

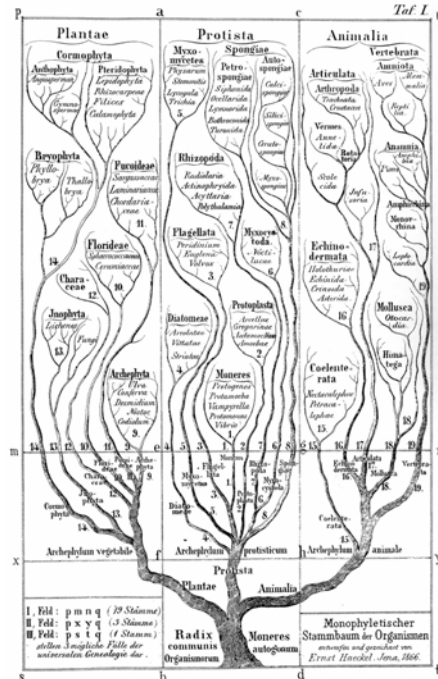
# Cours de Phylogénie et évolution moléculaire

10-14 septembre 2007

*Jan Pawlowski*

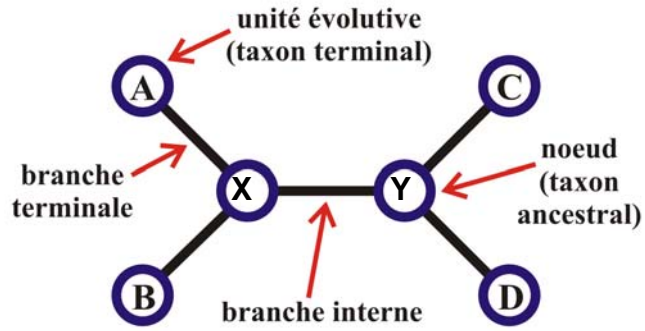
Cours de  
Phylogénie  
et évolution  
moléculaire

10-14  
septembre  
2007



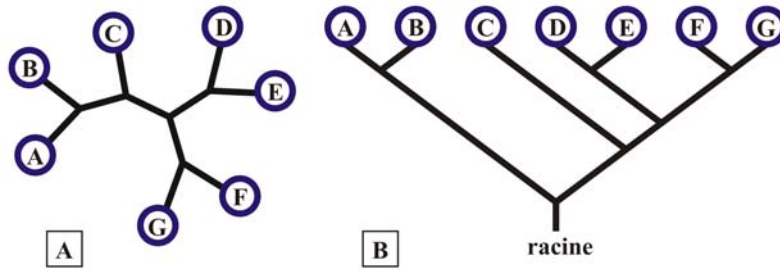
Ernst Haeckel, 1866  
Monophyletische  
Stammbaum der  
Organismen

## Un arbre phylogénétique



10-14  
septembre  
2007

## Un arbre phylogénétique



Un arbre non enraciné

Un arbre enraciné

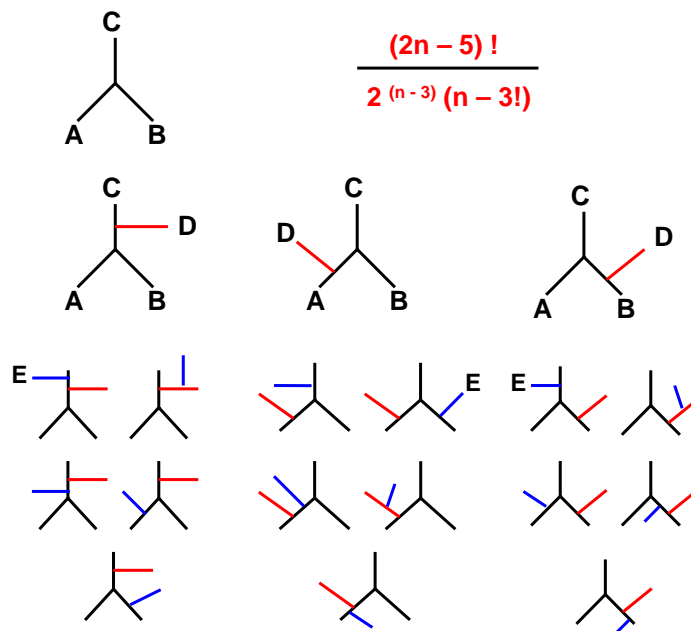
10-14  
septembre  
2007

Le nombre d'arbres augmente très rapidement avec le nombre de taxa:

| Nb de taxa | Nb d'arbres non enracinés | Nb d'arbres enracinés |
|------------|---------------------------|-----------------------|
| 3          | 1                         | 3                     |
| 4          | 3                         | 15                    |
| 5          | 15                        | 105                   |
| 6          | 105                       | 945                   |
| 7          | 945                       | 10 395                |
| 8          | 10 395                    | 135 135               |
| 9          | 135 135                   | 2 027 025             |
| 10         | 2 027 025                 | 34 459 425            |

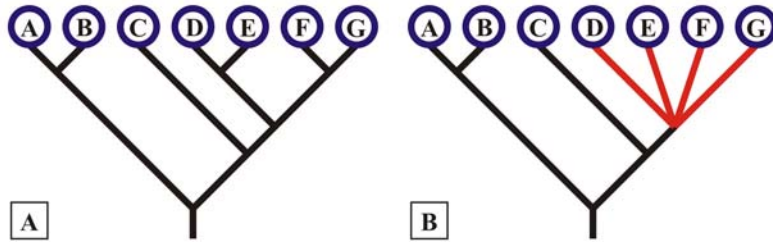
*Casse-tête d'un phylogénéticien:  
une histoire et des millions d'arbres possibles*

Cours de  
Phylogénie  
et évolution  
moléculaire



10-14  
septembre  
2007

## Un arbre phylogénétique



Un arbre totalement résolu

Un arbre partiellement irrésolu

Lorsque chaque nœud d'un arbre est le point de jonction de trois branches, on dit que l'arbre est totalement résolu, c'est-à-dire que les relations évolutives entre tous les taxa sont connues. Un arbre dans lequel plus de trois branches partent d'un ou plusieurs nœud(s) est irrésolu, ce qui indique que certaines de relations entre les taxa restent inconnues.

10-14  
septembre  
2007

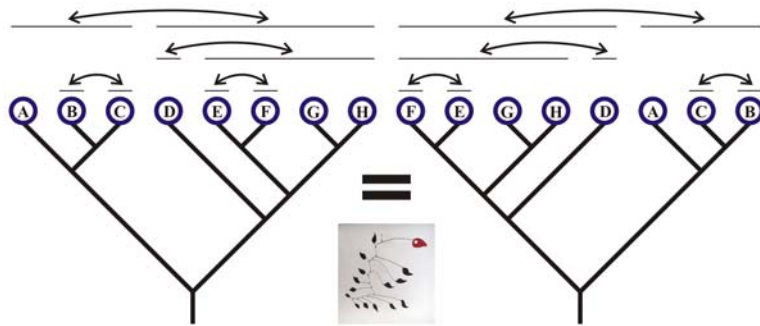
## Un arbre phylogénétique



Faire pivoter les branches d'un arbre les unes par rapport aux autres ne modifie pas l'information phylogénétique qu'il contient.

10-14  
septembre  
2007

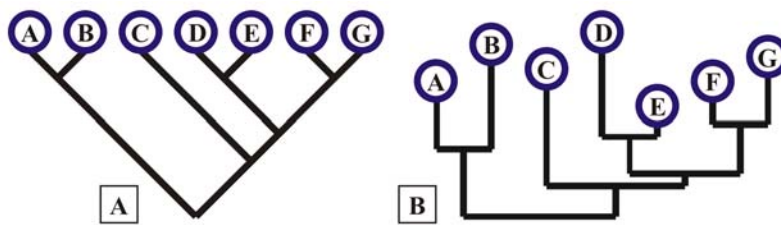
## Un arbre = un mobile



Faire pivoter les branches d'un arbre les unes par rapport aux autres ne modifie pas l'information phylogénétique qu'il contient.

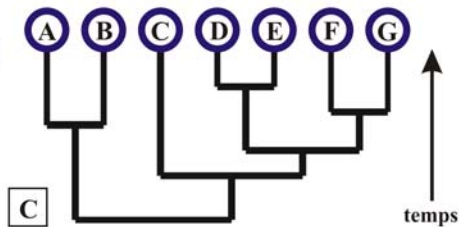
10-14  
septembre  
2007

## Un arbre phylogénétique



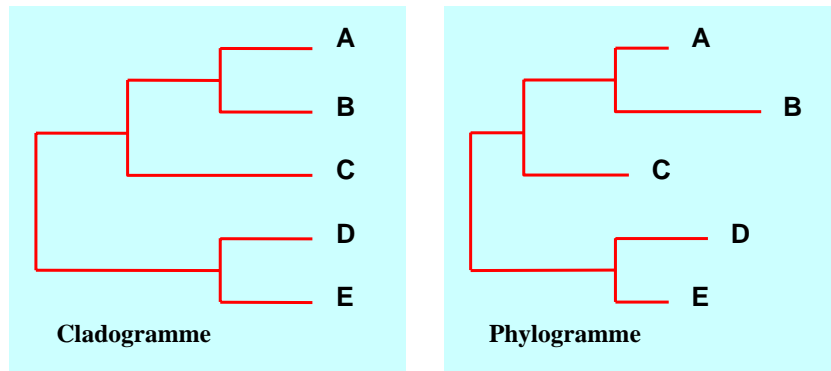
A  
Le cladogramme

B  
Le phylogramme



C  
L'arbre ultramétrique

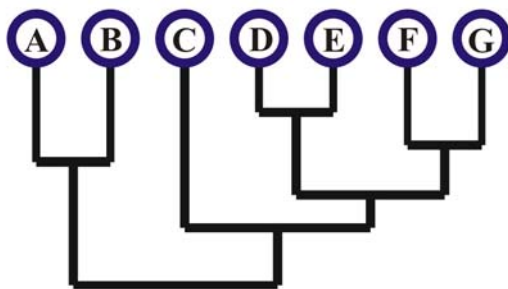
10-14  
septembre  
2007



**Le cladogramme** est un dendrogramme exprimant les relations phylogénétiques entre plusieurs taxa et construit à partir d'une analyse cladistique.

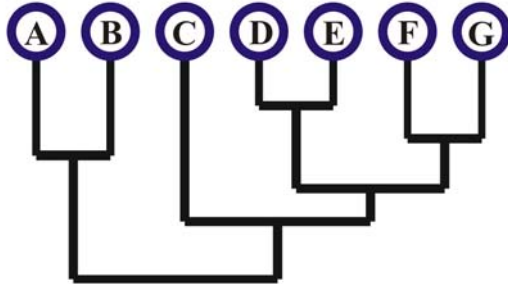
**Le phylogramme** est un dendrogramme exprimant les relations phylogénétiques entre plusieurs taxa et dont la longueur des branches est proportionnelle aux distances séparant les séquences, exprimées en nombre de substitutions par site.

## Un arbre phylogénétique



La formule écrite de cet arbre:  
((A,B),(C,((D,E),(F,G))))

## Un arbre phylogénétique



La formule écrite de cet arbre (format Newick):

**((A,B),(C,((D,E),(F,G))))**

**((A:0.1,B:0.1):0.5,(C:0.7,((D:0.1,E:0.1):0.4,(F:0.2,G:0.2))))**

10-14  
septembre  
2007

## Méthodes d'analyse phylogénétique

### Méthodes de distances

Séquences  
↓  
matrice de distances  
↓  
arbre phylogénétique

### Méthodes de caractères

Séquences  
↓  
arbre phylogénétique

## Méthodes d'analyse phylogénétique et algorithmes de choix de l'arbre

|                         | distance          | parcimonie                          | vraisemblance |
|-------------------------|-------------------|-------------------------------------|---------------|
| Approche agglomérative  | NJ<br>UPGMA       | "random sequence addition"          |               |
| Approche d'optimisation | Minimum Evolution | recherche exhaustive ou heuristique |               |

## Méthodes de distances

Ces méthodes permettent la construction d'un arbre à partir d'une matrice des distances séparant chaque paire de séquences.

### Etape 1: Construction de la matrice de distances

Les distances observées peuvent être corrigées en fonction d'un modèle d'évolution, notamment pour tenir compte des substitutions multiples à un même site et de vitesses d'évolution différentes entre sites.

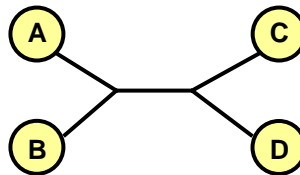
### Etape 2: Construction d'un arbre à partir de la matrice

Dans cette étape on reconstruit l'arbre dont la topologie et les longueurs de branches correspondent le mieux aux distances calculées à l'étape 1.

## Méthodes de distances

Propriétés des distances additives:

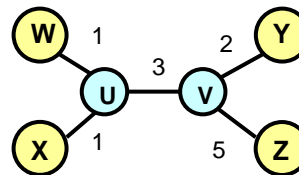
1. **Non négativité:**  $d(a,b) \geq 0$ ;  $d(a,b) = 0$  si et seulement si  $a=b$
2. **Symétrie:**  $d(a,b) = d(b,a)$
3. **Inégalité triangulaire:**  $d(a,c) \leq d(a,b) + d(b,c)$
4. **Condition des quatre points:**  $d(a,b) + d(c,d) \leq \max [d(a,c) + d(b,d), d(a,d) + d(b,c)]$



10-14  
septembre  
2007

Comment calculer la longueur des branches?

|   |   |   |   |
|---|---|---|---|
|   | X | Y | Z |
| W | 2 | 6 | 9 |
| X |   | 6 | 9 |
| Y |   |   | 7 |



$$WU = \frac{1}{2} (WY + WX - XY) = 1$$

$$XU = \frac{1}{2} (XY + XW - WY) = 1$$

$$YV = \frac{1}{2} (YX + YZ - ZX) = 2$$

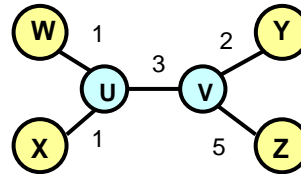
$$ZV = \frac{1}{2} (ZX + ZY - YX) = 5$$

$$UV = \frac{1}{4} [(WY + WZ + XY + XZ) - 2WX - 2YZ] = 3$$

10-14  
septembre  
2007

Comment calculer la longueur totale de l'arbre?

|   | X | Y | Z |
|---|---|---|---|
| W | 2 | 6 | 9 |
| X |   | 6 | 9 |
| Y |   |   | 7 |



La longueur totale d'un arbre est égale à la somme des distances entre les séquences:

$$S_1 = WU + XU + UV + YV + ZV$$

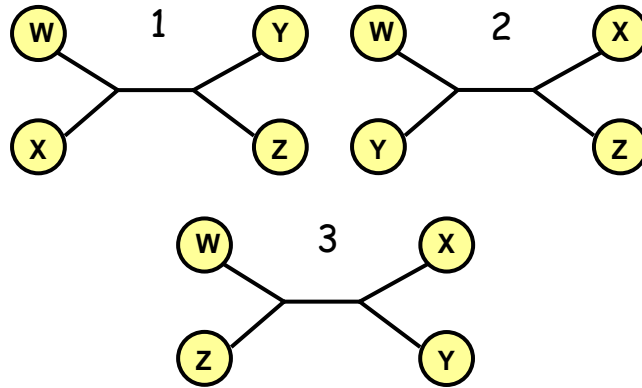
$$S_1 = \frac{1}{4} (WY + WZ + XY + XZ + 2WX + 2YZ) = 12$$

### Minimum evolution (ME)

C'est une méthode d'optimisation qui consiste à chercher un arbre le plus court parmi les arbres des distances.

1. Calculer la longueur totale de tous les arbres possibles.
2. Choisir l'arbre dont la longueur totale est la plus courte.

### Minimum evolution (ME)



$$S_1 = \frac{1}{4} (WY + WZ + XY + XZ + 2WX + 2YZ) = \mathbf{12}$$

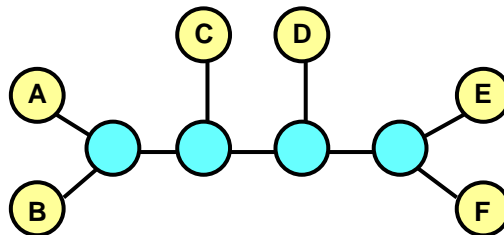
$$S_2 = \frac{1}{4} (WX + WZ + XY + YZ + 2WY + 2XZ) = 13.5$$

$$S_3 = \frac{1}{4} (WY + WX + ZX + ZY + 2WZ + 2XY) = 13.5$$

10-14  
septembre  
2007

### Neighbor-Joining (NJ)

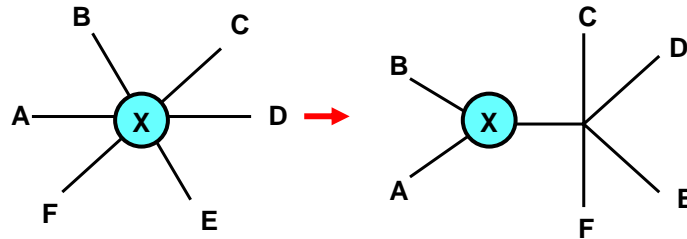
C'est une méthode agglomérative, considérée  
comme une version simplifiée de la méthode ME.



**Neighbors** are sequences connected by a  
single node in an unrooted tree (ex. 1-2, 5-6)

10-14  
septembre  
2007

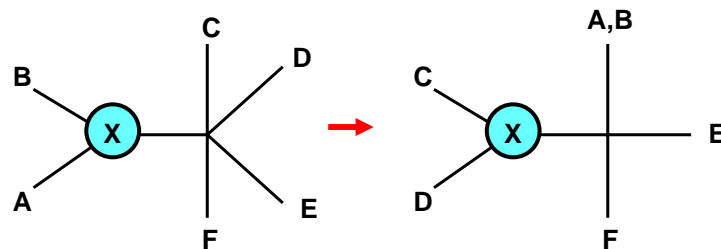
## Neighbor-Joining



1. Construct a star tree with all sequences
2. Choose two sequences and make them neighbors
3. Compute the length of the tree
4. Test other possible neighbor combinations
5. Identify the pair giving the shortest tree.

10-14  
septembre  
2007

## Neighbor-Joining



1. Choose two sequences and make them neighbors
2. Compute the length of the tree
3. Test other possible neighbor combinations
4. Identify the pair giving the shortest tree.
5. Collapse it into a composite sequence.
6. Repeat with remaining sequences until the totally resolved tree is produced

10-14  
septembre  
2007

|   |   |   |   |
|---|---|---|---|
|   | X | Y | Z |
| W | 2 | 6 | 9 |
| X |   | 6 | 9 |
| Y |   |   | 7 |

## Neighbor-Joining

1. A partir d'une matrice, calculer pour chaque séquence sa divergence nette par rapport à toutes les autres séquences:  $r(W)=17$ ,  $r(X)=17$ ,  $r(Y)=19$ ,  $r(Z)=25$
2. Créer une nouvelle matrice où la distance entre chaque paire de séquences est corrigée sur la base de leur divergence moyenne par rapport aux autres séquences.

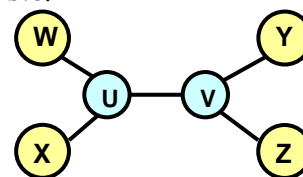
$$M_{ij} = d_{ij} - (r_i + r_j / N - 2)$$

|   |     |     |     |
|---|-----|-----|-----|
|   | X   | Y   | Z   |
| W | -15 | -12 | -12 |
| X |     | -12 | -12 |
| Y |     |     | -15 |

10-14  
septembre  
2007

## Neighbor-Joining

3. Choisir la paire de séquences dont la distance corrigée est la plus faible: WX ou YZ.
4. Définir la longueur des branches de deux séquences par rapport au nœud interne le plus proche (U), ex.  $d_{WU}$  et  $d_{XU}$ .
5. Recalculer la divergence nette entre le nœud U et chacune des séquences restantes.
6. Etablir une nouvelle matrice de distance.
7. Chercher une nouvelle paire de séquences au distance corrigée la plus faible.



10-14  
septembre  
2007

## Méthodes de distances

**Avantages** : Extrêmement rapides, permettent des analyses de grands jeux de données ; tiennent compte d'un modèle d'évolution pour corriger les estimations de distances.

**Défauts** : Traitent l'information contenue dans chaque site de manière globale ; sensibles aux différences de taux de substitution entre taxa; perdent d'information; confondent homologie et homoplasie.

## Méthode de parcimonie

Les méthodes de parcimonie reposent directement sur les principes de **la cladistique** qui ont été créées pour analyser des données morphologiques.

Le postulat de base est que **l'évolution est parcimonieuse**.

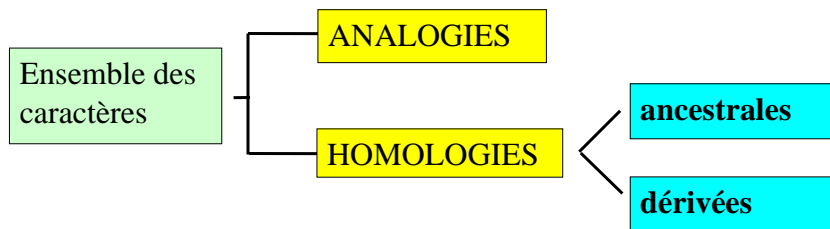
**Willi Hennig** (1913-1976)

Entomologiste allemand,  
auteur du livre  
« Phylogenetic  
systematics » (1966)

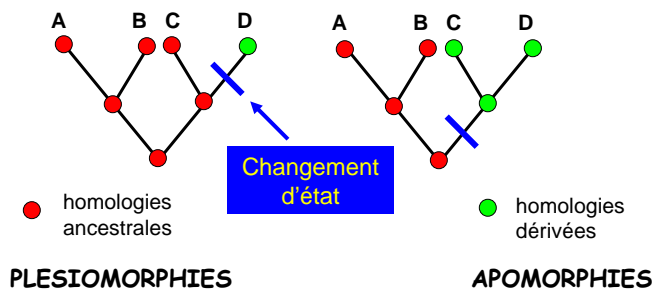


## Les principes de la cladistique I

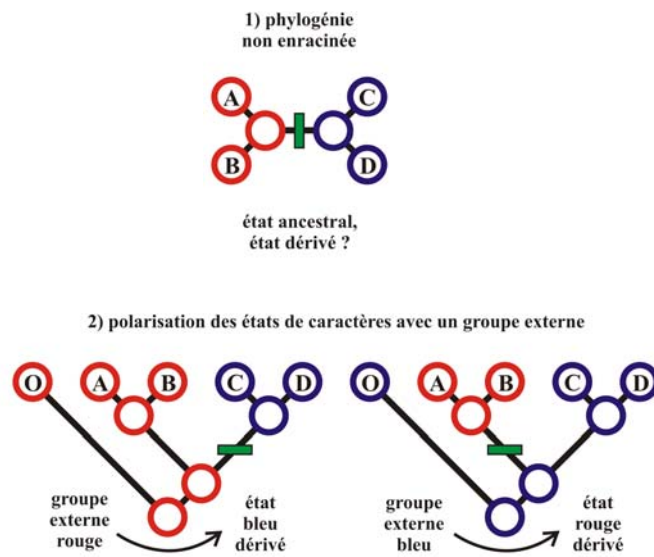
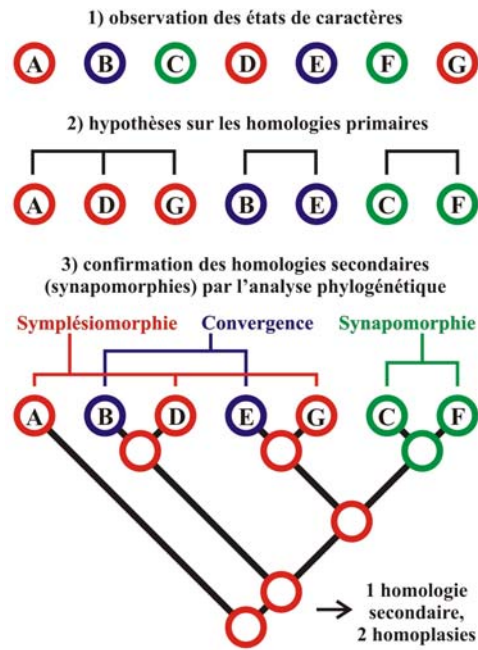
L'analyse cladistique vise à déterminer pour chaque caractère l'état ancestral et les états dérivés ainsi qu'à reconstituer les séquences de transformations évolutives de ces caractères.



## Les principes de la cladistique II

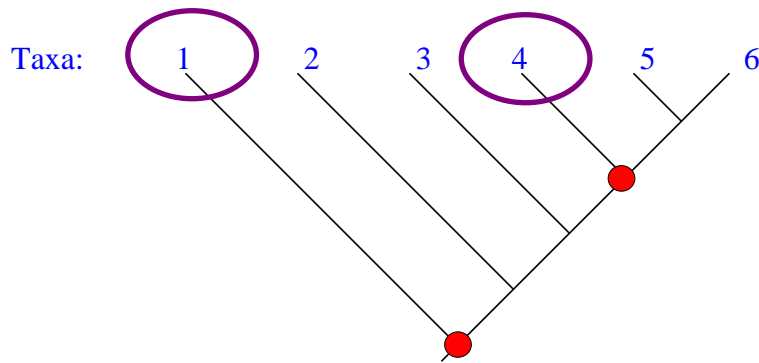


Les parentés entre les taxons étudiés sont identifiées sur la base des seuls états apomorphes partagés (**synapomorphies**).



### Les principes de la cladistique III

Groupe polyphylétique = Taxa 1 & 4



Les espèces du groupe **polyphylétique** ne partagent pas le même ancêtre commun le plus proche.

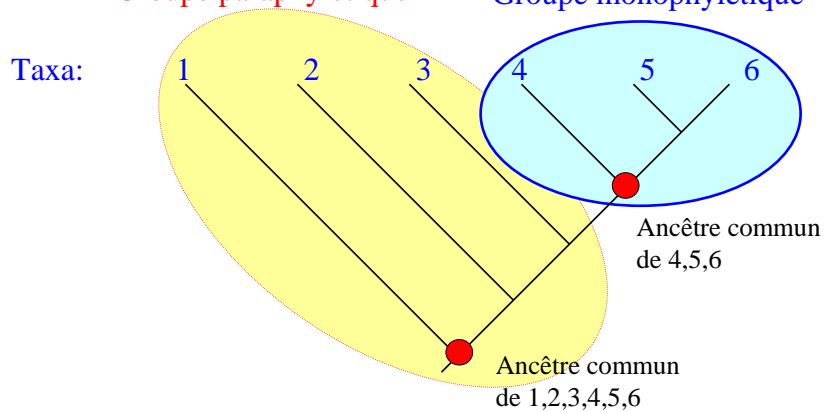
### Les principes de la cladistique III

un groupe **paraphylétique** comprend un taxon ancestral et une partie seulement de ses descendants

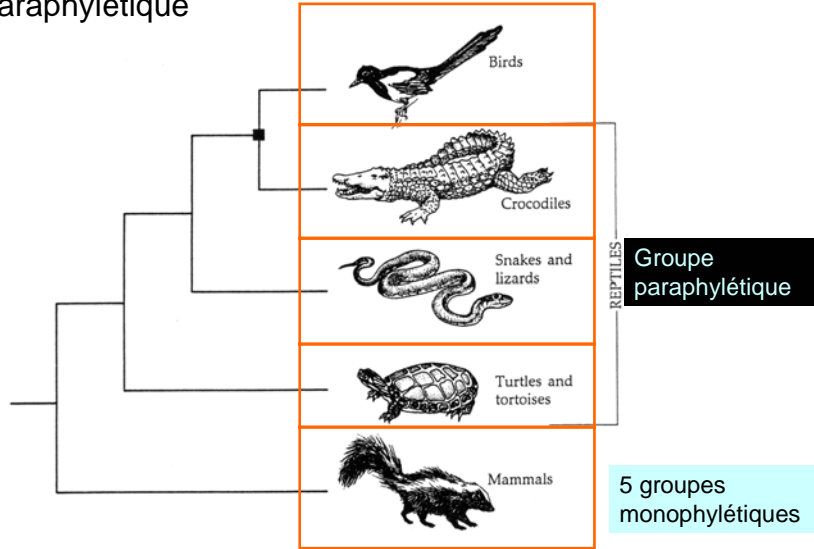
un groupe **monophylétique** comprend un taxon ancestral et tous ses descendants

Groupe paraphylétique

Groupe monophylétique

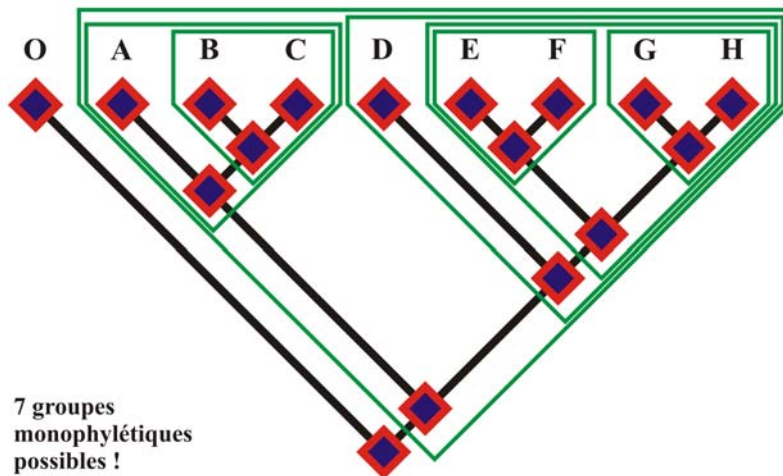


Les reptiles – exemple classique d'un groupe paraphylétique



Combien de groupes monophylétiques comprend l'arbre suivant:

**((A, (B,C)), (D, ((E, F), (G, H))))**



7 groupes monophylétiques possibles !

## Méthode de parcimonie

Les méthodes phylogénétiques basées sur le principe de parcimonie cherchent à reconstruire l'arbre qui minimise le nombre de changements d'état (le nombre de « pas ») nécessaires pour expliquer l'ensemble des données dont on dispose.

On part du principe que l'arbre phylogénétique le plus plausible est celui qui fait appel à la « quantité minimale d'évolution ».

La longueur de l'arbre parcimonieux  $L$  est égale à la somme du nombre de changements  $l_i$  pour chacun des  $k$  sites

$$L = \sum_{i=1}^k l_i$$

10-14  
septembre  
2007

## Méthode de parcimonie

Les méthodes phylogénétiques basées sur le principe de parcimonie cherchent à reconstruire l'arbre qui minimise le nombre de changements d'état (le nombre de « pas ») nécessaires pour expliquer l'ensemble des données dont on dispose.

On part du principe que l'arbre phylogénétique le plus plausible est celui qui fait appel à la « quantité minimale d'évolution ».

Dans la méthode de « weighted parcimonie » les caractères peuvent être pondérés en fonction de leur signification.

$$L = \sum_{i=1}^k w_i l_i$$

10-14  
septembre  
2007

|         |          |      |      |
|---------|----------|------|------|
| taxon 1 | ACGTCGTC | AAAA | GCAT |
| taxon 2 | ACGACGTC | GAAT | GCAT |
| taxon 3 | ACGTCGTC | GATT | GCGT |
| taxon 4 | AAGTCGTC | ACGT | GCAT |
| taxon 5 | ACGTCGTT | ACGT | GCAT |

site constant  
**site variable informatif**  
site variable non-informatif

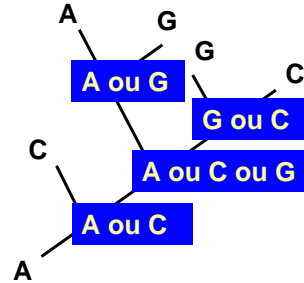
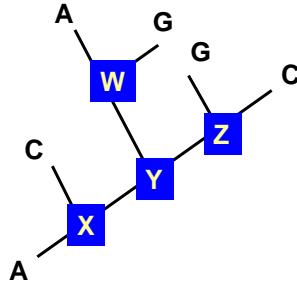
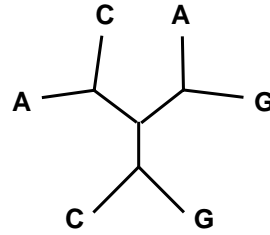
**Un site est informatif** uniquement s'il y a au moins deux types de nucléotides présents dans ce site et si chacun d'eux est représenté dans au moins deux séquences comparées.

Les sites parcimonieusement informatifs sont les seuls qui influencent le choix de l'arbre.

### Les étapes d'analyse de parcimonie (MP)

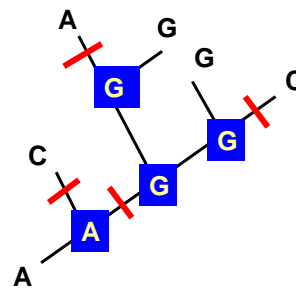
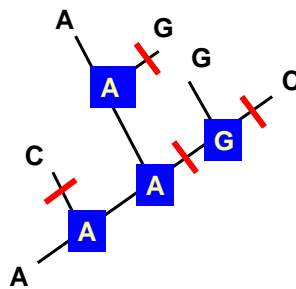
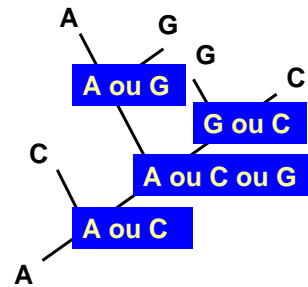
1. Identifier les sites informatifs
2. Inférer toutes les topologies d'arbres possibles pour les séquences données.
3. Calculer le nombre minimum de substitutions pour chaque site informatif.
4. Calculer la somme de changements pour chaque arbre.
5. Choisir la topologie de l'arbre qui nécessite le moins de changements.

Analyse de parcimonie:



10-14  
septembre  
2007

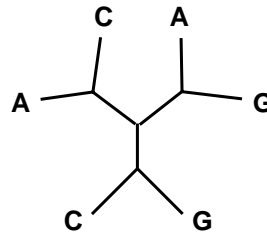
Analyse de parcimonie:



**4 changements d'état nécessaires**

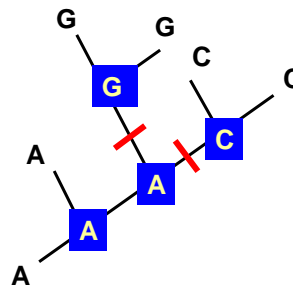
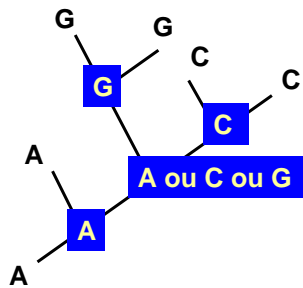
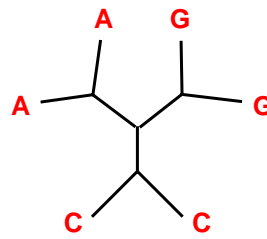
10-14  
septembre  
2007

Quel serait l'arbre le plus  
parcimonieux?



10-14  
septembre  
2007

Quel serait l'arbre le plus  
parcimonieux?



**2 changements d'état nécessaires**

10-14  
septembre  
2007

## Mesures de l'homoplasie

### *Consistency index (CI) - indice de cohérence*

Pour un site donné, l'indice de cohérence est égal au nombre minimum **M** de substitutions (c'est-à-dire le nombre total des substitutions moins un) divisé par le nombre de substitutions observés **S**.

$$CI = M / S$$

### *Homoplasy index (HI) - indice des homoplasies*

$$HI = 1 - CI$$

## Mesures de l'homoplasie

### *Retention index (RI) - indice de rétention*

$$RI = (G - S) / (G - M)$$

où **G** est égal à la somme de changements pour tous les caractères dans un arbre complètement polytomique.

### *Rescaled consistency index (RC)*

$$RC = CI \times RI$$

RC est défini comme le produit de l'indice de cohérence **CI** et de l'indice de rétention **RI**.

## Méthode de maximum de parcimonie (MP)

**Avantages** : Méthode de caractères relativement rapide ; permet de polariser les caractères.

**Défauts** : Ne tient compte que des sites informatifs, ne permet pas l'utilisation d'un modèle d'évolution, est très sensible aux différences de taux de substitution entre taxa.

## Méthodes d'optimisation

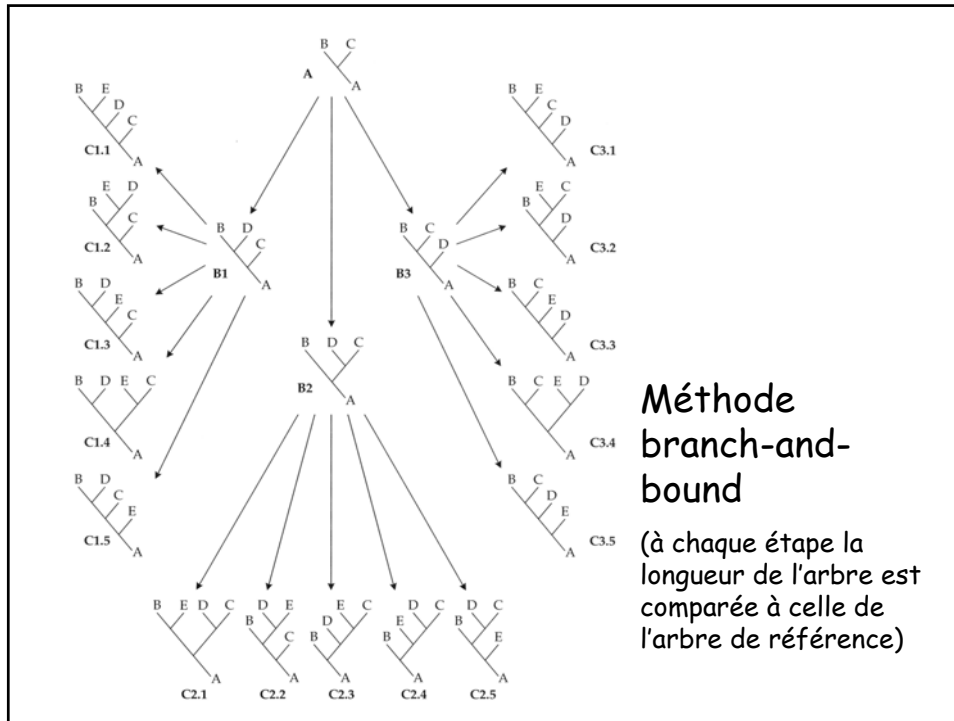
Ces méthodes explorent toutes les différentes topologies d'arbres possibles et choisissent la meilleure en fonction d'un critère donné

### 🌐 Recherche exhaustive

Pour un petit nombre de taxa (1 à 10), il est possible de réellement comparer toutes les topologies possibles et on est certain d'aboutir au meilleur arbre possible.

### 🌐 Recherche heuristique

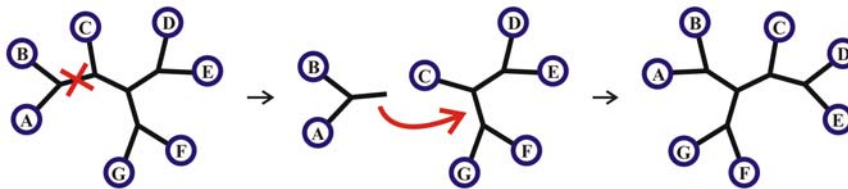
Pour plus de 10 taxa (la grande majorité des cas !), le nombre de topologies à tester est bien trop grand, et on doit recourir à un stratagème pour maximiser les chances de trouver le meilleur arbre sans pour autant devoir tester toutes les topologies possibles.



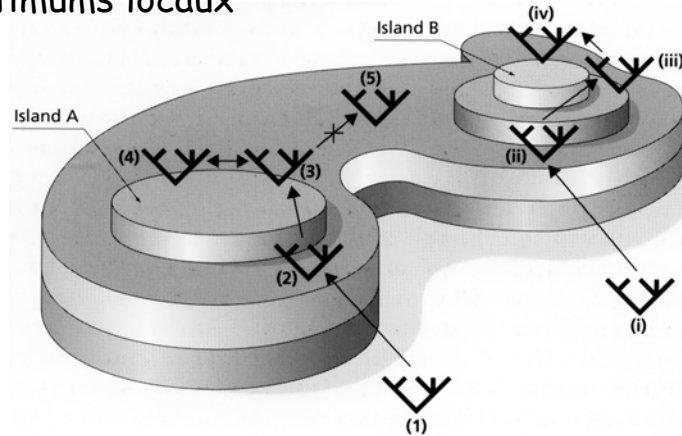
## Méthodes heuristiques

Dans une recherche heuristique, on utilise une topologie de départ (obtenue par « stepwise addition algorithm ») à laquelle on impose des réarrangements au hasard. Chacune des branches internes de l'arbre de départ est cassée et replacée à tous les endroits possibles de l'arbre (« **branch swapping** »). A chaque fois que le réarrangement mène à un arbre de meilleur score que l'arbre de départ, celui-ci est éliminé et l'arbre réarrangé devient le nouvel arbre de départ.

Tree bisection and reconnection (TBR)



## Méthodes heuristiques - le problème des optimums locaux



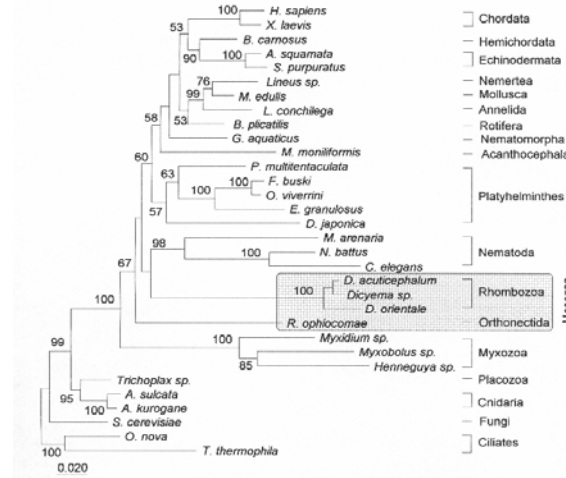
On choisit un arbre au hasard et on évalue son score, puis on crée un changement dans l'arbre et on évalue le nouveau score ; si ce nouveau score est meilleur que celui de l'arbre de départ, le nouvel arbre devient l'arbre de départ de l'étape suivante.

## Les critères de choix de la méthode phylogénétique:

1. Efficacité
2. Puissance
3. Consistance
4. Robustesse
5. Falsifiabilité

## Bootstrap

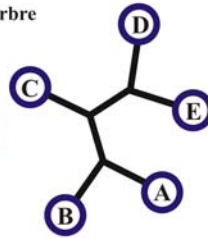
C'est la méthode la plus souvent utilisée pour tester la fiabilité des branches internes.



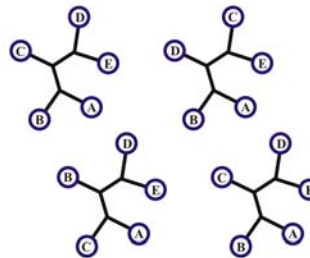
10-14  
septembre  
2007

1) construction du meilleur arbre à partir de l'alignement réel

|         |   |   |   |   |   |   |   |   |   |
|---------|---|---|---|---|---|---|---|---|---|
|         | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
| taxon A | C | G | G | T | G | A | C | T | A |
| taxon B | C | G | G | T | C | A | C | G | A |
| taxon C | C | T | G | T | A | G | C | T | G |
| taxon D | A | T | G | A | T | A | C | G | G |
| taxon E | A | T | G | A | C | G | C | G | G |



3) construction d'un arbre à partir de chaque alignement artificiel



2) génération de N alignements artificiels par tirage aléatoire avec remise à partir de l'alignement réel

|         |   |   |   |   |   |   |   |   |   |
|---------|---|---|---|---|---|---|---|---|---|
|         | 7 | 3 | 2 | 8 | 5 | 3 | 1 | 2 | 4 |
| taxon A | C | G | G | T | G | G | C | G | T |
| taxon B | C | G | G | G | C | G | C | G | T |
| taxon C | C | G | T | A | G | C | T | T |   |
| taxon D | C | G | T | G | T | G | A | T | A |
| taxon E | C | G | T | G | C | G | A | T | A |

|         |   |   |   |   |   |   |   |   |   |
|---------|---|---|---|---|---|---|---|---|---|
|         | 6 | 7 | 5 | 9 | 2 | 7 | 4 | 8 | 6 |
| taxon A | A | C | G | A | G | C | T | A |   |
| taxon B | A | C | C | A | G | C | T | G | A |
| taxon C | G | C | A | G | T | C | T | T | G |
| taxon D | A | C | T | G | T | C | A | G | A |
| taxon E | G | C | C | G | T | C | A | G | G |

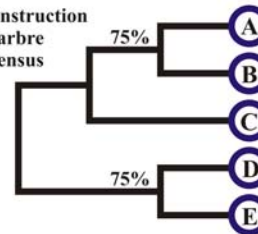
  

|         |   |   |   |   |   |   |   |   |   |
|---------|---|---|---|---|---|---|---|---|---|
|         | 1 | 8 | 7 | 5 | 9 | 4 | 1 | 8 | 3 |
| taxon A | C | T | C | G | A | T | C | T | G |
| taxon B | C | G | C | C | A | T | C | G | G |
| taxon C | C | T | C | A | G | T | C | T | G |
| taxon D | A | G | C | T | G | A | A | G | G |
| taxon E | A | G | C | C | G | A | A | G | G |

|         |   |   |   |   |   |   |   |   |   |
|---------|---|---|---|---|---|---|---|---|---|
|         | 4 | 9 | 6 | 1 | 6 | 3 | 5 | 4 | 2 |
| taxon A | T | A | A | C | A | G | G | T | G |
| taxon B | T | A | A | C | A | G | C | T | G |
| taxon C | T | G | G | C | G | G | A | T | T |
| taxon D | A | G | A | A | G | T | A | T |   |
| taxon E | A | G | G | A | G | C | A | T |   |

4) construction de l'arbre consensus



## Bootstrap

L'interprétation des valeurs de bootstrap est délicate.

- il est discutable quand les branches sont significativement soutenues;
- le bootstrap peut être « positively misleading » lorsque l'analyse converge vers une mauvaise topologie
- une bonne topologie peut n'est pas être soutenue si le signal phylogénétique est faible.

10-14  
septembre  
2007

## Molecular clock hypothesis

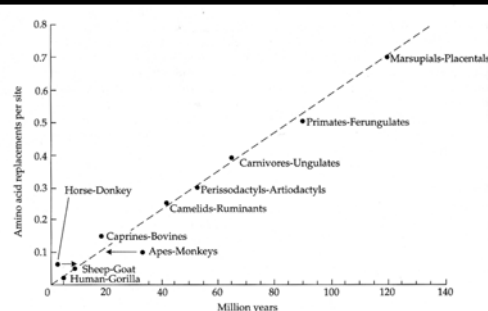
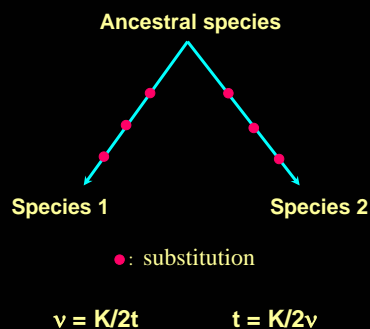
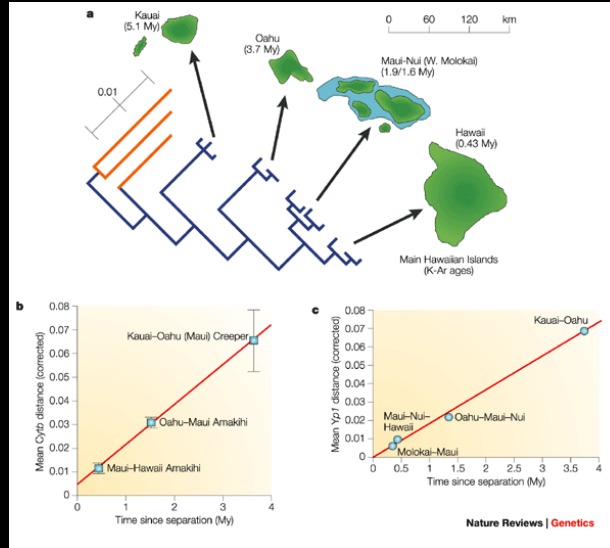


FIGURE 4.15 Number of amino acid replacements per amino acid site in a combined sequence consisting of hemoglobins  $\alpha$  and  $\beta$ , cytochrome  $c$ , and fibrinopeptide A among various mammalian groups plotted against geological estimates of divergence times. The dashed line represents the molecular clock expectation of equal rates of amino acid replacement in all evolutionary lineages. There are two large deviations of the observed values from the expected line. These deviations indicate a slowdown in evolution following the divergence between apes and monkeys, and an acceleration following the divergence between horse and donkey. However, these inferences are based on specific paleontological estimates of divergence times (33 million years for the ape-monkey split and 2 million years for the horse-donkey split), and if these time estimates are inaccurate (arrows), the deviation of these lineages from a strict molecular clock may not be significant. Modified from Langley and Fitch (1974).

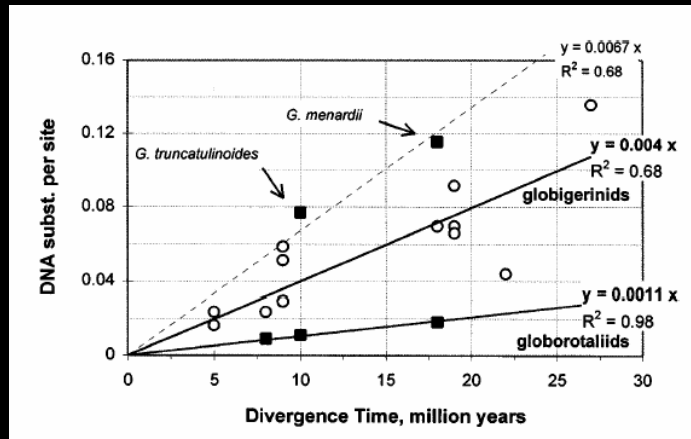
According to molecular clock hypothesis, for a given molecule, the rates of substitution are constant through the time (Zuckerlandl and Pauling, 1962)

## Local molecular clocks

Molecular clock and phylogeny of birds (*Hemignathus*) and insects (*Drosophila*) from Hawaiian Islands (Bromham and Penny, 2003)



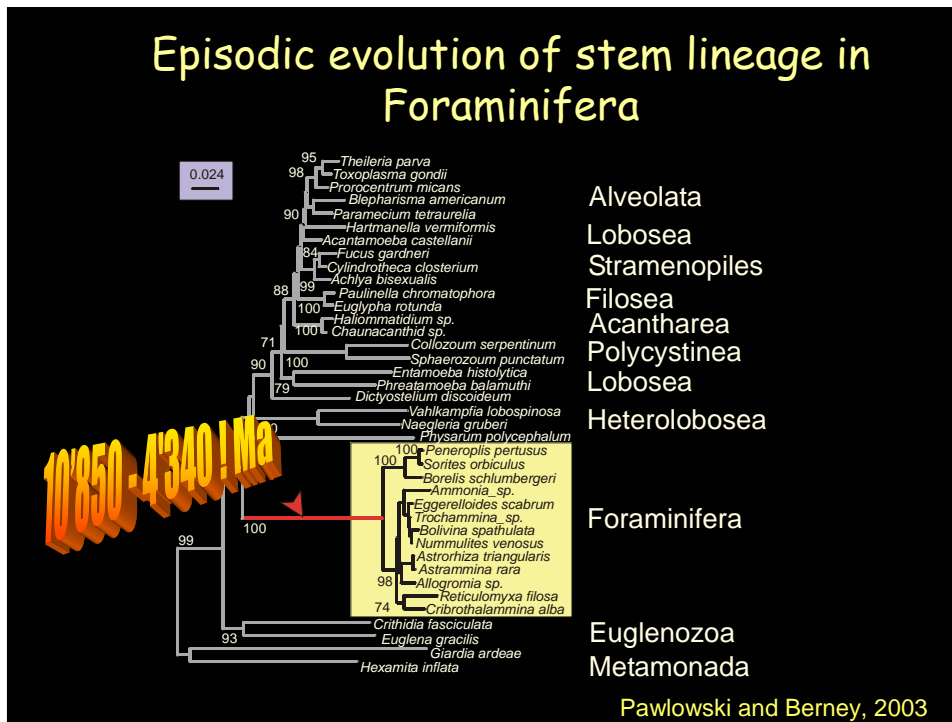
## Erratic clock in globorotaliid foraminifera



### Molecular versus Taxonomic Rates of Evolution in Planktonic Foraminifera

Colomban de Vargas<sup>1</sup> and Jan Pawlowski\*

## Episodic evolution of stem lineage in Foraminifera



**Relative rate test** : one of the most common ways of testing molecular clock hypothesis

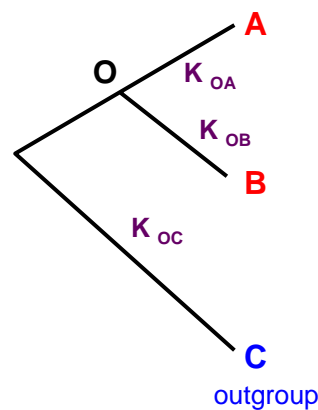
Test whether A and B have the same rate of substitution since their divergence (O)

$$d(AO) = d(BO)$$

Null hypothesis:

$$d(AC) = d(BC)$$

$$d(AC) - d(BC) = 0$$



## RRTree: Relative-Rate Tests between groups of sequences on a phylogenetic tree

Marc Robinson-Rechavi and Dorothee Huchon, *Bioinformatics*, 16:296-297 (2000)

RRTree is a user-friendly program for comparing substitution rates between lineages of protein or DNA sequences. Genetic diversity is taken into account through use of several sequences, and phylogenetic relations are integrated by topological weighting.

<http://pbil.univ-lyon1.fr/software/rrtree.html>

**Example:**  
Substitution rates  
in SSU rDNA of  
Amoebozoa

