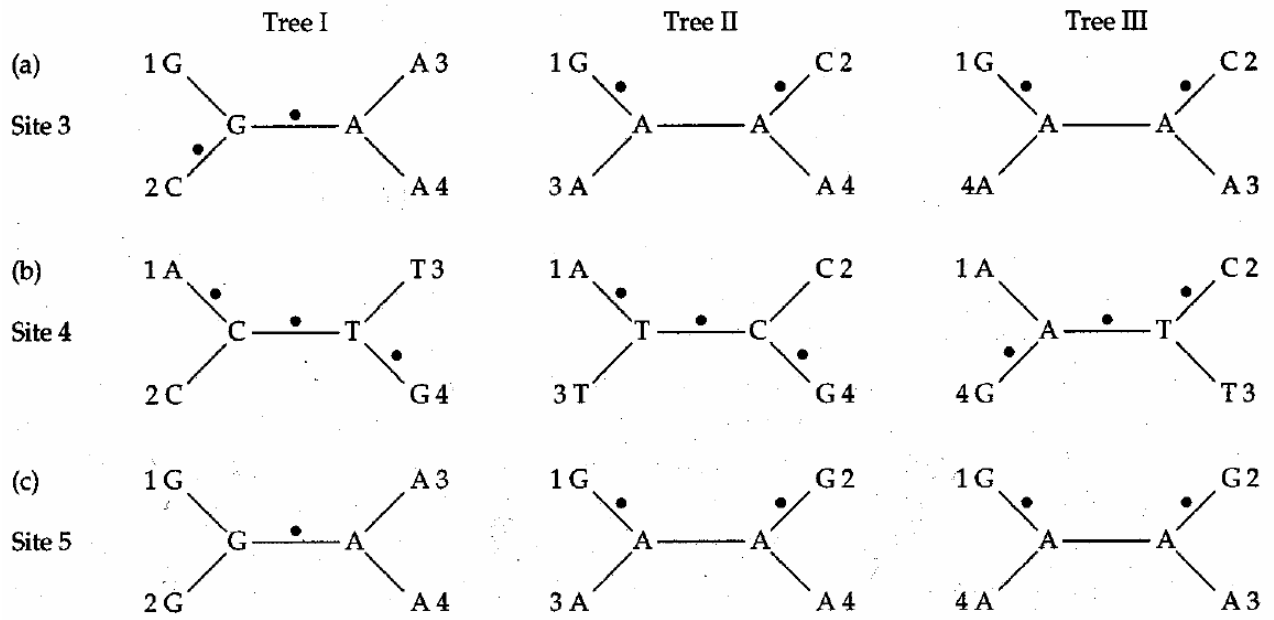
To be informative and help discriminate between trees, a site must have the same sequence character in at least two taxa (sites 5, 7 and 9)

Taxa	Sequence position (sites) and character								
	1	2	3	4	5	6	7	8	9
1	A	A	G	A	G	T	G	C	A
2	A	G	C	C	G	T	G	C	G
3	A	G	A	T	A	T	C	C	A
4	A	G	A	G	A	T	C	C	G

Adapted from Li and Graur 1991.

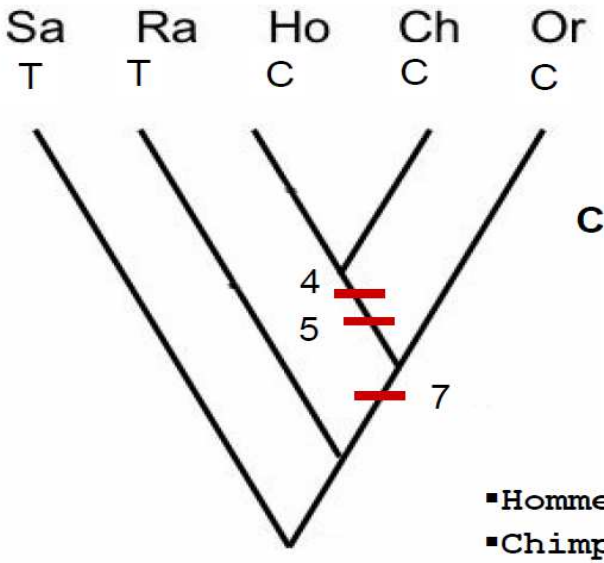
noninformative

informative



Construction of the cladogram

- Assessment of all possible trees
- Localisation of the transformations on the tree
- Choice of the most parcimonious tree



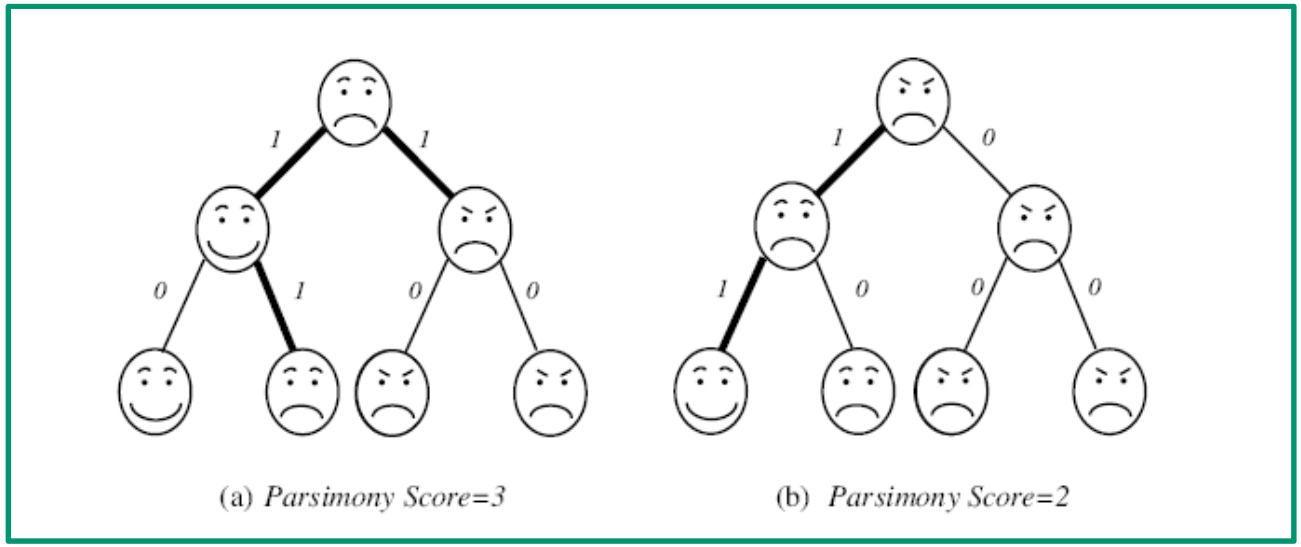
Cladogramme

Informative sites

	4	5	7		
▪Homme	TAC	TTC	CCG	CAC	TTC
▪Chimpanzé	---	---	--C	---	---
▪Orang-outang	---	GG-	--C	---	---
▪Rat surmulot	---	GG-	T-T	---	A-T
▪Saumon atlantique	---	GG-	T-C	---	-GG

Parsimony with morphological characters

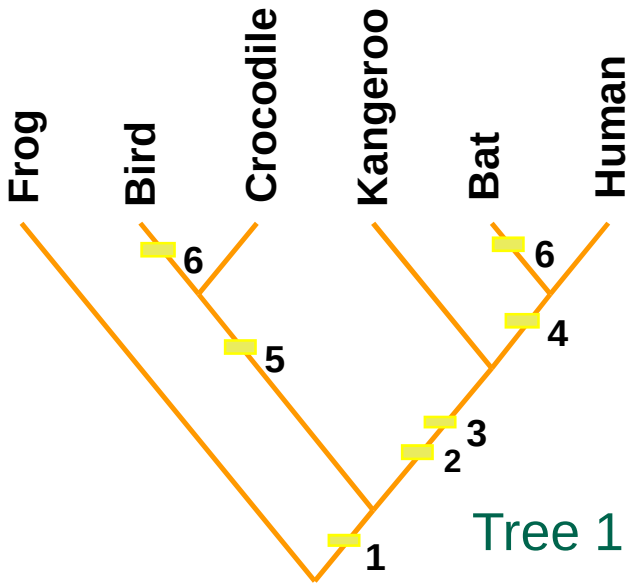
- A character alone does not say anything about the branching order within the "haves" and the "have-nots"
- We need a large set of characters that evolved at different point since time

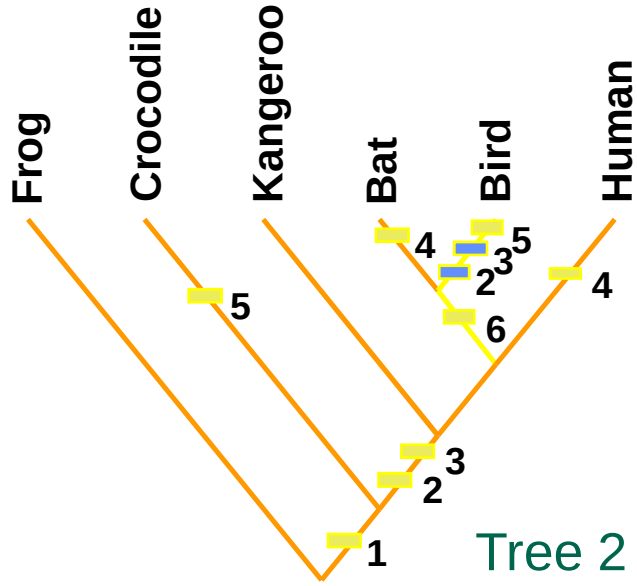


(a) *Parsimony Score=3* (b) *Parsimony Score=2*

- The sums over all characters is called the tree length
- The most parcimonious trees with the minimum tree length are selected

		CHARACTERS						
		1	2	3	4	5	6	
		amnion	hair	lactation	placenta	antorbital fenestra	wings	
TAXA	Frog	-	-	-	-	-	-	
	Bird	+	-	-	-	+	+	
	Crocodile	+	-	-	-	+	-	
	Kangeroo	+	+	+	-	-	-	
	Bat	+	+	+	+	-	+	
	Human	+	+	+	+	-	-	TREE LENGTH
FIT	Tree 1	1	1	1	1	1	2	7
	Tree 2	1	2	2	2	2	1	10



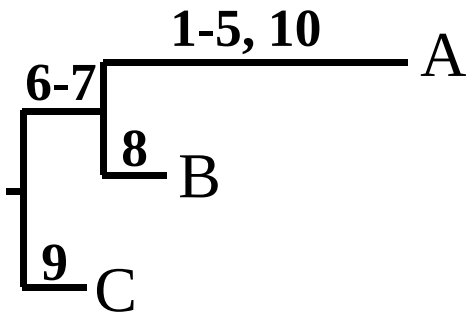


- Maximise congruence by looking for synapomorphies (derived character shared by ≥ 2 taxonomic groups)
- The shortest tree1 still requires homoplasy (convergence /reversion) : wings

The same data may give different trees

Cladistic method

	character									
	1	2	3	4	5	6	7	8	9	10
sequence										
A	1	1	1	1	1	1	1	0	0	1
B	0	0	0	0	0	1	1	1	0	0
C	0	0	0	0	0	0	0	0	1	0

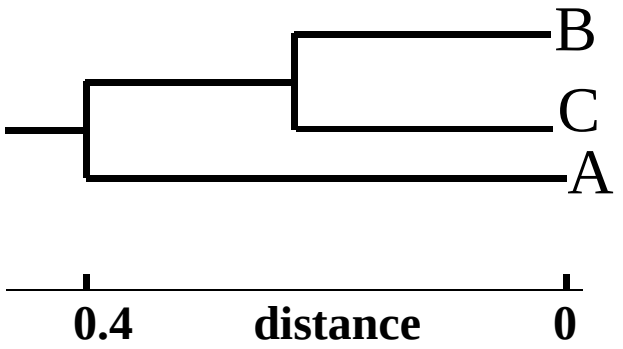


1 : evolved state

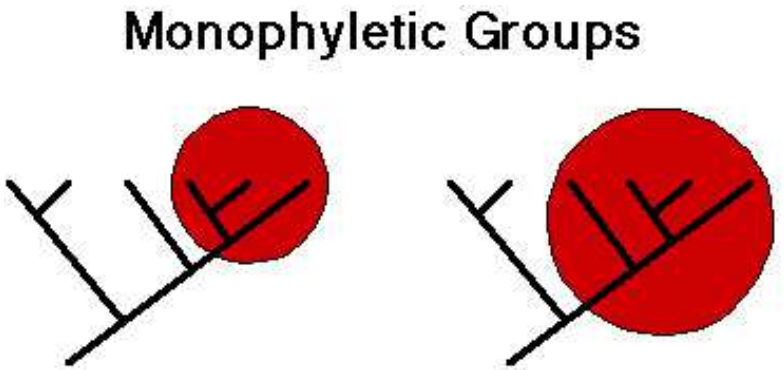
0 : primitive state

Distance method (UPGMA) (% difference)

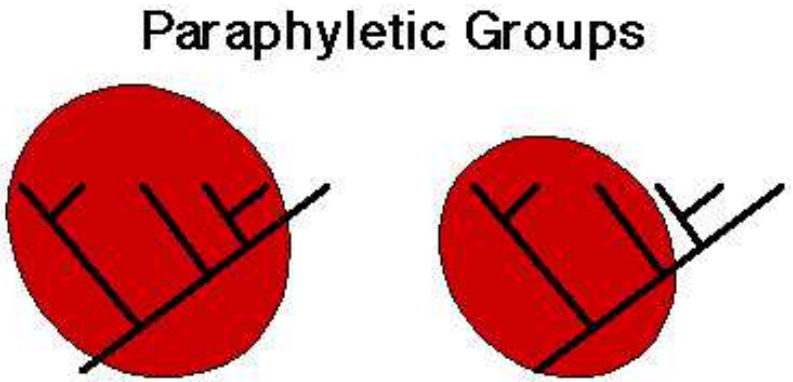
	A	B	C
A	0		
B	0.7	0	
C	0.9	0.4	0



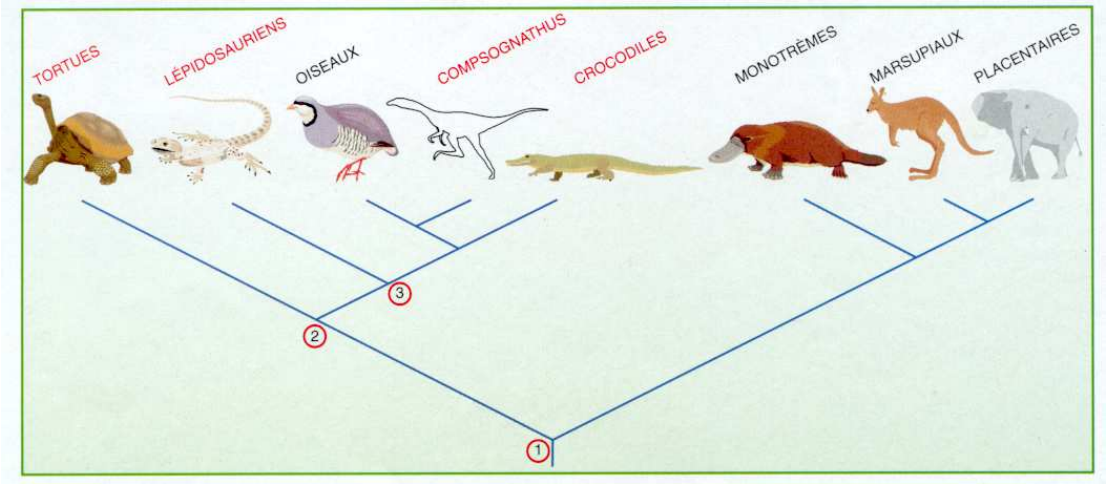
- Clades contain all the OTUs derived from a single common ancestor, and which includes all the descendents of that ancestor



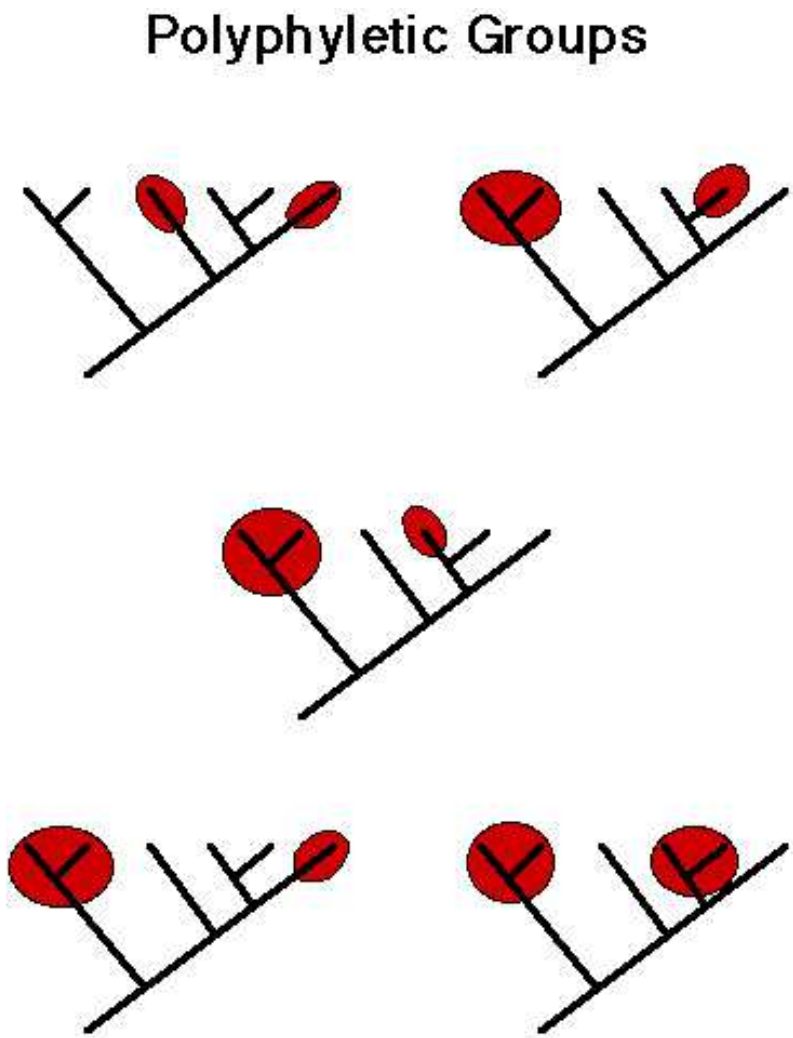
- Paraphyletic groups contain species (taxa) which are derived from a common ancestor but the groups do not include all descent taxa of the same common ancestor



Reptiles as a paraphyletic group



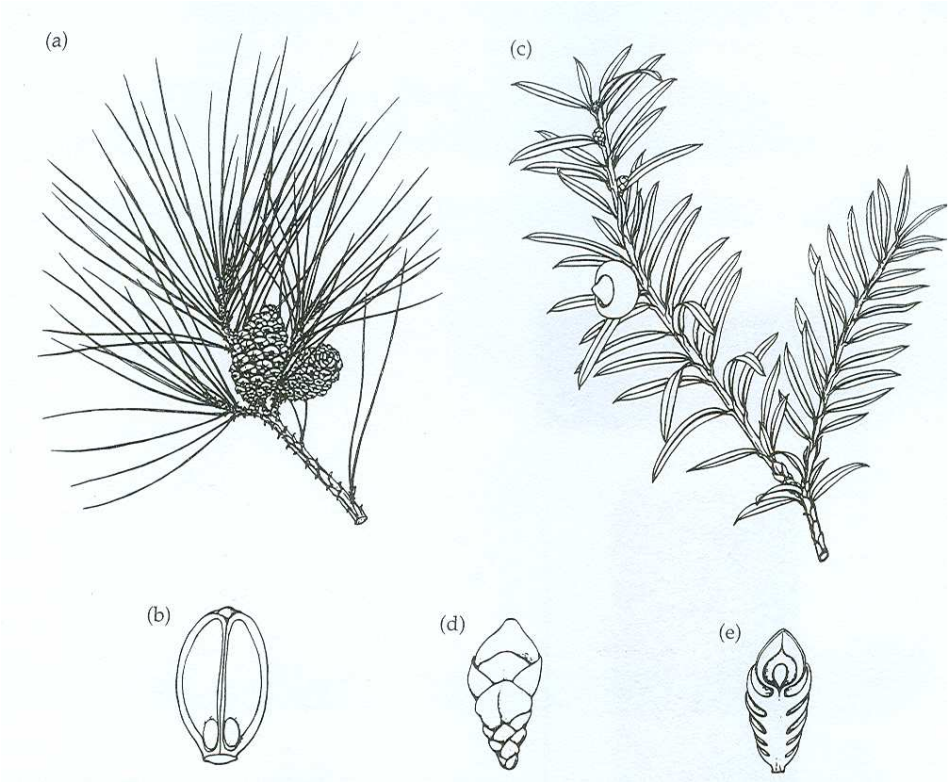
Polyphyletic groups contain species (taxa) which are derived from more than one common ancestor



Phylogenetics analysis of conifers

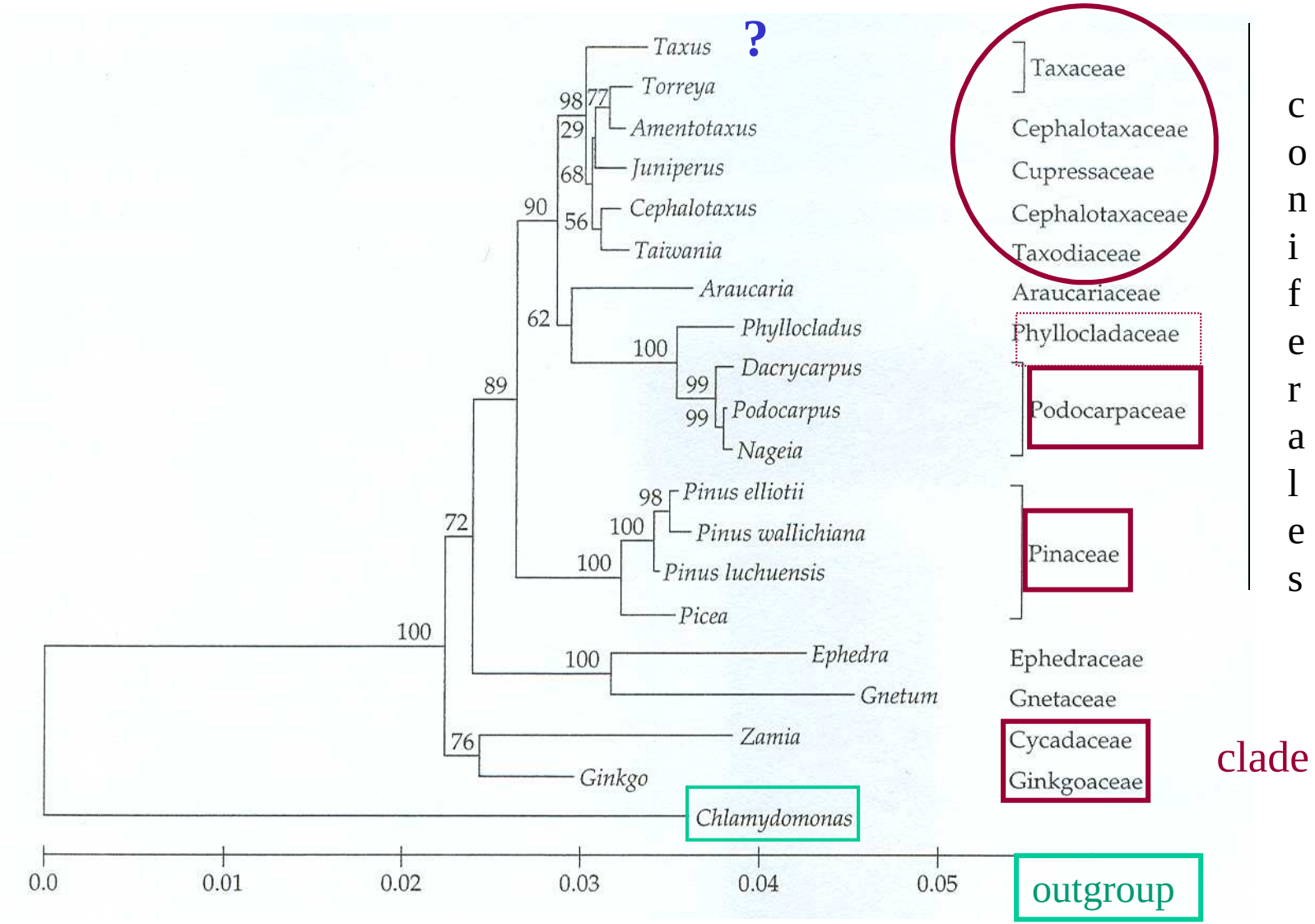
The pine Pinus luchuensis: (a) branch with mature female cones and (b) the adaxial view of ovuliferous scale dissected from cone.

The yew Taxus mairei: (c) branch with mature fruit and ovuliferous branchlet, (d) the lateral view of ovuliferous branchlet and (e) a longitudinal section showing the ovule



Question: Are the yews conifers or not?

Phylogeny of gymnosperms inferred from 18S rRNA/DNA sequences



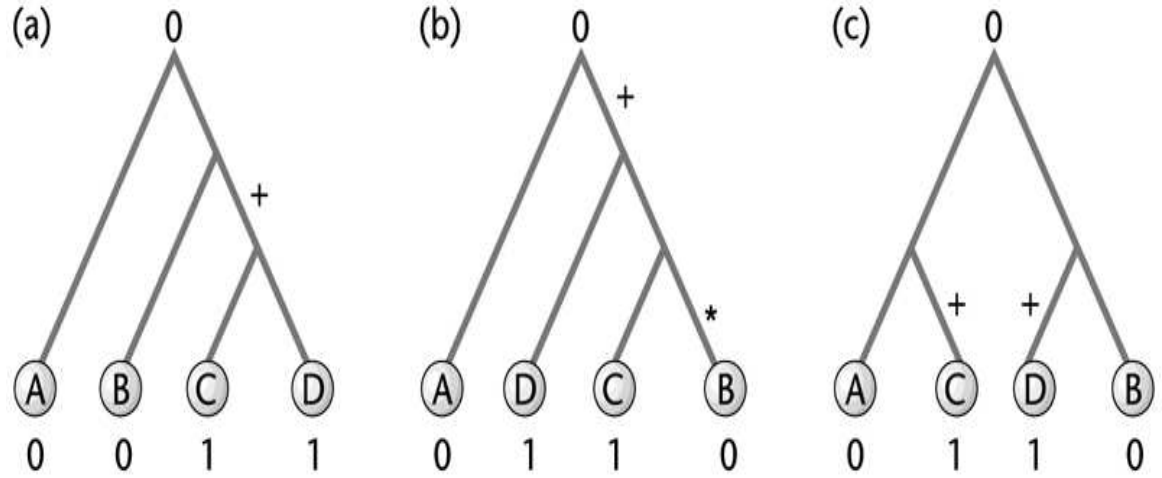
The yew belongs to the coniferales, a paraphyletic group

2. Algorithms used in molecular phylogenetics

2.1. Maximum parsimony

- Character-state approach
  - Define a set of binary character states, 0 and 1, where 0 represents the ancestral state and 1 represents a derived state
  - Find a tree giving the simplest possible explanation of the data

The parsimony criterion says that tree (a) is to be preferred.
Note homoplasy in b (reversion) and c (convergence)

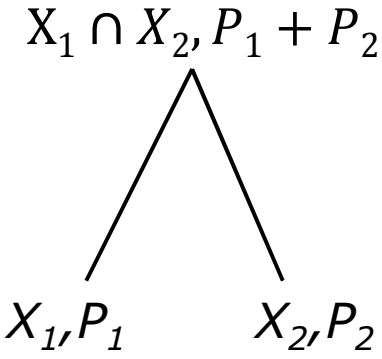


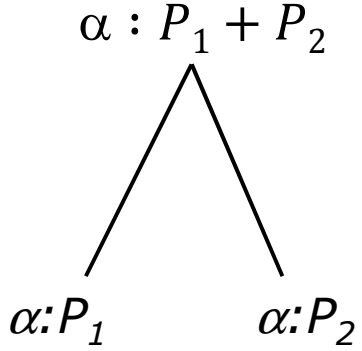
These trees are unrooted !

- Fitch algorithm

Goal: calculate the minimal number of changes that are consistent with all the leaves of a tree

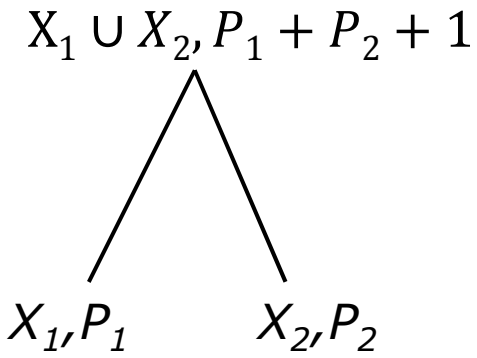
1st case: $X_1 \cap X_2$ *is not empty*

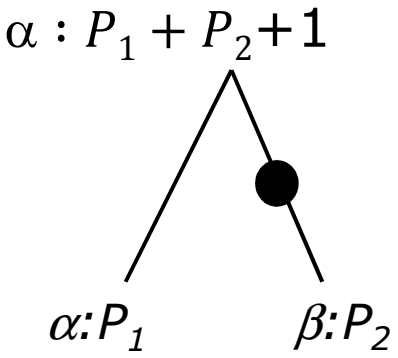


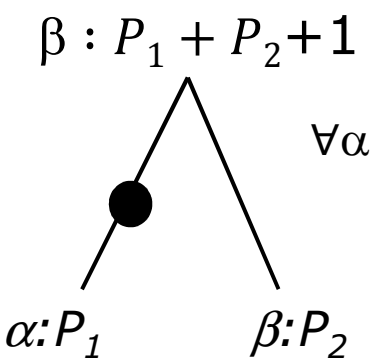


$\forall a \in X1 \cap X2$

2nd case: $X_1 \cap X_2$ *is empty*





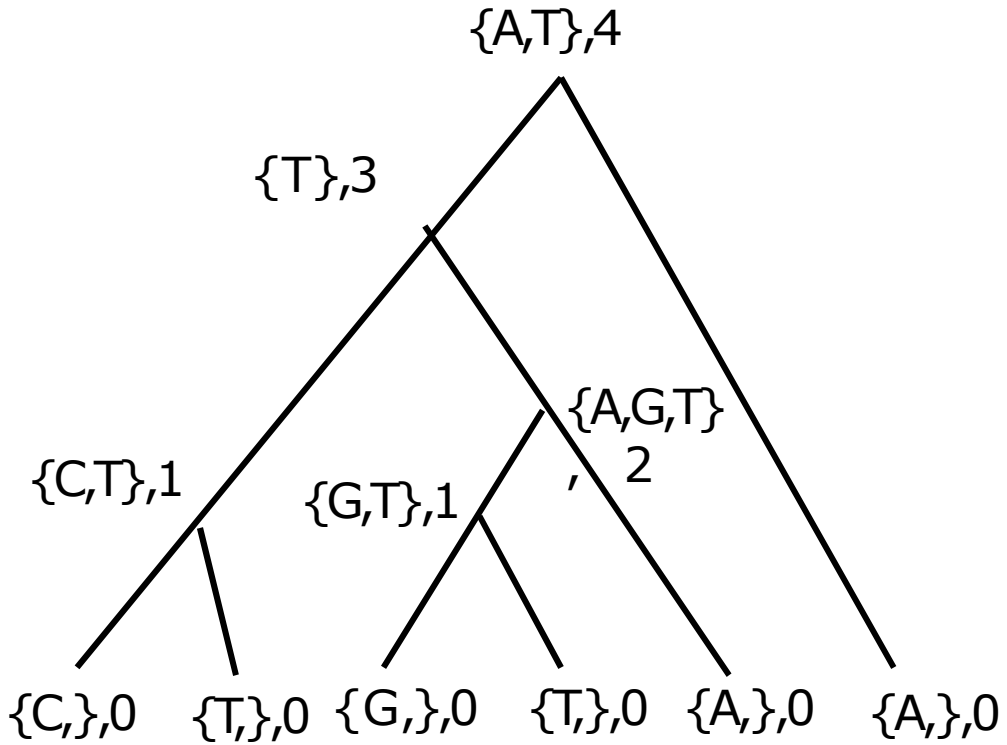


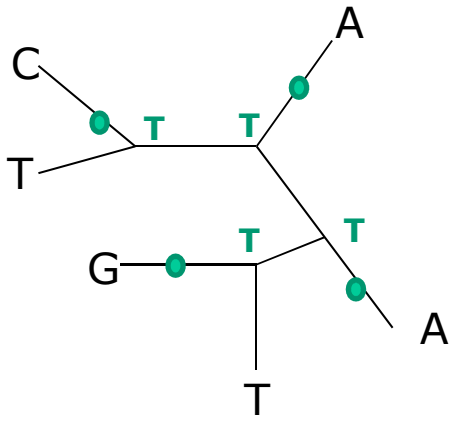
$\forall \alpha \in X_1, \forall \beta \in X_2$

- Fitch algorithm

$$\text{If } X_1 \cap X_2 \neq \emptyset, then\ X = X_1 \cap X_2\ and\ P = P_1 + P_2$$

$$\text{If } X_1 \cap X_2 = \emptyset, then\ X = X_1 \cup X_2\ and\ P = P_1 + P_2 + 1$$





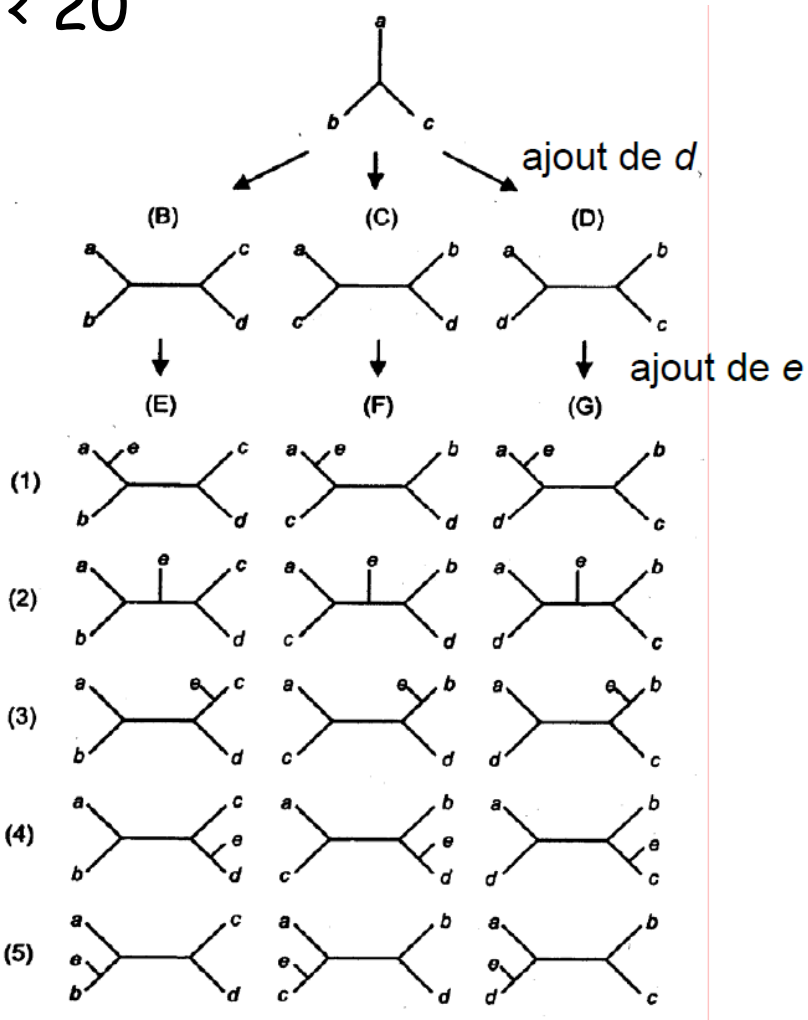
Possible cladogram with 4 steps

Search of alternative tree topologies

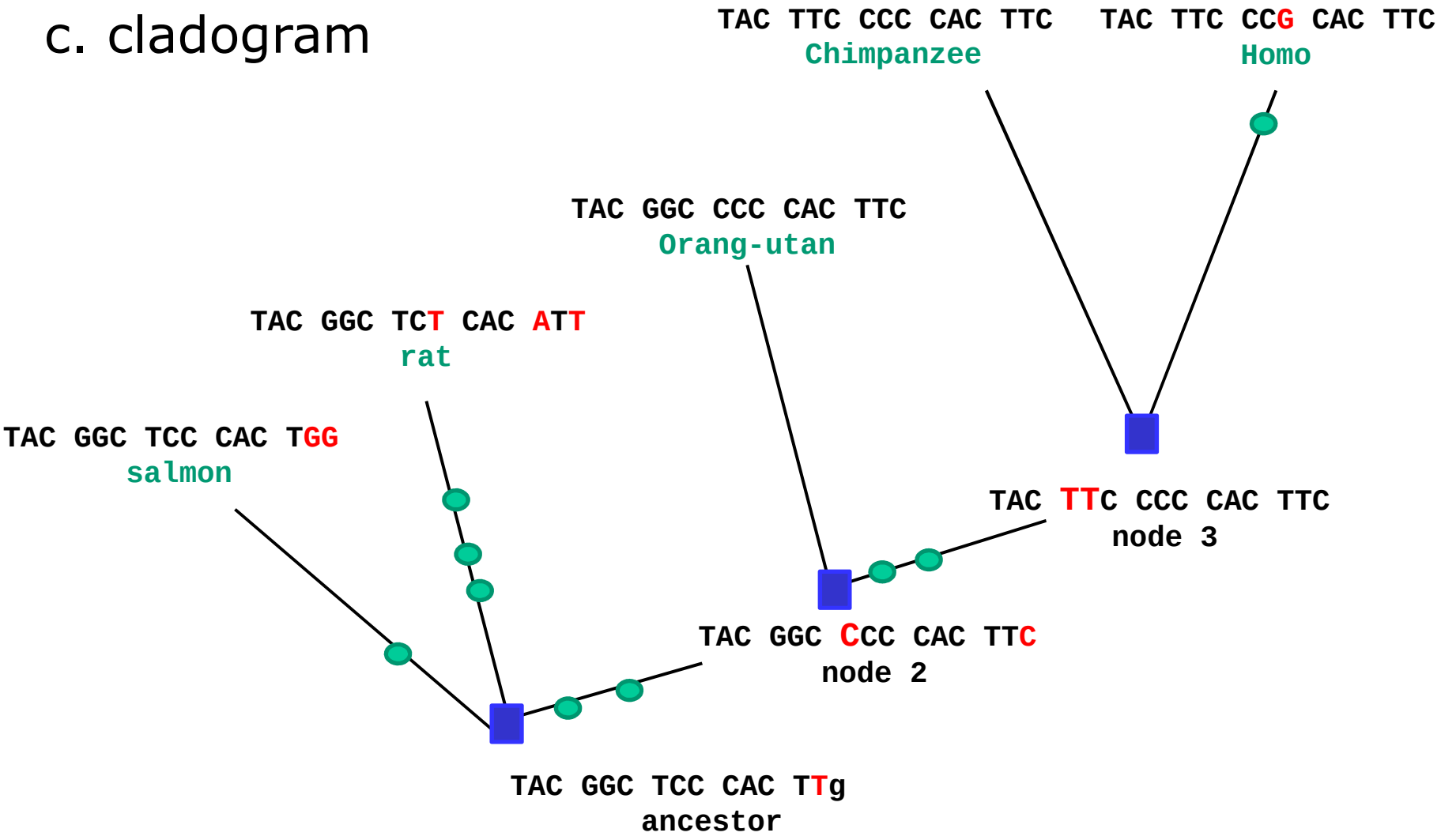
- Parsimony analysis is time consuming: m < 10
- Alternative approaches
  - branch-and-bound method m < 20

1. Start with a 3 OTU-tree (arbitrary order)
2. Add one more OTU to every branch of the tree and keep the position with the lowest cost
3. Assess the nearest neighbor trees and keep the tree with the lowest score
4. Go back to step 2

Repeat the whole procedure changing the order of the sequences to be added



c. cladogram



9 steps

Conclusions

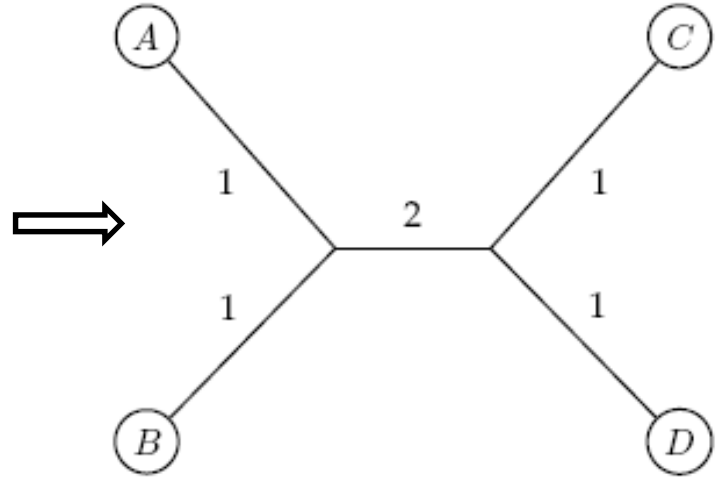
- Not suitable for more than 20 sequences without heuristics
- One or more most parsimonious trees → consensus tree
- Does not seem to depend on an explicit model of evolution
- Gives both trees and associated hypotheses of character evolution
- Branch lengths do not have an unique meaning
- Programs PAUP and DNAPARS and PROTPARS (PHYLIP)

2.2 Distance matrix methods

Goal: Reconstruct an evolutionary tree from a distance matrix

- Input: N x N distance matrix D_{ij} (obtained from an alignment and correcting for multiple substitutions)
- Output: weighted tree T_{ij} with n leaves fitting D_{ij}

	A	B	C	D
A	0	2	4	4
B	2	0	4	4
C	4	4	0	2
D	4	4	2	0



Most common distance methods

- The UPGMA method (molecular clock, rooted tree)

  Change rate along the branches of a tree is constant and the distance matrix is ultrametric

- The Fitch-Margoliash method (no clock, unrooted tree)

  Change rates are variable (different branch lengths)

  Least-squares fit of alternative trees

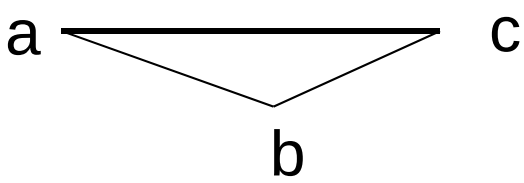
- The Neighbor-Joining method of Satou and Nei (no clock, unrooted tree)

  Similar to the FM method except that the choice as to which sequences to pair at the very first step is determined so to give one tree only

Ultrametric distance matrix

- A tree defines a metric; i.e., the pairwise distances between all pairs of leaves have properties such as:

1. $d(a,a)=0$, $d(b,b) = 0$ and $d(a,b) \geq 0$
2. If $a \neq b$, then $d(a,b) > 0$
3. $d(a,b) = d(b,a)$
4. $d(a,b) \leq d(a,c) + d(b,c)$



$\rightarrow d(a,b) \leq \max\{d(a,c),d(b,c)\}$

This is generally defined as the three point condition

- An ultrametric matrix fits a rooted tree
  - ✓ all distances from the root to a leaf are the same
  - ✓ satisfies the molecular clock hypothesis

Unweighted Pair Group Method with Arithmetic mean (UPGMA)

1. Begin with N x N symmetric distance matrix
2. At each step, the two closest taxa are selected as neighbors
3. The height of the least common ancestor of any pair of leaves is half the distance between the leaves
4. Calculate a new distance matrix between the new composite taxon and all existing taxa
   - Reduces size of the distance matrix
5. Go to step 2. Continue until all taxa have been merged into a single cluster

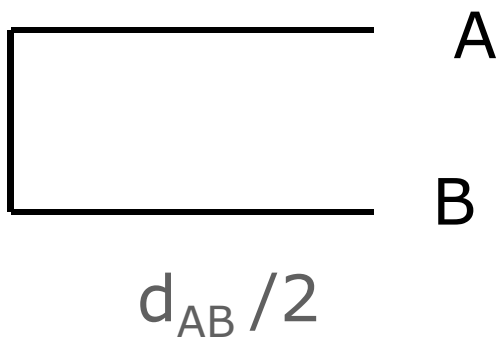
UPGMA algorithm

Let us consider a case of four OTUs,

- d_{AB} has the lowest value

	OTU		
OTU	A	B	C
B	d_{AB}		
C	d_{AC}	d_{BC}	
D	d_{AD}	d_{BD}	d_{CD}

Tree at step 1

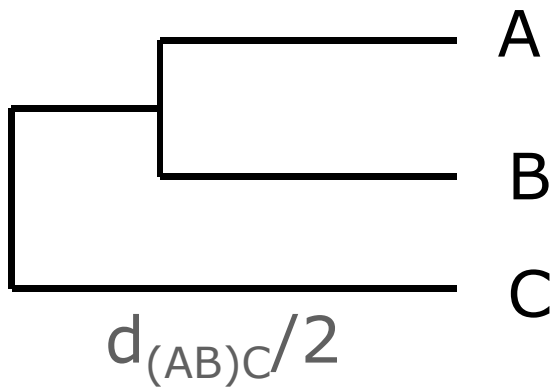


- After the first clustering, A and B are treated as a single composite OTU and a new distance matrix is calculated.
- $d_{(AB)C}$ turns out the be the smallest distance

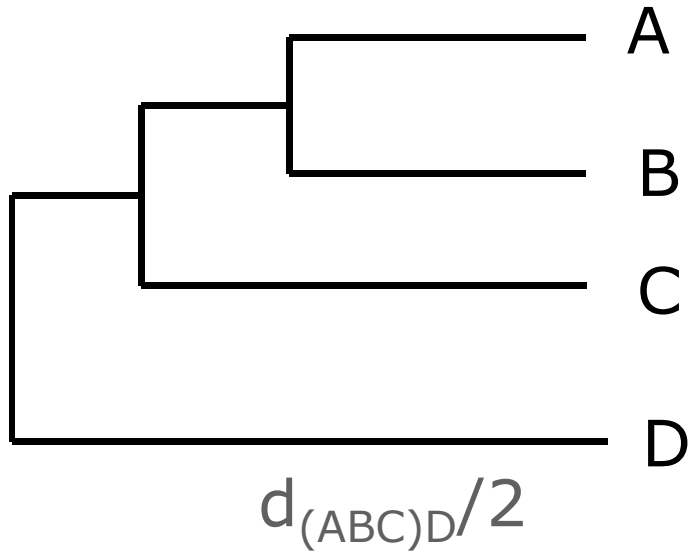
	OTU	
OTU	(AB)	C
C	$d_{(AB)C}$	
D	$d_{(AB)D}$	d_{CD}

$d_{(AB)C} = (d_{AC} + d_{BC})/2$ and $d_{(AB)D} = (d_{AD} + d_{BD})/2$

Tree at step 2



Tree at step 3



$d_{(ABC)D} = (d_{AD} + d_{BD} + d_{CD})/3$

- The distance between two composite OTUs is computed as the arithmetic mean of the pairwise distances between the constituent OTUs of the two composite OTUS

$$d_{(ij)(mn)} = (d_{im} + d_{in} + d_{jm} + d_{jn})/4$$

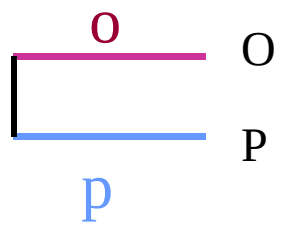
In general: $d_{XY} = \sum_{i,j} d_{ij}/(nXnY)$

Distance matrix

OTU	M	N	O	P
M	0.0	15	26	28
N		0.0	29	31
O			0.0	12
P				0.0

First cycle:

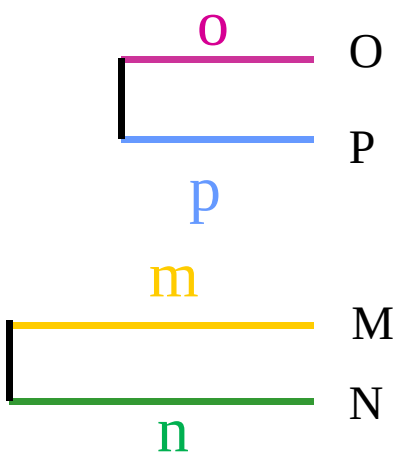
- select the two species x_i, x_j with the minimal distance
- join the species with edges length equal to $d(x_i,x_j)/2$



o = p = 6

- Second Cycle

OTU	M	N	(O,P)
M	0	15	27 (26+28)/2
N		0	30 (29+31)/2
(O,P)			0

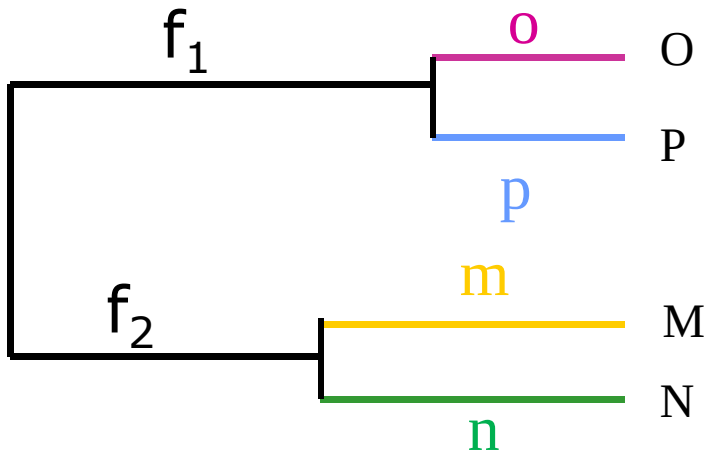


o = p = 6

m = n = 7.5

- Last cycle

OTU	(M,N)	(O,P)
(M,N)	0.0	28,5 (26+28+29+31)/4
(O,P)		0.0



$f_2 + m = 28,5/2$

$f_2 = 14.25 - 7.5 = 6.75$

$f_1 = 8.25$

Note that the tree distances are different from the observed distances

D	N	O	P
M	15	26	28
N	0	29	31
O		0	12

Observed distances

T	N	O	P
M	15	28.5	28.5
N	0	28.5	28.5
O		0	12

Tree distances

UPGMA : an example

We obtained the wrong tree because D is not ultrametric !

Consider the triple, {M,N,O}

- 15 ≤max(26,29)
- 29 ≤max(15,26)
- 26 ≤max(15,29)

D	N	O	P
M	15	26	28
N	0	29	31
O		0	12

However the matrix is additive D as the 4-point condition is satisfied and there exists a unique tree topology with branch lengths such that T[i,j] = D[i,j]

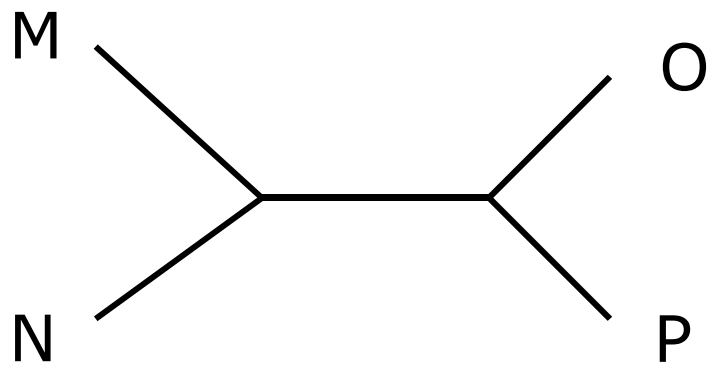
In real life, observed distance matrix is never additive

A matrix is additive if it satisfies the 4-point condition

- The 4-point condition holds for all quartets

$d(a,b) + d(c,d) \leq \max \{d(a,c) + d(b,d), d(a,d) + d(b,c)\}$

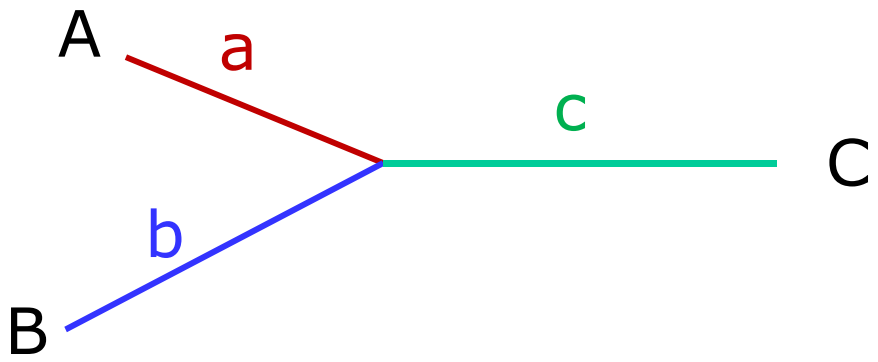
$d(a,b) + d(c,d) \leq d(a,c) + d(b,d) = d(a,d) + d(b,c)$



MO + NP = 26 + 31 = 57
MN + OP = 15 + 12 = 27
MP + NO = 28 + 29 = 57

Fitch and Margoliash method

- No molecular clock is considered, therefore the branches linking the leaves to the shared node might have different lengths
- Branch lengths is inferred from a set of 3 pairwise distances



$$a+c = d_{AC}$$
$$b+c = d_{BC}$$
$$a+b = d_{AB}$$

The solution is:

$$a = \tfrac{1}{2}(d_{AB} + d_{AC} - d_{BC})$$
$$b = \tfrac{1}{2}(d_{AB} + d_{BC} - d_{AC})$$
$$c = \tfrac{1}{2}(d_{AC} + d_{BC} - d_{AB})$$

- Example of the F-M method

OTU	M	N	O	P
M	0.0	15	26	28
N		0.0	29	31
O			0.0	12
P				0.0

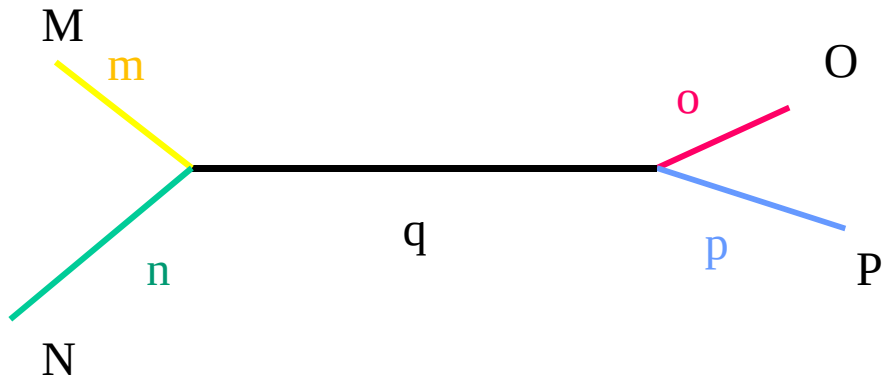
1. Select the two species with the minimal distance and calculate the mean distance to the remaining taxa

New matrix with the remaining taxa combined

	O	P	(M,N)
O	0.0	12	27.5
P		0.0	29.5
(M,N)			0.0

2. Determine branch lengths *o* and *p*

	O	P	(M,N)
O	0.0	12	27.5
P		0.0	29.5
(M,N)			0.0



$d_{OP} = o+p = 12$ (1)

$d_{O(M,N)} = (o+q+m+o+q+n)/2 = 27.5$ (2)

$d_{P(M,N)} = (p+q+m+p+q+n)/2 = 29.5$ (3)

Substract (2) from (3), p - o = 2 (4)

Add (1) to (4), 2p = 14, p = 7 and o = 5

3. Treat the O and P taxa as a single composite taxa (O,P) and calculate a new distance table

OTU	M	N	(O,P)
M	0.0	15	27
N		0.0	30
(O,P)			0.0

4. Determine branch lengths *m* and *n*

$d_{MN} = m+n = 15$ (1)

$d_{M(O,P)} = (m+q+o+m+q+p)/2 = 27$ (2)

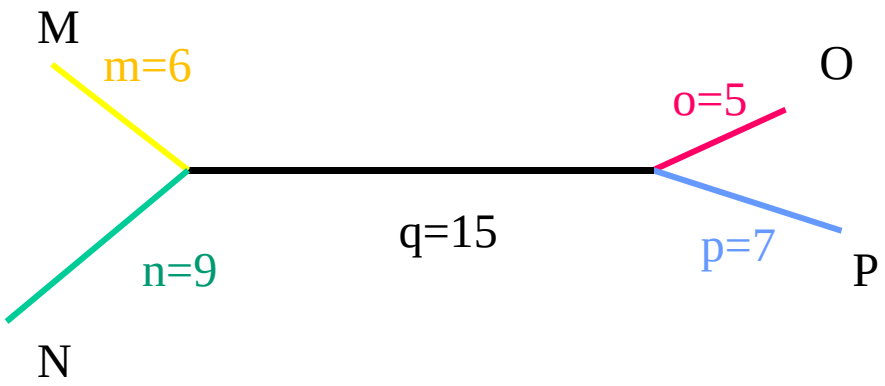
$d_{N(O,P)} = (n+q+o+n+q+p)/2 = 30$ (3)

Substract (2) from (3), n - m = 3 (4)

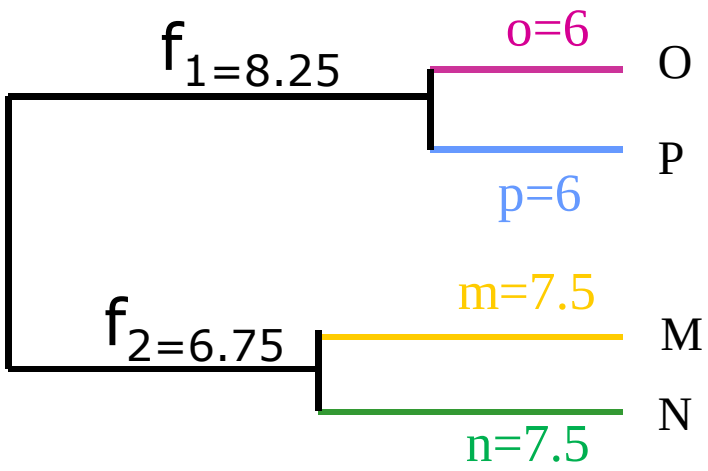
Add (1) to (4), 2n = 18, n = 9 and m = 6

Tree comparison

F-M



UPGMA



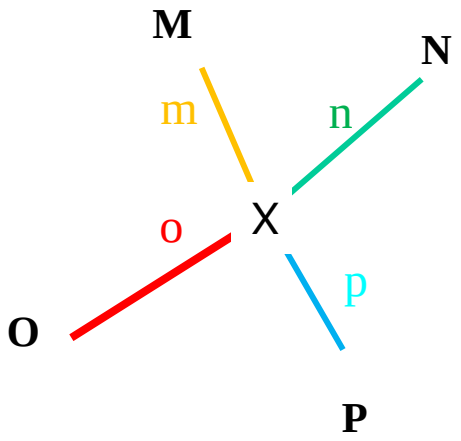
- If D is not additive, we look for a tree T that best fits the original data using a least squares method :

$$\Sigma_{i,j} (T[ij] - D[ij])^2$$

- The FM method repeats the process starting with all possible pairs of sequences M and O, N and O, M and P,...)
- The percent change from the predicted to the original distances is determined for each sequence pair
- These values are squared and summed over all possible pairs
- This sum divided by the numbers of pairs n(n-1)/2 less one (the number of degrees of freedom) provides the square of the percent standard deviation of the result

The Neighbor-joining method (Saitou and Nei, 1987)

1. Begin by placing all the taxa in a star-like topology.



$$\sum_{i=1}^{m} L_{iX} = \frac{1}{m-1} \sum_{i<j}^{m} d_{ij}$$

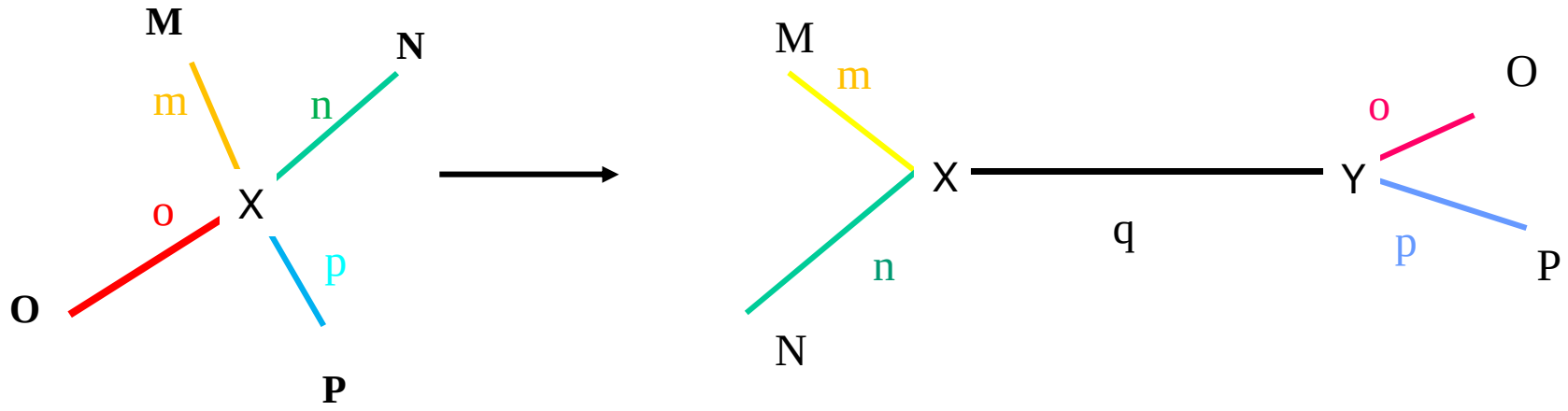
L_{ix}= branch length
d_{ij}=distance
m= number of taxa

The sum of branch lengths is 141/3 = 78.5

In summing the 6 distances = 15 + 26 + … + 12 = 141, each branch m,n,o,p, is counted 3 times

2.Decompose the star-like tree by combining pairs of sequences

- O and P are connected to remaining taxa via an internal branch, q. Calculate the sum of the branch lengths



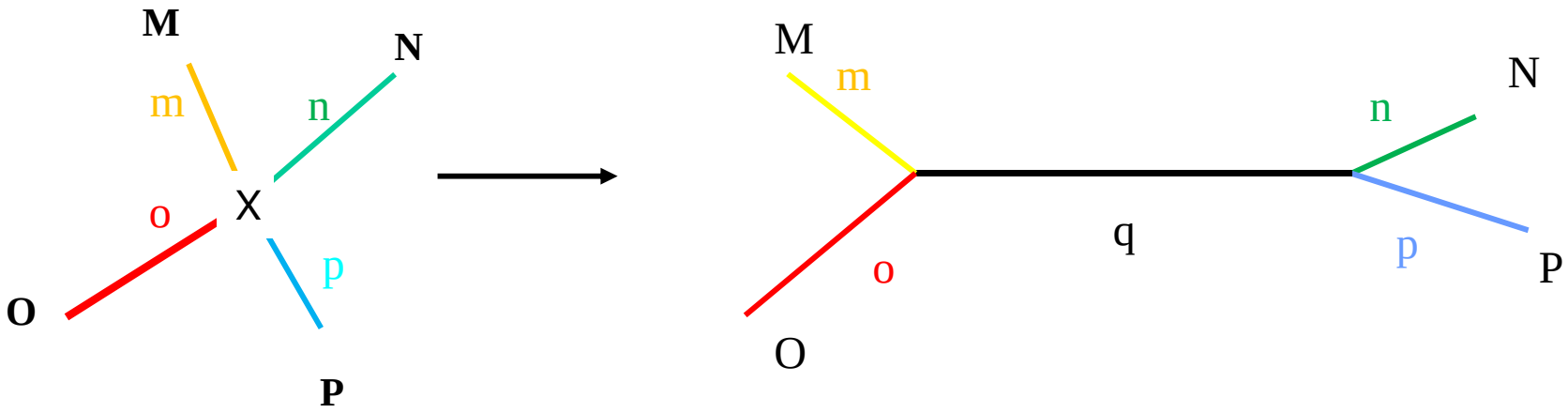
Sum of branch lengths (Σ_{OP}) = m + n + q + o + p

= (26+29+28+31)/4 +12/2 + 15/2 = 42

$\Sigma_{OP}=[\Sigma(d_{iO} + d_{iP})/2(m-2)] + d_{OP}/2 + \Sigma d_{ij} / m-2$ with i, j ≠ O,P and i < j

- Repeat the calculation for each combination :

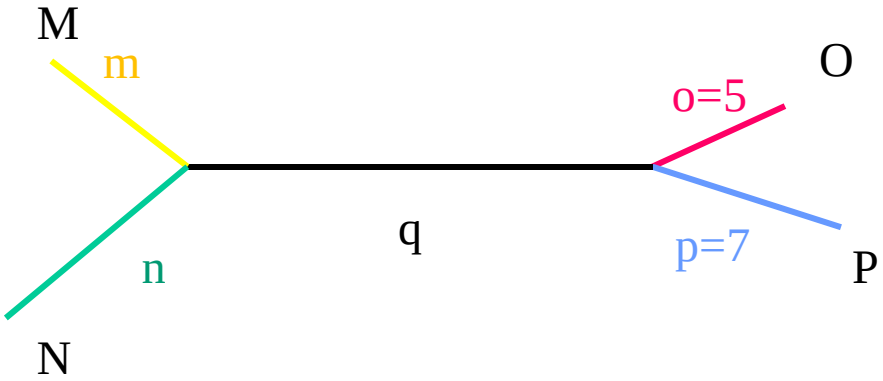
Σ_{MO} = 49.5 and Σ_{MP} = 49.5 (symetrical topology)
Σ_{MN}= 42



and select the neighbors reducing the total branch lengths to the largest extent.

3. Once the choice of neighbors O, P has been made, the branch lengths *o* and *p* are calculated by the F-M method

	O	P	(M,N)
O	0.0	12	27.5
P		0.0	29.5
(M,N)			0.0



$d_{OP} = o+p = 12$ (1)

$d_{O(M,N)} = (o+q+m+o+q+n)/2 = 27.5$ (2)

$d_{P(M,N)} = (p+q+m+p+q+n)/2 = 29.5$ (3)

Substract (2) from (3), $p - o = 2$ (4)

Add (1) to (4), $2p = 14 \Rightarrow p = 7$ and $o = 5$

4. A new table of distance with O and P forming a single composite sequence is produced.

The N-J algorithm is used to find the next pair and the F-M algorithm is then used to find the next branch lengths

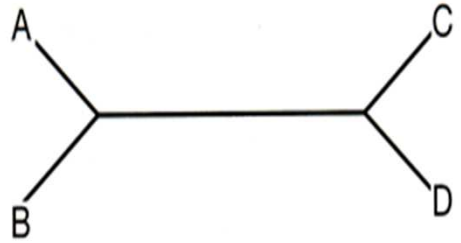
The cycle is repeated until the corrected branched tree and the branch distances on that tree have been identified

2.3 Maximum likelihood method

A. Sequences

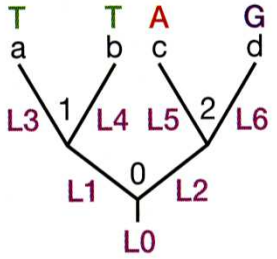
sequence a	ACGCGTTGGG
sequence b	ACGCGTTGGG
sequence c	ACGCAATGAA
sequence d	ACACAGGGAA

B. An unrooted phylogenetic tree for the sequences A-D.

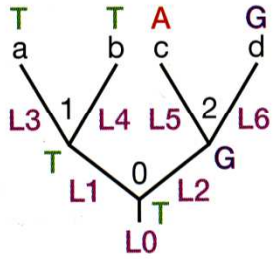


This method uses probability calculations to find a tree that best accounts for the variation in the sequences
It uses explicit evolutionary models

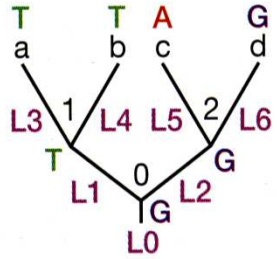
C. A rooted phylogenetic tree for the sequences A-D showing the bases for one set of aligned sequence positions in A.



D. A rooted phylogenetic tree showing one set of base assignments to nodes 0, 1 and 2.



E. A rooted phylogenetic tree showing a second set of base assignments to nodes 0, 1 and 2.



F. L(Tree) = L(Tree1) + L(Tree2) + ... + L(Tree64)