Introduction

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

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

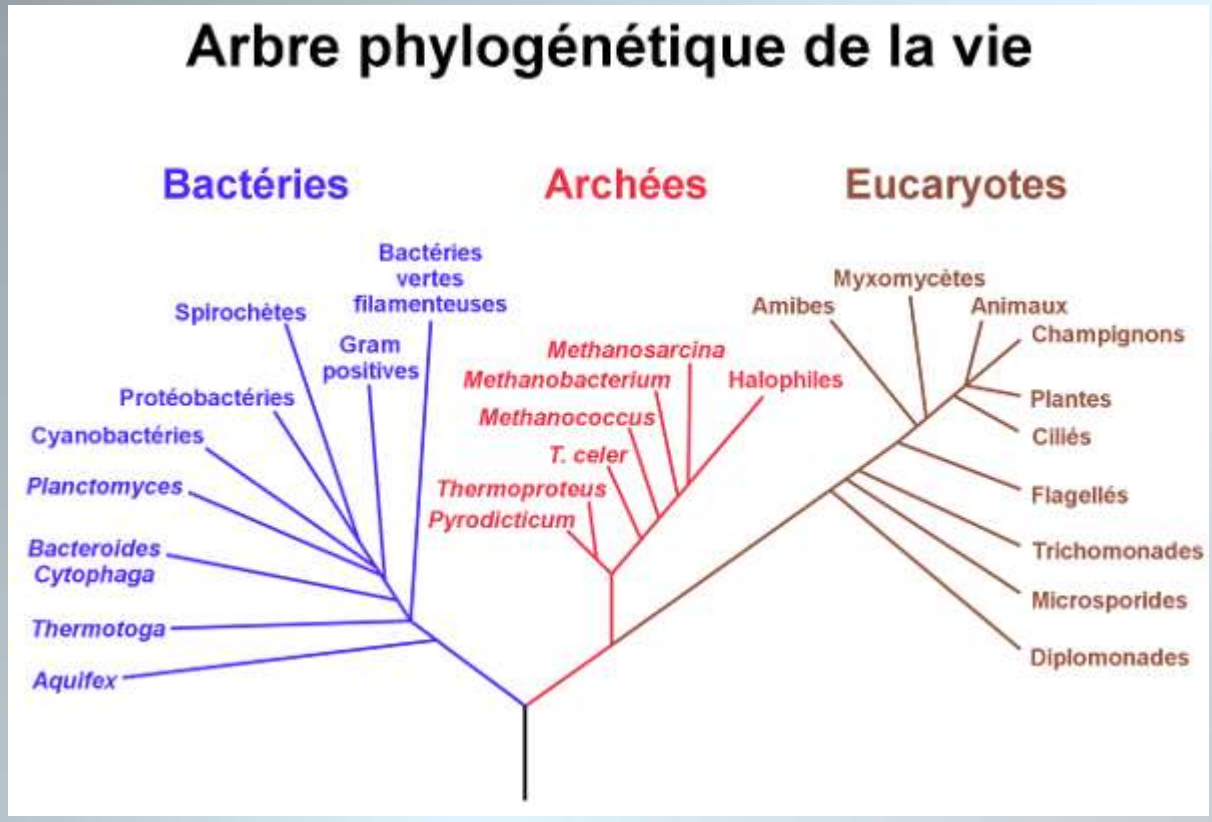
Premières analyses faites en 1965 par E. Zuckerkandl et L. Pauling montrant que la phylogénie des vertébrés était a peu près identique quand elle était basée sur la comparaison de séquences protéiques ou sur des données morphologiques, anatomiques et paléontologiques.

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

Découverte du troisième domaine du vivant par Carl Woese en 1977 par l'analyse phylogénétique des séquences d'ARNr 16S.



Extrait de L'évolution du vivant expliquée à ma boulangère (2009) Virginie Nepoux (http://www.ilv-bibliotheca.net/librairie/levolution_du_vivant_expliquee_a_ma_boulangere.html).

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 structures (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 dirons que deux gènes sont **homologues** s'ils ont divergé à partir d'une séquence ancêtre commune.

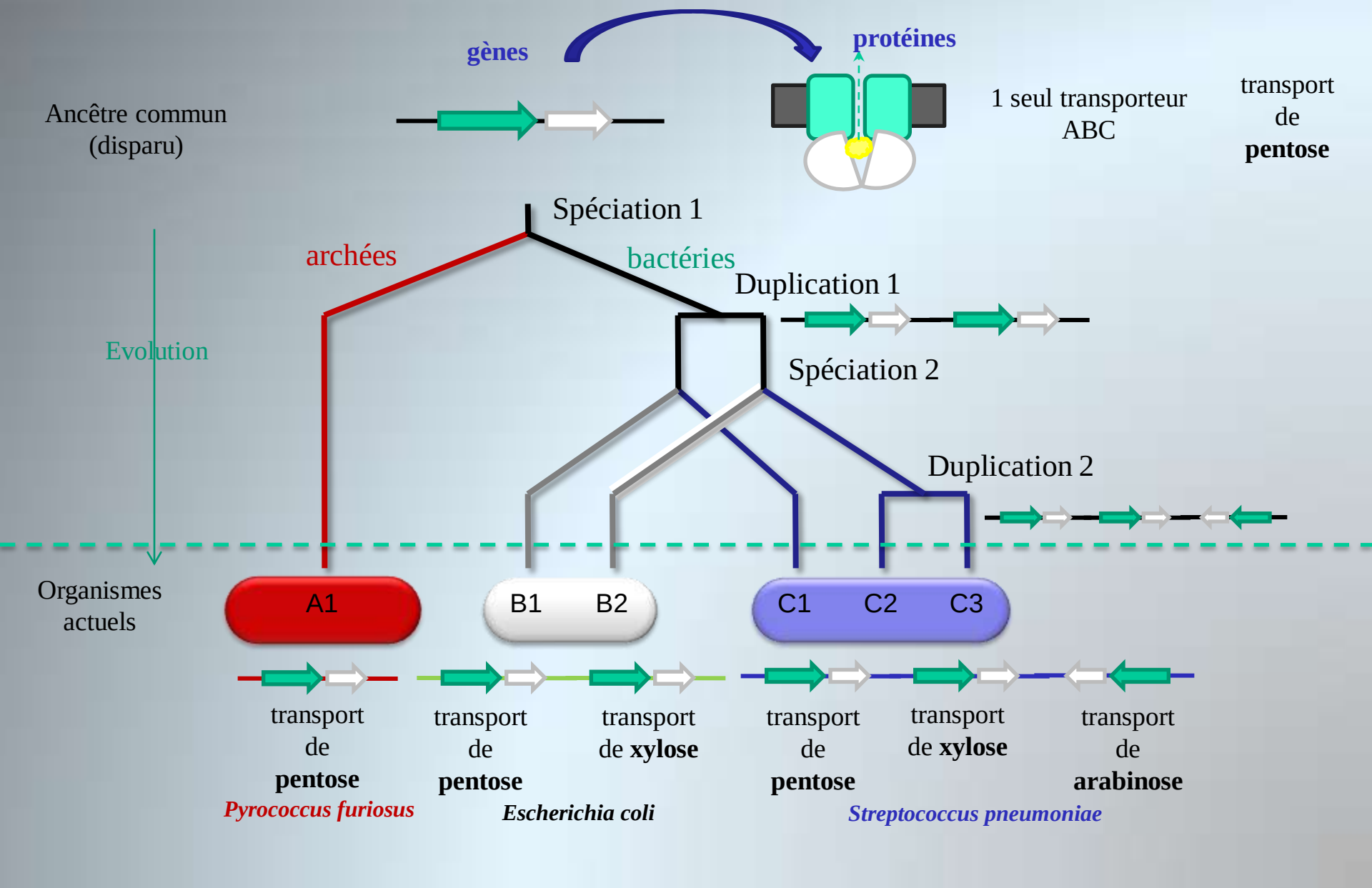
Définition insuffisante pour reconstruction de l'histoire évolutive car plusieurs mécanismes possibles pour dériver d'une séquence ancêtre.

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

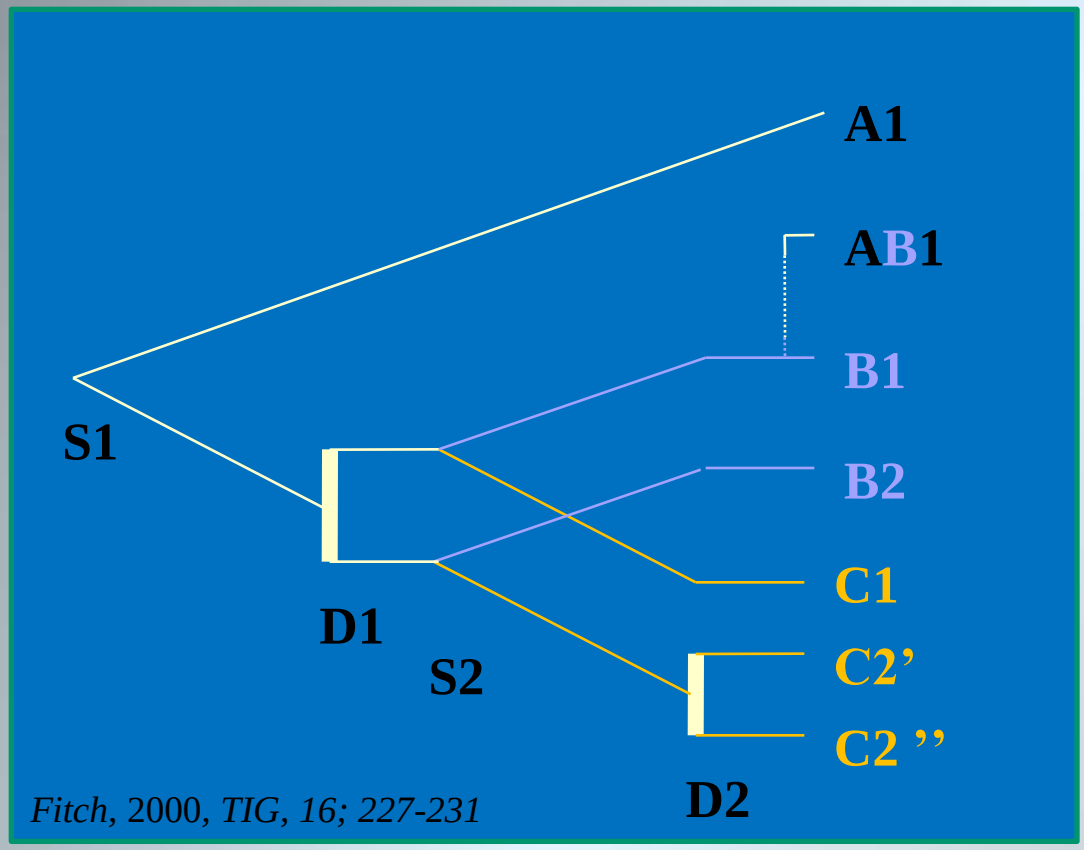
Paralogie : deux gènes sont **paralogues** si leur divergence est due à la duplication du gène ancêtre.

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

Illustration

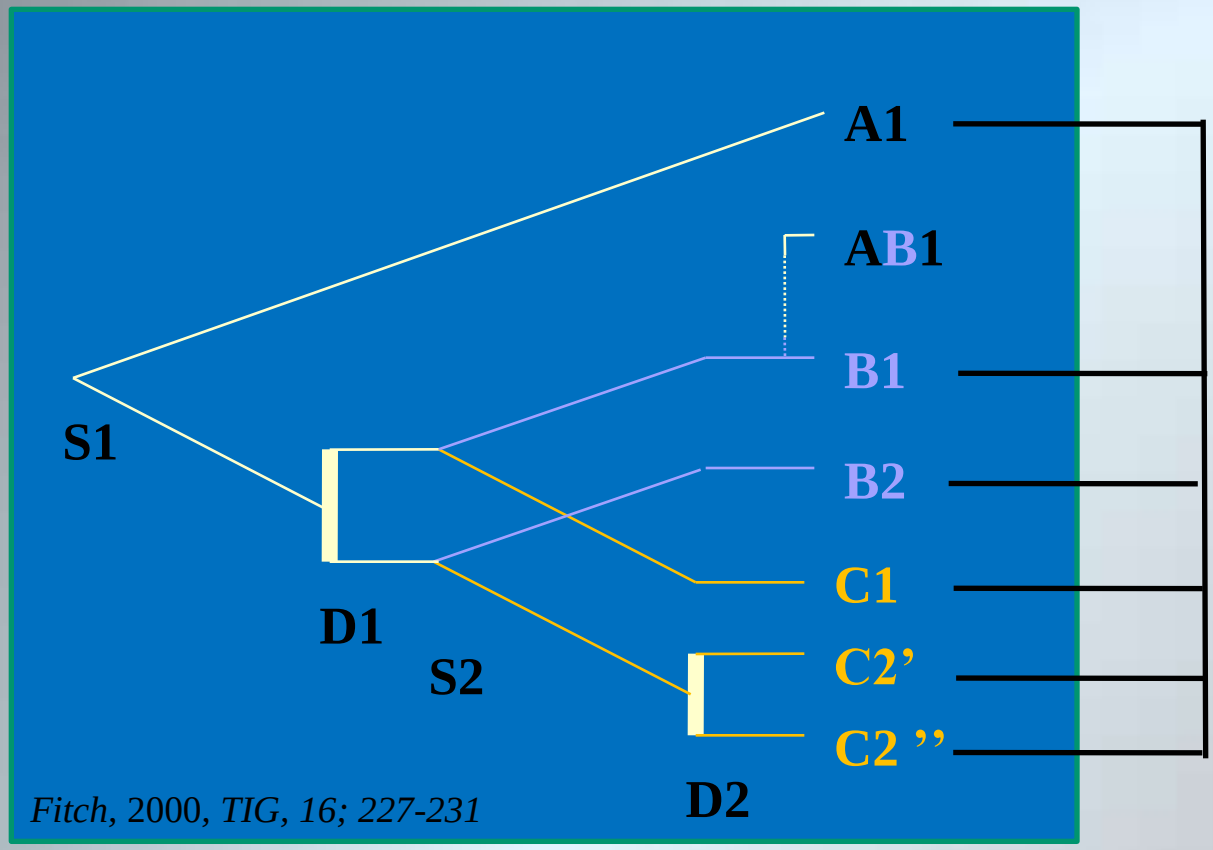


Notions de base, définitions



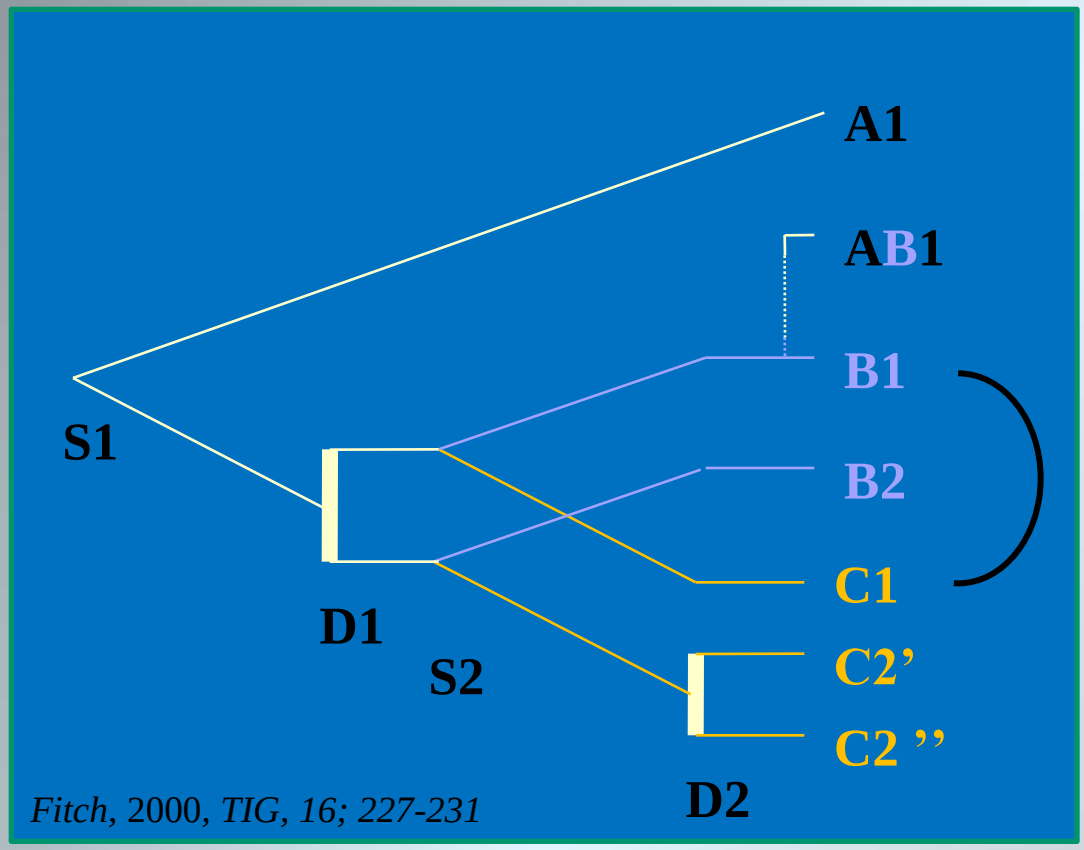
Fitch, 2000, *TIG*, *16; 227-231*

Notions de base, définitions



Orthologues à A1

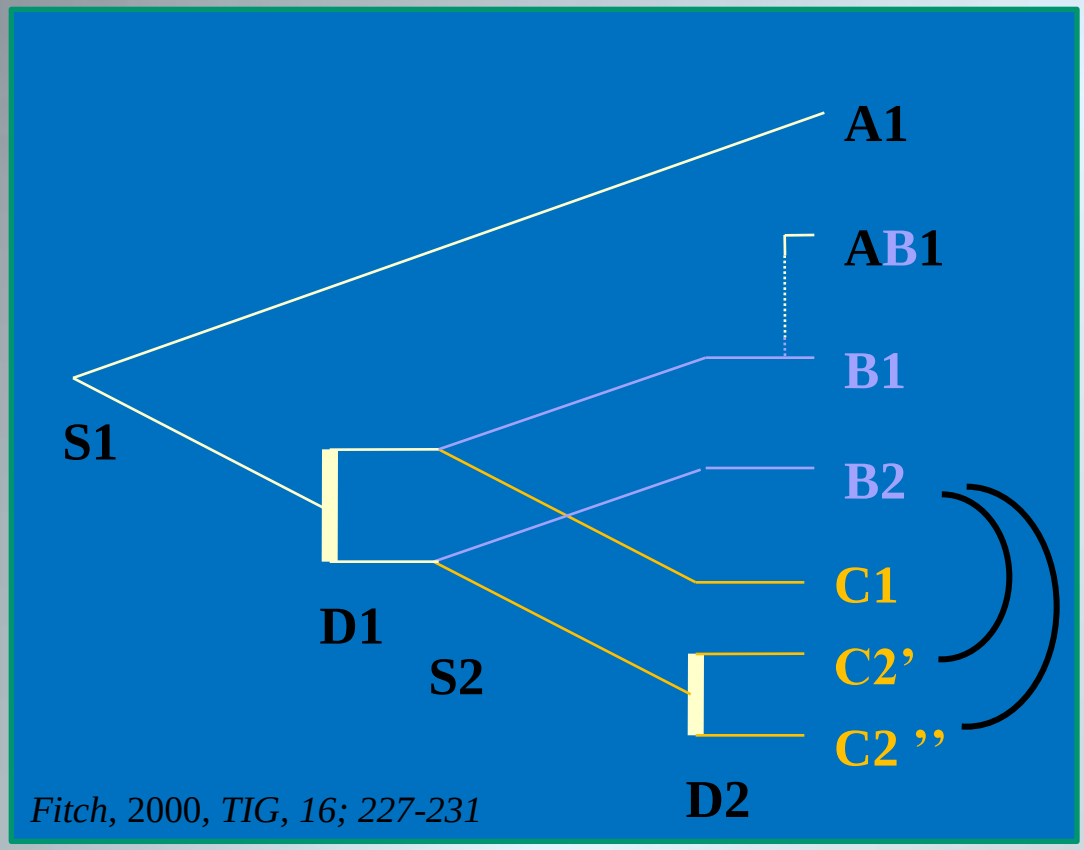
Notions de base, définitions



Fitch, 2000, *TIG*, *16; 227-231*

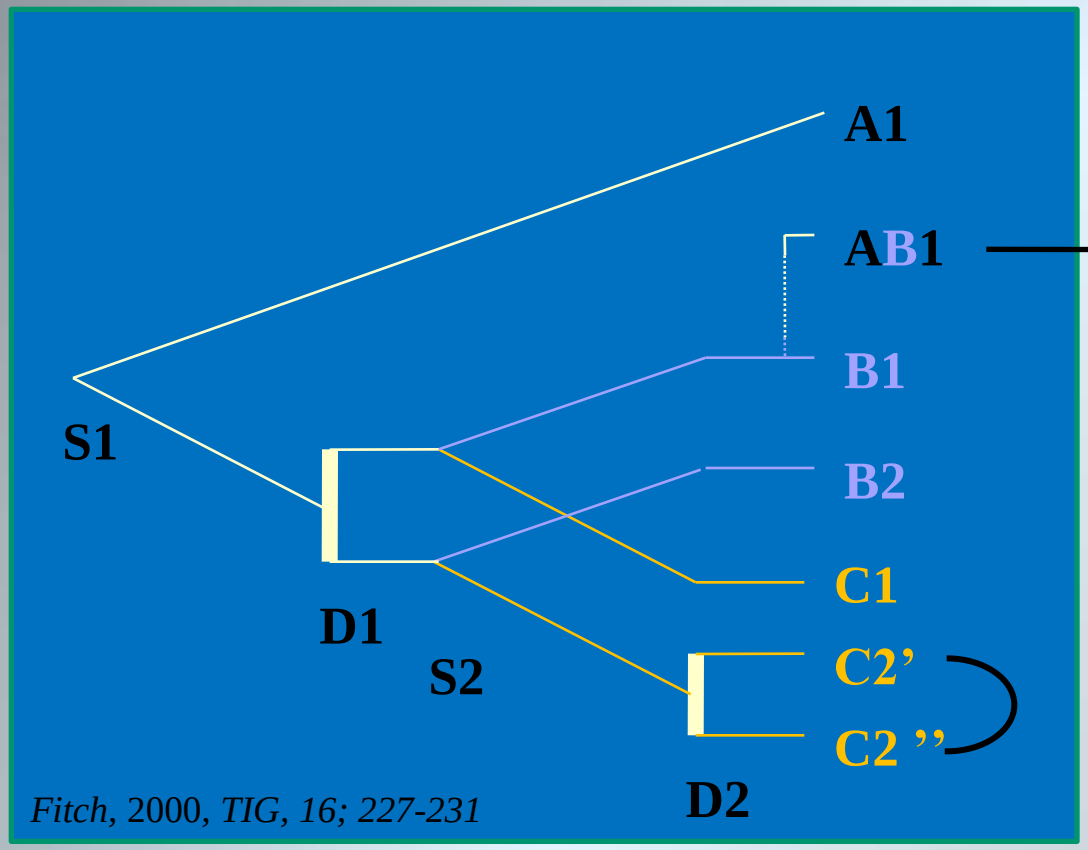
Orthologues

Notions de base, définitions



Orthologues

Notions de base, définitions



Fitch, 2000, *TIG*, *16; 227-231*

transfert horizontal

paralogues

Notions de base, définitions

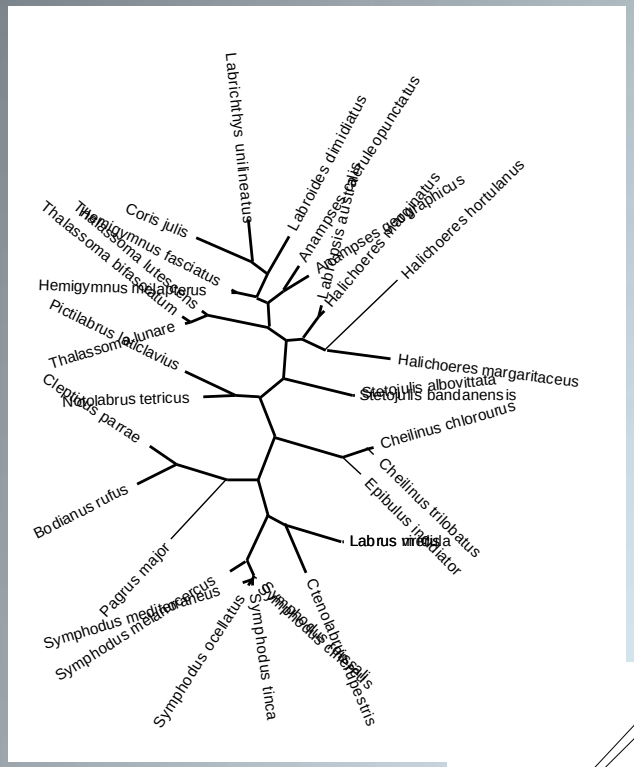
Pourquoi la comparaison de séquences :

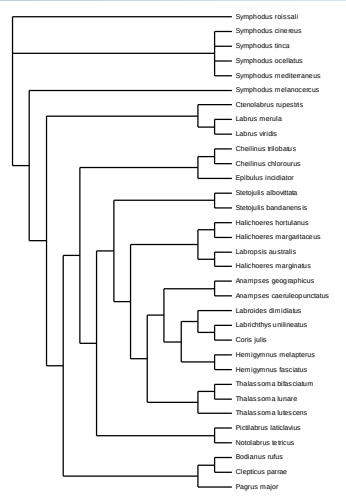
Hypothèse : si deux ou plusieurs séquences possèdent des résidus conservés (bases ou acides aminés), cela signifie qu'elles ont une histoire évolutive commune. Elles ont évolué à partir d'une séquence ancêtre commune.

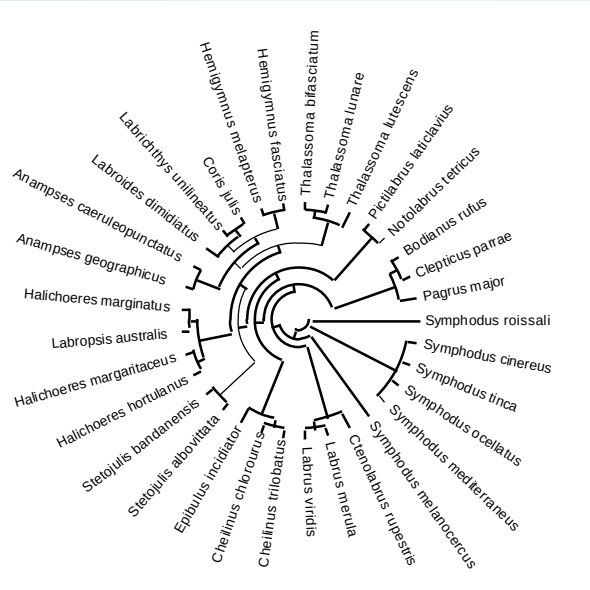
Le pourcentage de similarité entre deux séquences est considéré comme reflétant la distance évolutive existant entre ces deux séquences. Les différences observées sont dues à l'accumulation de mutations au cours du temps. Les mutations prises en compte sont les substitutions et les insertions/délétions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

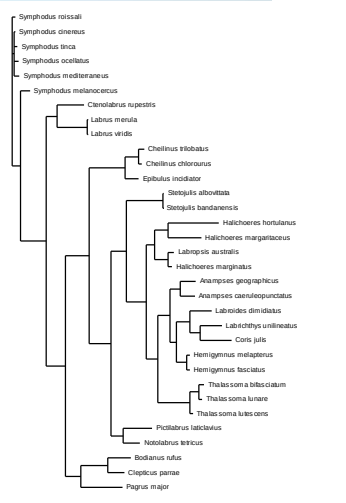
→ Alignement multiple

Notions de base : arbres phylogénétiques



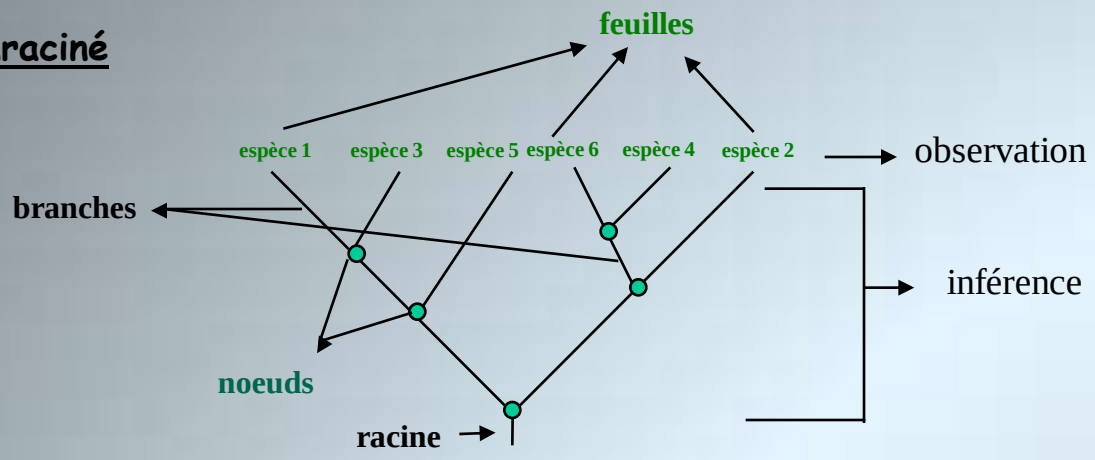






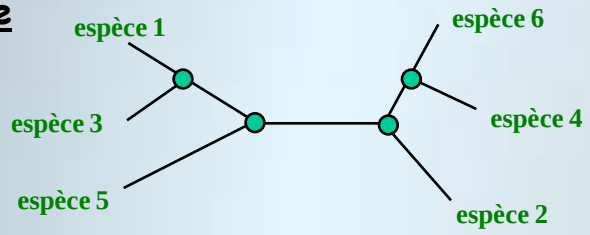
Notions de base : arbres phylogénétiques

Arbre enraciné



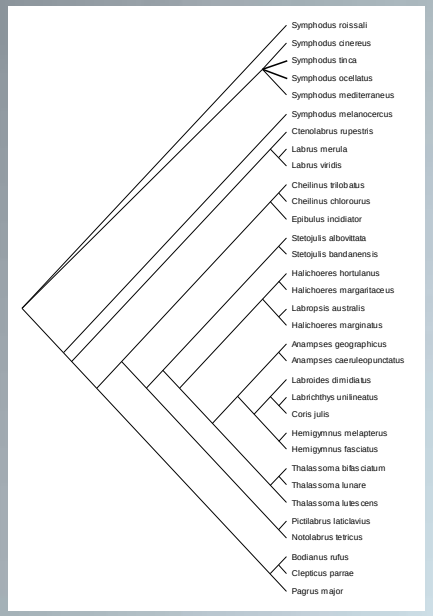
- Les sommets externes sont appelés **feuilles**. C'est la seule partie basée sur l'observation.
- Les sommets internes sont appelés **nœuds**. Ils représentent l'ancêtre commun hypothétique dans le sens où leur existence n'est pas fondée sur l'observation mais sur le processus de reconstruction.
- La 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** définit l'origine commune des espèces traitées. Les liens entre nœuds et feuilles sont orientés, on part de la racine et on remonte aux feuilles.

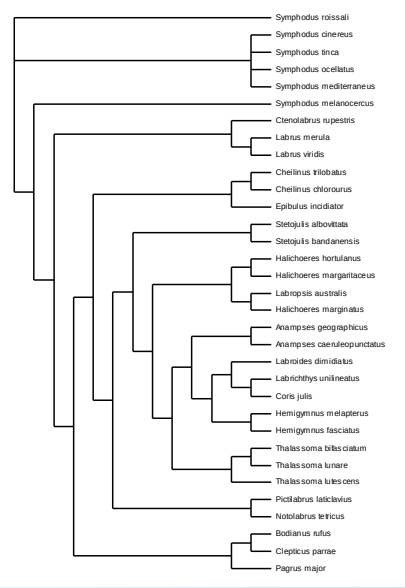
Arbre sans racine



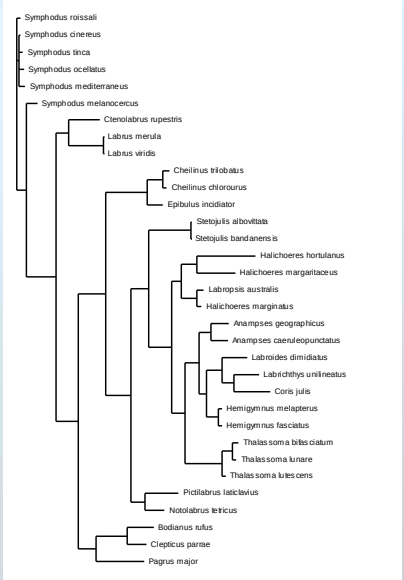
Dans un arbre sans racine, les liens entre nœuds ne sont pas orientés et un seul et unique chemin permet de passer d'un sommet à l'autre.

Notions de base : arbres phylogénétiques





Arbre ultramétrique



Arbre additif

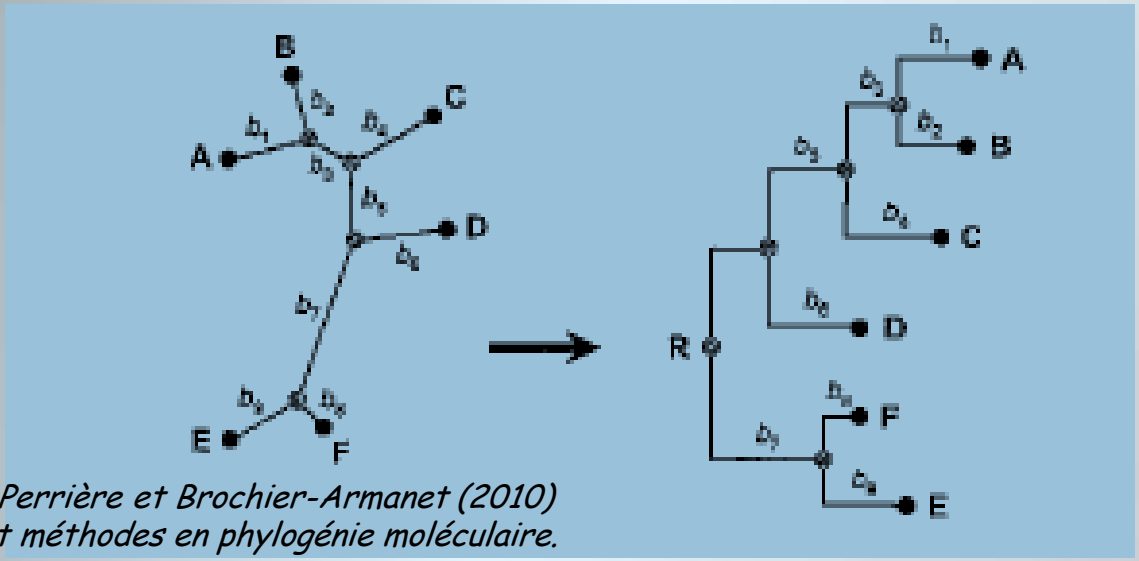
Cladogramme : pas de longueur de branche (feuilles sous un même nœud appelée clade)

Phylogramme : longueurs de branches

Notions de base : arbres racinés et non racinés

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

- enraciner un arbre :
  - Racinemenent au barycentre : ne nécessite pas de connaissances *a 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 à toutes les feuilles. Fait l'hypothèse de <u>l'horloge moléculaire</u> : 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.

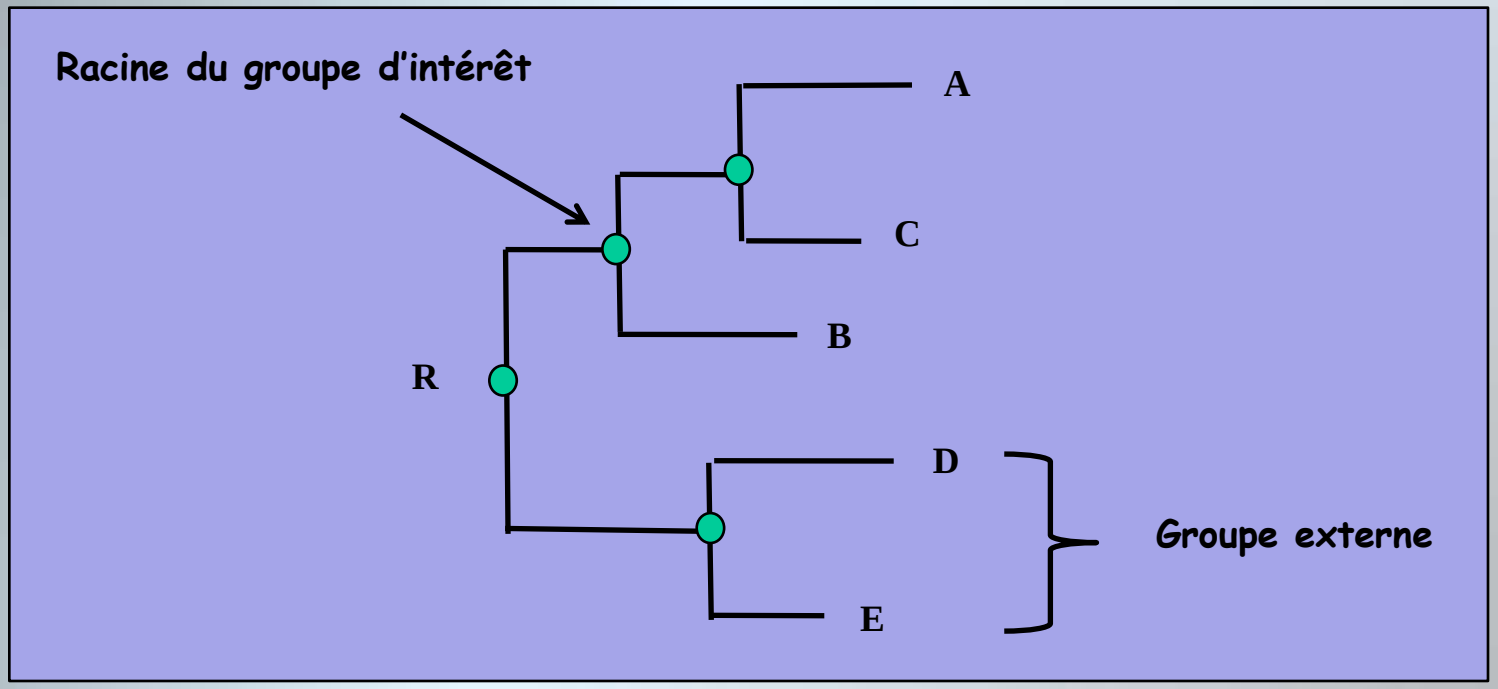


Extrait de Perrière et Brochier-Armanet (2010) Concepts et méthodes en phylogénie moléculaire.

Notions de base : arbres racinés et non racinés

- enraciner un arbre :
  - La méthode du <u>groupe externe</u> : 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.



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 :

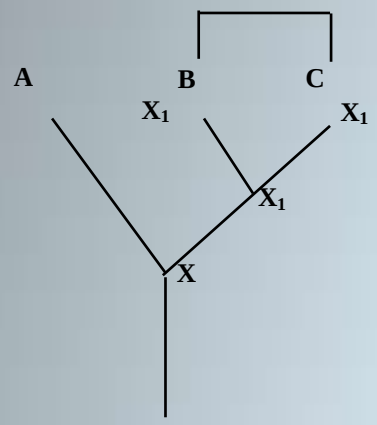
- les phénéticiens adeptes de la « taxonomie numérique ». Les liens entre les taxons ne peuvent être fondés que sur la base d'une similitude globale exprimée à partir de matrices de calcul de distances. Dans le cas des séquences, à partir d'un alignement multiple, on calculera les distances entre les 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 similitude

Il peut être divisé en :

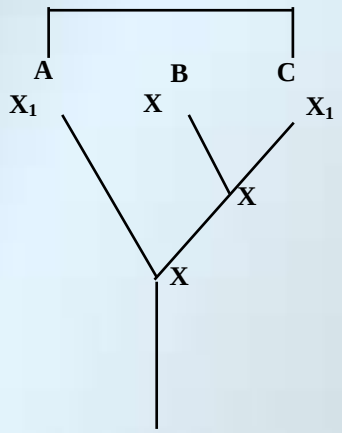
- homologie = similitude héritée d'un ancêtre commun
- homoplasie = similitude non héritée d'un ancêtre commun et qui est subdivisée en :
  - convergence : apparition indépendante dans deux espèces d'un même état dérivé de caractère
  - réversion : apparition d'un état de caractère ayant la forme ancestrale

homologie



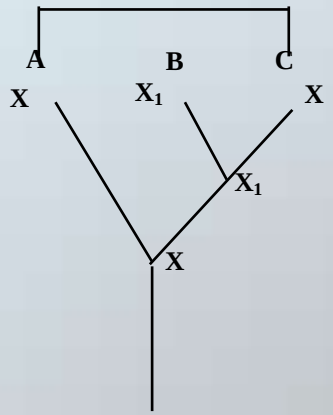
L'état de caractère X_1 présent chez B et C est hérité de l'ancêtre de B et C.

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.

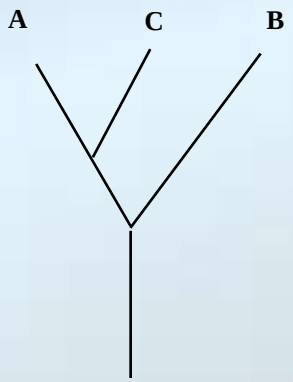
réversion



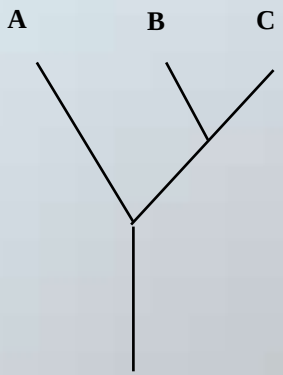
L'état de caractère X présent chez A et C n'est pas hérité d'un ancêtre commun. Chez A, il correspond à l'état ancestral et chez C il est secondairement transformé (retour à la forme ancestrale).

Le concept de similitude

Arbre qui sera déduit à partir des valeurs des états de caractères :



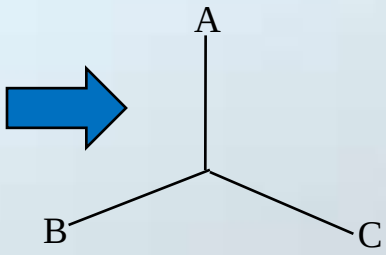
Or dans les deux cas, convergence ou réversion, le vrai arbre est :



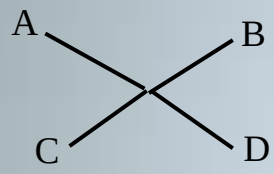
Trouver l'arbre

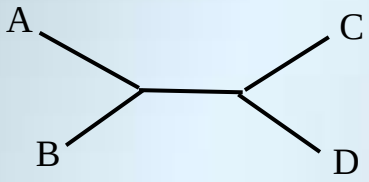
<u>Problème</u> : un seul arbre vrai, l'arbre évolutif
Comment le distinguer dans tous les arbres possibles

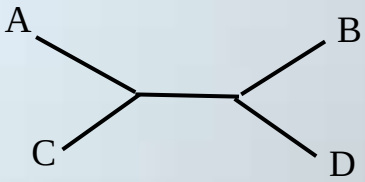
Si trois OTU : un seul arbre non raciné et trois racinés

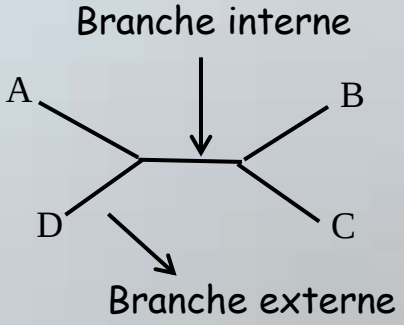


Si quatre OTU : quatre arbres non enracinés dont trois résolus









4 branches = 4 racines possibles

5 branches : 5 racines possibles

Total : 19 arbres enracinés possibles

Trouver l'arbre

Nombre de topologies d'arbres non racinés 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

Méthodes de reconstruction phylogénétique

Quatre familles principales de méthodes :

- Parcimonie : à partir d'un ensemble de caractères choisis.
- Méthodes de distance : à partir de distances établies sur un ensemble de caractères.
- Méthodes du maximum de vraisemblance : à partir des probabilités de l'apparition des transformations d'un état de caractères en un autre.
- Approche bayésienne

Principe général d'une méthode de distance

Alignement de séquences

```
CAAACAGCGTT---GGCTCTCTA
AAAATAACACCaacATGCAAATG
AAAACAGCACCaacGTGCAAATG
AAAACAGCACCaacGTGCAAATG
```

Mesure des distances évolutives

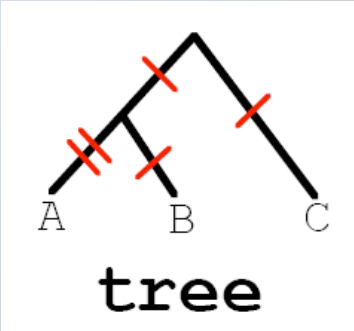
Matrice des distances évolutives entre paires de séquences

	A	B	C
A	0		
B	3	0	
C	4	3	0

distance matrix

Calcul d'un arbre à partir des distances

Arbre (non enraciné)



Calcul des distances entre deux séquences

La distance *d* séparant deux séquences est définie comme le **nombre moyen de substitutions par site** qui s'est produit depuis que ces deux séquences ont divergé de leur ancêtre commun.

Matrice de distance dans le cas de 5 séquences. Elle est symétrique par rapport à la diagonale et la distance d'une séquence à elle-même est nulle.

	1	2	3	4	5
1	0				
2	d_{12}	0			
3	d_{13}	d_{23}	0		
4	d_{14}	d_{24}	d_{34}	0	
5	d_{15}	d_{25}	d_{35}	d_{45}	0

Calcul des distances entre deux séquences

<u>Divergence observée ou *p*-distance</u> : la plus simple

On compte le nombre *s* de substitutions observées entre deux séquences alignées que l'on rapporte au nombre de sites homologues *n* alignés, donc :

$$p = \frac{s}{n}$$

Proportional (*p*) Distance

DNA Site

Species	1	2	3	4	5	6	7	8	9	10
I	A	T	A	T	A	C	G	T	A	T
II	A	T	G	T	A	C	G	T	A	T
III	G	T	A	-	A	C	G	T	G	C
IV	G	C	G	T	A	T	G	C	A	C

$$p = \frac{\text{\# differences}}{\text{\# sites}}$$

	I	II	III	IV
I	-	0.1	0.4	0.6
II		-	0.5	0.5
III			-	0.6
IV				-

facile à calculer mais quand les séquences ne sont pas proches (issues d'organismes distants dans l'évolution), elle sous-estime les distances évolutives.

Cause : l'existence de substitutions multiples. Phénomène plus critique pour les séquences d'acides nucléiques car possèdent un alphabet plus pauvre que les séquences protéiques : quatre lettres au lieu de 20.

Calcul des distances entre deux séquences nucléiques

Substitutions multiples

Séquence1 **GAAAAG**
Séquence2 **ATGAAG**

Type de substitution	Séquence 1	Séquence 2	Nombre de substitutions observé	Nombre de substitutions réel
Substitution unique (simple)	G	G ➤ A	1	1
Substitutions multiples	A	A ➤ C ➤ T	1	2
Substitutions coïncidentes au même site	T ➤ A	T ➤ G	1	2
Substitutions parallèles	T ➤ A	T ➤ A	0	1
Substitutions convergentes	C ➤ G ➤ A	C ➤ A	0	3
Substitution réverse (inverse)	G ➤T ➤ G	G	0	2

Calcul des distances entre deux séquences nucléiques

Pour tenter de corriger le biais du aux mutations multiples, des hypothèses sont faites sur la façon dont les bases se sont substituées à un locus donné

- Construction d'un modèle évolutif

Il existe plusieurs distances pour rendre compte de l'évolution des séquences

La plus simple : distance de Jukes et Cantor mais vision simplificatrice de l'évolution des séquences :

- tous les sites évoluent indépendamment et selon le même processus
- toutes les substitutions sont équiprobables

On utilise la loi de Poisson pour prendre en compte les changements multiples (donc des reversions)

$$d = -\frac{3}{4}\log(1-\frac{4}{3}p)$$

Calcul des distances entre deux séquences nucléiques

Autre distance : distance de Kimura à 2 paramètres, vision un peu moins simplificatrice :

- tous les sites évoluent indépendamment et selon le même processus
- les substitutions se produisent suivant deux taux distincts, l'un pour les transitions, l'autre pour les transversions, les transitions étant plus fréquentes (transition = A<->G ou T<->C)

$$d = -\frac{1}{2}\log[(1-2p-q)\sqrt{1-2q}]$$

Avec :

p fréquence observée des transitions
q fréquence observée des transversions

Pour les autres distances : Inferring Phylogenies,J. Felsenstein, Sinauer Associates Inc. Publishers Sunderland, Massachutsetts
Perrière et Brochier-Armanet (2010) Concepts et méthodes en phylogénie moléculaire.

Calcul des distances entre deux séquences protéiques

- Séquences protéiques fréquemment utilisées en phylogénie moléculaire car plus appropriées quand les analyses comportent des séquences issues de lignées séparées par de grandes distances évolutives ou quand les séquences évoluent rapidement (au niveau ADN perte du signal phylogénétique car les sites sont dits saturés, *i.e.*, ont subi de nombreuses substitutions multiples).

- Egalement plusieurs modèles pour estimer la distance entre deux séquences

<u>Modèle de Poisson</u>

- Première estimation meilleure que la p-distance repose sur le concept de distribution de Poisson.
- Hypothèses :
  - tous les sites évoluent indépendamment et suivant un même processus
  - toutes les substitutions sont équiprobables
  - le taux de réversion est négligeable

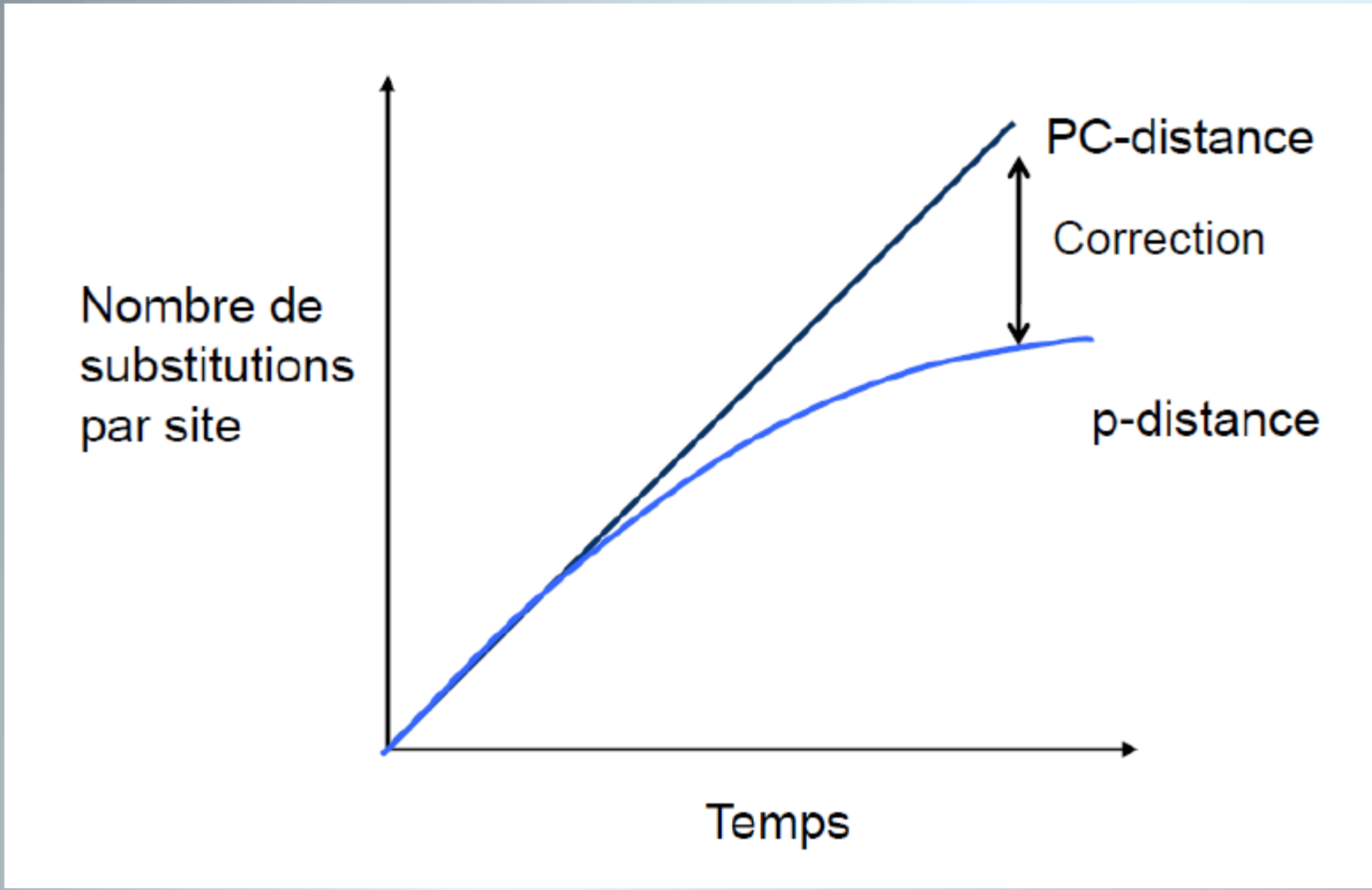
$$\boxed{d = -Log(1-p)}$$

correction de Poisson

Valeur de *p* estimée par la *p*-distance

Calcul des distances entre deux séquences protéiques

Relation entre la *p*-distance et la distance corrigée de Poisson



Calcul des distances entre deux séquences protéiques

Modèle de Poisson

- Cependant vision très simplificatrice car en particulier :
  - taux de substitutions plus ou moins élevé en fonction de l'importance fonctionnelle du site
  - présence aussi de substitution parallèle et de réversion donc on va sous-estimer la distance entre deux séquences
  - ne peut être utilisée que si séquences globalement peu divergentes

Donc autres modèles ont été développés.

- Le modèle PAM
- Le modèle JTT
- Le modèle WAG
- Le modèle LG

• modèles les plus performants étant ceux bâtis sur le plus grand nombre de données (car estimation des taux de substitutions pour tous les modèles).
• modèles WAG et LG supérieurs aux modèles PAM et JTT.
• si distances évolutives faibles, on peut utiliser le modèle de Poisson ou de Kimura car aussi bons résultats. Donc si même résultat avec deux modèles, utiliser le plus simple car plus rapide.

Choix d'un modèle évolutif

Des méthodes permettant de tester l'adéquation du modèle aux données existent mais souvent le choix du modèle est du fait de l'utilisateur et de ses connaissances.

<u>Quelques règles simples :</u>

- construction d'une phylogénie à partir de :
  - séquences très distantes dans l'évolution : utilisation des séquences protéiques.
  - séquences proches dans l'évolution : utilisation des séquences acides nucléiques voir travailler uniquement sur les positions synonymes.

- Utilisation de séquences nucléiques : grand nombre de modèles
  - critère important : le degré de divergence entre les séquences.
  - pas toujours pertinent d'utiliser les modèles avec beaucoup de paramètres :
    - ❖ si les séquences sont courtes ou trop similaires les estimations des paramètres sont mauvaises.
    - ❖ modèle arrivant à saturation plus rapidement donc si séquences très divergentes, fréquemment impossible de calculer les distances.
  - donc si même résultat avec deux modèles, utiliser le plus simple

Choix d'une méthode de reconstruction d'arbre

Méthode de parcimonie : cherche à minimiser le nombre de changements. Plus adéquate pour des données de présence/absence. Pas conseillée pour la reconstruction phylogénétique à partir des séquences

- On peut avoir plusieurs topologies s'expliquant par le même nombre de changements : famille d'arbres.

- Dépend de l'ordre dans lequel sont ajoutées les séquences pour la construction de l'arbre. On n'obtiendra pas forcément le même arbre si on change l'ordre des séquences. Pour pallier ce problème, heuristique de réarrangement des branches. En répétant plusieurs fois l'opération, on peut trouver l'arbre le plus parcimonieux.

- permet d'inférer l'état des caractères ancêtres.

- critique principale de la parcimonie : méthode non robuste pouvant conduire sous certaines conditions à des résultats erronés (démontré par Felsenstein (1978, Syst. Zool, 27, 401-10)).

Choix d'une méthode de reconstruction d'arbre

Méthode de distance : **la Neighbor Joining (NJ)**

Constitue une approximation du minimum d'évolution (critère d'optimisation).

Principe général du minimum d'évolution : Examine toutes les topologies, calcule la somme de la longueur des branches de chacune d'entre elles et retient celle qui minimise la somme des longueurs des branches (arbre de longueur minimum).

Des variantes avec pondérations ont également été développées comme la BIONJ (Gascuel, 1997). La BIONJ apporte des améliorations évidentes surtout quand les séquences sont fortement divergentes et/ou quand elles présentent des vitesses d'évolution différentes.

Conclusion :

- Méthode performante car bon équilibre entre rapidité et efficacité. Elle peut être appliquée sur des très grands jeux de données. Robuste car ne dépend pas de l'ordre des séquences.
- ne fait pas l'hypothèse de l'horloge moléculaire (n'impose pas aux distances estimées d'être ultra-métriques).
- Souvent utilisée pour chercher des arbres qui vont servir de point de départ pour des méthodes plus coûteuses en temps de calcul comme la méthode du maximum de vraisemblance (cf. ce chapitre du cours).
- Elle peut être appliquée sur n'importe quel type de distances évolutives.
- Problème pouvant être rencontré quand deux paires de voisins différentes donnent des arbres minimums de même longueur. Dans ce cas, tirage au hasard d'une solution. Situation pas fréquemment rencontrée.
- Ne donne pas d'informations sur les états de caractères de l'ancêtre commun.

Choix d'une méthode de reconstruction d'arbre

Méthode du maximum de vraisemblance :

On a un grand nombre de scénarios évolutifs possibles. Cependant, certains d'entre eux sont plus susceptibles que d'autres de produire les séquences actuelles.

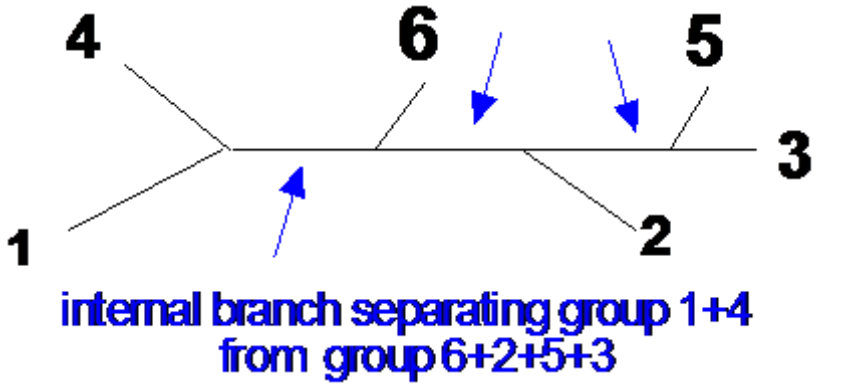
Le but des méthodes de maximum de vraisemblance est d'identifier ces scénarios, c'est-à-dire de trouver les valeurs des paramètres du modèle Θ qui maximisent la probabilité d'observer les séquences actuelles.

- C'est la méthode la mieux justifiée sur le plan théorique.
- Des expériences de simulation de séquences ont montré que cette méthode est supérieure aux autres dans la plupart des cas.
- Mais c'est une méthode très lourde en calculs.
- Il est presque toujours impossible d'évaluer tous les arbres possibles car ils sont trop nombreux. Une exploration partielle de l'ensemble des arbres est réalisée en utilisant des méthodes de réarrangements locales ou globales similaires à celles utilisées en parcimonie.

- Deux méthodes principales : TREE-PUZZLE et PhyML.

Fiabilité des arbres phylogénétiques: le bootstrap

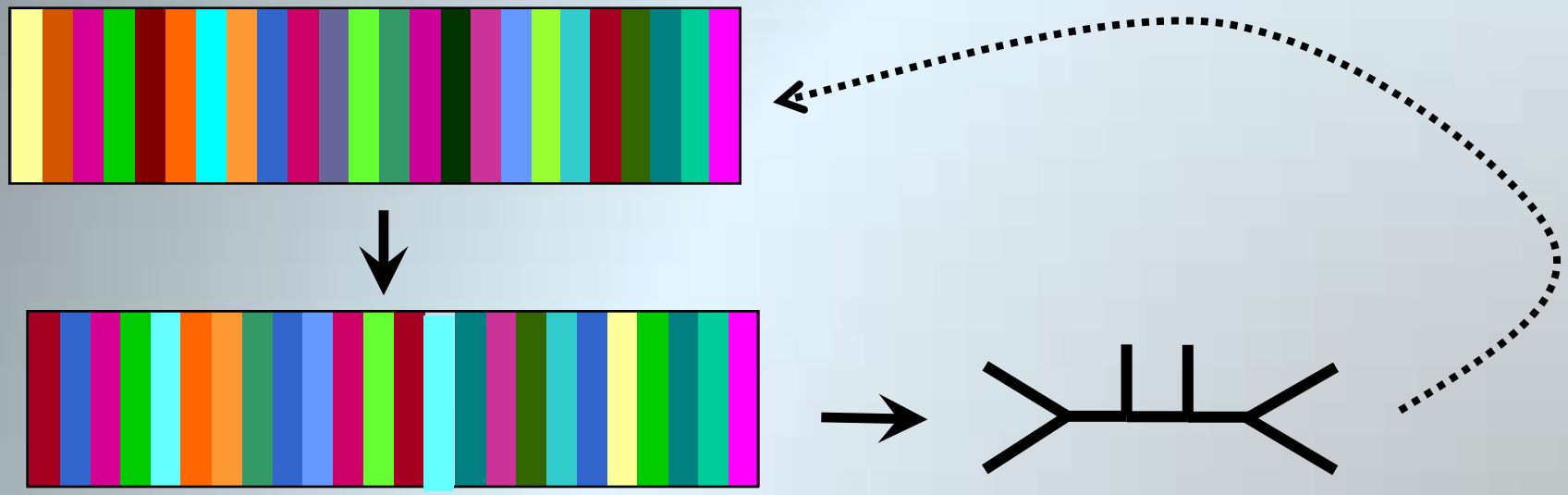
- L'information phylogénétique contenue dans un arbre non raciné réside entièrement dans ses branches internes.



- La forme de l'arbre est déterminée par la liste des branches internes.
- Evaluer la fiabilité d'un arbre = évaluer celle de chaque branche interne.

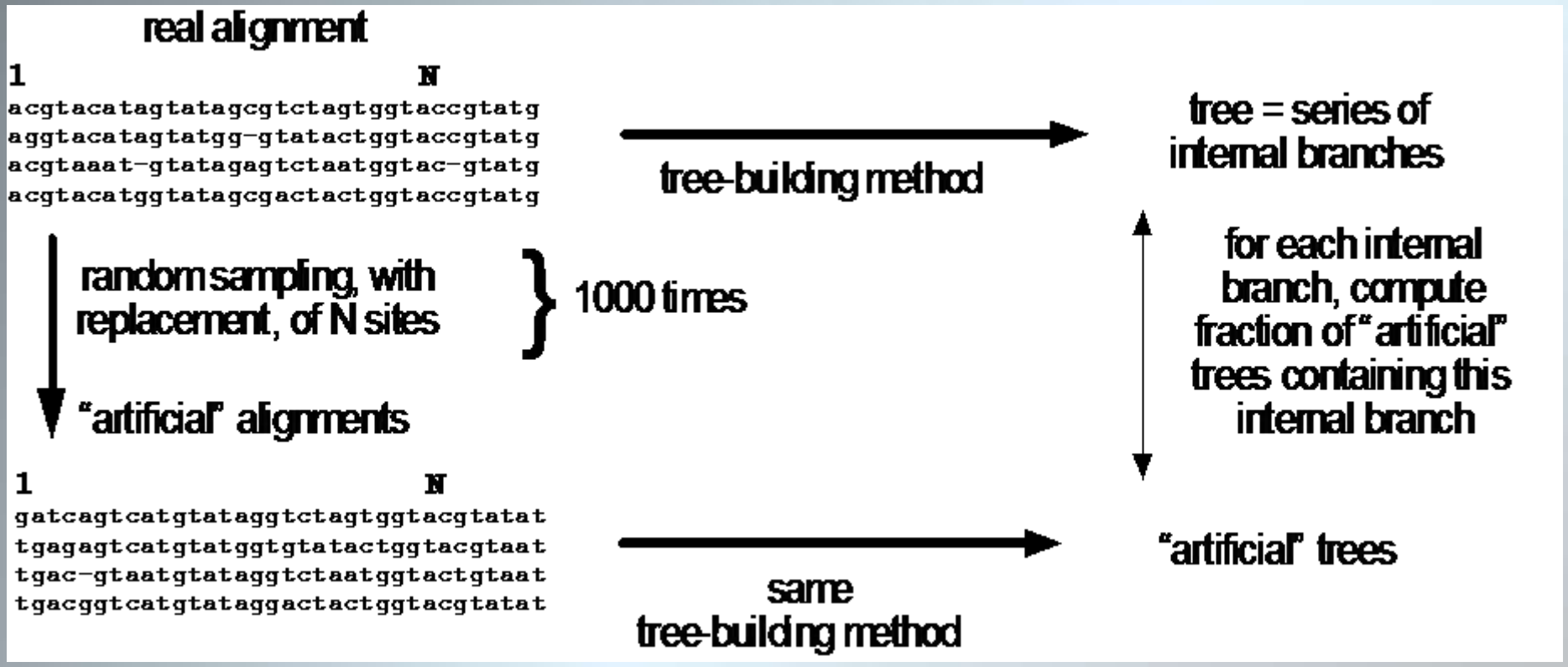
Le bootstrap : procédure

- Tirage avec remise de *n* positions parmi *n* positions
  - Le tirage avec remise de positions, en respectant l'effectif original, revient à conférer un poids aléatoire aux positions
- Construire l'arbre phylogénétique
- Répéter 1) et 2) un grand nombre de fois (1000)



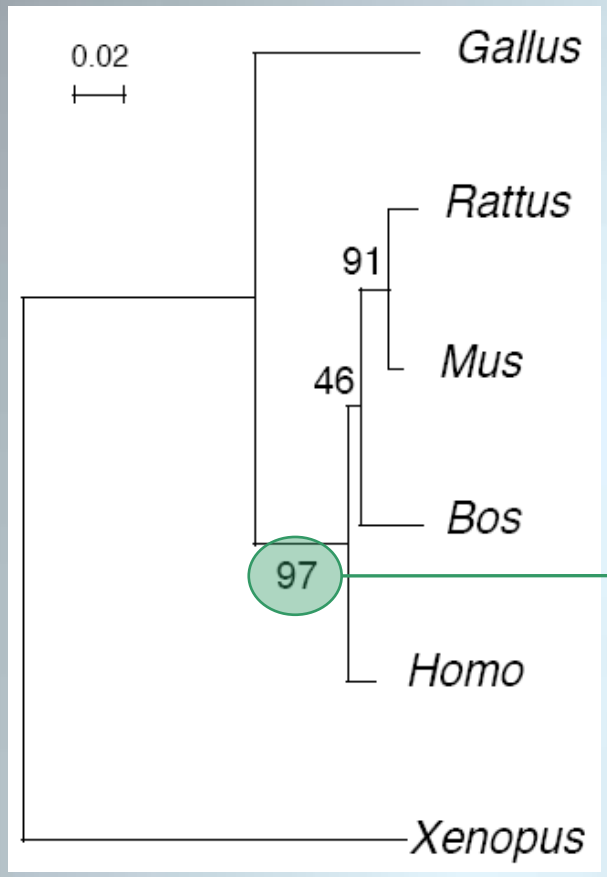
Pour chaque arbre reconstruit on compte le nombre de fois que l'on observe la branche interne. Le support de la branche est exprimé en pourcentage de réplications. Si la branche est observée dans tous les arbres, la valeur du bootstrap est égale à 100.

Le bootstrap : procédure



Le bootstrap

Test individuellement la validité de chaque branche interne de l'arbre. Pour cela, on calcule le pourcentage de fois où chaque branche interne de l'arbre de départ se retrouve dans les arbres construits par rééchantillonage. Ce pourcentage correspond à la valeur du bootstrap.



Estimation statistique de la **confiance** à accorder à une branche

97% des 1000 arbres contenaient cette branche

Le bootstrap : interprétation

Problème beaucoup discuté .

De manière générale, une faible valeur de bootstrap indique que la quantité d'information supportant la bipartition induite par une branche interne est faible.

Quel seuil ?

Si on applique les critères standards utilisés en statistique, il ne faudrait considérer comme valide que les branches ayant un support de bootstrap ≥ 95% (sinon la branche n'existe pas).
Des travaux ont montré que ce seuil était trop élevé, notamment ceux de Hillis et Bull (1993, Syst. Biol., 42, 182-92) qui à l'aide de simulations ont montré que des supports de 70% pouvaient correspondre à des groupements significatifs.
Cependant résultat pas généralisable à toutes les analyses.