Introduction

Phylogénie : reconstruire l'histoire évolutive des espèces. Trouver des liens de parenté. Constitue une hypothèse.

Evolution moléculaire : étude de la modification du génotype causée par les mutations et qui peuvent parfois être visibles au niveau du phénotype.

Reconstruction d'arbres phylogénétiques en comparant l'information génétique présente dans le génome des êtres vivants.

Discipline relativement récente : années 1960 avec l'apparition des premières séquences.

Apport important pour la reconstruction de l'arbre du vivant car avant utilisation de caractères morphologiques, physiologiques et biochimiques, au pouvoir de résolution plus faible notamment pour les micro-organismes.

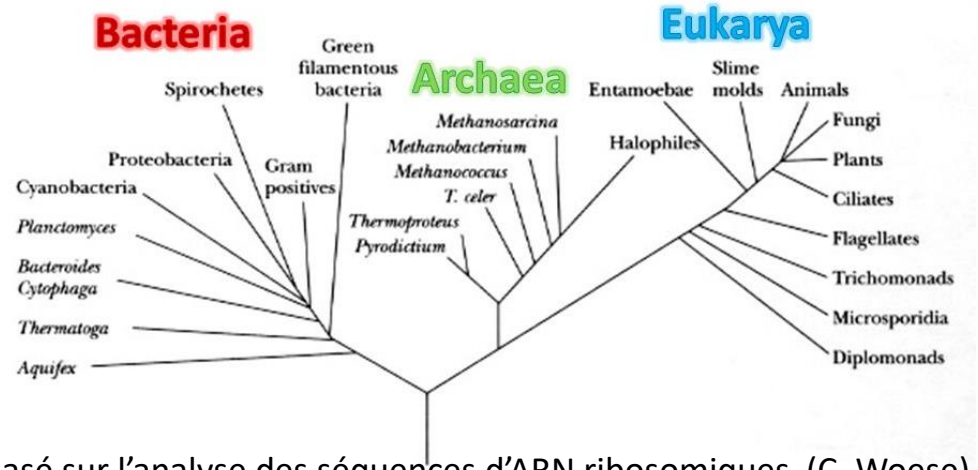
Introduction

Premières analyses faites en 1965 par E. Zuckerkandl et L. Pauling montrant que la phylogénie des vertébrés était a peu près identique quand elle était basée sur la comparaison de séquences protéiques ou sur des données morphologiques, anatomiques et paléontologiques.

Fitch et Margoliash, 2 ans plus tard, ont établi une phylogénie des vertébrés à peu près identique par comparaison des protéines du cytochrome C.

A. Wilson, grâce à l'analyse de nombreuses séquences protéines, a pu montrer que la divergence entre l'homme et les grands singes d'Afrique (chimpanzé et gorille) ne daterait que de 5 à 10 millions d'années et non de 30 millions d'années comme prédit par de nombreux paléontologues.

Introduction : les trois domaines du vivant

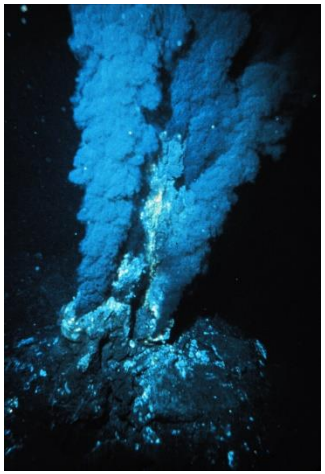


Basé sur l'analyse des séquences d'ARN ribosomiques (C. Woese)

Les archaea ont été tout d'abord découvertes dans des environnements extrêmes comme :

- les sources hydrothermales (Thermococcales, Archeoglobales)
- les sources chaudes volcaniques
- les environnements avec des concentrations en sel très élevées (Halobacteriales)

Aujourd'hui découvertes dans de nombreux biotopes pas forcément extrêmes comme le sol, les océans (très nombreuses), la flore intestinale …



Cheminée hydrothermale dans l'océan Atlantique

Submarine geothermal vent, the habitat of *N. equitans*

L'image représente " Grand Prismatic Spring" du parc national de Yellowstone

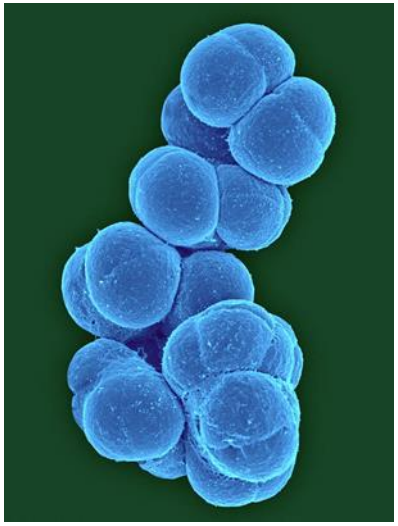
Mer morte, presque huit fois plus salée que les océans (275 g/l de NaCl). Habitat de *Haloferax volcanii*

Archaea

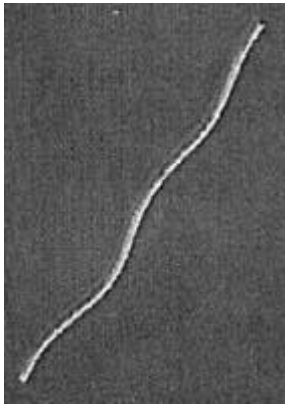
Les archaea :

• composition chimique de la paroi différente de celle des bactéries (plusieurs types de composition existent)

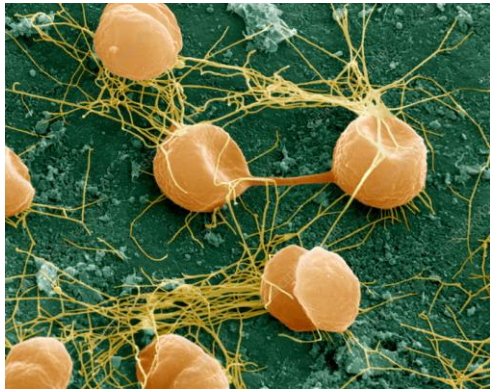
• phospholipides membranaires ont des caractéristiques spécifiques ne se retrouvant ni chez les bactéries, ni chez les eucaryotes



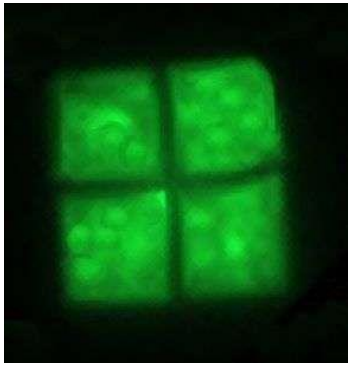
Methanosarcina barkeri



Methanobacterium thermoautotrophicum

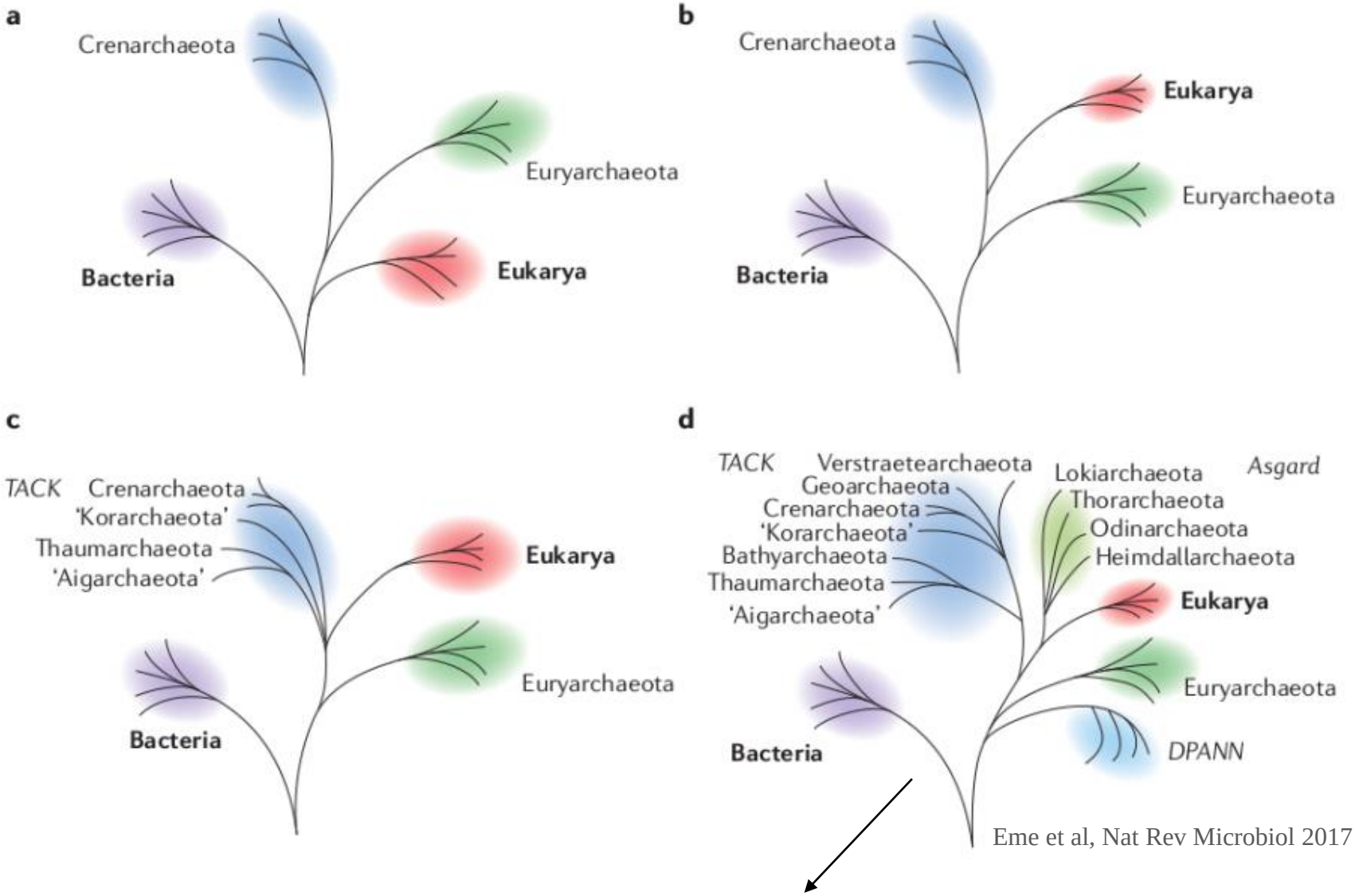


Pyrococcus furiosis



Haloquadratum walsbyi

Archaea and the evolution of the tree of life



Eme et al, Nat Rev Microbiol 2017

Arbre « Eocyte », hypothèse proposée par J. A. Lake en 1984 en se basant sur la comparaison de la structure ribosomale chez les bactéries, archées et eucaryotes (Proc. Natl Acad. Sci. USA 81, 3786–3790 (1984).

Métagénomique

La **métagénomique** est une méthode d'étude du contenu génétique d'échantillons obtenus à partir de prélèvements réalisés dans des environnements naturels complexes (ex : intestin, océan, sols, air, etc.) par opposition à des échantillons cultivés en laboratoire).

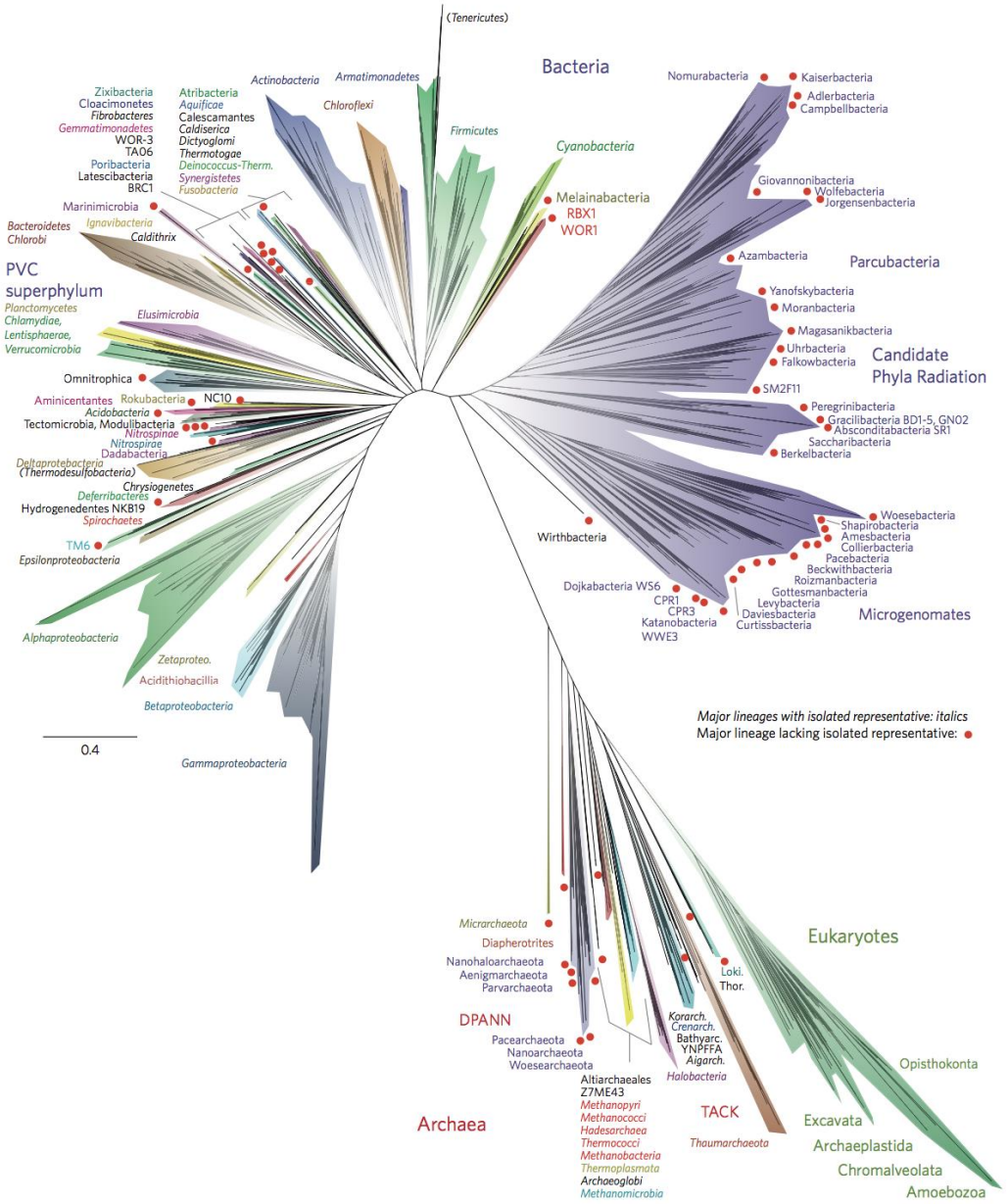
Cette approche, via le séquençage direct de l'ADN présent dans l'échantillon, permet une description génomique du contenu de l'échantillon mais offre aussi un aperçu du potentiel fonctionnel d'un environnement.

Préfixe « méta » → « *ce qui vient après* » : la métagénomique vient après la génomique.

Extrait de Wikipedia

Exemple : études des communautés microbiennes présentes dans ce cours d'eau recevant le drainage acide de mines de charbon en surface.

Métagénomique : nouvel arbre du vivant



Les phyla marqués par des points rouges ont été identifiés par métagénomique et ne possède pas de représentant qui ont été isolé.

Extrait de Hug *et al.* (2016) *Nature Microbiology.* ***1*** *(5): 16048*

Introduction

Aujourd'hui l'évolution moléculaire utilisée non seulement par les spécialistes de la phylogénie mais aussi par de nombreux biologistes désirant mieux analyser leurs séquences, comprendre l'évolution de leur fonction, analyser l'histoire des duplications etc….

Pour cela il faut entre autre connaître :

- les différents modèles évolutifs qui ont été proposés
- les différentes méthodes de reconstruction d'arbres qui ont été développées
- apprendre à analyser les arbres obtenus

Notions de base, définitions

Homologie :

Deux structure (ou deux caractères) sont dits homologues si elles dérivent d'une structure unique présente chez l'ancêtre commun aux organismes qui les portent. Ces structures ont donc une origine évolutive commune mais peuvent présenter des variations suite à une évolution indépendante.
Donc nous diront que deux gènes sont **homologues** s'ils ont divergé à partir d'une séquence ancêtre commune.

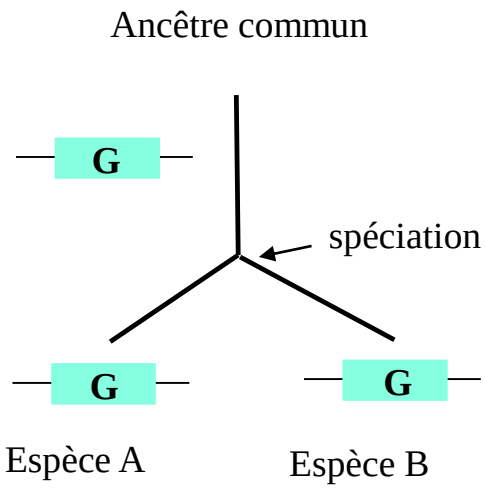
Définition insuffisante pour reconstruction de l'histoire évolutive car plusieurs mécanismes possibles pour dériver d'une séquence ancêtre.

Orthologie : deux gènes sont **orthologues** si leur divergence est due à la spéciation (le gène ancêtre commun se trouvait dans l'organisme ancêtre).

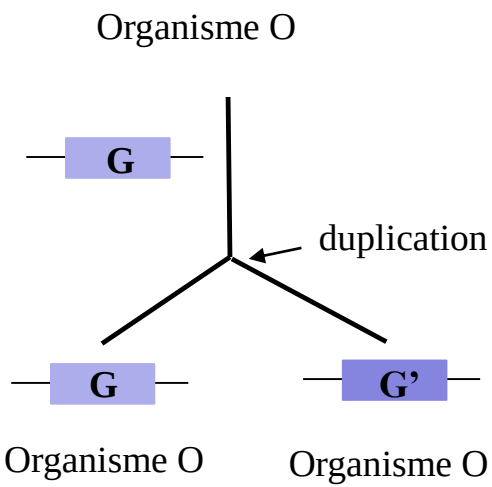
Paralogie : deux gènes sont **paralogues** si leur divergence est due à la duplication du gène ancêtre.

Xénologie : deux gènes sont xénologues si l'un d'entre eux a été acquis par transfert horizontal

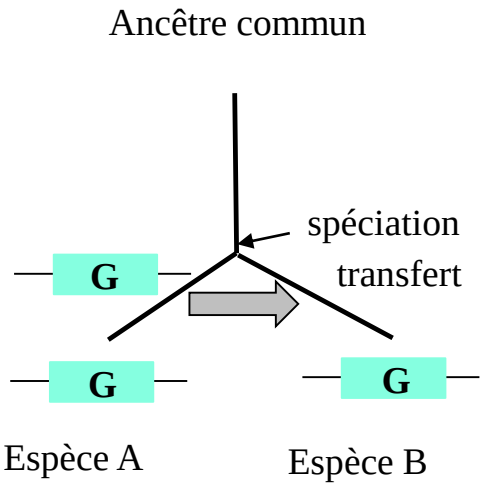
Notions de base, définitions



Génes orthologues



Génes paralogues



Génes xénologues

Notions de base, définitions

Pourquoi la comparaison de séquences :

Hypothèse 1: si deux ou plusieurs séquences possèdent des résidus conservés (bases ou acides aminés), cela signifie qu'elles ont une histoire évolutive commune. Elles ont évoluées à partir d'une séquence ancêtre commune.

Hypothèse 2 : si deux séquences sont homologues, alors elles doivent avoir des fonctions similaires.

Le pourcentage de similarité entre deux séquences est considéré comme reflétant la distance évolutive existant entre ces deux séquences. Les différences observées sont dues à l'accumulation de mutations au cours du temps. Les mutations prises en compte sont les substitutions et les insertions/délétions (indels).

Alignement multiple

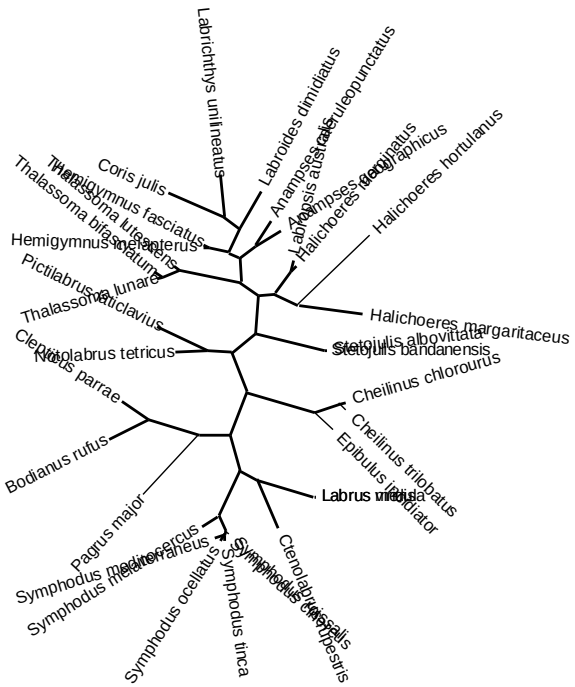
Notions de base, définitions

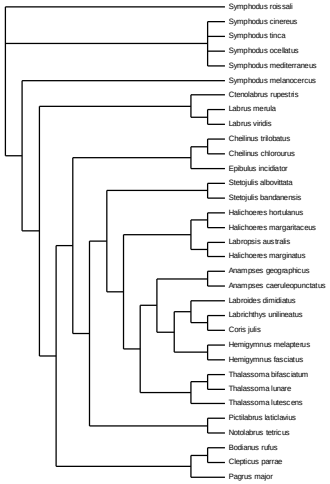
Importance de la prédiction des gènes (protéines) orthologues :

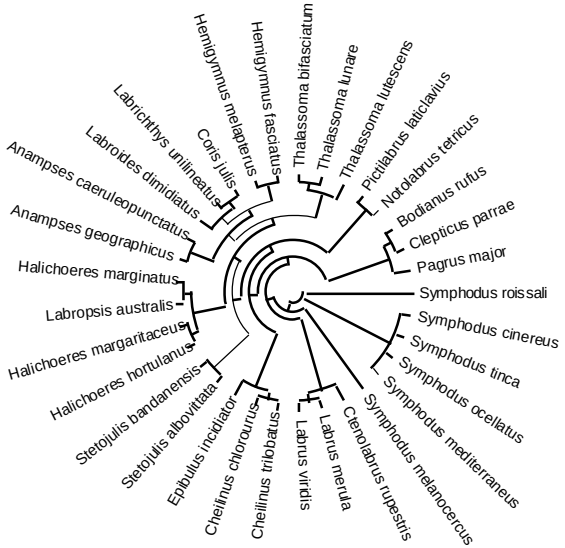
- Pour l'inférence fonctionnelle lors de l'analyse d'une famille de protéines car des protéines orthologues sont supposées avoir conservé la même fonction. En effet, la paralogie conduisant à la redondance fonctionnelle si les gènes dupliqués ne sont pas éliminés au cours de l'évolution permettrait entre autre la néo- fonctionnalisation (apparition de nouvelles fonctions)

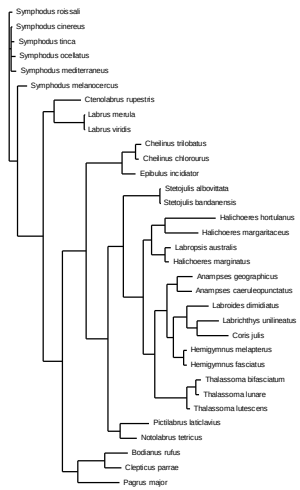
- Pour établir les relations évolutives entre les espèces car seuls les gènes/protéines orthologues récapitulent les relations entre les espèces. Les gènes paralogues et xénologues ne reflètent pas ces liens de parenté

Notions de base : arbres phylogénétiques



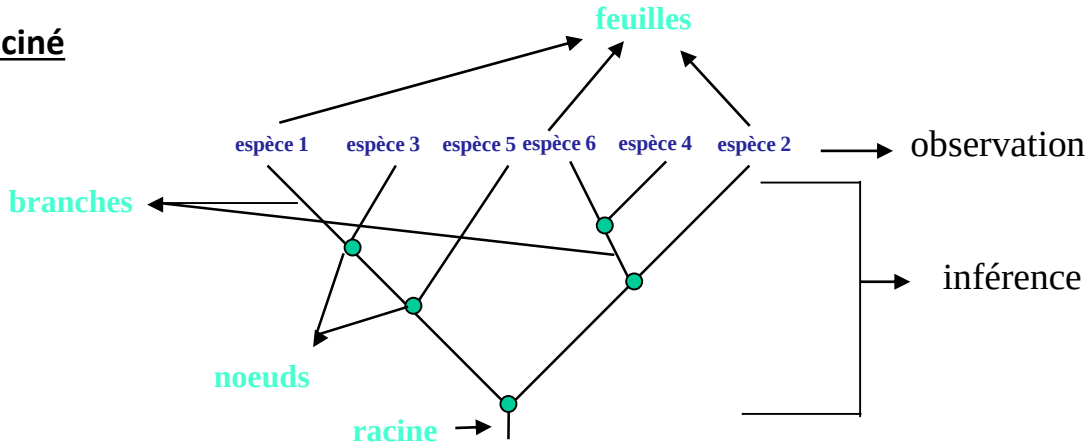






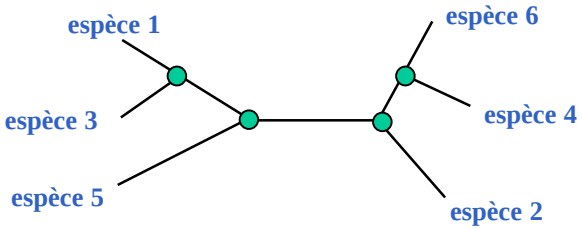
Notions de base : arbres phylogénétiques

Arbre enraciné



• Les sommets externes sont appelés **feuilles**. C'est la seule partie basée sur l'observation.
• Les sommets internes sont appelés **nœuds**. Ils représentent l'ancêtre commun hypothétique dans le sens où leur existence n'est pas fondée sur l'observation mais sur le processus de reconstruction.
• La relation entre deux nœuds est appelée **branche**. Les branches peuvent être valuées, c'est à dire que l'on peut leur associer une mesure (ex: une distance, une quantité d'évolution, un nombre de mutations) qui dépend de la méthode de reconstruction utilisée. Elles donnent une estimation de la divergence entre les nœuds.
• La **racine** définie l'origine commune des espèces traitées. Les liens entre nœuds et feuilles sont orientés, on part de la racine et on remonte aux feuilles.

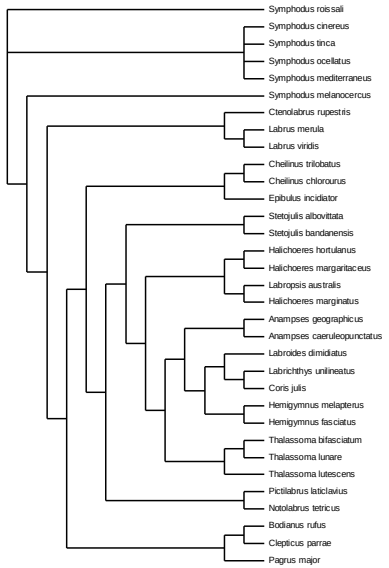
Arbre sans racine



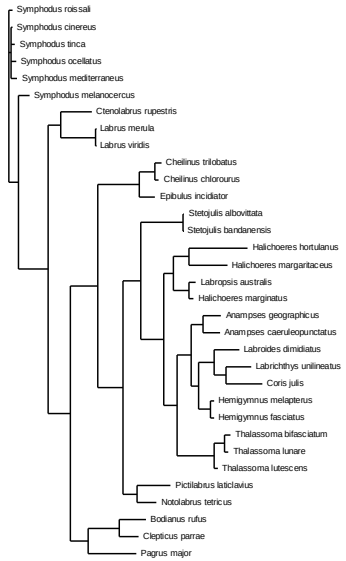
Dans un arbre sans racine, les liens entre nœuds ne sont pas orientés et un seul et unique chemin permet de passer d'un sommet à l'autre.

Notions de base : arbres phylogénétiques

Cladogramme : pas de longueurs de branches
(feuilles sous un même nœud appelée clade)



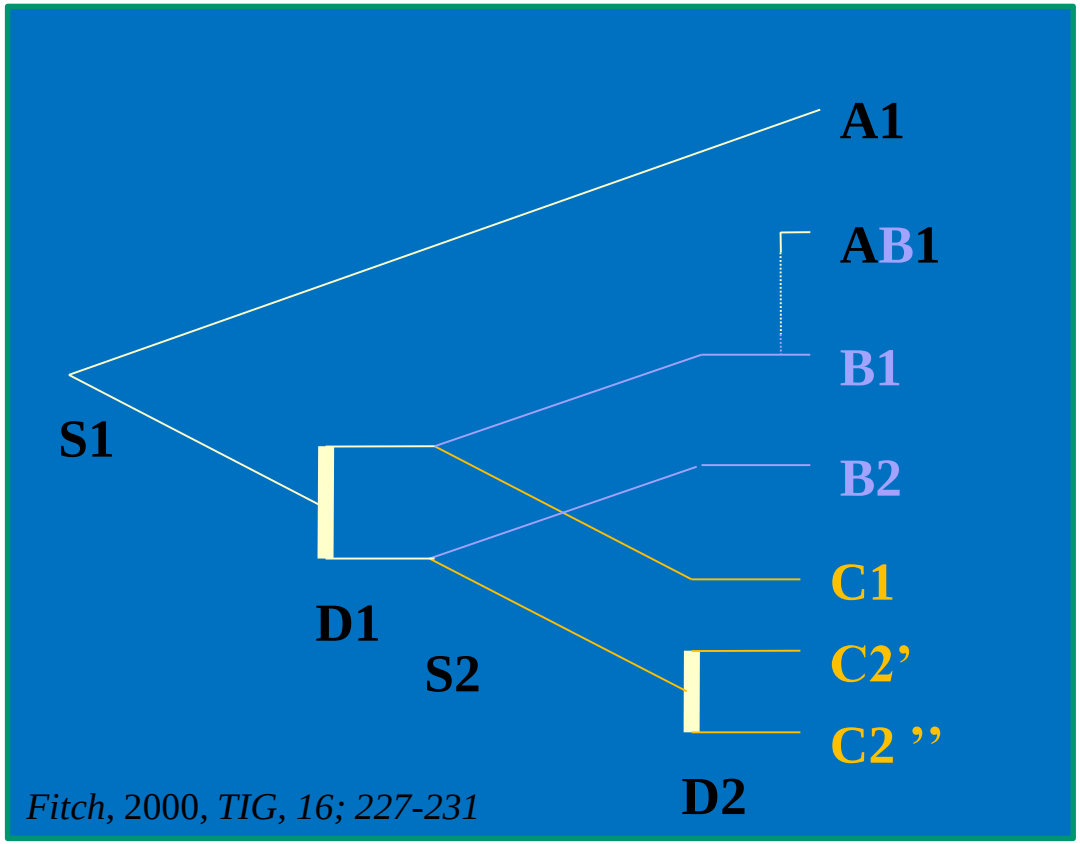
Arbre ultramétrique



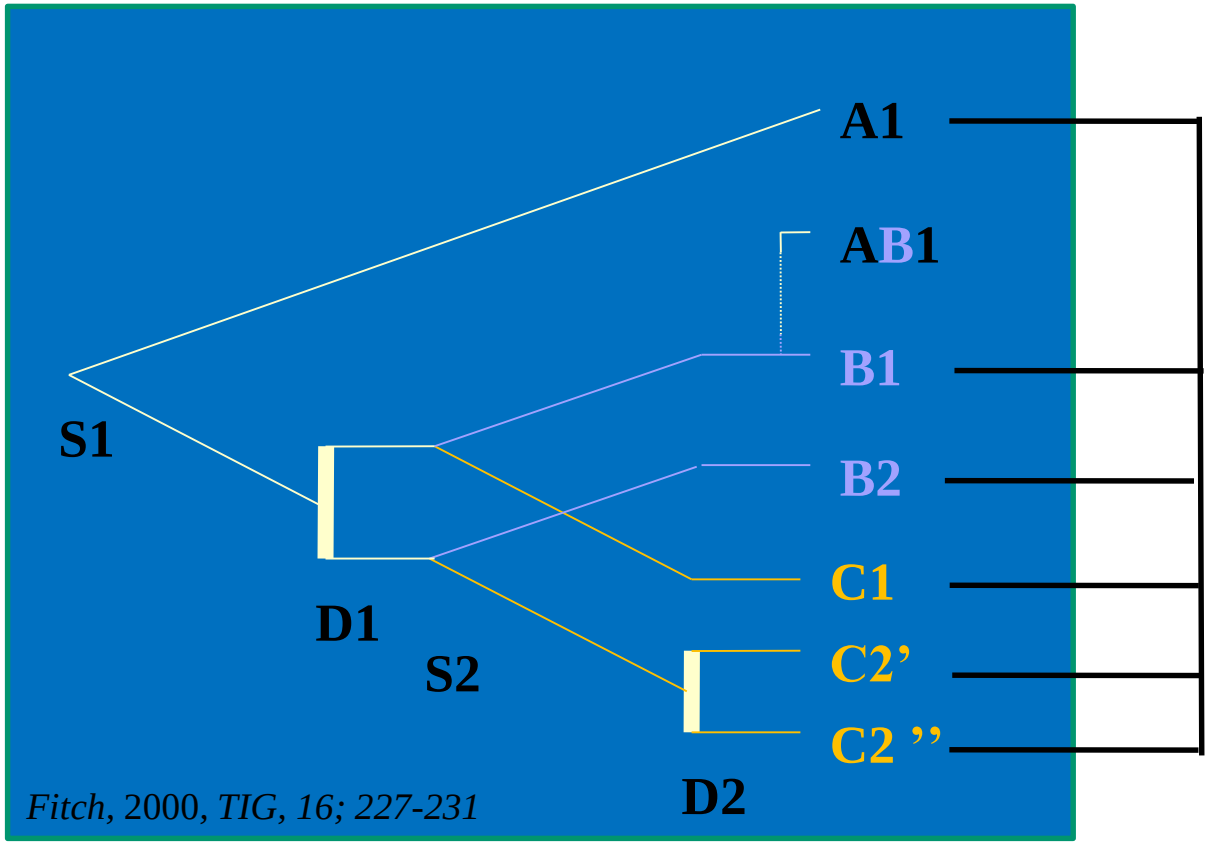
Arbre additif

Phylogramme ou dendrogramme : longueurs de branches

Notions de base: arbres phylogénétiques

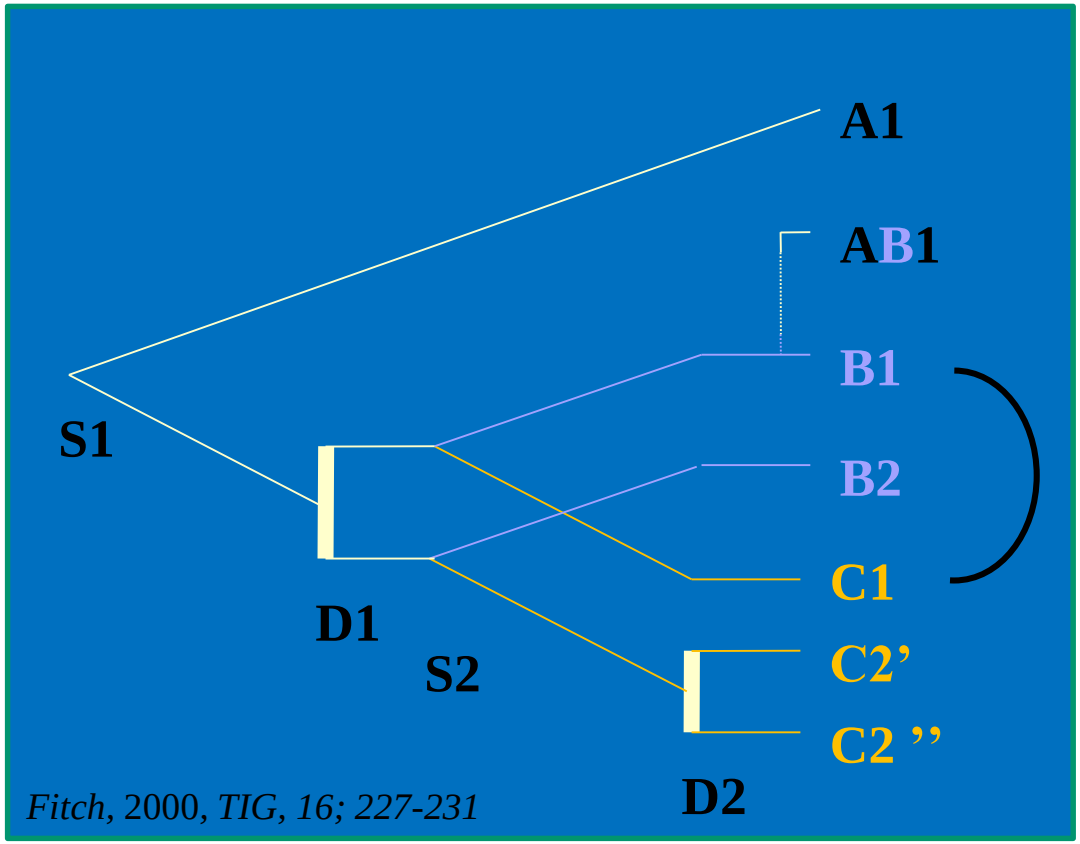


Notions de base: arbres phylogénétiques



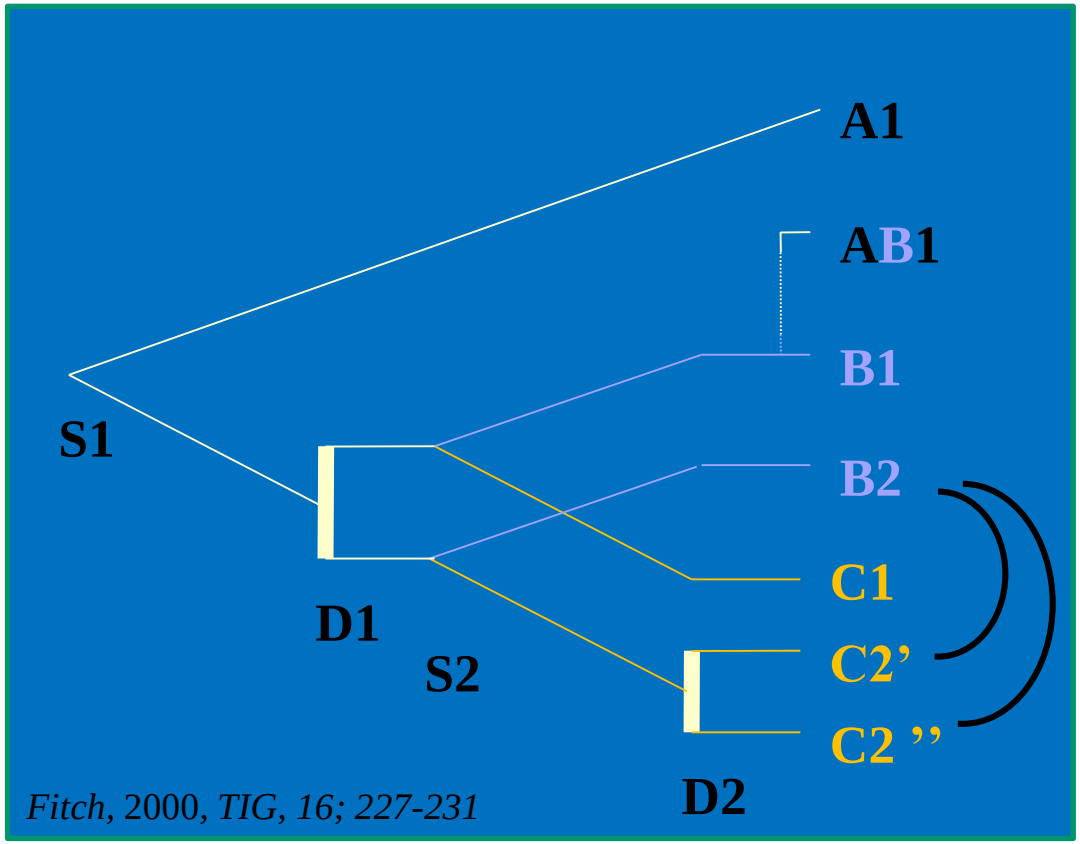
Orthologues à A1

Notions de base: arbres phylogénétiques



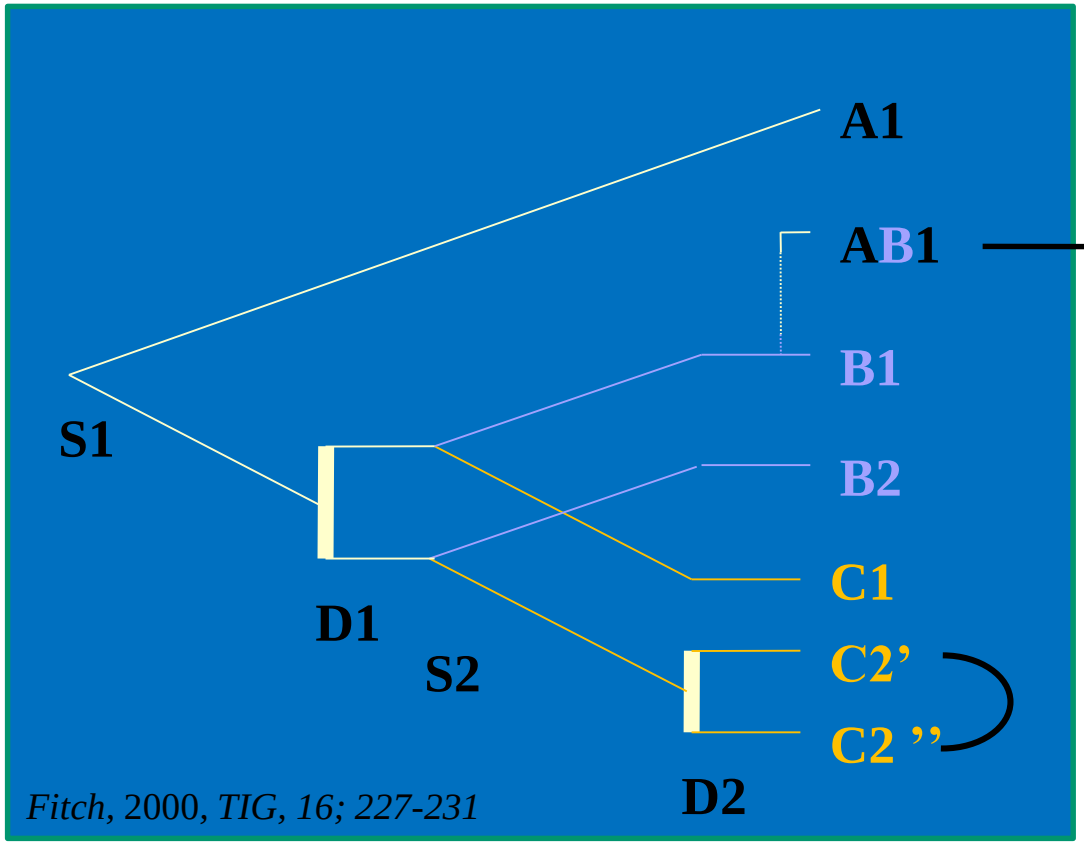
Orthologues

Notions de base: arbres phylogénétiques



Orthologues

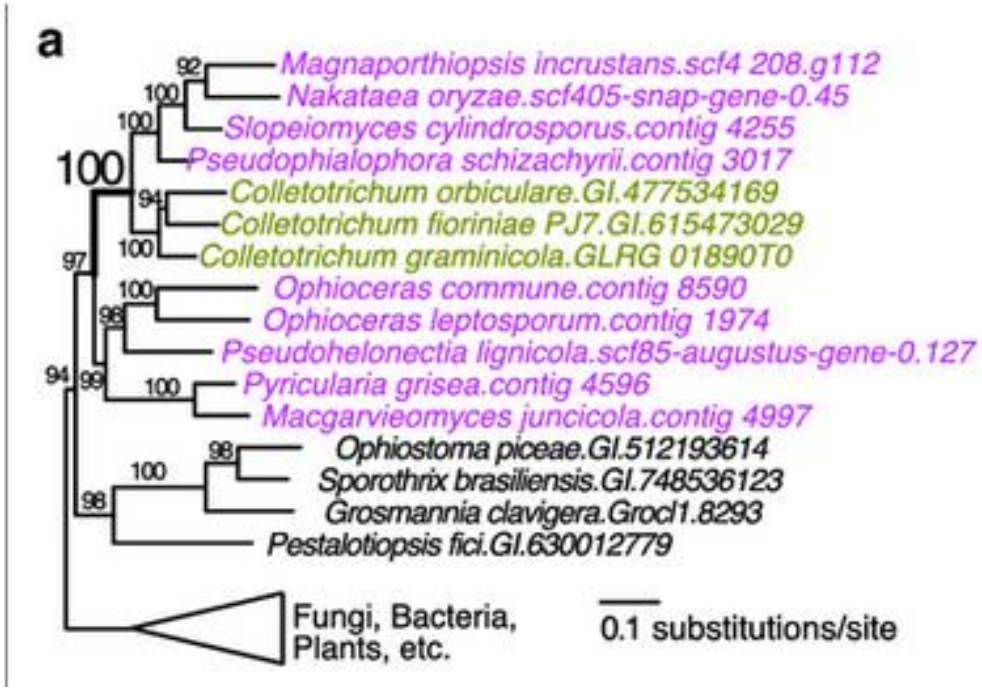
Notions de base: arbres phylogénétiques



transfert horizontal

paralogues

Notions de base: arbres phylogénétiques



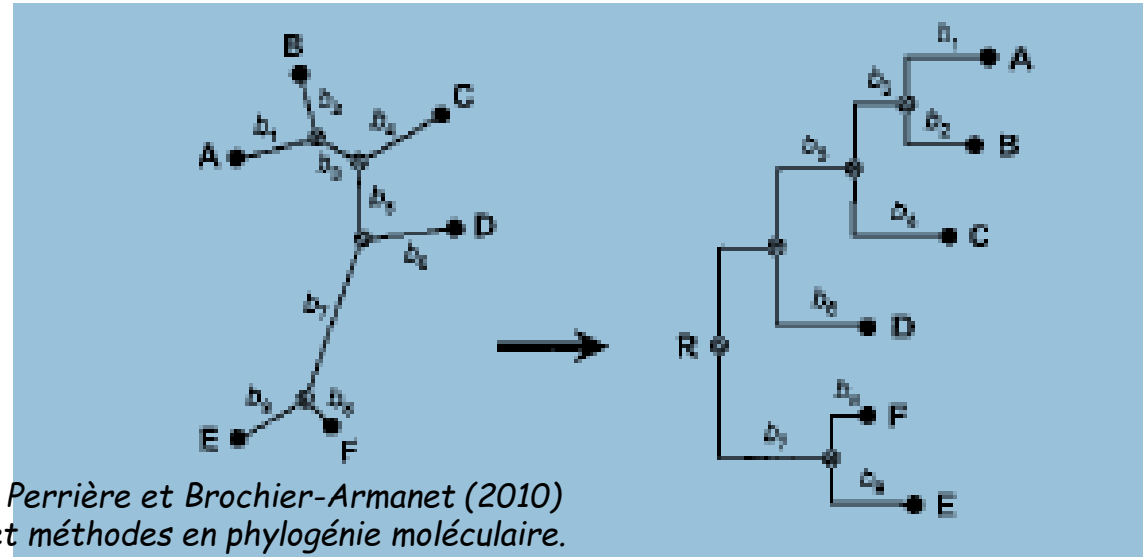
Extrait de Qiu *et al.,* BMC Biology (2016) 14:41

Cet arbre montre un exemple de transfert de genes entre des Magnaporthales aux Colletotrichum, deux genres de champignons ascomycètes

Notions de base : arbres racinés et non racinés

La plupart des méthodes produisent des arbres non racinés car elles détectent des différences entre séquences mais n'ont aucun moyen d'orienter temporellement ces différences.

- enraciner un arbre :
  - Enracinemenent au barycentre : ne nécessite pas de connaissances *à priori* . Positionne la racine au milieu du chemin séparant les deux groupes de feuilles les plus éloignés. La racine est donc le point de l'arbre équidistant de toutes les feuilles. Fait l'hypothèse de l'horloge moléculaire : on suppose que toutes les séquences ont évolué à la même vitesse depuis leur divergence de leur ancêtre commun. Attention, ici on fait une hypothèse très lourde qui est rarement vérifiée par les données.

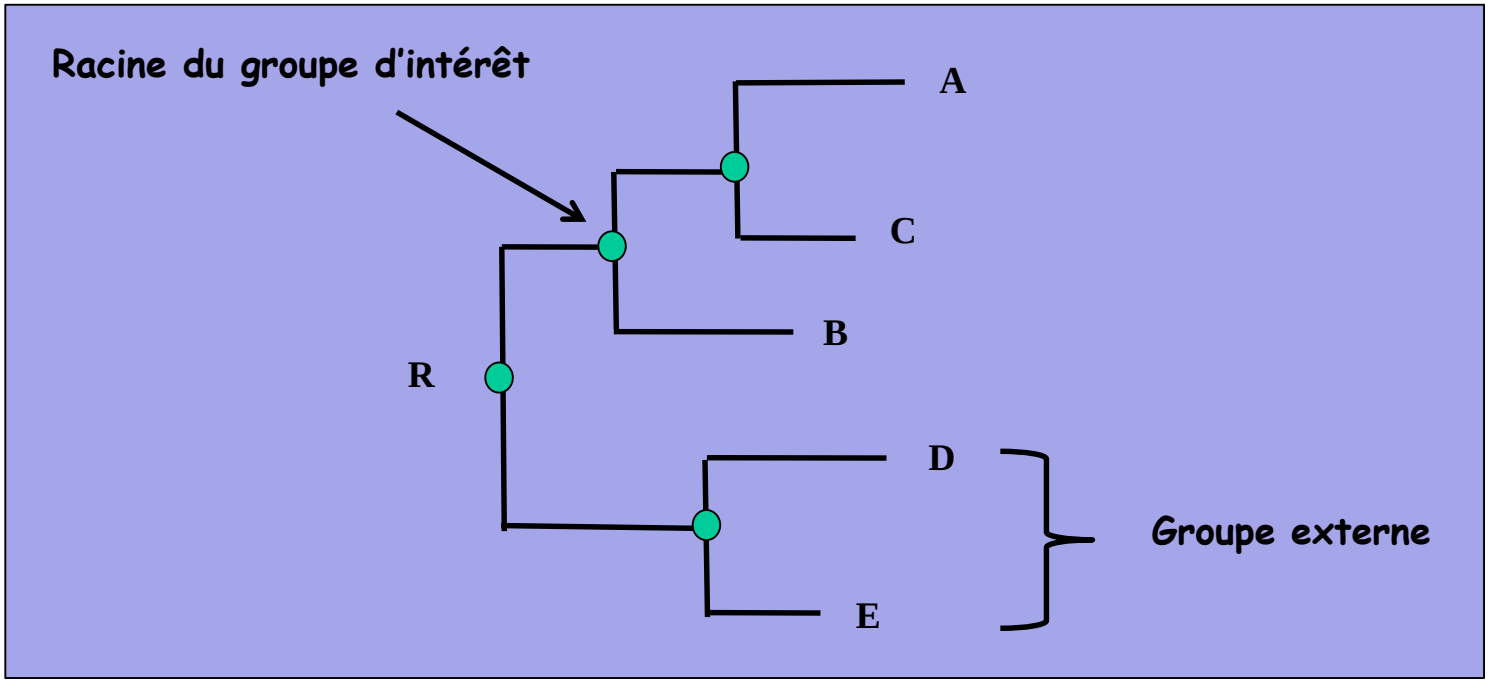


Extrait de Perrière et Brochier-Armanet (2010)
Concepts et méthodes en phylogénie moléculaire.

Notions de base : arbres racinés et non racinés

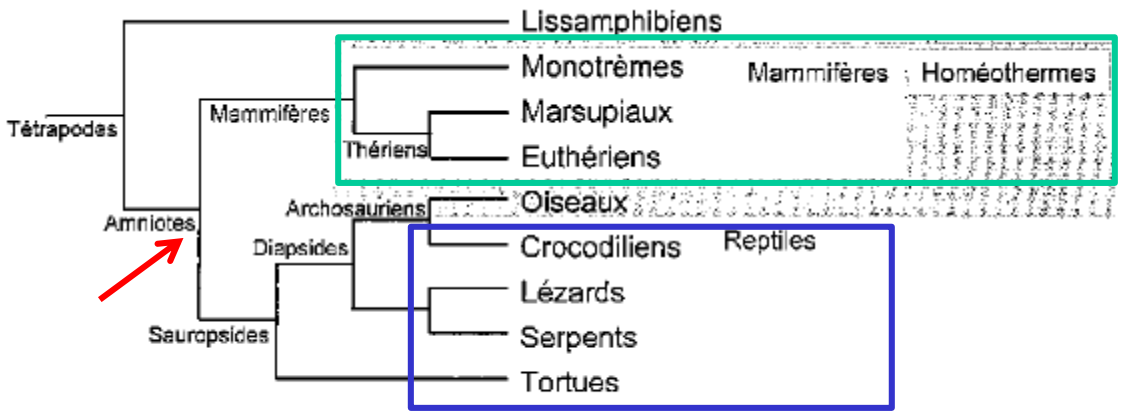
- enraciner un arbre :
  - La méthode du groupe externe : inclure un groupe de séquences connues *a priori* comme externes au groupe d'intérêt; la racine est alors sur la branche qui relie le groupe externe aux autres séquences. Séquences connues comme ayant a priori divergé avant le groupe d'intérêt.

  Problème : choix du groupe externe, qui doit être le plus proche possible du groupe d'intérêt.



Notions de base : vocabulaire

Exemple de la phylogénie des Tétrapodes



Extrait de Perrière et Brochier-Armanet (2010) Concepts et méthodes en phylogénie moléculaire.

Le Groupe des Mammifères est *monophylétique* car l'ensemble des feuilles sont les descendants d'un même ancêtre.

Le Groupe des Reptiles (Crocodiliens, Lézards, Serpents et Tortues) est *paraphylétique* car les oiseaux qui sont des descendants de l'ancêtre des Reptiles ne font pas partie de ce groupe (donc paraphylétique quand une partie des descendants d'un même ancêtre n'est pas présent dans le même groupe que les autres)

Les Tétrapodes à sang chaud (Mammifères et Oiseaux) forment un groupe *polyphylétique* car leur ancêtre commun, celui des Amniotes, n'est pas à sang chaud et donc pas inclus dans le groupe.

Notions de base : les caractères

Caractères : Organismes composés de différents caractères
Chaque position alignée d'un alignement multiple

Ces caractères prennent des formes différentes selon les taxons : elles sont appelées *états de caractères*

L'état du caractère peut être soit **ancestral** (présent chez l'ancêtre commun des Operational Taxonomic Unit (OTU) analysées, soit **dérivé** (observables que dans certains OTU)

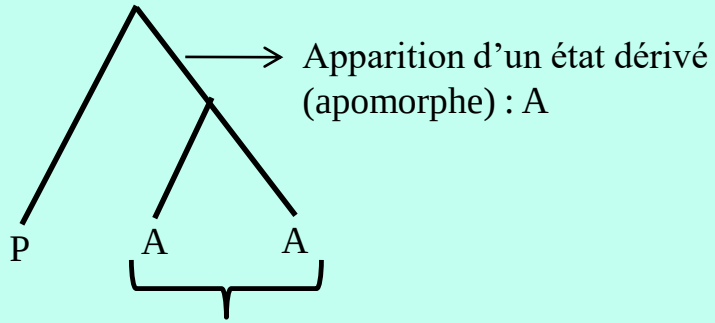
- état ancestral est dit *plésiomorphe*
- état dérivé est dit *apomorphe*
- le partage d'un état dérivé ou apomorphe par plusieurs OTU est appelé *synapomorphie*
- le partage d'un état ancestral ou plésiomorphe par plusieurs OTU est appelé *symplésiomorphie*
- un état dérivé porté que par une seule OTU est appelé *autapomorphie*

L'inférence phylogénétique se fait à partir des différences entre états de caractères

- On cherche à établir le lien entre ancêtre et descendant par la présence/absence d'un état de caractère
- On cherche l'apparition de nouveaux états de caractères dans les descendants

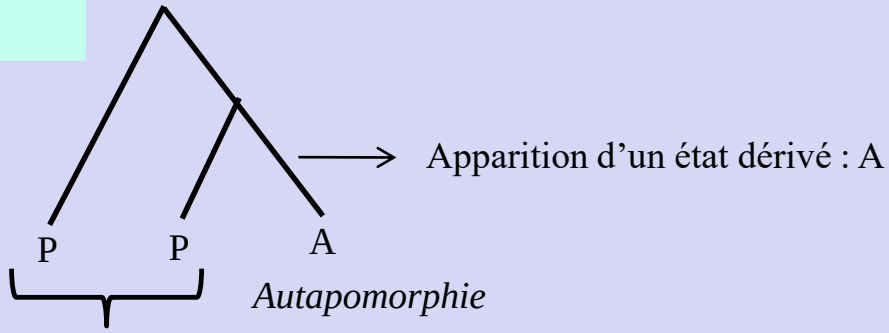
Notions de base : les caractères

état ancestral (plésiomorphe P)



Synapomorphie

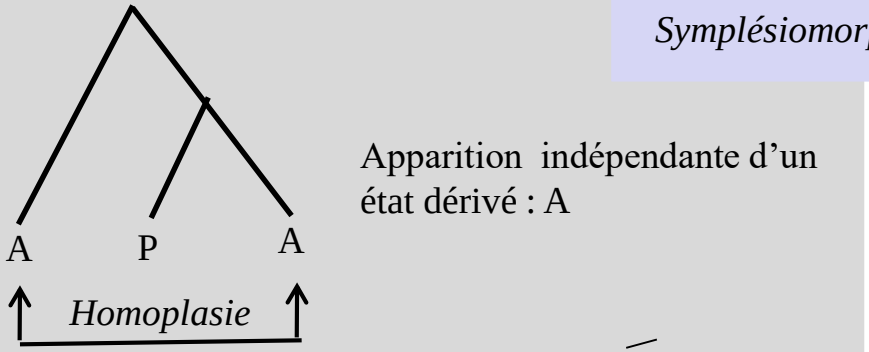
état ancestral (P)



Autapomorphie

Symplésiomorphie

état ancestral (P)



Apparition indépendante d'un état dérivé : A

Reconstruction phylogénétique : deux écoles

A partir de l'observation des états des caractères, il va falloir reconstruire l'arbre et interpréter les ressemblances.

Deux écoles :

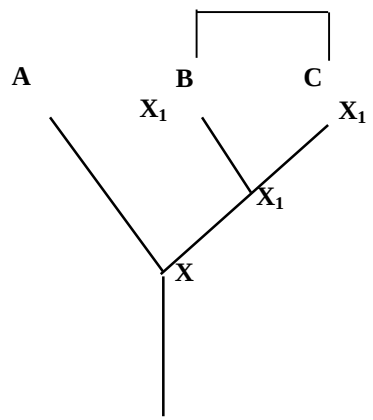
- les phénéticiens adeptes de la « taxonomie numérique ». Les liens entre les taxons ne peuvent être fondés que sur la base d'une similitude globale exprimée à partir de matrices de calcul de distances. Dans le cas des séquences, à partir d'un alignement multiple, on calculera les distances entre les séquences prises deux à deux en prenant en compte toutes les positions alignées sans indels.
  L'analyse phénétique se fonde sur l'analyse du plus grand nombre de caractères.
- Les cladistes préfèrent élaborer des phylogénies à partir d'un ensemble préalablement choisi de caractères.

Le concept de similarité

Il peut être divisé en :

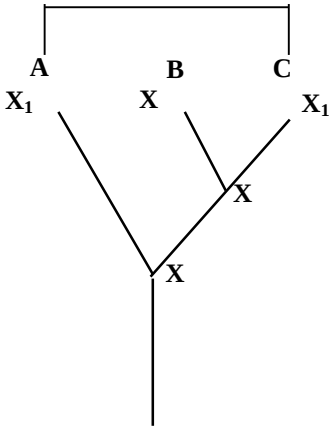
- homologie similarité héritée d'un ancêtre commun
- homoplasie similarité non héritée d'un ancêtre commun et qui est subdivisée en :
  - convergence : apparition indépendante dans deux espèces d'un même état dérivé de caractère
  - réversion : apparition d'un état de caractère ayant la forme ancestrale

homologie



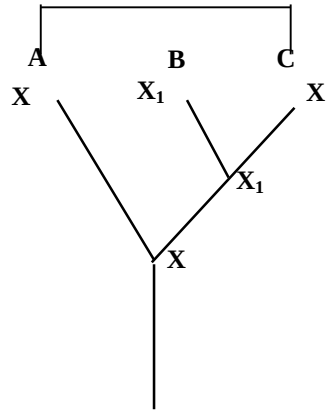
L'état de caractère X_1 présent chez B et C est hérité de l'ancêtre à B et C.

convergence



L'état de caractère X_1 présent chez A et C n'est pas hérité d'un ancêtre commun. Il est apparu deux fois de façon indépendante.

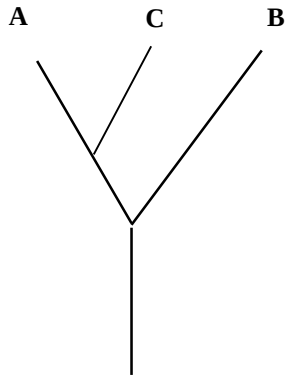
réversion



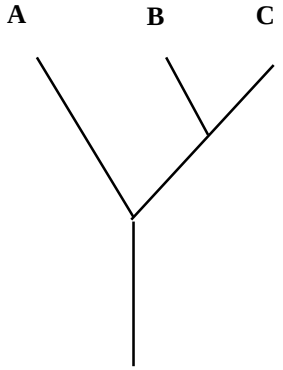
L'état de caractère X présent chez A et C n'est pas hérité d'un ancêtre commun. Chez A, il correspond à l'état ancestral et chez C il est secondairement transformé (retour à la forme ancestrale).

Le concept de similarité

Arbre qui sera déduit à partir des valeurs des états de caractères dans
Le cas convergence ou réversion:



Or dans les deux cas, convergence ou réversion, le vrai arbre est :



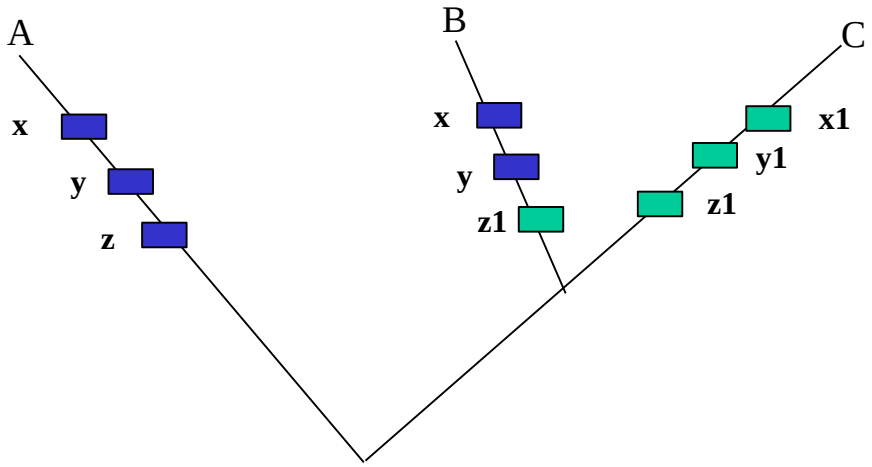
Etats dérivés- Etats primitifs

Cladistes : états dérivés plutôt qu'états primitifs des caractères homologues

Trois espèces et 3 caractères x, y et z

État primitif x, y et z

État dérivé x1, y1 et z1



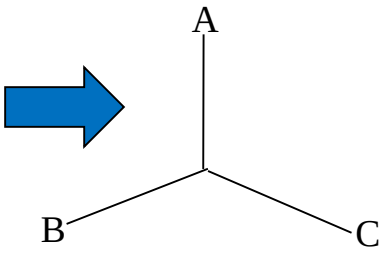
Le fait que B et C possèdent le même état dérivé z1 est plus informatif que le fait que A et B possèdent les mêmes états primitifs x et y

Pour trouver les états dérivés et primitifs, on considère une espèce éloignée et on voit si l'état du caractère est partagé ou pas. Si partagé : état primitif.

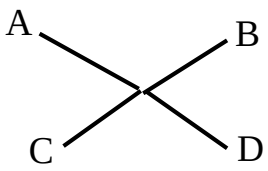
Trouver l'arbre

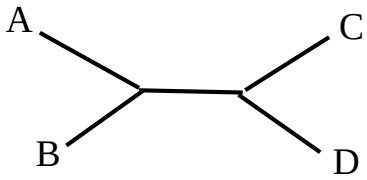
Problème : un seul arbre vrai, l'arbre évolutif
Comment le distinguer dans tous les arbres possibles

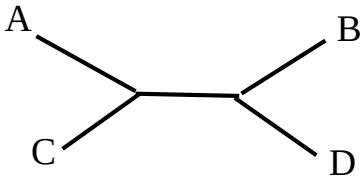
Si trois OTU : un seul arbre non raciné et trois racinés

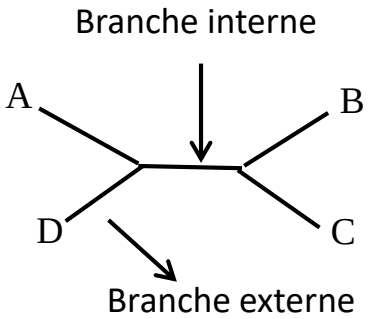


Si quatre OTU : quatre arbres non enracinés dont trois résolus









4 branches = 4 racines possibles

5 branches : 5 racines possibles

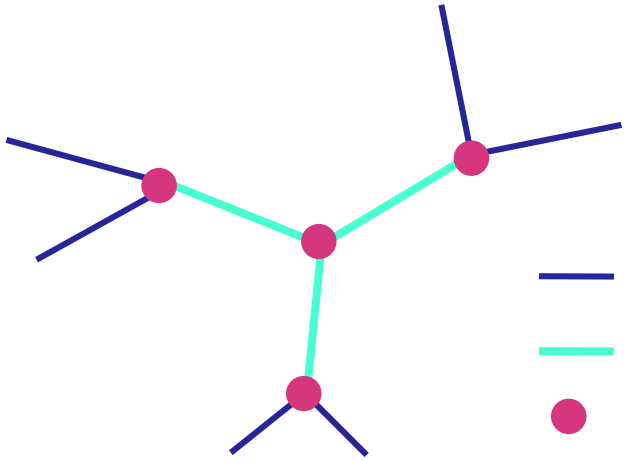
Total : 19 arbres enracinés possibles

Trouver l'arbre

Si à la place de vouloir placer une racine dans l'exemple précédent, on voulait ajouter une 5ème OTU, on aurait également 19 possibilités (sur chacune des branches) donc 19 arbres possibles.

Le calcul du nombre d'arbres non enracinés possibles présentant 3 segments par nœuds internes repose sur le raisonnement récursif suivant (Edwards et Cavalli-Sforza, 1964) :
Un arbre composé de n OTU possède :

- *n* branches externes (une pour chaque feuille de l'arbre)
- *n-3* branches internes
- *n-2* nœuds internes



Exemple avec 6 OTU

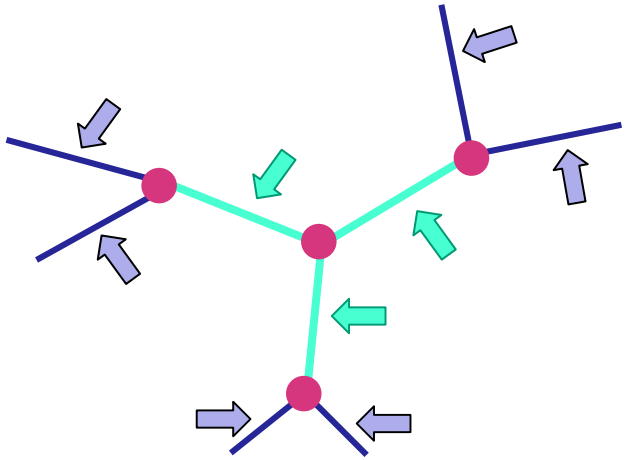
Branche externe : *n* = 6

Branche interne : 3 -> *n* - 3

Noeud interne : 4 -> *n* - 2

Trouver l'arbre

Si on veut rajouter une nouvelle OTU, on peut soit la positionner sur une branche interne ou une branche externe. On a donc *n+(n-3)*, soit *2n-3* possibilités.



Si T_{n-1} est le nombre d'arbres non enracinés possibles pour (*n-1*) OTU, ce nombre sera pour *n* OTU :

$$T_n = T_{n-1} \times (2(n-1)-3) = T_{n-1} \times (2n-5)$$

On peut donc écrire :

$$Tn = \prod_{k=3}^{n} (2k-5)$$

Trouver l'arbre

Nombre de topologies d'arbres non racinées binaires pour *n* taxons

$$Tn = \prod_{k=3}^{n} (2k-5)$$

Arbre binaire = d'un ancêtre, seuls deux organismes peuvent diverger

$$N_{arbres} = 3.5.7...(2n-5) = \frac{(2n-5)!}{2^{n-3}(n-3)!}$$

n	N_{arbres}
4	3
5	15
6	105
7	945
...	...
10	2.027.025
...	...
20	~ 2 x10^{20}

Construire un arbre d'évolution de **10 espèces** revient à **réfuter 2.027.024** cas possibles

[redacted]

Trouver l'arbre

Nombre de topologies d'arbres racinées binaires pour *n* taxons s'obtient en suivant le même raisonnement, on a alors :

$$Tn_r = \frac{(2n-3)!}{2^{n-2}(n-2)!}$$

Soit pour *n* = 10 : 34 459 425 arbres racinés possibles.

La recherche de l'arbre vrai par énumération de tous les arbres possibles devient irréalisable pour des grandes valeurs de *n* (> 10).

Donc développement de stratégie efficace pour trouver cet arbre.

Mais comment identifie-t-on l'arbre vrai?

Hypothèse : on recherche l'arbre le plus parcimonieux ou le plus vraisemblable.

Méthodes de reconstruction phylogénétique

Quatre familles principales de méthodes :

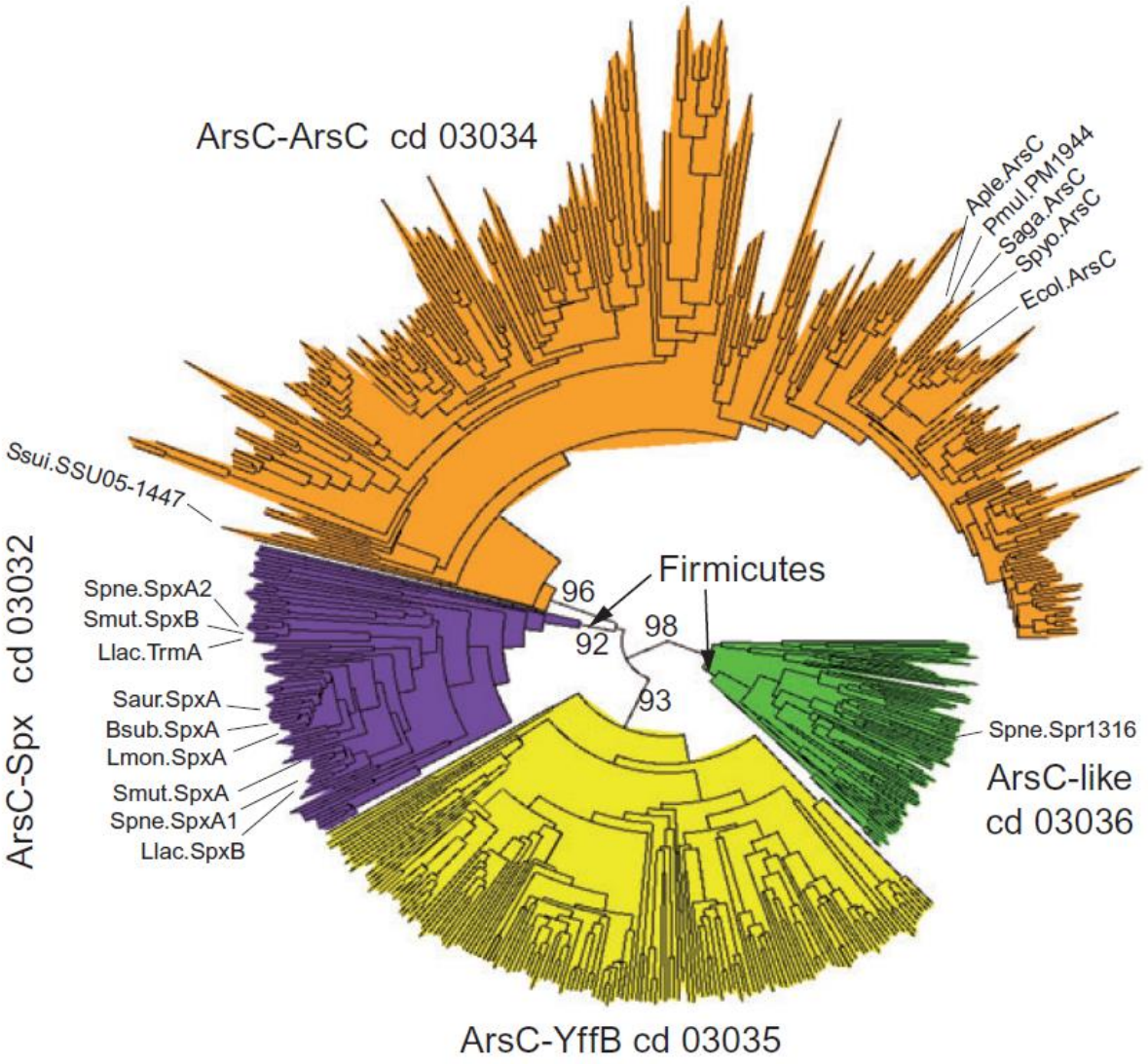
- Parcimonie : à partir d'un ensemble de caractères choisis. Recherche l'arbre qui minimise le nombre de changements permettant d'expliquer les données.

- Méthodes de distance : à partir de distances établies sur un ensemble de caractères recherche l'arbre qui représente au mieux les distances évolutives entre les données.

- Méthodes statistiques : recherche l'arbre le plus vraisemblable en fonction du modèle évolutif considéré :
  - ✓ Méthodes du maximum de vraisemblance : à partir des probabilités de l'apparition des transformation d'un état de caractères en un autre.
  - ✓ Approche bayésienne

Exemple d ’analyse d’une famille multigénique

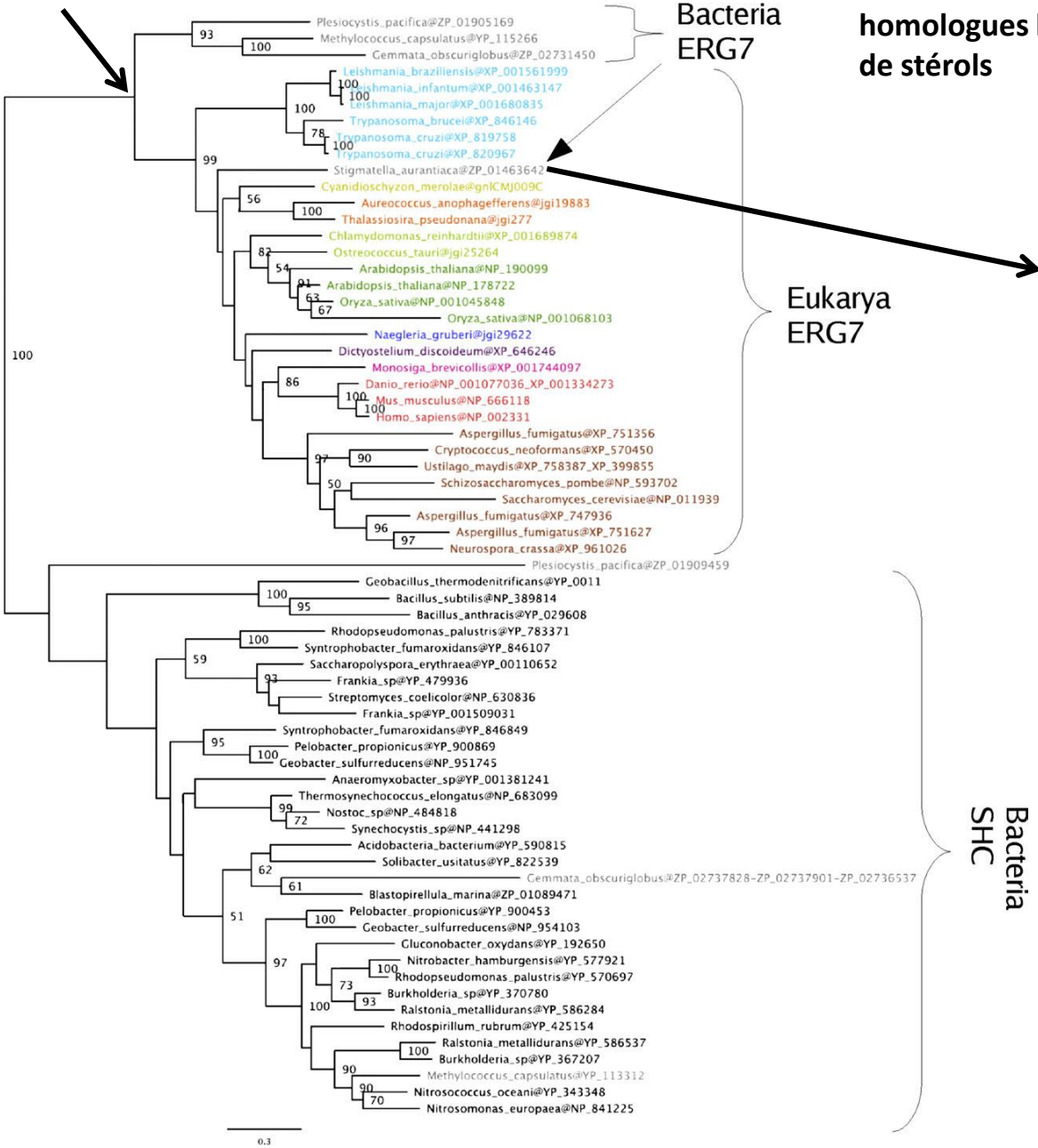
Protéines présentant une similarité avec le domaine COG1393 (arsenate reductase and related proteins)



(Extrait de Turlan *et al.,* 2009, Molecular Biology 73:492-506)

Ancêtre commun aux séquences ERG7

Arbre obtenu sur les protéines ERG7 et SHC, protéines homologues bactériennes non impliquées dans la synthèse de stérols



Localisation de la séquence bactérienne de *Stigmatella aurantiaca* avec les séquences eucaryotes : une indication de l'acquisition de cette séquence par la bactérie au travers d'un transfert horizontal d'une séquence provenant d'un génome eucaryote.

Identification de la séquence ERG7 dans seulement 4 génomes bactériens : en faveur de l'acquisition de cette séquence par ces génomes via un transfert horizontal dont la source serait un génome eucaryote.
Hypothèse alternative : la séquence du gène ERG7 était présente dans l'ensemble des génomes procaryotes, au moins ceux possédant SHC et qu'ensuite elle ait été perdue par la majorité de ces génomes excepté les quatre génomes en question.
Conséquence : un grand nombre de pertes indépendantes
Hypothèse la plus parcimonieuse : acquisition par HGT. De plus, la position de la séquence de *S. aurantiaca* indique clairement l'acquisition horizontale du gène.

(Extrait de Desmond and Gribaldo, 2009, Genome Biol. Evol., 364-81)

Exemple : Evolution des protéines RNase J

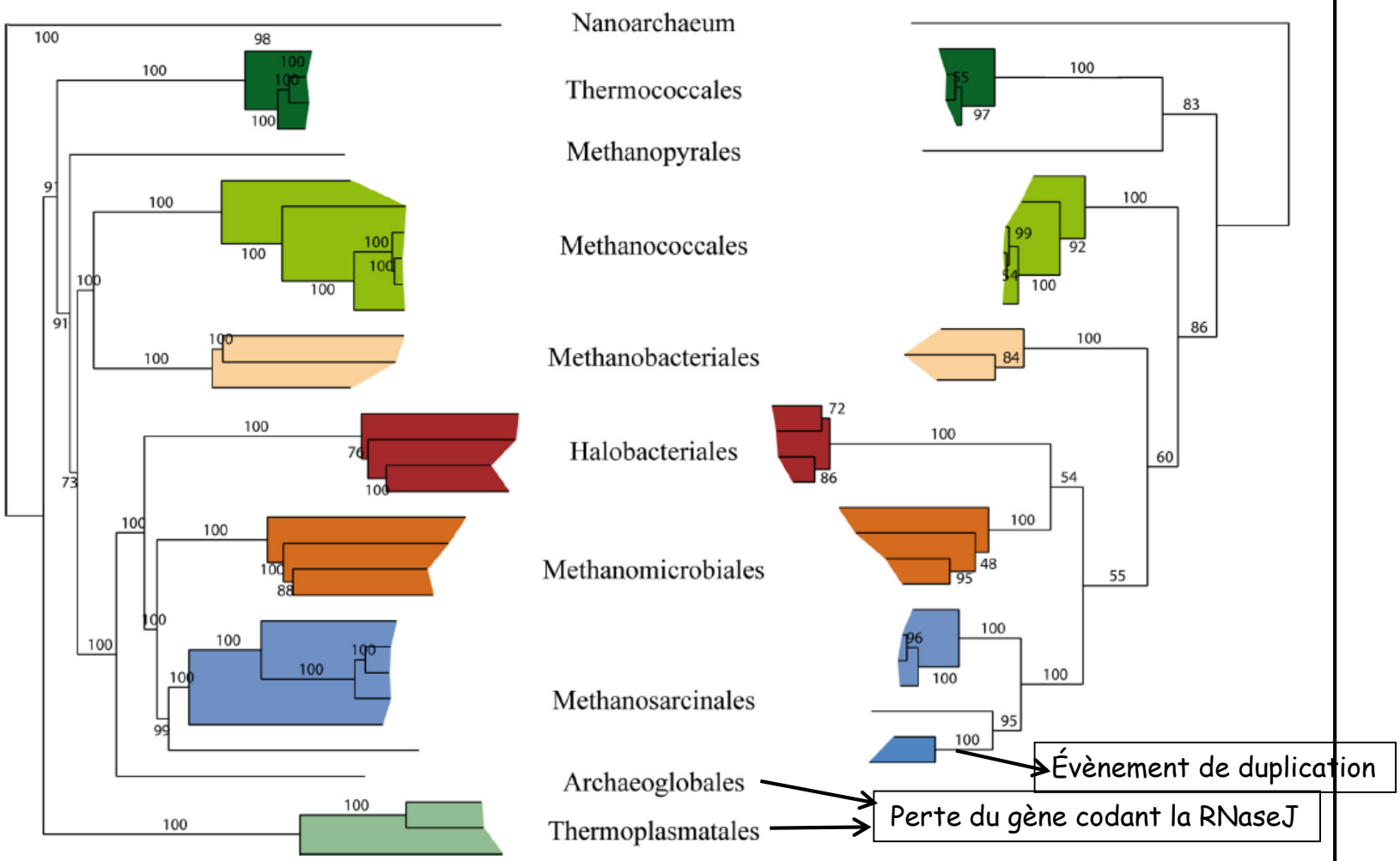


FIGURE 7. **Phylogenetic trees of the Euryarchaeota and euryarchaeal RNase J.** The tree of the Euryarchaeota was constructed from a concatenated sequence of 70 ubiquitous proteins (see "Experimental Procedures"). The euryarchaeal RNase J tree was constructed from sequences that were available at the time of the analysis (supplemental Fig. S1). The trees are arbitrarily rooted. Bootstrap values are indicated at the branch points.

(Extrait de Clouet d'Orval *et al.*, 2010, J. Biol. Chem. 285:17574-583)