Phylogénie moléculaire

O. Lecompte
Laboratoire de Bioinformatique et Génomique Intégratives – IGBMC

odile.lecompte@igbmc.fr

La phylogénie moléculaire

- Pour quoi faire ?
  - retracer l'histoire évolutive d'une famille de gènes
  - reconstruire les relations évolutives entre espèces
    ex : arbre du vivant
  - classer une nouvelle espèce
    ex : souche virale

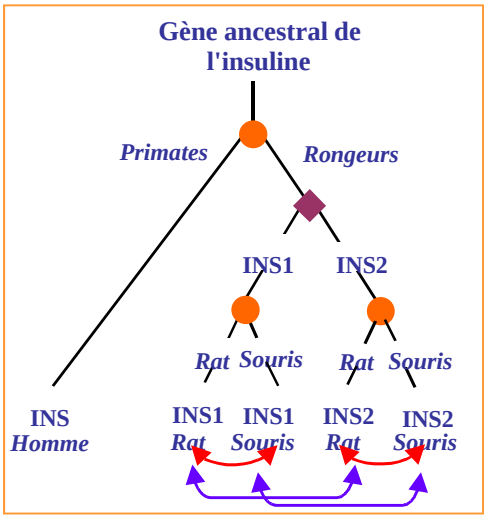
Homologie, orthologie, paralogie

Homologie :
2 gènes sont homologues s'ils ont un ancêtre commun

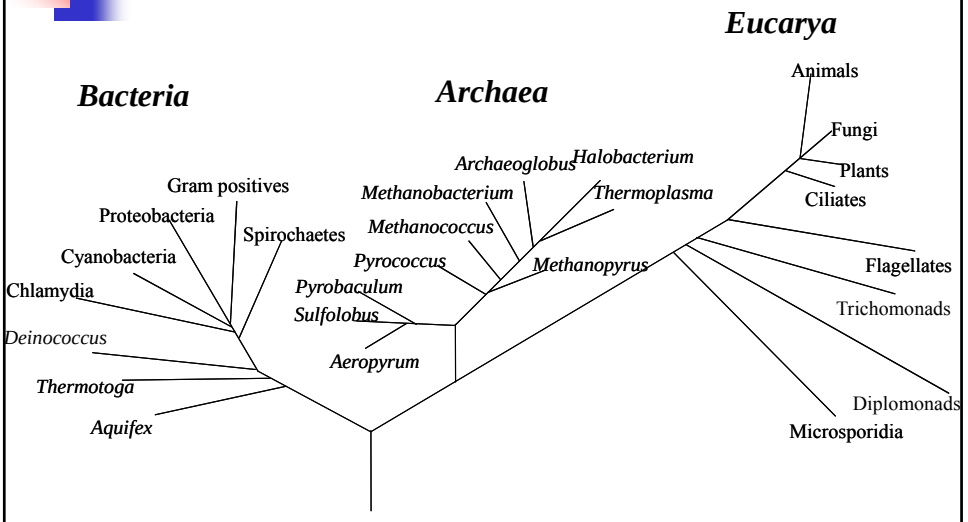
Orthologie :
2 gènes sont orthologues s'ils ont divergé à la suite d'un événement de spéciation

Paralogie :
2 gènes sont paralogues s'ils ont divergé à la suite d'un événement de duplication

Spéciation
Duplication



Arbre du vivant



La phylogénie moléculaire

- Pour quoi faire ?
  - retracer l'histoire évolutive d'une famille de gènes
  - reconstruire les relations évolutives entre espèces
    ex : arbre du vivant
  - classer une nouvelle espèce
    ex : souche virale
- Comment ?
  1. *Aligner correctement les séquences nucléiques ou protéiques*
  2. *Appliquer une méthode de génération d'arbres*
  3. *Évaluer statistiquement la robustesse des arbres*

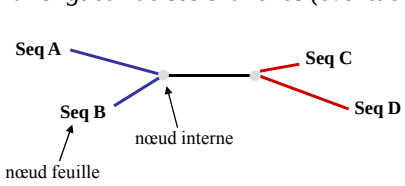
Phylogénie moléculaire

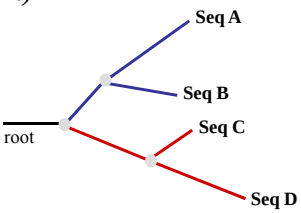
- Notions de base en phylogénie
- Principales méthodes de construction d'arbres
  - Méthodes basées sur les distances
  - Méthodes basées sur les séquences
- Evaluer la fiabilité d'un arbre
- Les limites de la phylogénie
- Quelques programmes

Notions de base : terminologie

Un arbre phylogénétique est caractérisé par :

- *sa topologie*
- *la longueur de ses branches (éventuellement)*





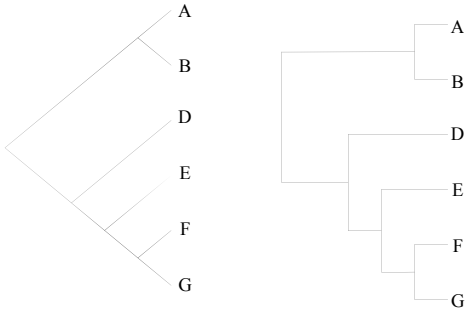
Noeud : *estimation de l'ancêtre commun des éléments appartenant à ce nœud*

Racine (root) : *ancêtre commun de tous les éléments de l'arbre.*

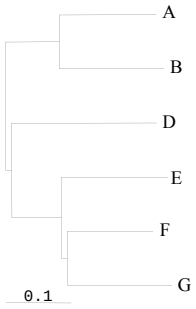
Un arbre peut avoir ou non une racine.

Cladogrammes

(longueur des branches non significative)

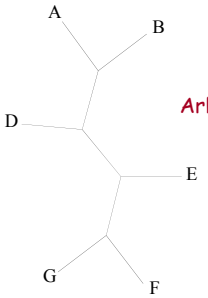


Phylogramme

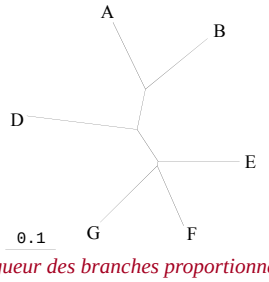


(longueur des branches proportionnelle au nombre de substitutions)

Arbres sans racine (networks)



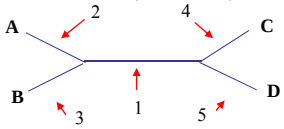
(longueur des branches non significative)



(longueur des branches proportionnelle au nombre de substitutions)

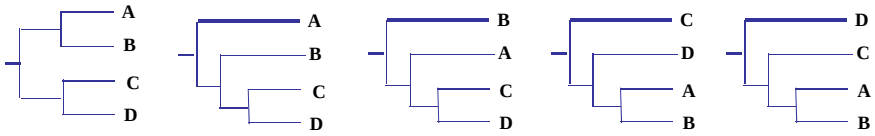
Notions de base : racine

Pour un arbre sans racine (unrooted), il existe plusieurs arbres avec racine



→ ***Position de la racine***

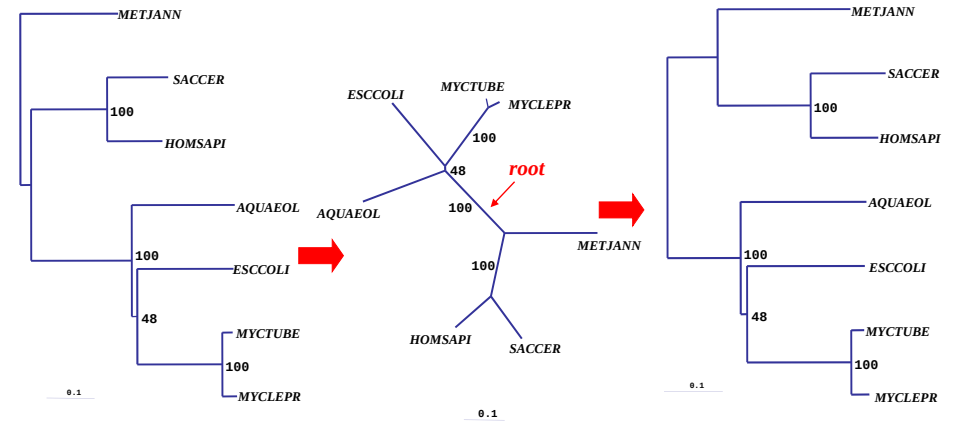
La racine permet d'orienter l'arbre.



Dans la plupart des programmes, la position de la racine est choisie de manière arbitraire :
- « midpoint rooting » (racine placée au milieu de la plus longue branche)
- « outgroup rooting »

L'utilisateur peut définir la ou les séquences constituant l 'outgroup pour enraciner l'arbre.
Ces séquences doivent être éloignées des autres séquences tout en étant homologues.

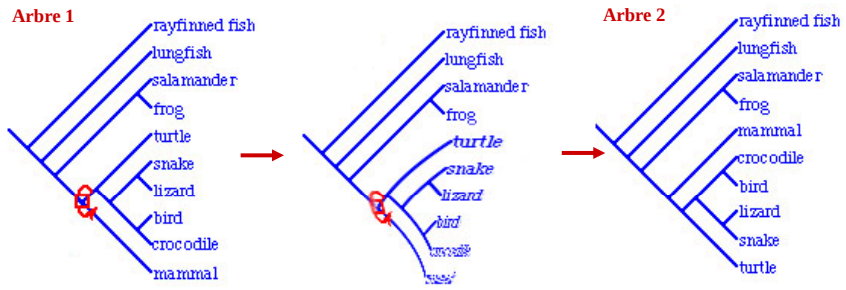
Notions de base : racine



Rooted tree Unrooted tree Rooted tree

Notions de base : ordre des branches

L'ordre des branches appartenant à un même nœud n'a aucune importance.
La rotation autour d 'un nœud ne change rien à la topologie de l'arbre.



Arbre 1 = Arbre 2

Phylogénie moléculaire

- Notions de base en phylogénie
- Principales méthodes de construction d'arbres
  - Méthodes basées sur les distances
  - Méthodes basées sur les séquences
- Evaluer la fiabilité d'un arbre
- Les limites de la phylogénie
- Quelques programmes

Méthodes de construction d'arbres

- Méthodes basées sur les distances
  - Calcul des distances entre paires de séquences
    => matrice de distances
  - Regroupement des séquences
    - UPGMA
    - Neighbor-Joining

- facile, rapide
- les séquences ne sont pas considérées en tant que telles

Calcul des distances

Distance observée : nombre moyen de substitutions par site

$$\text{Dist observée} = \frac{\text{Nb substitutions}}{\text{Nb sites considérés}}$$

- Sites considérés :
  - *la 3ème base du codon peut ne pas être prise en compte.*
  - *les positions comportant des gaps sont généralement éliminées*

2 possibilités :

```
Seq1 MAIKKIISRSNSGIHNATVI
Seq2 MPIKK-ISRSNSGIHHSTVI
Seq3 MPIKKIISRSNTGI-HSTVI
```

Nb total de sites = 20

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

Global gap removal : 18 sites considérés
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

Correction des distances

Si le temps de divergence entre deux séquences augmente, la probabilité d'avoir plusieurs substitutions à un même site augmente également.

=> Le nombre de substitutions observées sous-estime le nombre réel de substitutions entre des séquences éloignées.

	Séquence1	Séquence2	Substitutions observées	Substitutions réelles
Substitution unique	C	C=>A	1	1
Substitution multiples	C	C=>A=>T	1	2
Substitutions coïncidentes	C=>G	C=>A	1	2
Substitutions parallèles	C=>A	C=>A	0	2
Substitutions convergentes	C=>A	C=>T=>A	0	3
Substitutions réverses	C	C=>T=>C	0	2

=> Nombreuses méthodes pour tenter d'estimer la distance réelle entre séquences

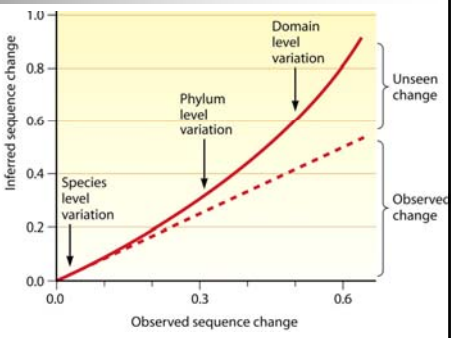
Correction des distances

Modèle de Jukes-Cantor (JC)

- Les 4 bases ont la même fréquence
- Transversions et transitions sont équiprobables

$d = -3/4 \ln (1-4/3 D)$

avec D distance observée



Modèle de Kimura à 2 paramètres (K2P)

- Les 4 bases ont la même fréquence
- Transversions et transitions ont un taux différent

$d = -\ ½ \ln(1-2P-Q) – ¼ \ln(1-2Q)$

avec P nb moyen de transitions
Q nb moyen de transversions

Modèle de Tamura-Nei

- prise en compte de la fréquence des 4 bases
- α1 taux de transition entre purines, α2 taux de transition entre pyrimidine, β taux de transversion

UPGMA

UPGMA = Unweighted Pair Group Method with Arithmetic Mean

UPGMA est le plus simple des algorithmes de clustering

Hypothèse :

le taux de mutation est le même dans toutes les lignées (horloge moléculaire)

Méthode

- Regroupement des 2 séquences les plus proches
- Le noeud est positionné à la distance d de chacune des séquences

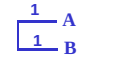
  d= (dist seq1,seq2)/2
- Calcul de la distance entre le nouveau groupe et les autres séquences :

  dist (seq1,seq2),seqx = (dist seq1,seqx + dist seq2,seqx) / 2
- etc...

UPGMA

Matrice de distances | Regroupement | Calcul d 'une nouvelle matrice de distances

	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

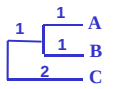



```
dist(A,B),C=(dist AC+dist BC)/2 = 4
dist(A,B),D=(dist AD+dist BD)/2 = 6
dist(A,B),E=(dist AE+dist BE)/2 = 6
dist(A,B),F=(dist AF+dist BF)/2 = 8
```

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

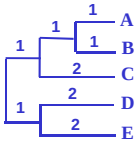

```
dist(D,E),(A,B)=(dist D(A,B)+dist E(A,B))/2 = 6
dist(D,E),C    =(dist DC    +dist EC    )/2 = 6
dist(D,E),F    =(dist DF    +dist EF    )/2 = 8
```

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



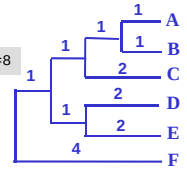

```
dist(AB,C),(D,E)=(dist (A,B)(D,E)+dist C(D,E))/2 = 6
dist(AB,C),F    =(dist (A,B)F    +dist CF    )/2 = 8
```

	(AB,C)	(D,E)
(D,E)	6	
F	8	8




```
dist(ABC,DE),F=(dist (AB,C)F+dist(D,E)F)/2=8
```

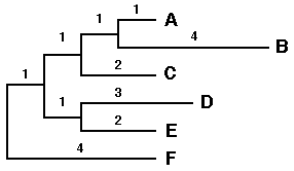
	(ABC,DE)
F	8



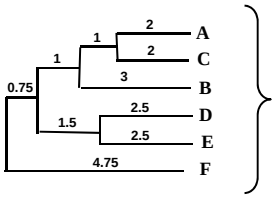
UPGMA

Problème majeur :
Si les taux de mutation diffèrent suivant les branches, la méthode UPGMA peut conduire à une topologie complètement erronée.

Ex: le taux de mutation de B est beaucoup plus important que celui de A



Arbre obtenu par la méthode UPGMA



Topologie fausse !!

Matrice initiale

	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 par 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)

- Référence:

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

- Méthode:
  - *Additivité des distances : la distance entre une paire de noeuds est égale à la somme de la longueur des branches qui les séparent*
  - *Critère d'agglomération : évolution minimale*

    Le choix des 2 unités taxonomiques à rassembler est fait de telle sorte qu'il minimise la somme des longueurs de toutes les branches de l'arbre
  - *Le calcul de la longueur des branches prend en compte la divergence moyenne de chacune des séquences avec toutes les autres*

    => autorise des taux de mutation différents suivant les branches

Neighbor-Joining (NJ)

- Estimation de la somme des longueurs des branches pour chaque regroupement possible (S_{12},S_{13},…)

=> choix du regroupement qui minimise la somme des longueurs des branches

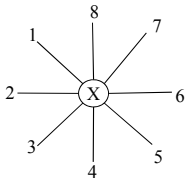
- Comment calculer la somme des longueurs des branches ?

Pour un arbre en étoile, la somme des longueurs des N branches :

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

Avec D_{ij} distance entre OTUi et OTUj

L_{iX} longueur entre le nœud i et le nœud X



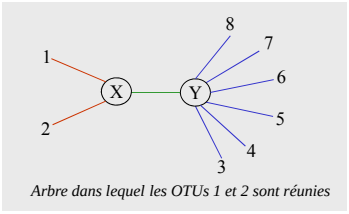
Arbre en étoile

Neighbor-Joining (NJ)

Pour un arbre dans lequel les OTUs 1 et 2 ont été réunies, la somme des longueurs des branches :

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

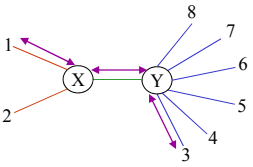
$$S_{12} = D_{12} + L_{XY} + \left(\sum_{3\le i<j}^{N} D_{ij}\right) / (N-3)$$



Arbre dans lequel les OTUs 1 et 2 sont réunies

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

Dans l'exemple :
L_{1X} et L_{2X} ont été comptées 6 fois chacune.
Les longueurs L_{3Y}, …, L_{8Y} ont été comptées 2 fois chacune.
La longueur L_{XY} a été comptée 12 fois.



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

Neighbor-Joining (NJ)

estimation de la somme des longueurs des branches pour chaque regroupement possible (S_{12},S_{13},…)

	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

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

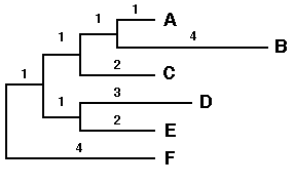
Application :

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

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

…

Regroupement AB

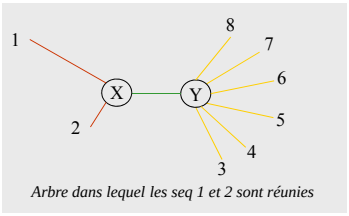


Neighbor-Joining (NJ)

La longueur des branches est estimée par la méthode de Fitch et Margoliash's (67)

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

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



Arbre dans lequel les seq 1 et 2 sont réunies

D1Z représente la distance moyenne de 1 avec les autres séquences

D2Z représente la distance moyenne de 2 avec les autres séquences

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

=> autorise des taux de mutation différents suivant les branches

Méthodes de construction d'arbres

- Méthodes basées sur les distances
  - UPGMA
  - Neighbor-Joining

  => facile, rapide

  => les séquences ne sont pas considérées en tant que telles

- Méthodes basées directement sur les séquences
  - Chacune des positions des séquences est considérée comme un caractère.
    - parcimonie (maximum parsimony)
    - maximum de vraisemblance (maximum likelihood)

  => temps de calcul très long

Maximum de parcimonie

Utilisé historiquement pour l'étude de caractères morphologiques

➔ on favorise le scénario évolutif qui demande le moins d'événements

Application aux séquences:

une position de l'alignement = un caractère

recherche des arbres les plus parcimonieux, c-a-d ceux dont la topologie demande le minimum de substitutions (arbre le plus court)

Méthodes :

1) Décompte du nb min de substitutions pour chaque topologie possible

   => permet la reconstruction de la séquence ancestrale

2) Décompte des sites informatifs : choix de la topologie favorisé par le plus de sites

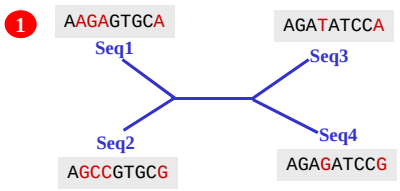
Recherche de l'arbre le plus court

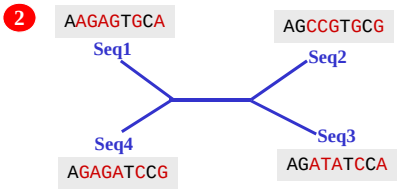
Méthodes exhaustives : nombre de substitutions pour toutes les topologies

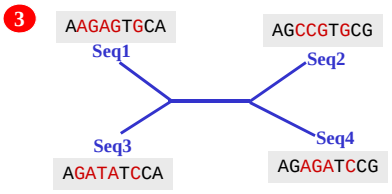
4 séquences

```
Seq1 AAGAGTGCA
Seq2 AGCCGTGCG
Seq3 AGATATCCA
Seq4 AGAGATCCG
```

nb de topologies possibles (arbres sans racine) : 3



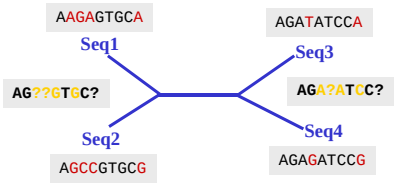




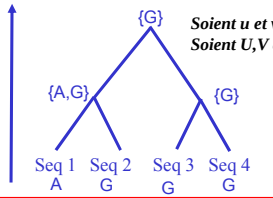
Recherche de l'arbre le plus court

Estimation du nombre min de substitutions pour une topologie donnée (Méthode de Fitch)

1) Enraciner l'arbre (n'importe où)



2) Passage des feuilles vers la racine :

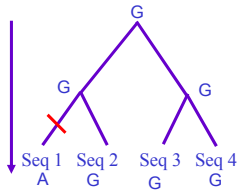


Soient u et v les nœuds fils du noeud n
Soient U,V et N les ensembles de nucléotides associés à ces noeuds

N= U∩V si U ∩V ≠ø
N= U∪V sinon

3) Passage de la racine vers les feuilles:

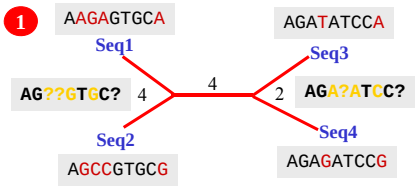
- choix d'un nucléotide x dans l'ensemble N de la racine n
- pour le nœud fils u, choix d'un nucléotide :
  x si x ∈ U
  un nucléotide quelconque de l'ensemble U sinon



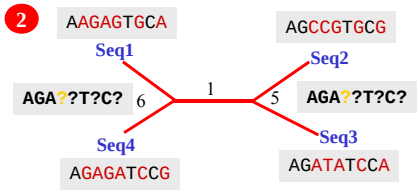
1 substitution

Recherche de l'arbre le plus court

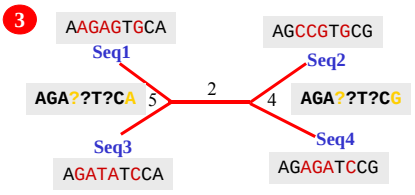
Comparaison de la « longueur » de l'arbre pour les différentes topologies



10 substitutions



12 substitutions



11 substitutions

Recherche de l'arbre le plus court

Problème : nb de topologies possibles !!!

n = nb de séquences

Nb arbres non enracinés = $\frac{(2n-5)!}{2^{n-3}(n-3)!}$

Nb de séquences	Nb d'arbres non rootés possibles
3	1
4	3
5	15
6	105
7	945
8	10 395
9	135 135
10	2 027 025
50	$>3.10^{74}$

Différentes approches

- Exhaustives : méthodes exactes, solution optimale assurée mais très lent
- Branch-and-bound : recherche non exhaustive mais solution optimale assurée
- Heuristiques : solution optimale non garantie

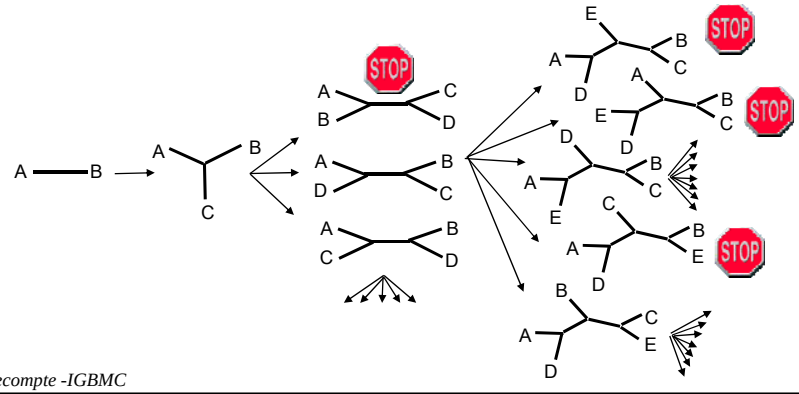
Recherche de l'arbre le plus court

Branch-and-bound Method (Hendy & Penny, 1982)

algorithme exact qui garantit la solution optimale sans nécessiter de recherche exhaustive

1) ajout des séquences une à une (suivant l'ordre de l'alignement)
calcul du nb de substitutions pour une topologie

2) exploration des différentes topologies
si nb de substitutions au cours de l'ajout > nb trouvé pour la meilleure topologie ⇨ stop



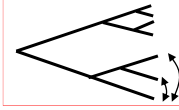
Recherche de l'arbre le plus court

Méthodes heuristiques

permettent de traiter des séquences plus longues et plus nombreuses mais ne garantissent pas la solution optimale

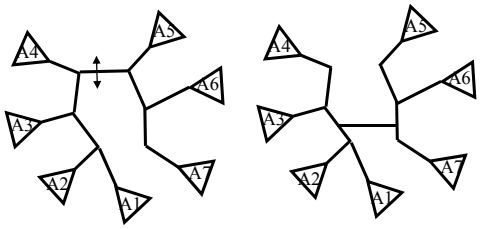
1) construction d'un arbre initial par addition progressive des espèces par exemple

2) l'arbre initial est ensuite réarrangé pour diminuer sa longueur *(branch swapping)*

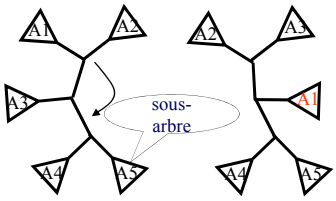


réarrangement local
(nearest neighbor interchanges)

réarrangement par bissection et reconnexion
(tree bisection and reconnection)



réarrangement global
(subtree pruning and regrafting)



l'ordre initial des séquences modifie le résultat

☛ ***faire plusieurs essais en modifiant l'ordre des séquences : option « jumble »***

Maximum de Parcimonie

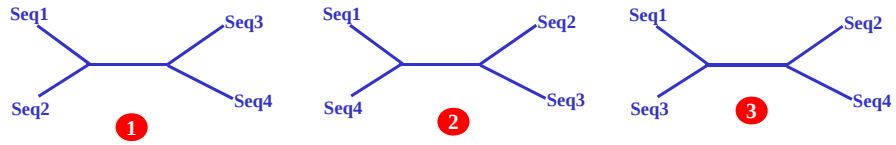
Variante :

- recherche des topologies possibles
- décompte des sites informatifs uniquement

Site informatif : au moins 2 résidus différents présents chacun au moins 2 fois

=> sites qui favorisent une topologie

```
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
```



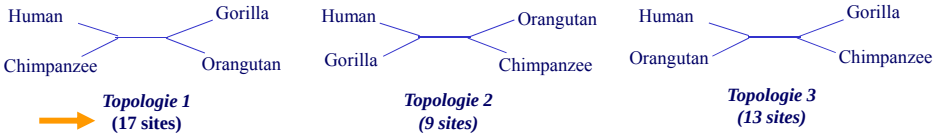
Méthode de parcimonie

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 21 2716 21772	2 1 1 38133	1 1 65717	1 1 63264	67719	63243	11124	11112	11311	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		

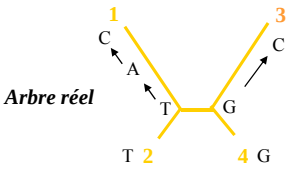
[a] The 46 observed patterns are arranged columnwise in the order of occurrence in the data, for example, the first two patterns, AAAA and GGGG, are observed at 222 and 71 sites, respectively. The three tree topologies supported by the ''informative'' patterns are T_1 = ((H,C),G,O), T_2 = ((H,G),C,O), 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)

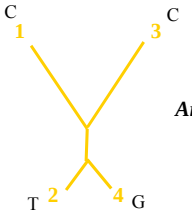


Maximum de parcimonie

Problème d'attraction des longues branches



Arbre réel



Arbre parcimonieux

Si des séquences évoluent à des vitesses très différentes, la probabilité d'avoir des substitutions convergentes est significative dans les longues branches

=> le regroupement apparemment le plus parcimonieux conduira à une fausse topologie

- pas de correction pour les substitutions multiples
- peut aboutir à plusieurs arbres ex-æquo
- méthode relativement lente, inutilisable pour un grand nombre de séquences
- pas d'information sur la longueur des branches (en général)

Maximum de vraisemblance (Maximum likelihood)

=> probabilité d'observer un arbre donné

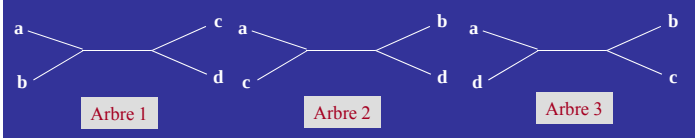
- Comme pour la méthode de parcimonie :
  - chaque colonne est considérée comme un caractère
  - Tous les arbres possibles sont estimés
  - plus l'arbre demande de mutations, plus sa probabilité est faible
    => les arbres demandant peu de mutations seront privilégiés
- Différences avec méthode de parcimonie:
  - Utilisation d'un modèle explicite d'évolution
  - Autorise des taux de substitutions variables suivant les branches
- La méthode peut être utilisée pour estimer un arbre

Maximum de vraisemblance

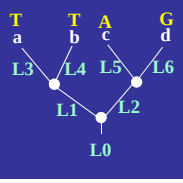
Alignement de 4 séquences

	1 2 3 4
Seq a	TTGC...
Seq b	TTGC...
Seq c	ATAC...
Seq d	GTAC...

3 arbres sans racines possibles



Arbre 1

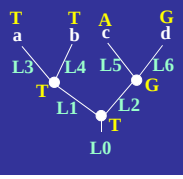


<u>Possibilités d'attribution des bases aux noeuds pour la position 1 :</u>
- 3 noeuds internes
- 4 bases possibles à chaque noeud

Nb de possibilités: 4 x 4 x 4 = 64

Maximum de vraisemblance

Exemple de possibilité :



<u>Estimation de la vraisemblance de l'arbre 1 pour la colonne 1 :</u>
somme des probabilités de chacune des 64 possibilités

<u>Estimation de la vraisemblance de l'arbre 1 :</u>
somme des probabilités obtenues pour chacune des colonnes

$$\ln L = \sum_{i=1}^{m} L_i$$

Le calcul est réalisé pour chaque arbre possible (3 ici).
L'arbre avec le maximum de vraisemblance est retenu.

<u>Exemple de modèle évolutif :</u>

L0 = fréquence de T ~ 0.25
L2 = probabilité de transversion de T=>G
L5 = probabilité de transition de G=>A
L1, L3, L4, L6 = ~1

Probabilité d'avoir cette possibilité pour la position 1 :
L= L0 x L1 x L2 x L3 x L4 x L5 x L6

Phylogénie moléculaire

- Notions de base en phylogénie
- Principales méthodes de construction d'arbres
  - Méthodes basées sur les distances
  - Méthodes basées sur les séquences
- Evaluer la fiabilité d'un arbre
- Les limites de la phylogénie
- Quelques programmes

Evaluer la fiabilité d'un arbre

- But :

  Estimer par une méthode statistique la fiabilité de la topologie de l'arbre
- Exemple : Méthode du bootstrap

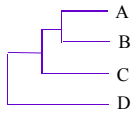
  On construit n pseudo-alignements par échantillonnage aléatoire des colonnes de l'alignement initial
  - chaque colonne de l'alignement initial peut être utilisée 0, 1, ou plusieurs fois
  - les pseudo-alignements ont la même longueur que l'alignement initial
  - le nombre de pseudo-alignements doit être suffisamment élevé pour que le test soit significatif (n ≥ nombre de colonnes)
  - Pour chaque pseudo-alignement, on construit un arbre.
  - Pour chaque branche de l'arbre initial, on indique le nombre de fois où cette branche a été retrouvée dans les n arbres.

Evaluer la fiabilité d'un arbre

Alignement d'origine

```
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
```

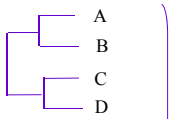
	A	B	C
B	2		
C	3	3	
D	6	4	4



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
```

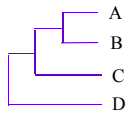
	A	B	C
B	1		
C	6	5	
D	8	7	4



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
```

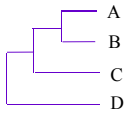
	A	B	C
B	2		
C	4	2	
D	7	5	3

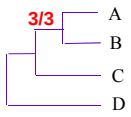


3

```
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
```

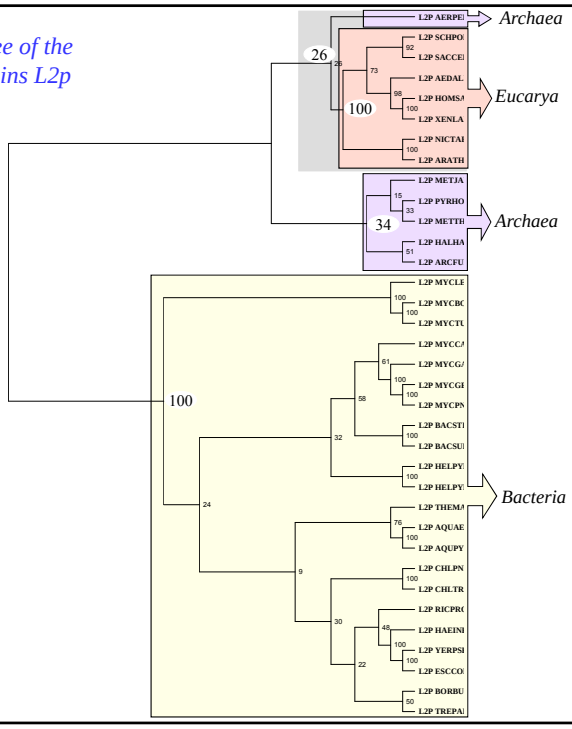
	A	B	C
B	1		
C	3	2	
D	6	5	3





Echantillonages aléatoires

Phylogenetic tree of the ribosomal proteins L2p



Phylogénie moléculaire

- Notions de base en phylogénie
- Alignement de séquences
- Principales méthodes de construction d'arbres
  - Méthodes basées sur les distances
  - Méthodes basées sur les séquences
- Evaluer la fiabilité d'un arbre
- Les limites de la phylogénie
- Quelques programmes

Les limites de la phylogénie moléculaire

Etapes

1. *Aligner correctement les séquences*
2. *Appliquer une méthode de génération d'arbres*
3. *Évaluer statistiquement la robustesse des arbres*

La qualité de l'alignement est cruciale pour la qualité de l'arbre !!

Les arbres peuvent varier suivant les régions sélectionnées.

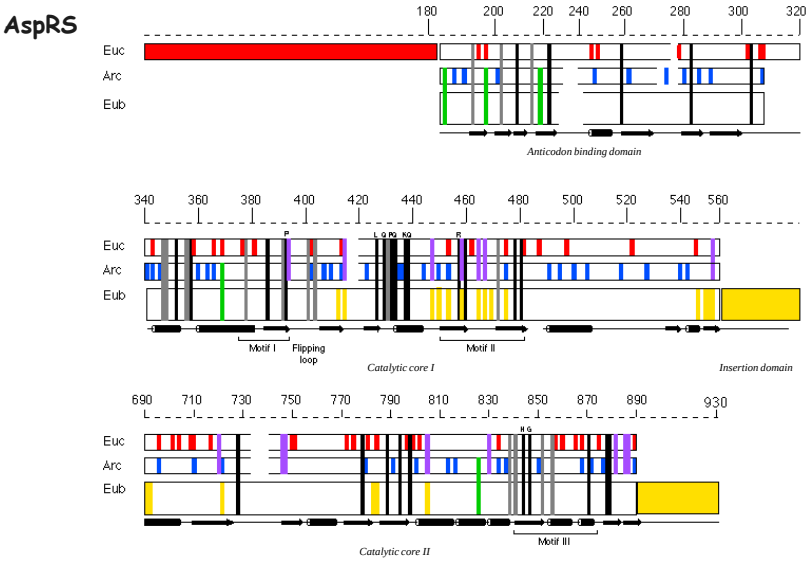
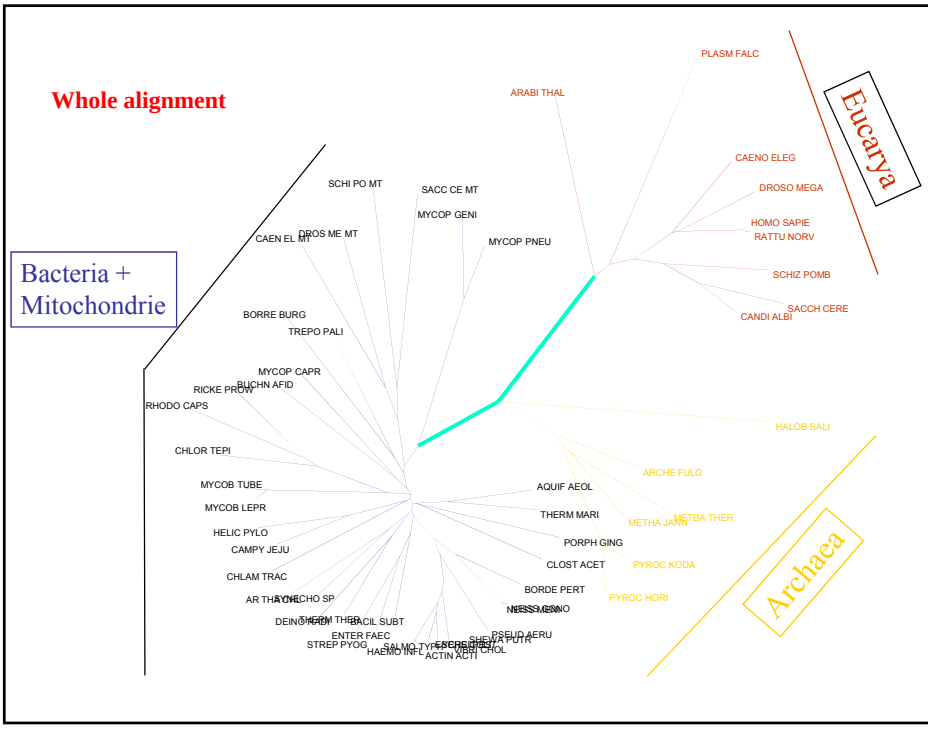
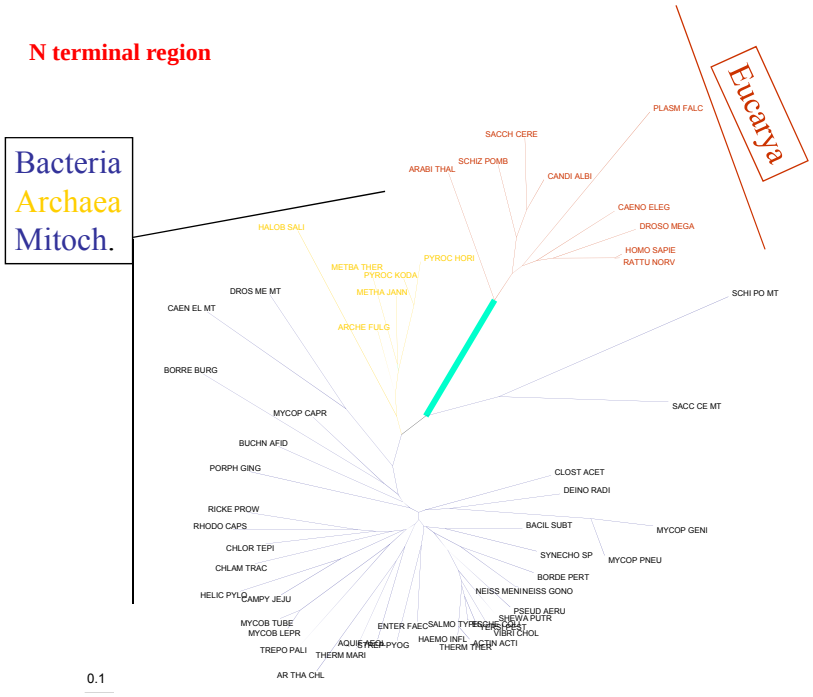


Figure 7 : Représentation schématique des conservations présentes dans les Asp tRNA synthétases après analyse OrdAli.
Colonnes noires et grises, résidus strictement conservés ou présents dans au moins 80% des séquences.
Colonnes rouges, bleues, jaunes, régions strictement conservées chez les Eucaryotes, Archaea et Bactéries respectivement.
Colonnes vertes et violettes, régions communes aux Eucaryotes et Archaea et aux Archaea et bactéries respectivement.



N terminal region



Les limites de la phylogénie

Une phylogénie est une hypothèse sur :

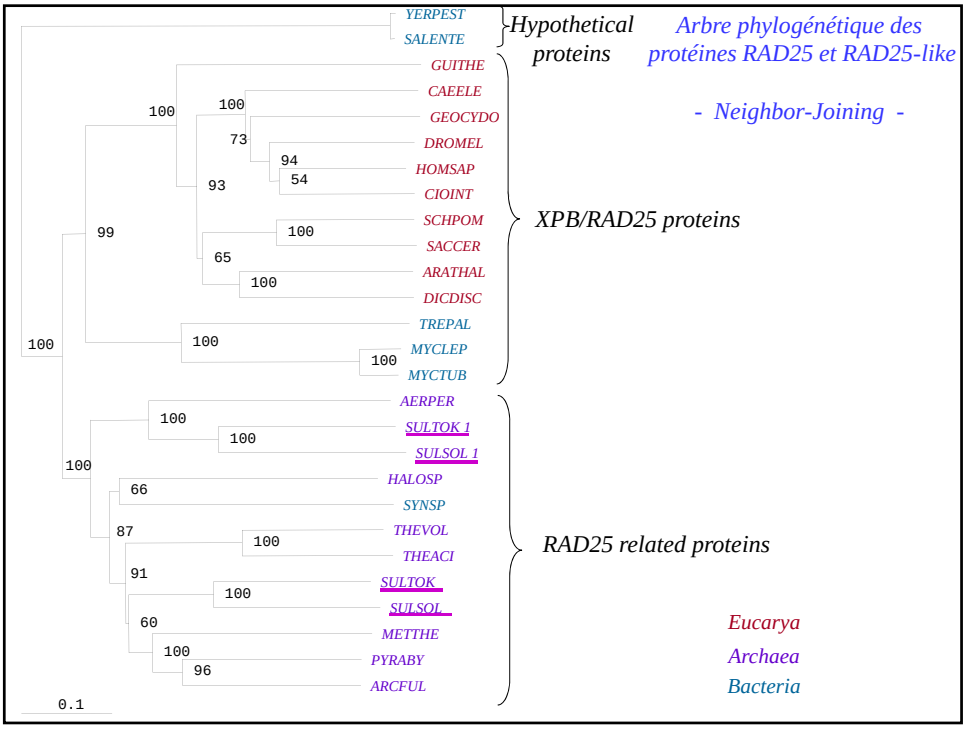
- les relations de parenté entre espèces
- les divergences et les duplications

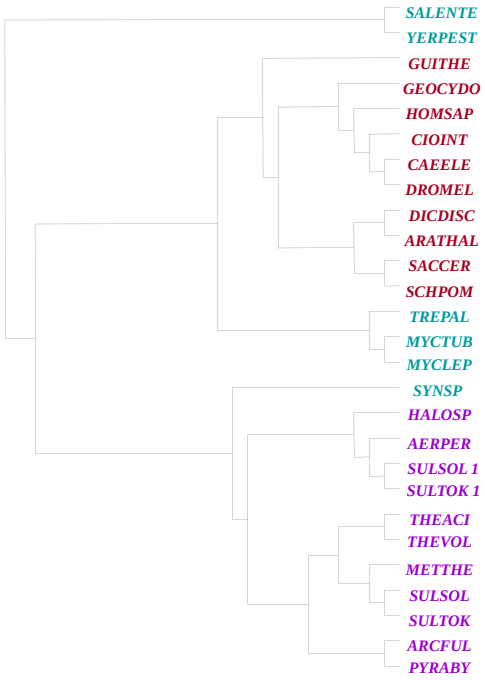
et éventuellement sur :

- la direction des changements entre espèces (arbre enraciné)
- les états ancestraux
- le taux d'évolution

Les limites de la phylogénie moléculaire

- Aucun algorithme n'est parfait
  - Il n'est jamais certain que l'arbre obtenu soit l'arbre réel !
  - Les mêmes données peuvent aboutir à des arbres différents suivant l'algorithme utilisé
- L'histoire évolutive des gènes n'est pas toujours transposable aux espèces
  - tous les gènes n'évoluent pas à la même vitesse (pressions de sélection différentes)
  - transferts horizontaux
  - paralogie





Arbre phylogénétique des protéines RAD25 et RAD25-like

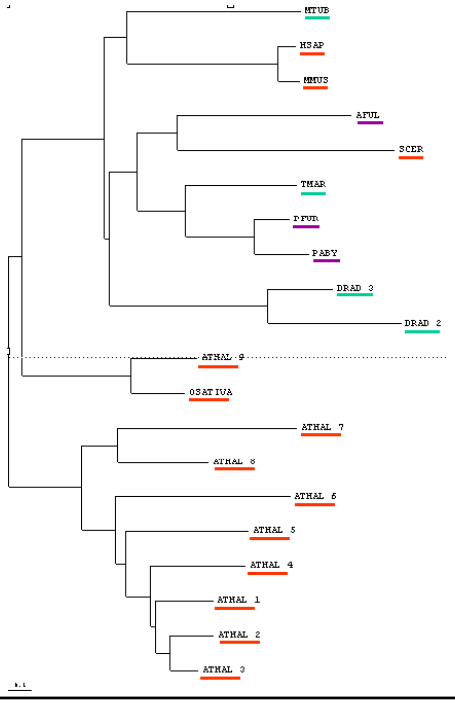
- Maximum de parcimonie -

Attention:
Aucune information sur la longueur des branches

Eucarya
Archaea
Bacteria

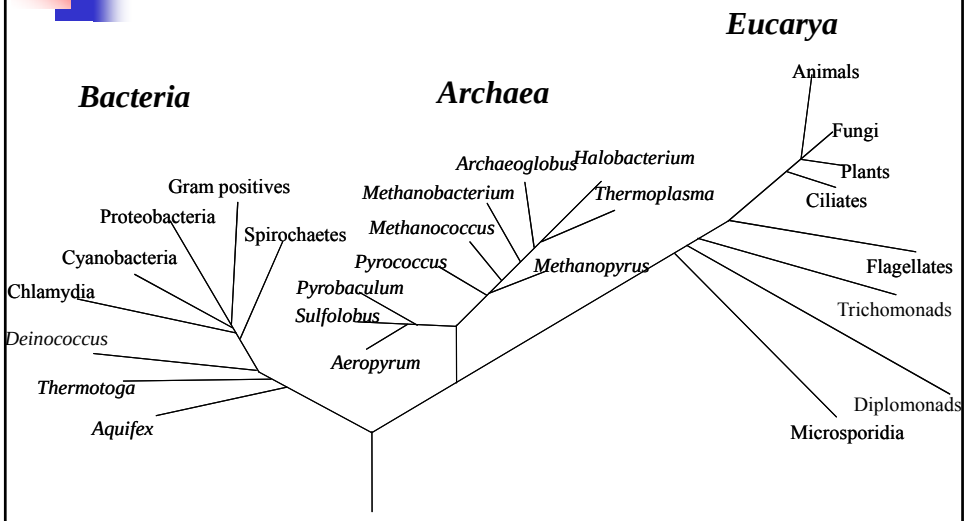
Une protéine opérationnelle : la lysophospholipase

- duplications
- pertes
- transferts

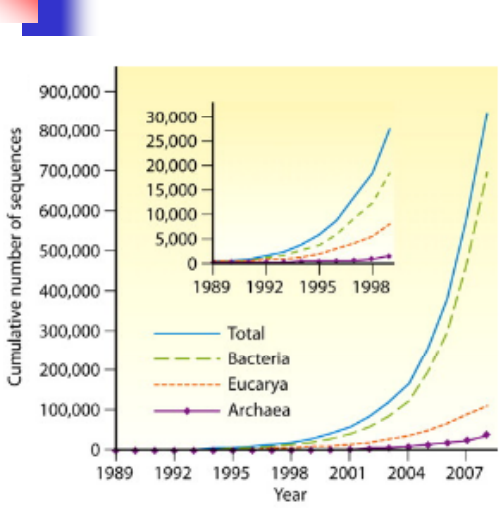


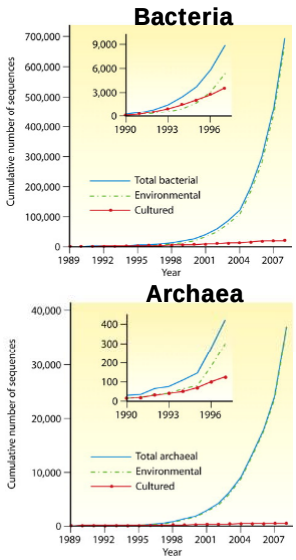
Bactéries
Archaea
Eucaryotes

Arbre du vivant

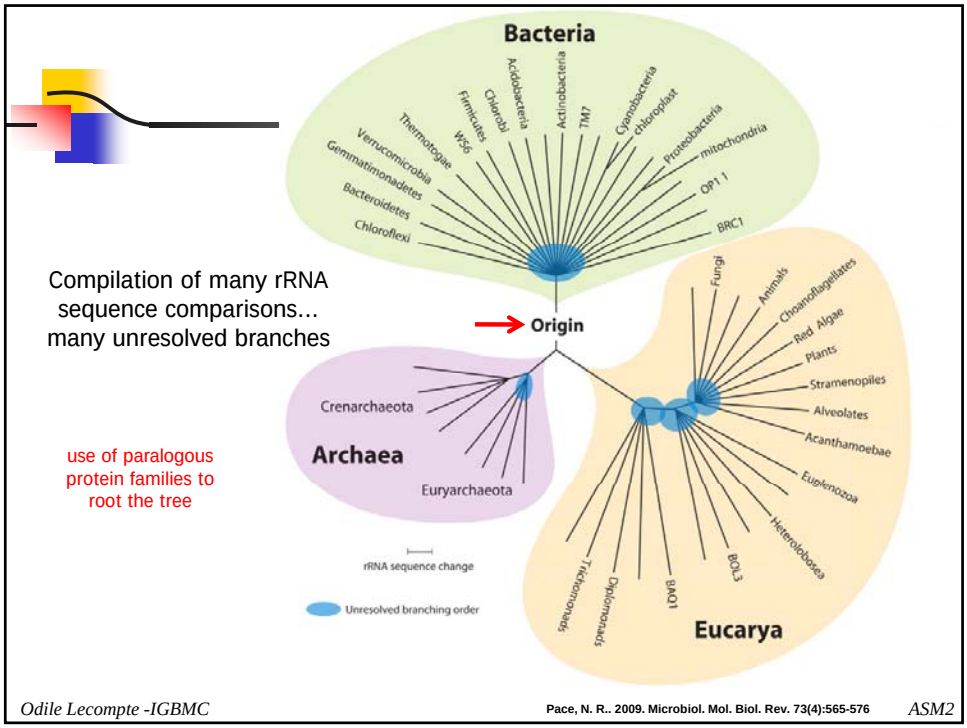


Chronological accumulation of SSU rRNA sequences





Pace, N. R.. 2009. Microbiol. Mol. Biol. Rev. 73(4):565-576

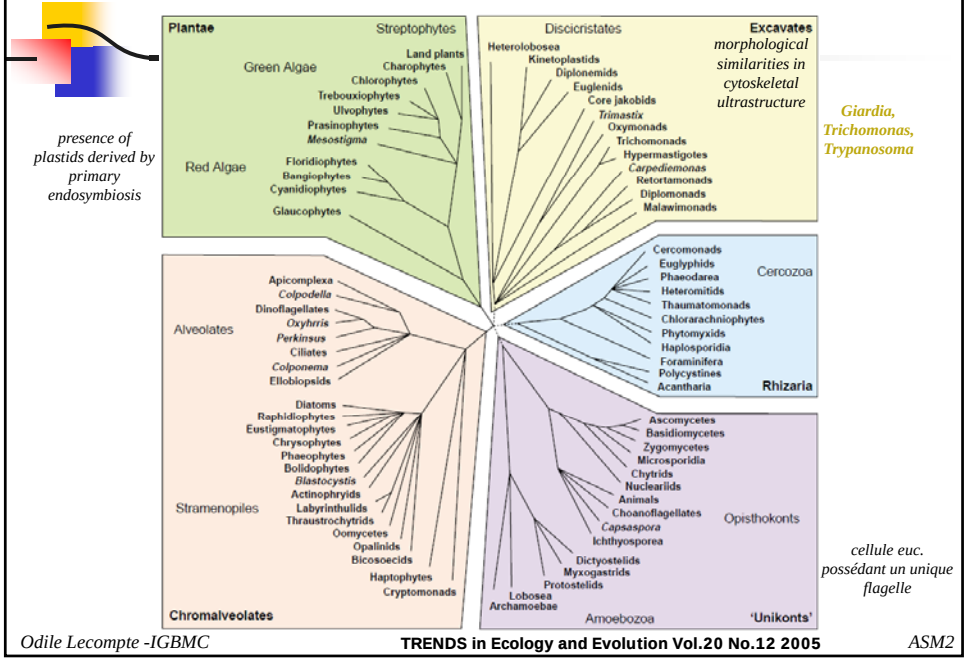


Compilation of many rRNA sequence comparisons... many unresolved branches

use of paralogous protein families to root the tree

Pace, N. R.. 2009. Microbiol. Mol. Biol. Rev. 73(4):565-576

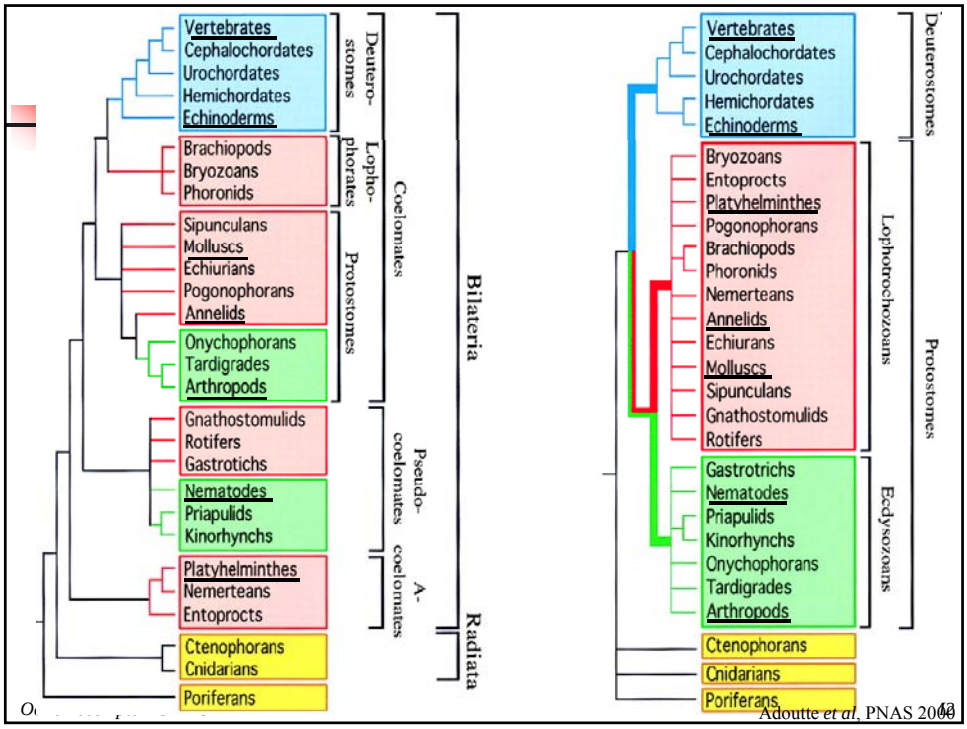
from various types of data (molecular phylogenies and other molecular characters, as well as morphological and biochemical evidence)



presence of plastids derived by primary endosymbiosis

cellule euc. possédant un unique flagelle

TRENDS in Ecology and Evolution Vol.20 No.12 2005

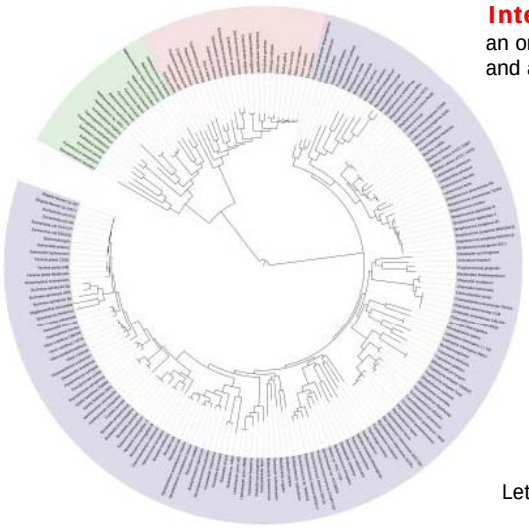


Adoutte *et al*, PNAS 2000

Quelques programmes...

- Les ensembles logiciels :
  - Phylip (très complet)
    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 (interface graphique)
    http://pbil.univ-lyon1.fr/software/phylowin.html
- Visualisation, manipulation
  - Njplot, baobab, treeedit, phylodendron...
  - Treeview (http://taxonomy.zoology.gla.ac.uk/rod/treeview.html)
  - Treedyn (http://www.treedyn.org/)

Quelques programmes...

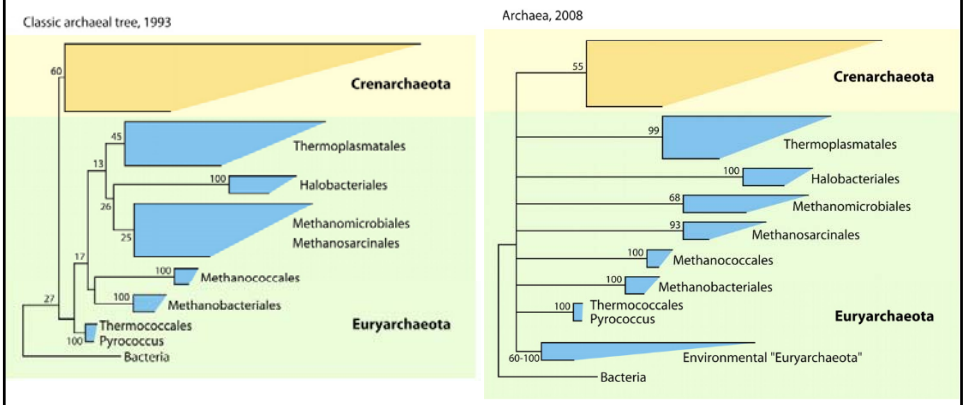


Interactive Tree Of Life (iTOL)
an online tool for phylogenetic tree display and annotation.

http://itol.embl.de/

Letunic I, Bork P. Bioinformatics 2007

Odile Lecompte -IGBMC *ASM2*


Classic archaeal tree, 1993
60
Crenarchaeota
45
Thermoplasmatales
13
100
Halobacteriales
26
Methanomicrobiales
25
Methanosarcinales
17
100
Methanococcales
27
100
Methanobacteriales
Thermococcales
100
Pyrococcus
Bacteria
Euryarchaeota
Archaea, 2008
55
Crenarchaeota
99
Thermoplasmatales
100
Halobacteriales
68
Methanomicrobiales
93
Methanosarcinales
100
Methanococcales
100
Methanobacteriales
100
Thermococcales
Pyrococcus
Euryarchaeota
60-100
Environmental "Euryarchaeota"
Bacteria