# Support de cours
# Introduction à la reconstruction phylogénétique

# Introduction

Phylogénie : reconstruire l'histoire évolutive des espèces. Trouver des liens de parenté. Le résultat est une hypothèse de l'histoire évolutive des espèces. Elle ne correspond toujours à l'évolution réelle de ces espèces !

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

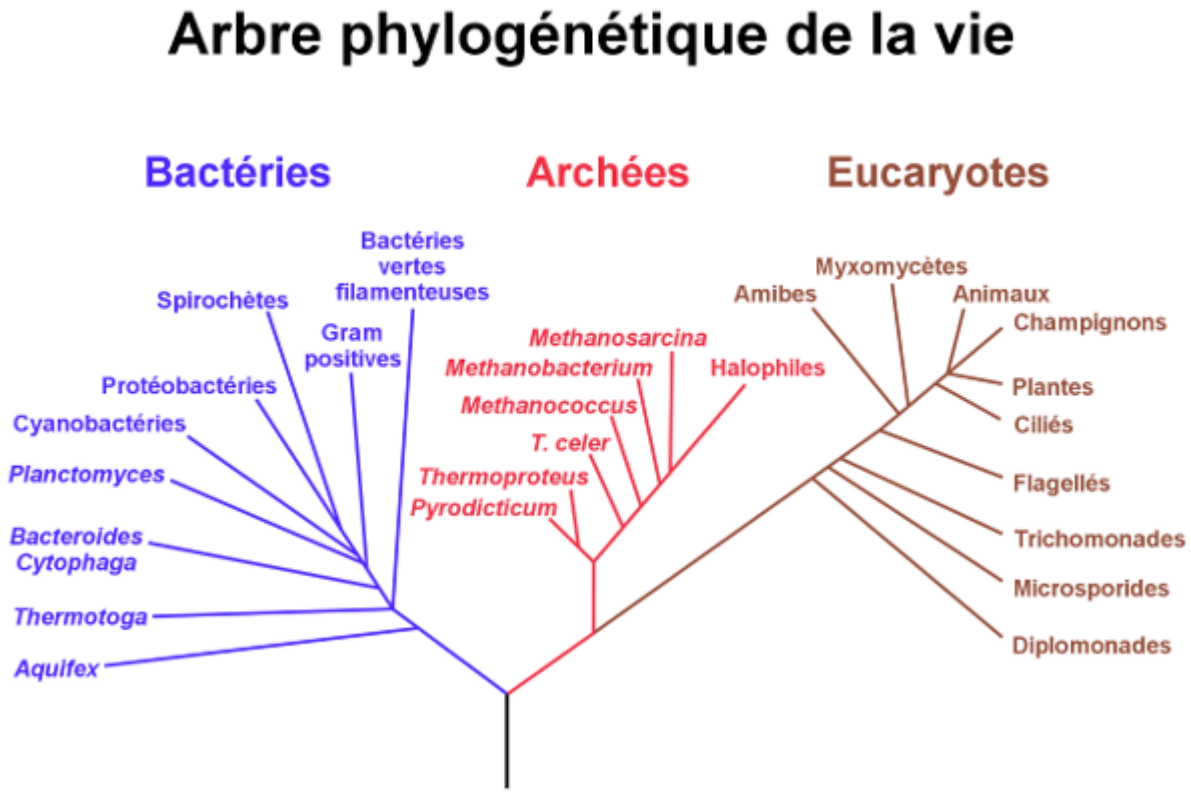
Premières analyses faites en 1965 par E. Zuckerkandl et L. Pauling  montrant que la phylogénie des vertébrés était a peu près identique quand elle était basée sur la comparaison de séquences protéiques ou sur des données morphologiques, anatomiques et paléontologiques.

Fitch et Margoliash, 2 ans plus tard, on établi une phylogénie des vertébrés à peu près identiques par comparaisons 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

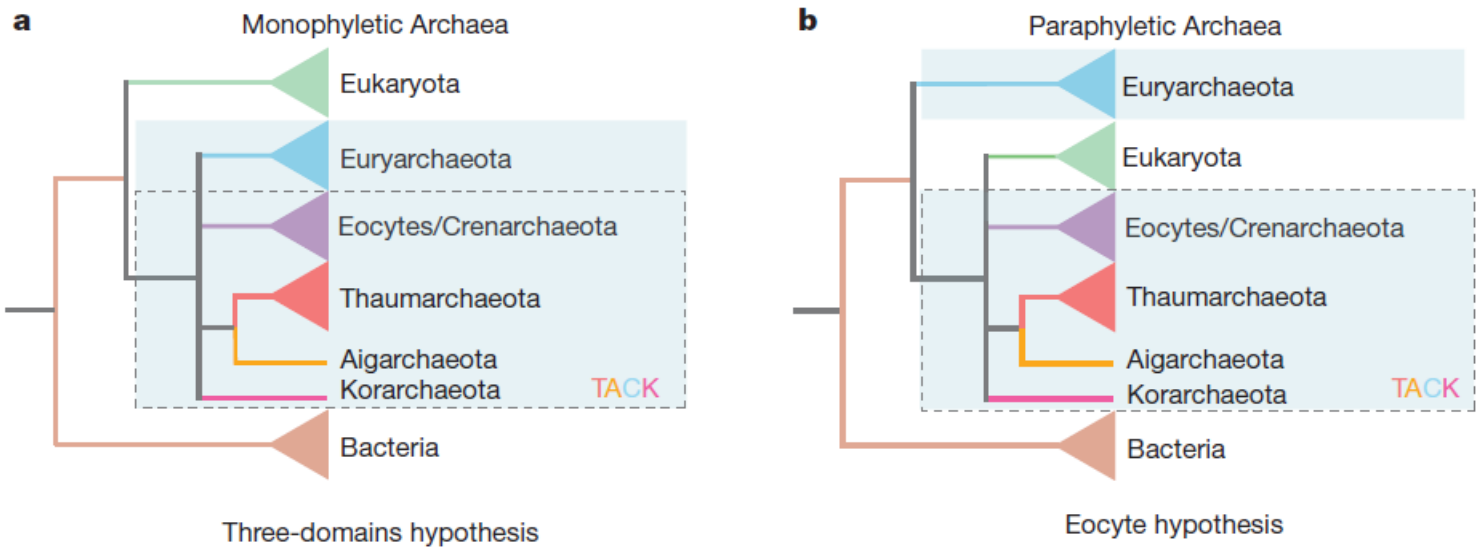
Découverte du troisième domaine du vivant par Carl Woese en 1977 par l'analyse phylogénétique des séquences d'ARNr 16S.




*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

Deux hypothèses en discussion sur l'origine des eucaryotes




*Extrait de Williams et al. (2013) Nature, 504, 231-236)*

# 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.

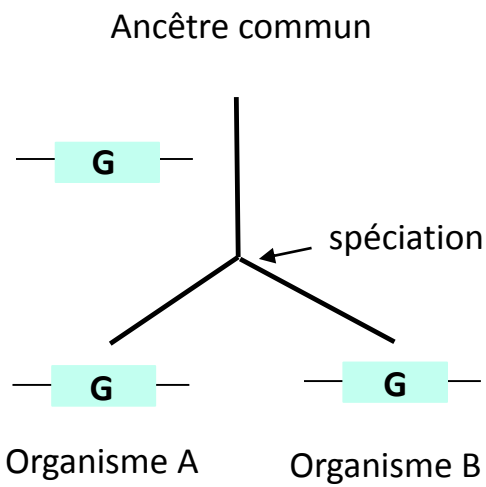
Définition insuffisante pour reconstruction de l'histoire évolutive car plusieurs mécanismes possibles pour dériver d'une séquence ancêtre.

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

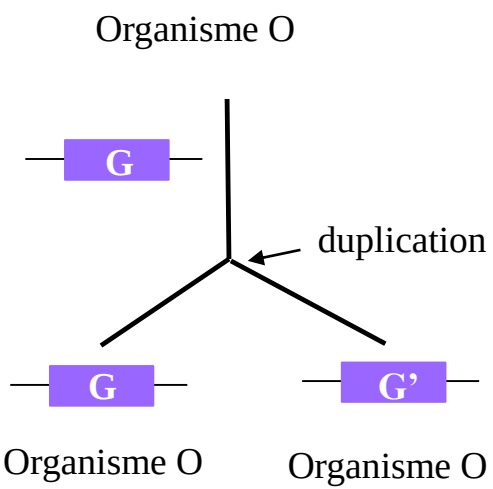
Paralogie : deux gènes sont **paralogues** si leur divergence est due à la duplication du gène ancêtre.

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

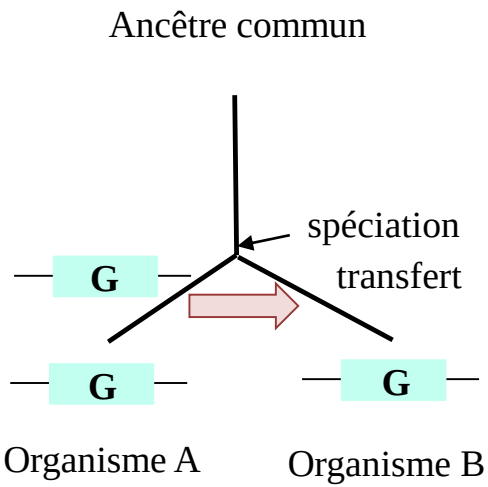
# Notions de base : définitions




**Génes orthologues**

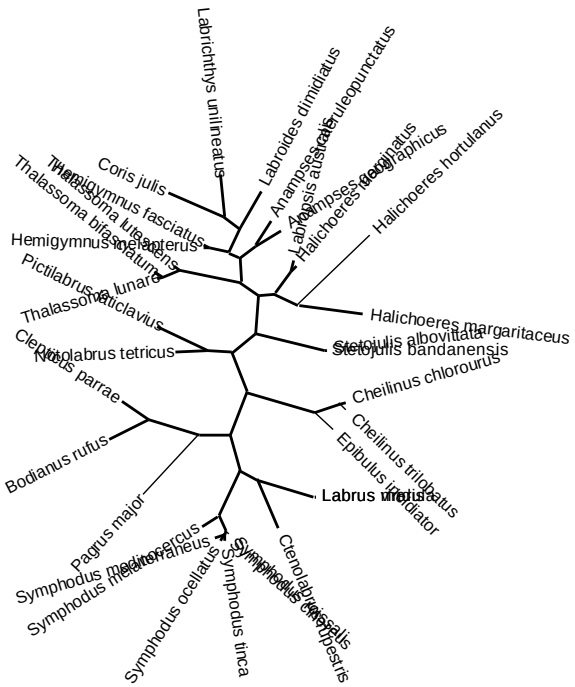


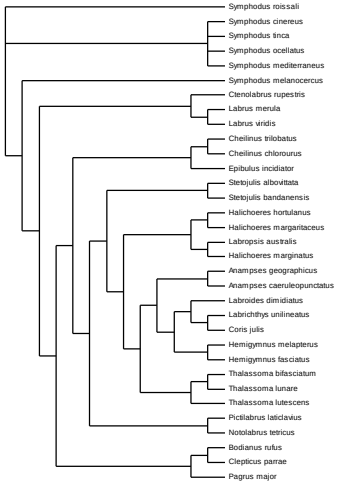

**Génes paralogues**

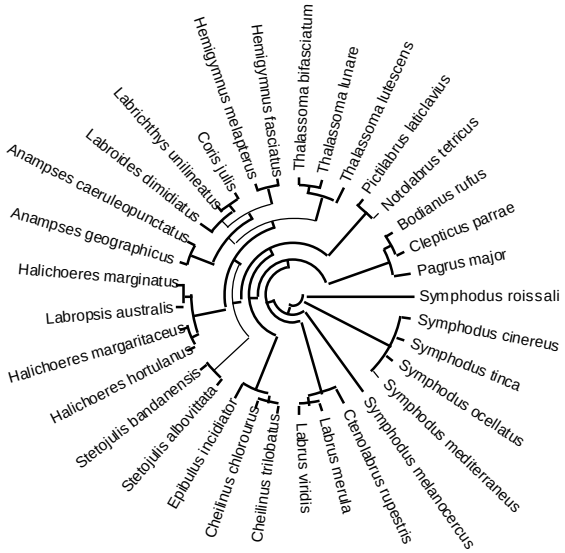



**Génes xénologues**

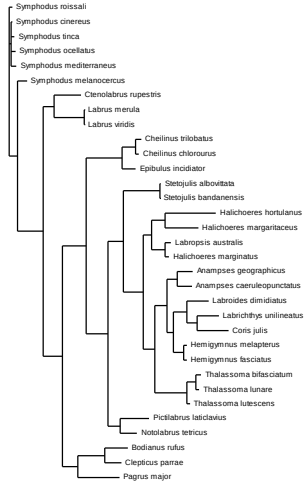
# Notions de base : arbres phylogénétiques



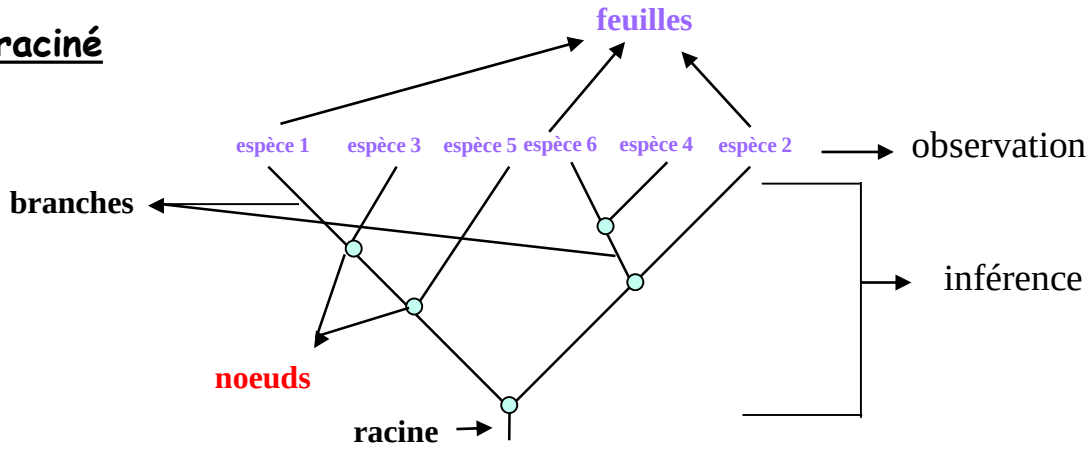







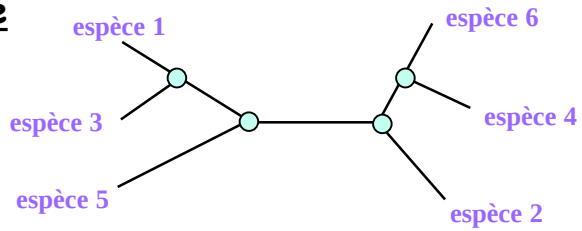
# Notions de base : arbres phylogénétiques

**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.

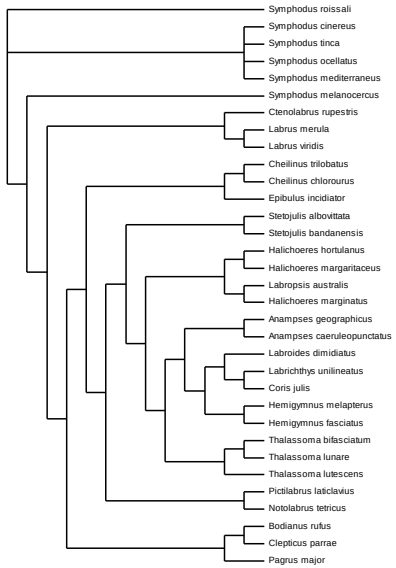
**Arbre sans racine**



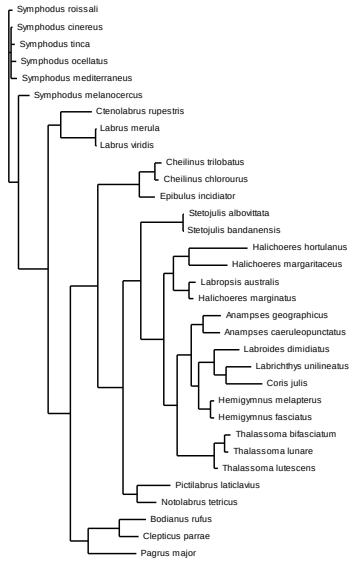

Dans un arbre sans racine, les liens entre nœuds ne sont pas orientés et un seul et unique chemin permet de passer d'un sommet à l'autre.

# Notions de base : arbres phylogénétiques


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



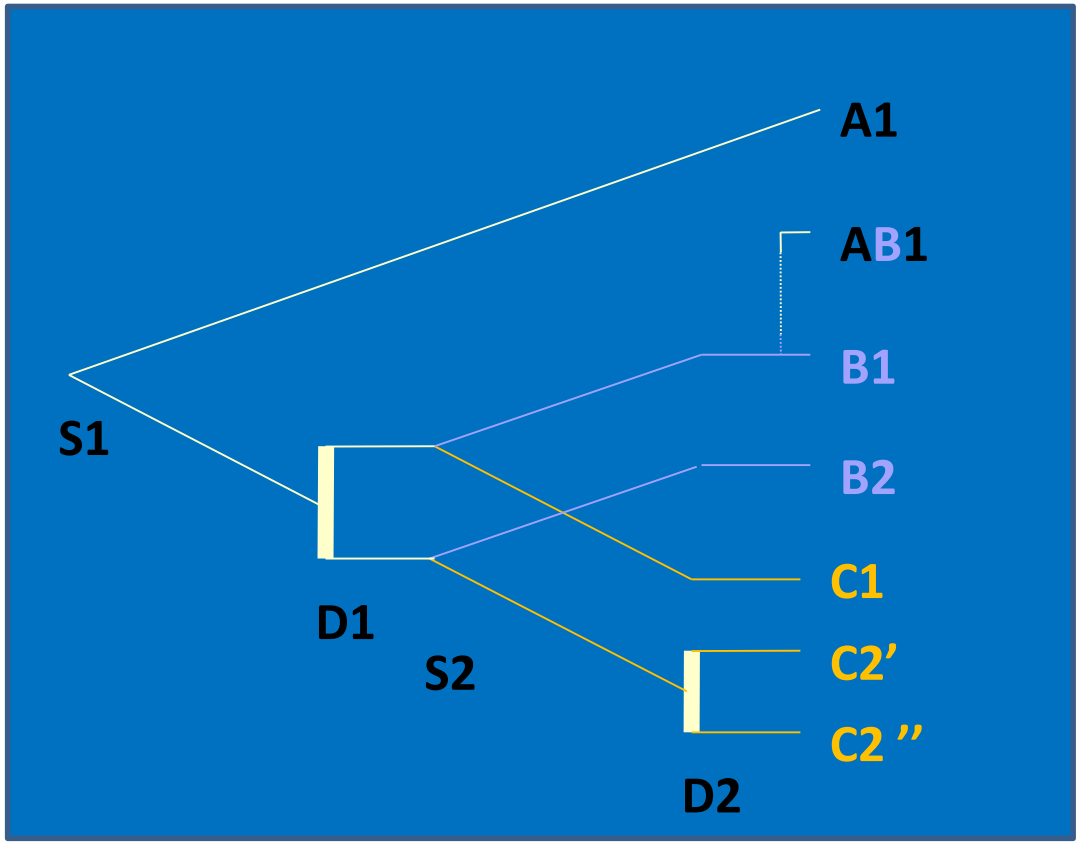

**Arbre ultramétrique**




**Arbre additif**

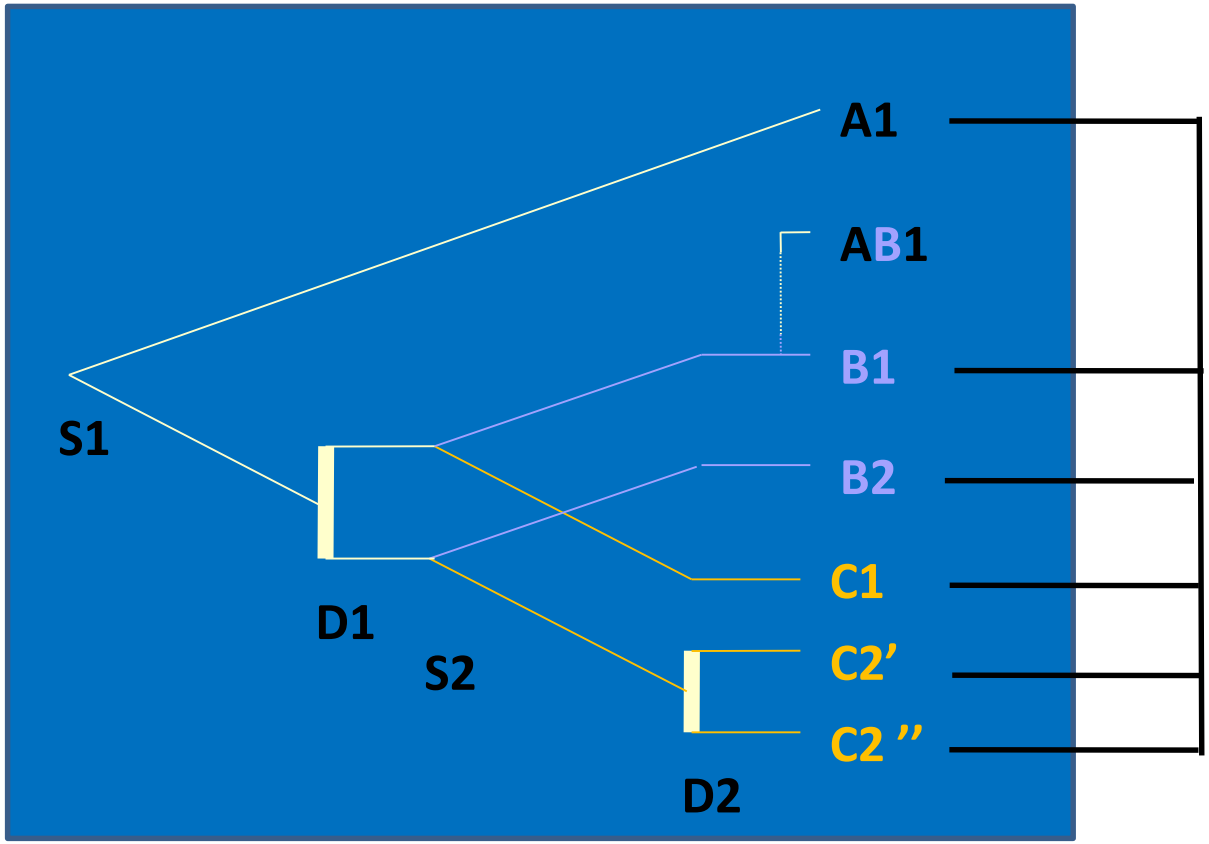
Phylogramme : longueurs de branches

# Notions de base : arbres phylogénétiques




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

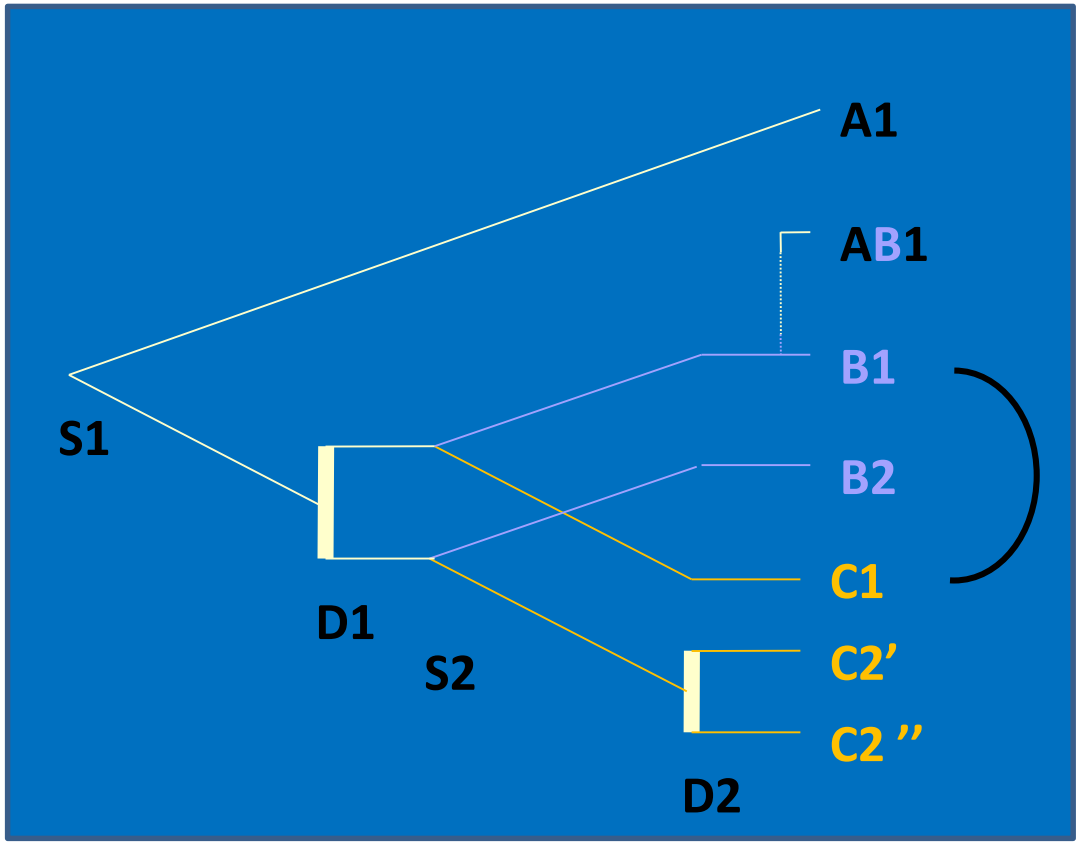
# Notions de base : arbres phylogénétiques




Orthologues à A1

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

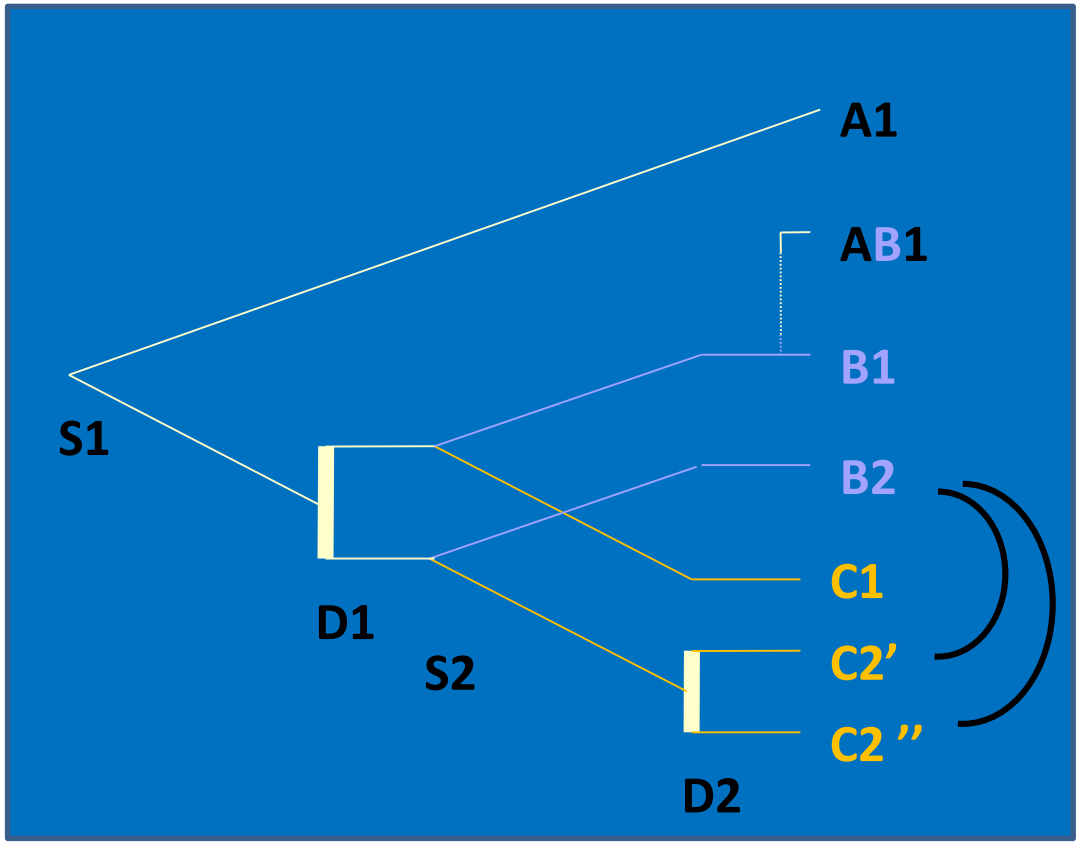
# Notions de base : arbres phylogénétiques




Orthologues

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

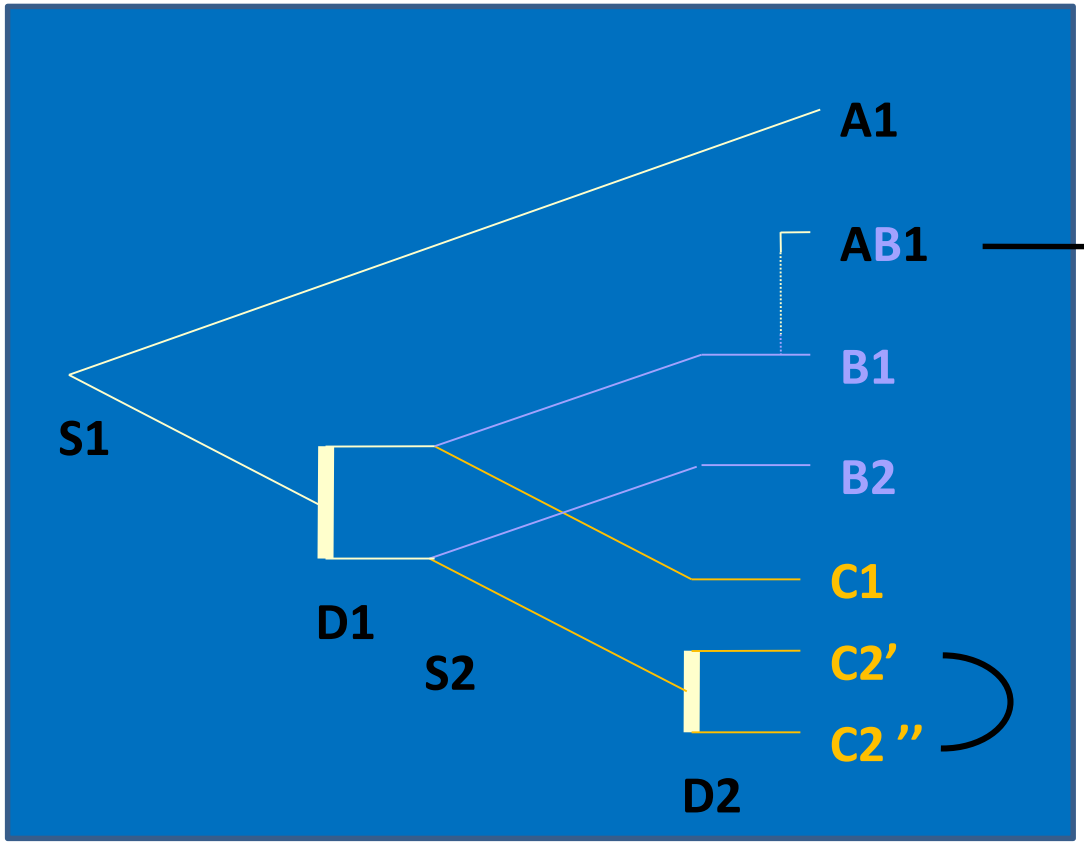
# Notions de base : arbres phylogénétiques




Orthologues

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

# Notions de base : arbres phylogénétiques




transfert horizontal

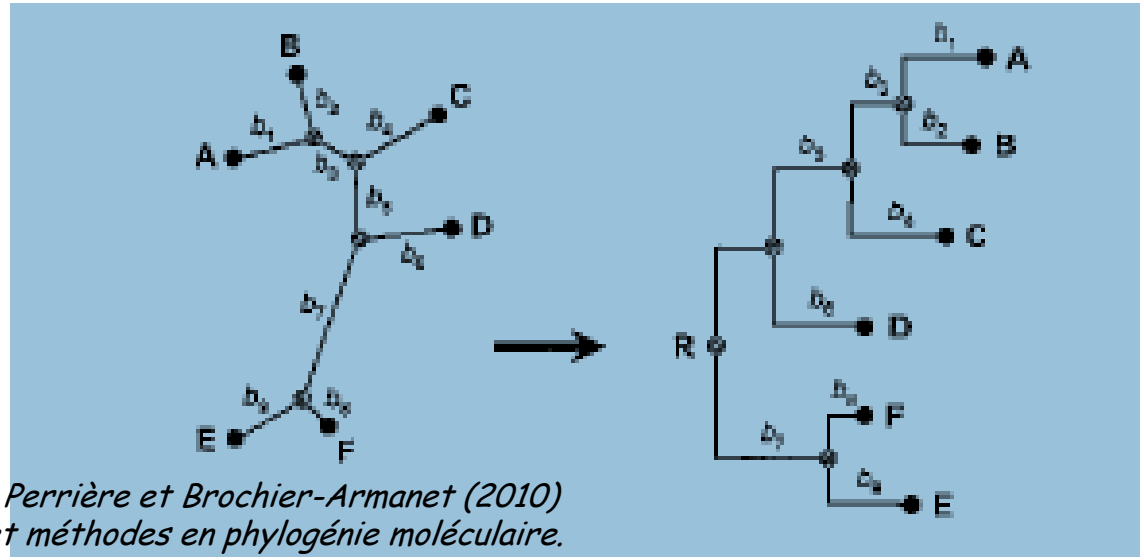
paralogues

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

# 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 :
  - Racinemenent 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.

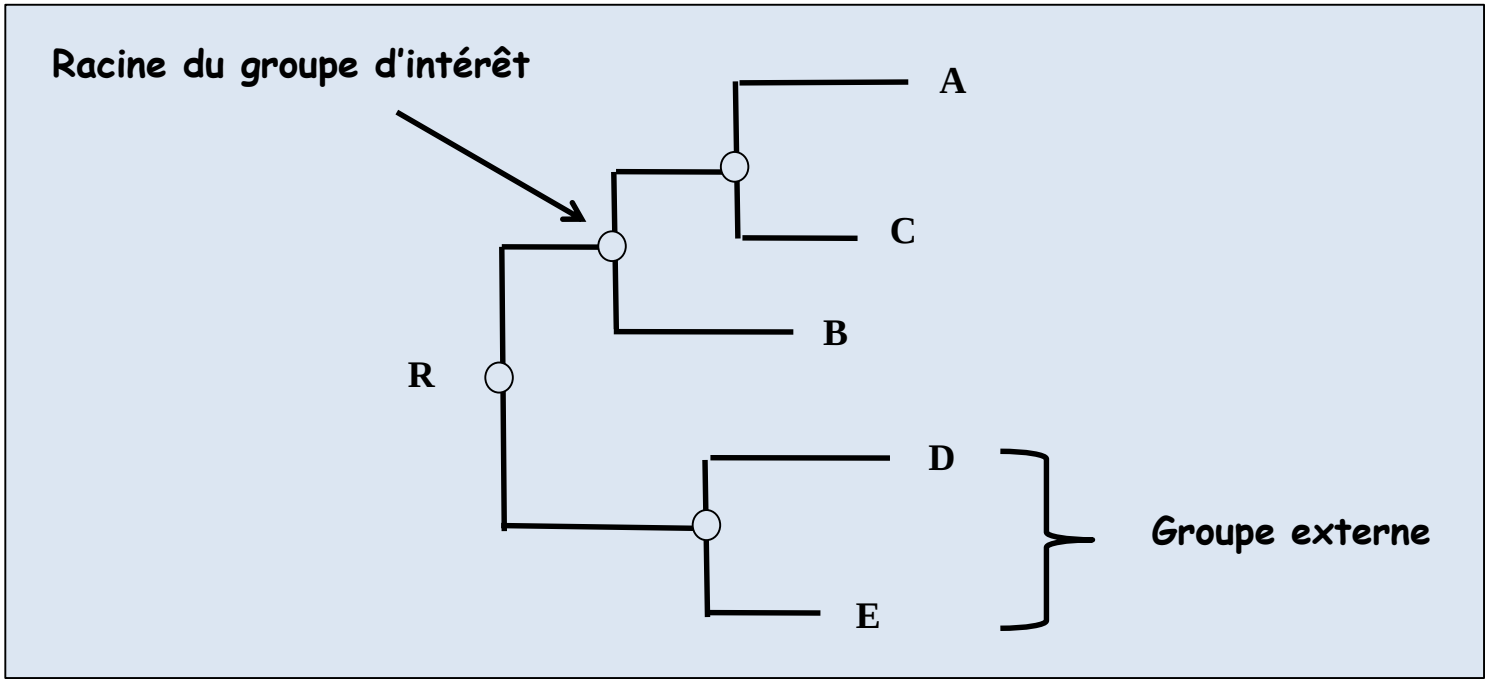



*Extrait de Perrière et Brochier-Armanet (2010)*
*Concepts et méthodes en phylogénie moléculaire.*

# Notions de base : arbres racinés et non racinés

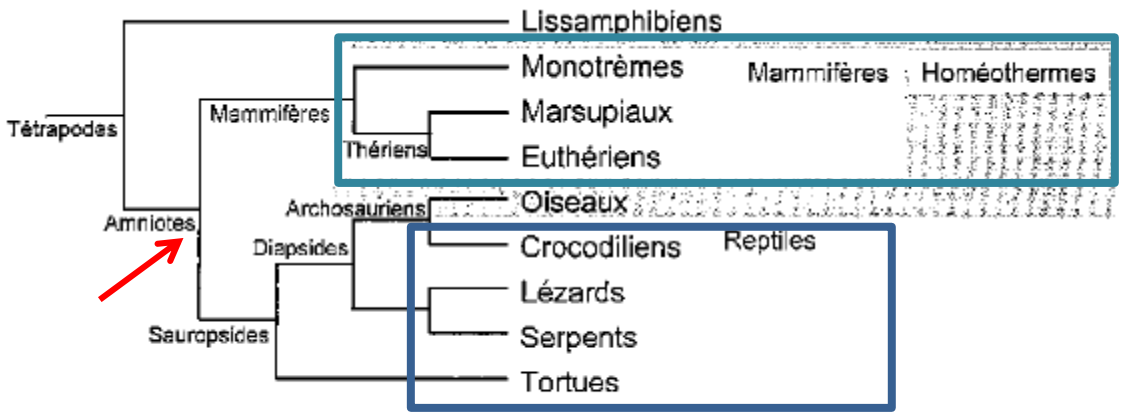
- enraciner un arbre :
  - La méthode du groupe externe : inclure un groupe de séquences connues *a priori* comme externes au groupe d'intérêt; la racine est alors sur la branche qui relie le groupe externe aux autres séquences. Séquences connues comme ayant a priori divergé avant le groupe d'intérêt.

  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 : vocabulaire

Caractères : Organismes composés de différentes caractéristiques
Chaque position alignée d'un alignement multiple

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

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

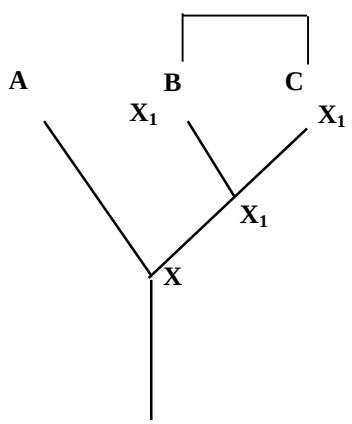
- On cherche à établir le lien entre ancêtre et descendant par la présence/absence d'un état de caractère
- On cherche l'apparition de nouveaux états de caractères dans les descendants

# Le concept de similarité

Il peut être divisé en :

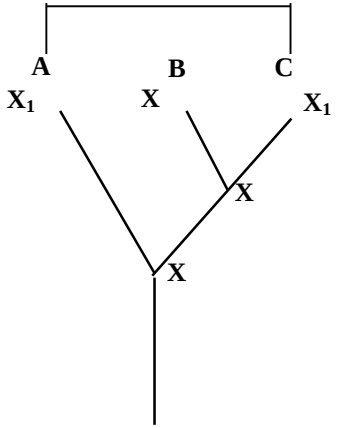
- homologie similarité héritée d'un ancêtre commun
- homoplasie similarité non héritée d'un ancêtre commun et qui est subdivisée en :
  - convergence : apparition indépendante dans deux espèces d'un même état dérivé de caractère
  - réversion : apparition d'un état de caractère ayant la forme ancestrale

**homologie**



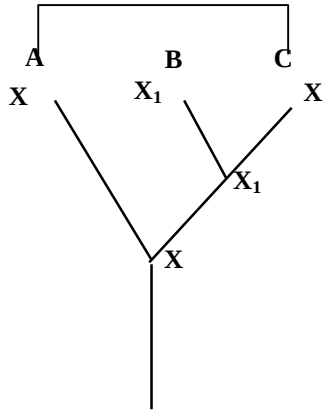

L'état de caractère $X_1$ présent chez B et C est hérité de l'ancêtre à B et C.

**convergence**



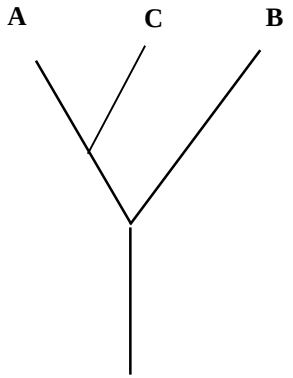

L'état de caractère $X_1$ présent chez A et C n'est pas hérité d'un ancêtre commun. Il est apparu deux fois de façon indépendante.

**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).

**arbre phylogénétique**




Arbre qui sera obtenu par reconstruction phylogénique à partir des valeurs des états de caractères dans le cas de la convergence ou de la réversion. Différent de l'arbre vrai

# 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 :

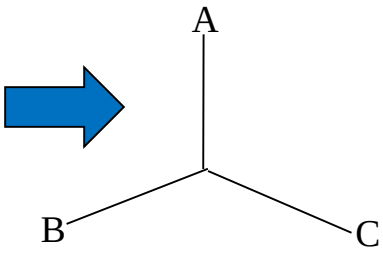
- les phénéticiens adeptes de la « taxonomie numérique ». Les liens entre les taxons ne peuvent être fondés que sur la base d'une similitude globale exprimée à partir de matrices de calcul de distances. Dans le cas des séquences, à partir d'un alignement multiple, on calculera les distances entre les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

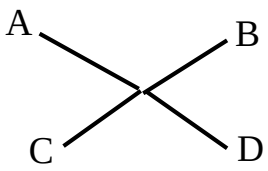
# Trouver l'arbre

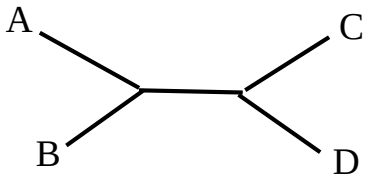
Problème : un seul arbre vrai, l'arbre évolutif
Comment le distinguer dans tous les arbres possibles

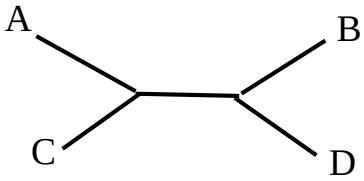
Si trois OTU : un seul arbre non raciné et trois racinés

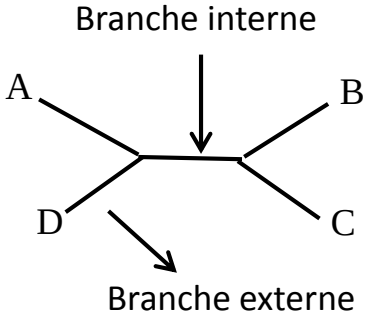



Si quatre OTU : quatre arbres non enracinés dont trois résolus




4 branches = 4 racines possibles

5 branches : 5 racines possibles

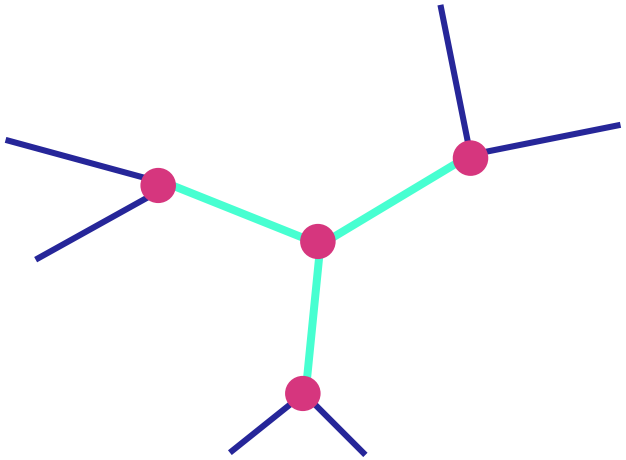
Total : 19 arbres enracinés possibles

# 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

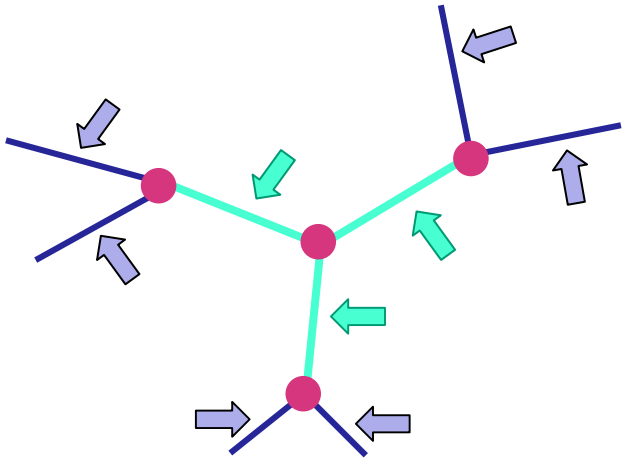
Branche externe : $n = 6$

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 x$10^{20}$ |

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

**50 taxons : 3 1074 (> atomes dans l'univers !!)**

# 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 :

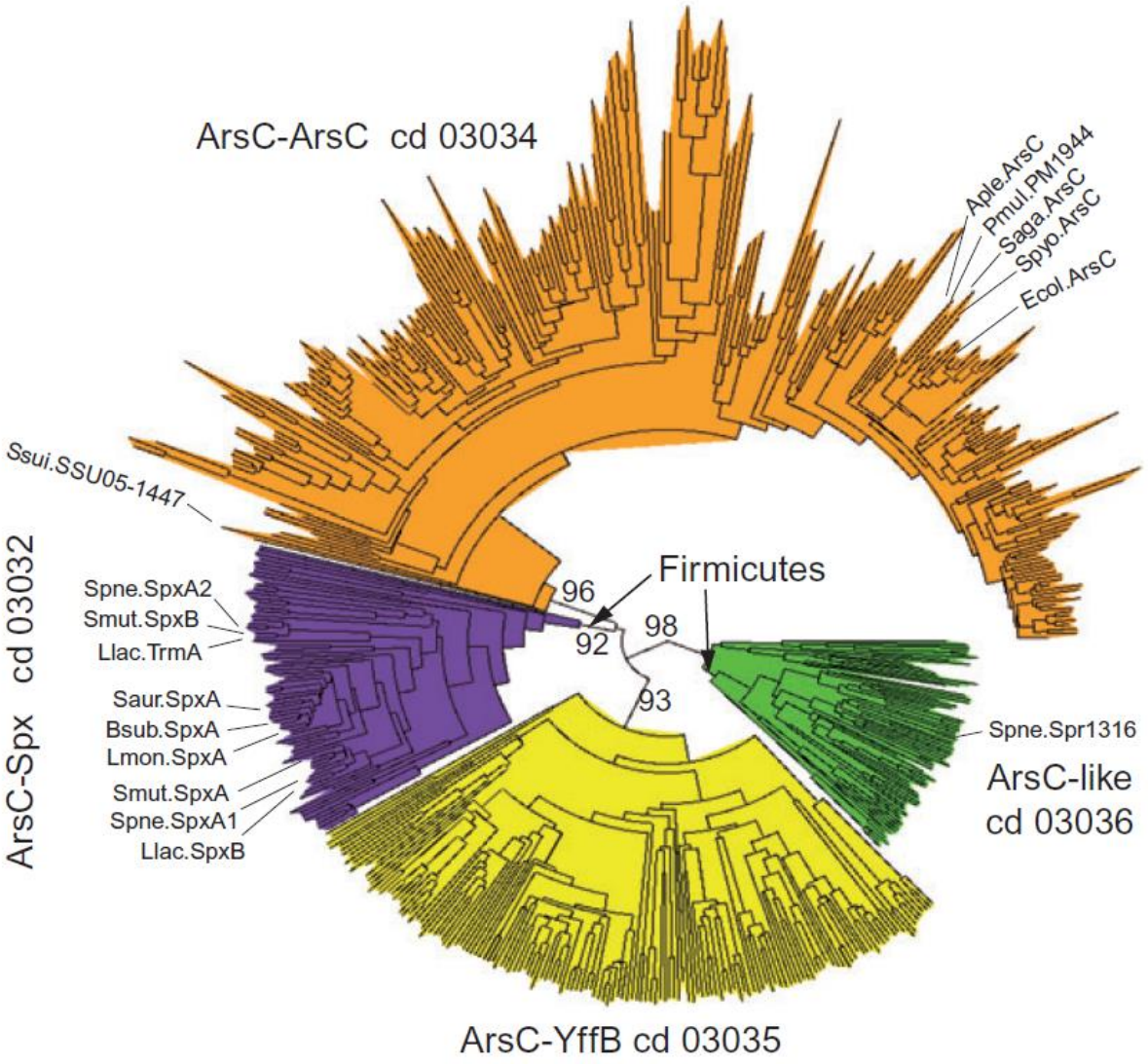
- Parcimonie : à partir d'un ensemble de caractères choisis. Recherche l'arbre qui minimise le nombre de changements permettant d'expliquer les données.

- 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

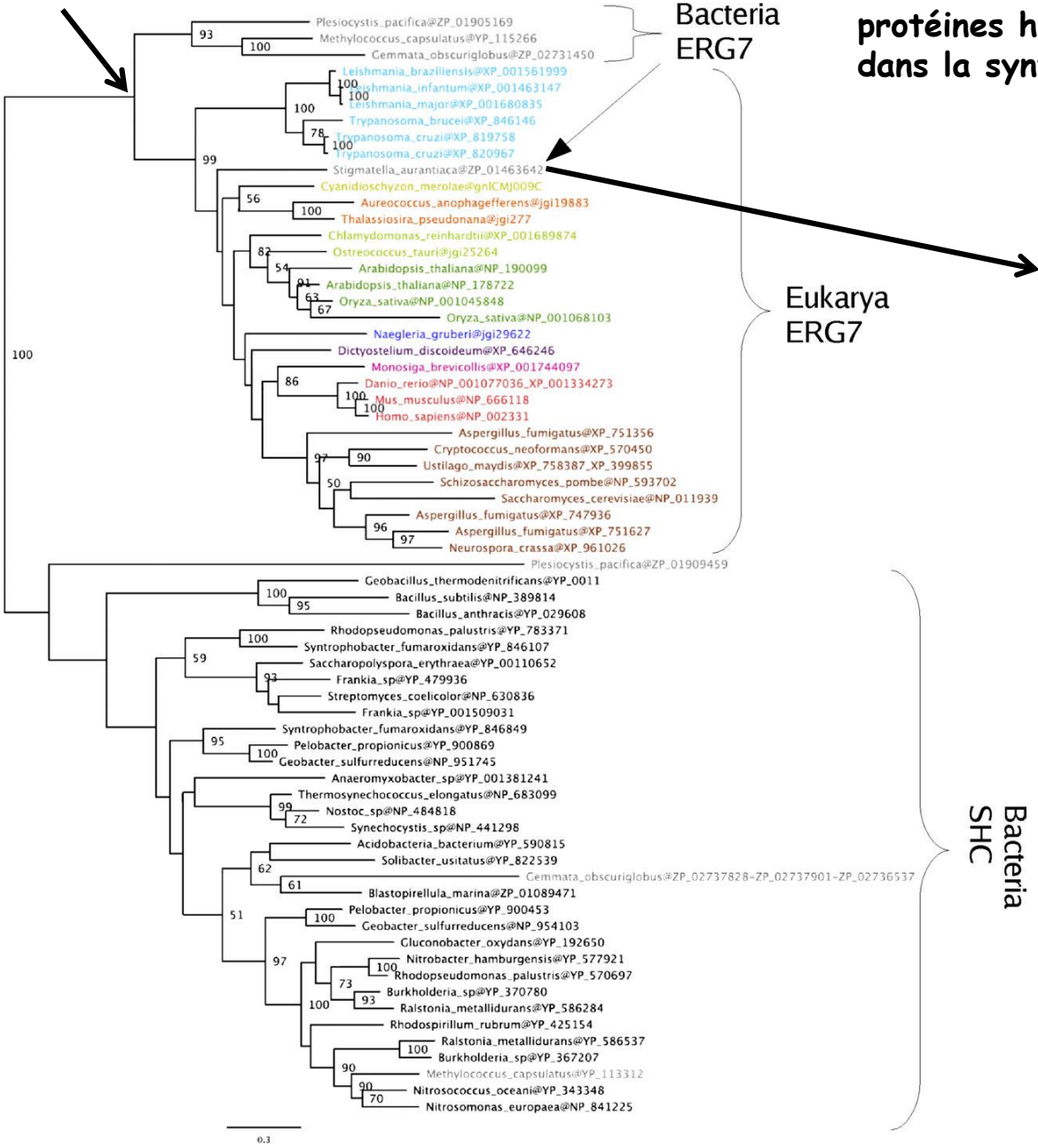
Protéines présentant une similarité avec le domaine COG1393 (arsenate reductase and related proteins)




**(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

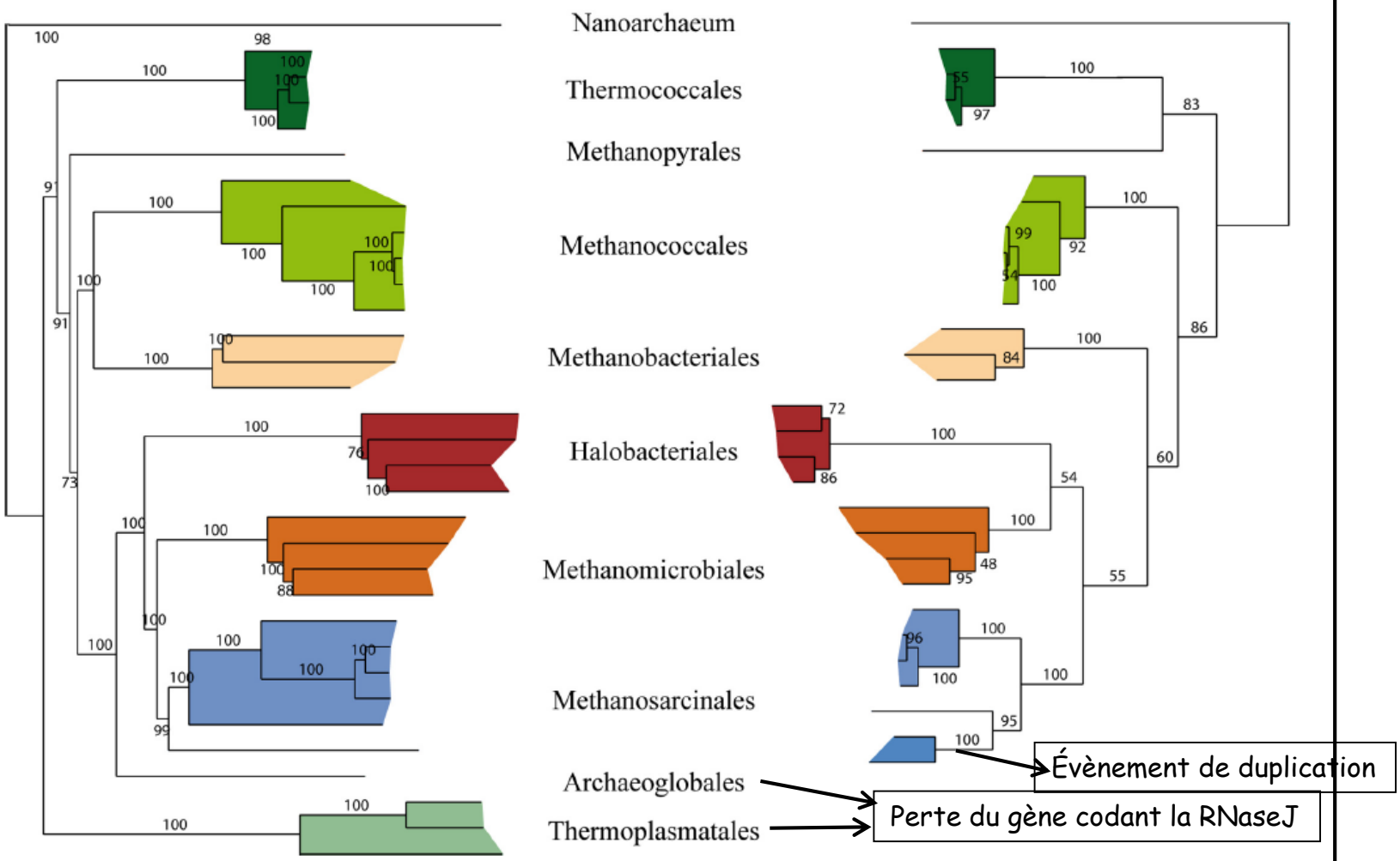
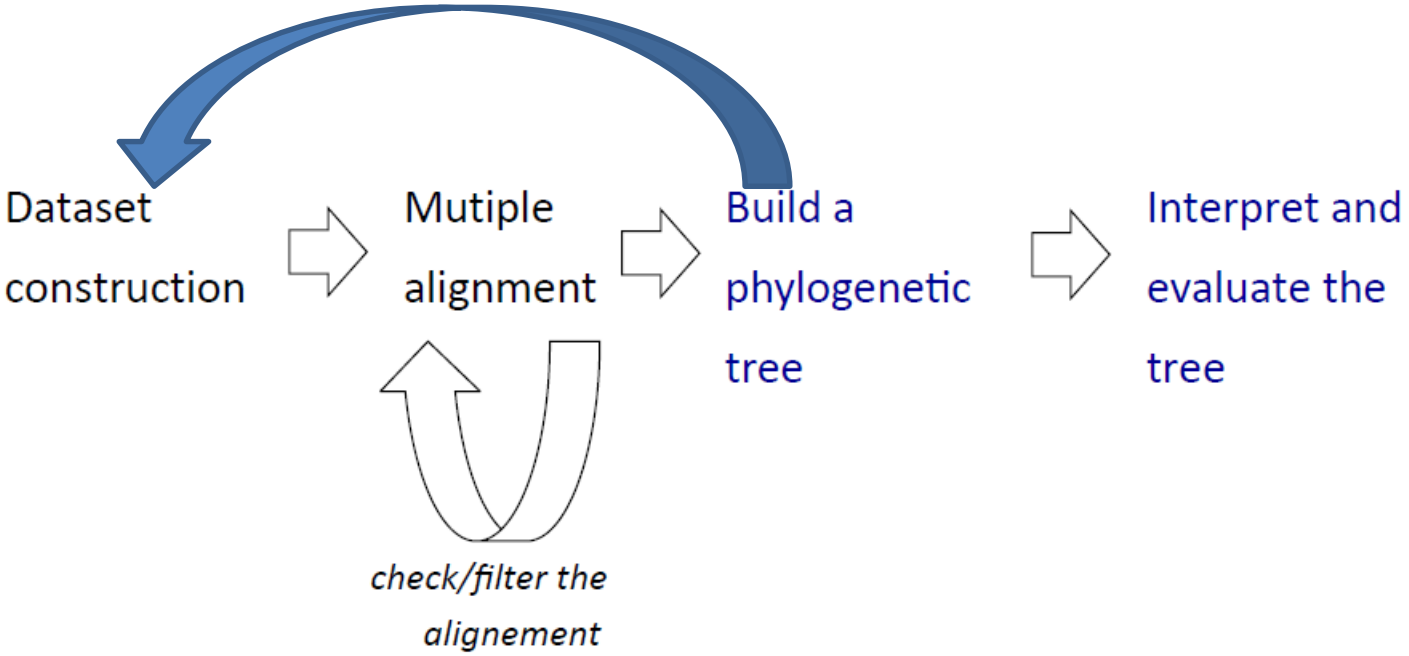



FIGURE 7. **Phylogenetic trees of the Euryarchaeota and euryarchaeal RNase J.** The tree of the Euryarchaeota was constructed from a concatenated sequence of 70 ubiquitous proteins (see "Experimental Procedures"). The euryarchaeal RNase J tree was constructed from sequences that were available at the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

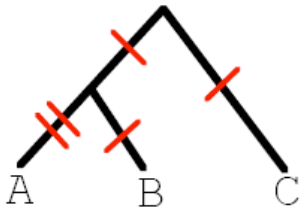
**(Extrait de Clouet d'Orval *et al.*, 2010, J. Biol. Chem. 285:17574-583)**

# Inférence phylogénétique

# Principales étapes d'une analyse phylogénétique

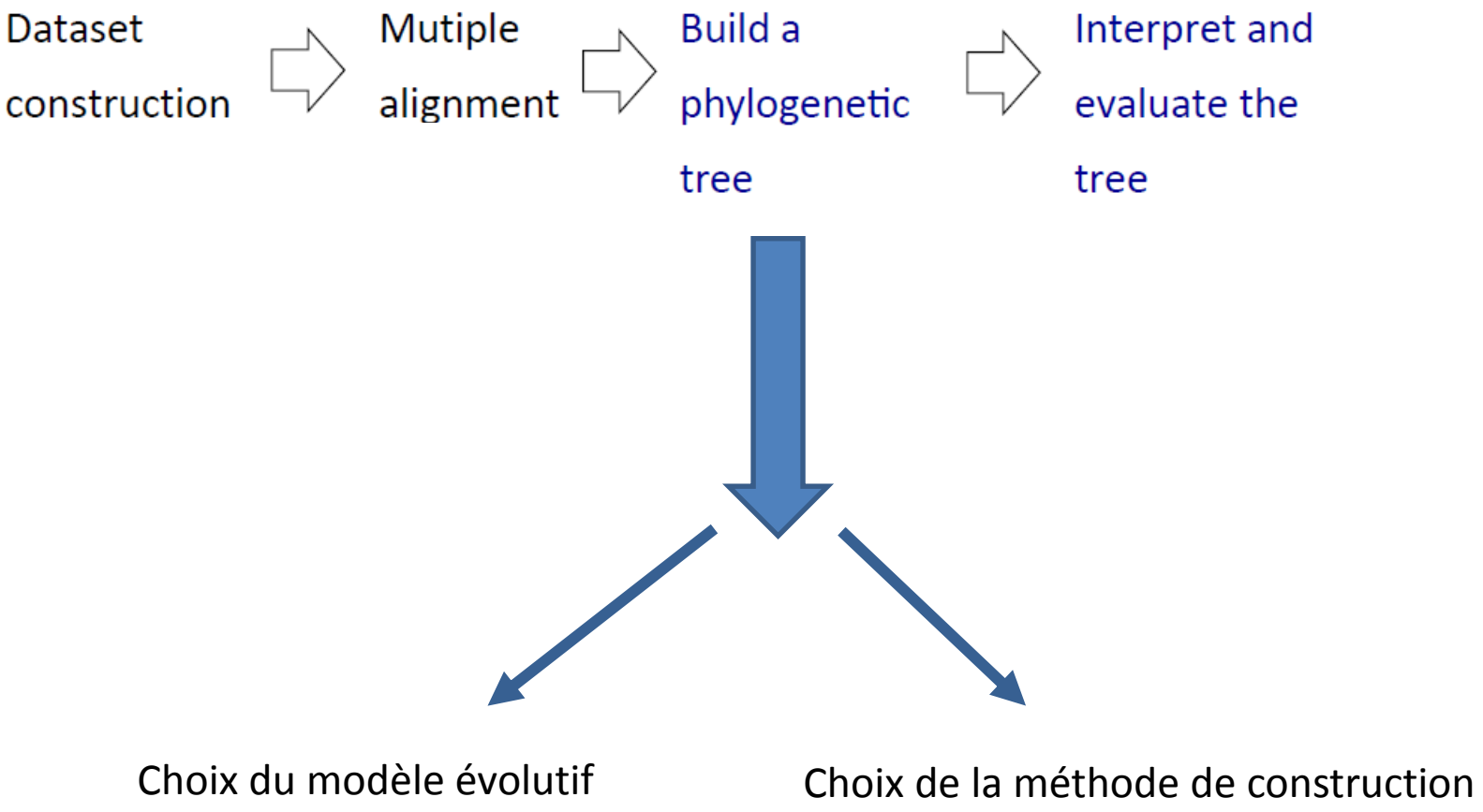



```
CAAACAGCGTT---GGCTCTCTA
AAAATAACACCaacATGCAAATG
AAAACAGCACCaacGTGCAAATG
AAAACAGCACCaacGTGCAAATG
```




tree

# Principales étapes d'une analyse phylogénétique



# Calcul d'une distance génétique (évolutive) entre deux séquences

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

Divergence observée ou *p*-distance : 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 |

$p$ = # differences / # 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 d'une distance génétique (évolutive) entre deux séquences

## 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 | 2 |
| Substitutions convergentes | C ➤ G ➤ A | C ➤ A | 0 | 3 |
| Substitution réverse (inverse) | G ➤T ➤ G | G | 0 | 2 |

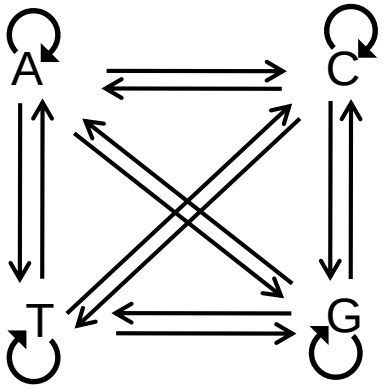
# Calcul d'une distance génétique (évolutive) entre deux séquences

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

- modéliser par un modèle de Markov en temps continu

Dans les modèles markoviens, l'information utile pour la prédiction du futur est contenue dans l'état présent du processus. Donc, l'état futur d'un site ne dépendra que de son état présents et pas des états passés.




Les substitutions à chaque site sont décrites par une chaîne de Markov dont les états correspondent aux quatre bases nucléotidiques et les probabilités de transitions sont données par les probabilités de passer d'un état à un autre ou de rester dans le même état.

L'évolution d'un site le long d'une branche d'un arbre phylogénétique est décrite par les probabilités de transition $p_{ij}$ d'un état initial *i* au nœud ancêtre à un état *j* au nœud fils.

# Calcul d'une distance génétique (évolutive) entre deux séquences

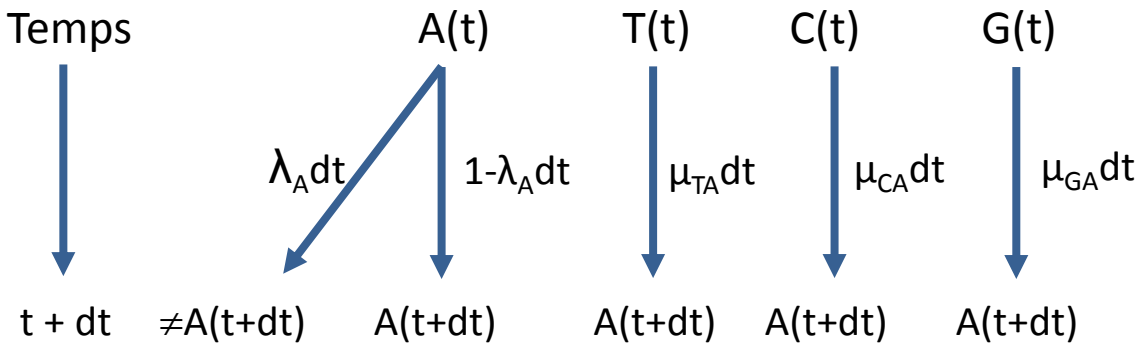
## Hypothèses liées au modèle markovien

1. **Homogénéité du processus** : les probabilités de substitution ne changent pas au cours du temps. Donc même processus applicable le long de toutes les branches de l'arbre.

   On peut donc définir :

   - Le taux de substitution instantané d'une base d'un état *i* vers un état *j* $\mu_{ij}$ ($i \neq j$).
   - Le taux de changement instantané d'un nucléotide dans l'état *i* vers un autre nucléotide $\lambda_i$ .

Au temps *t*, on a les probabilités suivantes des 4 bases : A(t), T(t), C(t) et G(t). On veut calculer la probabilité d'observer ces bases après un très court temps d'évolution *dt* . Exemple de la base A :




$$A(t+dt) = A(t) - A(t)\lambda_A dt + T(t)\mu_{TA} dt + C(t)\mu_{CA} dt + G(t)\mu_{GA} dt$$

# Calcul d'une distance génétique (évolutive) entre deux séquences

Si on fait le même raisonnement pour chacune des 4 bases on obtient le système de quatre équations différentielles linéaires :

$$
\begin{aligned}
A(t+dt) &= A(t) - A(t)\lambda_A dt + T(t)\mu_{TA}dt + C(t)\mu_{CA}dt + G(t)\mu_{GA}dt \\
T(t+dt) &= A(t)\mu_{AT}dt + T(t) - T(t)\lambda_T dt + C(t)\mu_{CT}dt + G(t)\mu_{GT}dt \\
C(t+dt) &= A(t)\mu_{AC}dt + T(t)\mu_{TC}dt + C(t) - C(t)\lambda_C dt + G(t)\mu_{GC}dt \\
G(t+dt) &= A(t)\mu_{AG}dt + T(t)\mu_{TG}dt + C(t)\mu_{CG}dt + G(t) - G(t)\lambda_G dt
\end{aligned}
$$

On peut en déduire la matrice M des taux de substitution instantanés:

$$
M = \begin{bmatrix}
-\lambda_A & \mu_{AT} & \mu_{AC} & \mu_{AG} \\
\mu_{TA} & -\lambda_T & \mu_{TC} & \mu_{TG} \\
\mu_{CA} & \mu_{CT} & -\lambda_C & \mu_{CG} \\
\mu_{GA} & \mu_{GT} & \mu_{GC} & -\lambda_G
\end{bmatrix}
$$

La différence entre les modèles d'évolution est liée à la définition des $\mu_{ij}$

La matrice M décrit les fréquences relatives des différents types de substitutions, seuls les rapports entre les valeurs de $\mu_{ij}$ sont informatifs (par exemple le rapport transitions/tranversions, la fréquence des bases à l'équilibre, etc) et participent à la description du modèle.

# Calcul d'une distance génétique (évolutive) entre deux séquences

## Hypothèses liées au modèle markovien

2. **Indépendance des sites :** les sites évoluent indépendamment les uns des autres. Hypothèse pas vérifiée dans de nombreux cas, notamment pour les ARN structuraux où pour maintenir la structure secondaire, il y a apparition de mutations compensatrices (coévolution des sites).

3. **Uniformité du processus :** tous les sites d'une séquence suivent le même processus, c'est-à-dire que les probabilités et taux de substitutions sont applicables à tous les sites. Conséquence, on suppose que les sites évoluent à la même vitesse. On sait que cette hypothèse est fausse mais elle utilisée dans la plupart des modèles d'évolution. Des améliorations dans les modèles ont été proposées pour prendre en compte l'existence de vitesses d'évolution différentes. La plus courante est l'utilisation de la distribution Gamma.

4. **Stationnarité :** les fréquences relatives des bases A, C, G et T sont à l'équilibre.

5. **Réversibilité :** quand l'équilibre est atteint, la quantité de changement de l'état *i* vers l'état *j* est égale à la quantité de changement de l'état *j* vers l'état *i*. On a $\mu_{AT}= \mu_{TA}$, $\mu_{AC}= \mu_{CA}$, $\mu_{AG}= \mu_{GA}$, $\mu_{CG}= \mu_{GC}$, $\mu_{CT}= \mu_{TC}$, $\mu_{GT}= \mu_{TG}$. Permet de simplifier les calculs, car les 12 paramètres non diagonaux de la matrice M peuvent être décrits par 9 paramètres, les 6 taux d'interchangeabilité ($\mu_{ij}$) et trois fréquences de bases à l'équilibre car la somme des fréquences des bases = 1 ($\sum_i \pi_i = 1$)

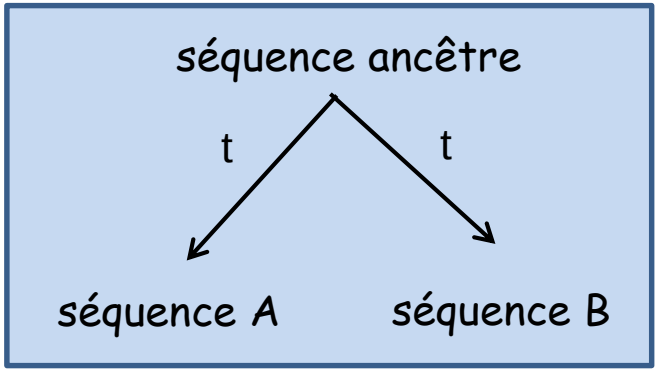
# Calcul d'une distance génétique (évolutive) entre deux séquences

## Distance évolutive entre deux séquences

Le nombre de substitution ayant eu lieu pendant un intervalle de temps *t* est donné par la multiplication du taux d'évolution global λ par *t*. Sous l'hypothèse de la stationnarité on a :

$$\lambda = \sum_i \pi_i \lambda_i \quad \text{avec} \quad \lambda_i = \sum_{i \neq j} \mu_{ij}$$

Quand on compare deux séquences homologues, le temps qui les sépare est non de *t* mais de 2*t*.




La distance évolutive séparant deux séquences est donc :

$$d = 2\sum_i \pi_i \lambda_i t$$

# Calcul d'une distance génétique (évolutive) entre deux séquences

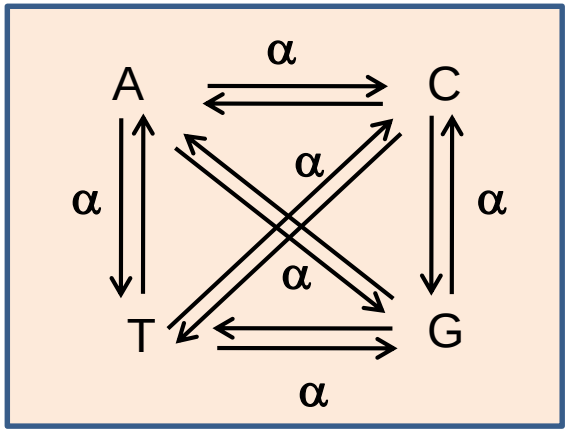
## Modèle de Jukes et Cantor (abrégé JC69)

- Modèle markovien le plus simple mais vision simplificatrice de l'évolution

- toutes les substitutions sont équiprobables donc un seul taux de substitution instantané $\alpha$ pour chacun des changements possibles (tous les $\mu_{ij}$ = a) et un seul taux de conservation global instantané 1-3$\alpha$ . La matrice M des taux de changements est donc simplifiée. On a :

$$M = \begin{bmatrix} -\lambda & \alpha & \alpha & \alpha \\ \alpha & -\lambda & \alpha & \alpha \\ \alpha & \alpha & -\lambda & \alpha \\ \alpha & \alpha & \alpha & -\lambda \end{bmatrix}$$

avec $\lambda = 3\alpha$




La matrice des probabilité de substitution P(t) peut donc s'écrire :

$$P(t) = \begin{bmatrix} q(t) & p(t) & p(t) & p(t) \\ p(t) & q(t) & p(t) & p(t) \\ p(t) & p(t) & q(t) & p(t) \\ p(t) & p(t) & p(t) & q(t) \end{bmatrix}$$

- q(*t*) probabilité qu'après un temps *t* le nucléotide reste inchangé
- p(*t*) probabilité qu'après un temps *t* il se soit substitué en un autre (passage de l'état *i* à l'état *j*).

# Calcul d'une distance génétique (évolutive) entre deux séquences

## Modèle de Jukes et Cantor (abrégé JC69)

Pour calculer la distance entre deux séquences, il faut que l'on trouve une relation entre *d* et la probabilité d'observer une substitution à un site qui est donnée par la distance observée ou *p*-distance.

## la distance de Jukes et Cantor est donnée par :

$$d = -\frac{3}{4} Log\left(1 - \frac{4}{3} p_{dist}\right)$$

Un facteur correcteur est donc apporter à la p-distance $p_{dist}$

Quand $p_{dist}$ = ¾ d → ∞ (Log 0 pas défini). Donc ce modèle n'est pas utilisable pour des séquences dont la distance observée est supérieure à 75%.

# Calcul d'une distance génétique (évolutive) entre deux séquences

```
Seq1    TCAAGTCAGGTTCGA
           |   |  ||
Seq2    TCCAGTTAGACTCGA
            | | |  | |
Seq3    TTCAATCAGGCCCGA
```

Distances observées

| | Seq1 | Seq2 | Seq3 |
|---|---|---|---|
| Seq2 | 0.266 | | |
| Seq3 | 0.333 | 0.333 | |

Distance observée

p-distance (seq1,seq2) = $\frac{4}{15}$ = 0.266

Distances évolutives Jukes et Cantor

| | Seq1 | Seq2 | Seq3 |
|---|---|---|---|
| Seq2 | 0.328 | | |
| Seq3 | 0.441 | 0.441 | |

Distance J&C

$$d = -\frac{3}{4} Log\left(1 - \frac{4}{3} p_{dist}\right)$$

$$d_{JC} = -\frac{3}{4} Log(1 - 0.266\frac{4}{3}) = 0.328$$

# Calcul d'une distance génétique (évolutive) entre deux séquences

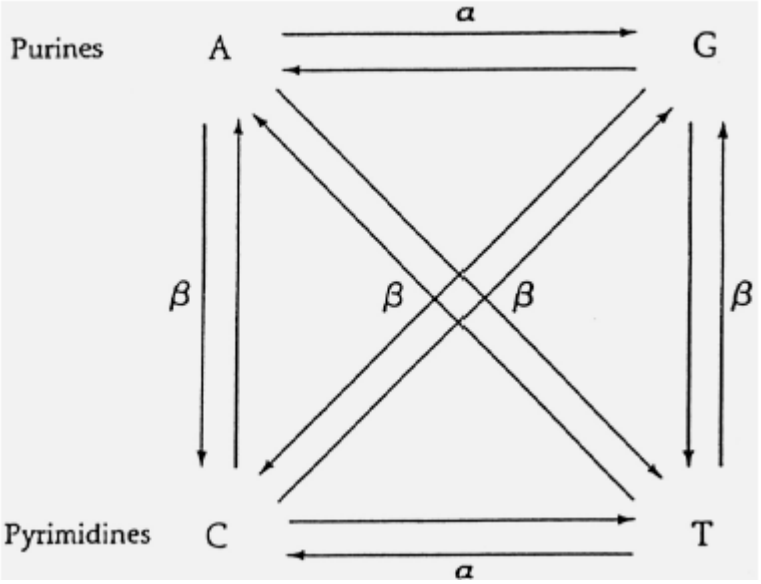
## Modèle de Kimura à deux paramètres (K80)

➢ Modèle moins simplificateur et biologiquement plus réaliste car il est observé que la fréquence des transitions est plus élevée que celle des transversions.

➢ 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). Les fréquences des bases à l'équilibre sont toutes égales à ¼, donc que le taux global de GC soit égal à ½. On a donc la matrice des taux de changements M suivante :

$$M = \begin{array}{c} \\ A \\ T \\ C \\ G \end{array} \begin{array}{c} \begin{array}{cccc} A & T & C & G \end{array} \\ \begin{bmatrix} -\lambda & \beta & \beta & \alpha \\ \beta & -\lambda & \alpha & \beta \\ \beta & \alpha & -\lambda & \beta \\ \alpha & \beta & \beta & -\lambda \end{bmatrix} \end{array}$$

avec $\alpha$ taux de transitions et β taux de tranversions
λ le taux instantané de changement pour une base quelconque avec $\lambda = 2\beta + \alpha$



# Calcul d'une distance génétique (évolutive) entre deux séquences

**Distance de Kimura à deux paramètres :**

$$d = -\frac{1}{2} Log(1-2p-q) - \frac{1}{4} Log(1-2q)$$

- *p* fréquence observée des transitions
- *q* fréquence observée des transversions

**Autres modèles :**

Il existe de nombreux autres modèles qui prennent en compte le taux de GC. Le modèle neutre le plus général est le modèle GTR (Generalised time-reversible, Tavaré 1986) dans lequel chacune des bases peut avoir des fréquences différentes et tous les types de mutations sont pris en compte.

Pour les autres modèles voir :
Concepts et méthodes en phylogénie moléculaire, Guy Perriére et Céline Brochier-Armanet, collection IRIS, Springer-Verlag France
Inferring Phylogenies,J. Felsenstein, Sinauer Associates Inc. Publishers Sunderland, Massachutsetts

# Distances synonymes et non synonymes

- Hypothèses des modèles précédents:

  Tous les sites évoluent indépendamment selon le même processus.
- Problème: dans les gènes protéiques, il existe deux classes de sites avec des taux d'évolution très différents.
  - substitutions non synonymes (changent l'acide aminé): lent
  - substitutions synonymes (ne changent pas l'acide aminé): rapide
- Solution: calculer deux distances évolutives
  - **$K_A$ ou $d_N$** = distance non-synonyme

    = nbr. substitutions non-synonymes / nbr. sites non-synonymes
  - **$K_S$ ou $d_S$** = distance synonyme

    = nbr. substitutions synonymes / nbr. sites synonymes

Si les séquences sont soumises à une sélection purificatrice, on attend un déficit de substitutions non synonymes : $d_N/d_S < 1$

Si les séquences sont soumises à une sélection positive, on attend un excès de substitutions non synonymes : $d_N/d_S > 1$

Si les séquences évoluent de façon neutre on aura : $d_N \approx d_S$

# 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). De même la distance observée sous estime la distance évolutive

- Egalement plusieurs modèles pour estimer la distance entre deux séquences

## Le plus simple : le modèle de Poisson

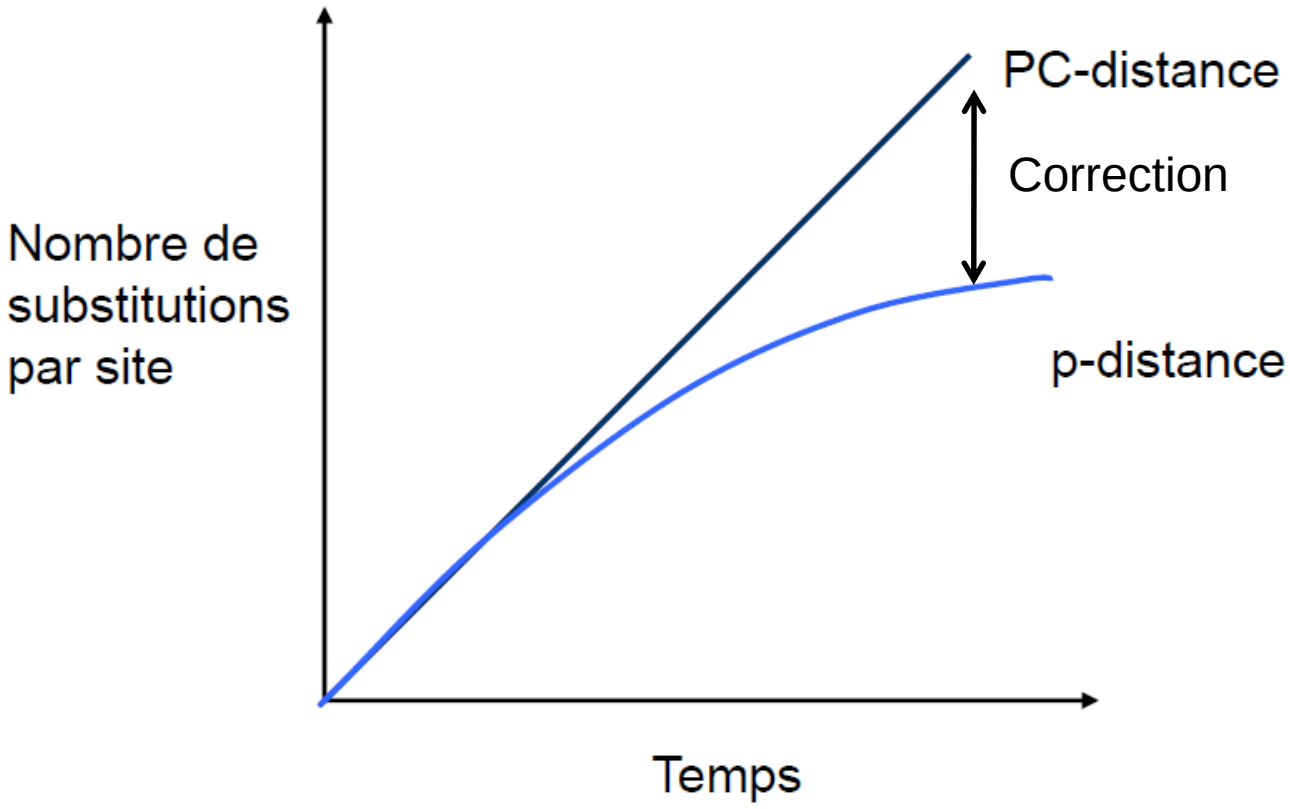
- Première estimation meilleure que la p-distance repose sur le concept de distribution de Poisson.
- Hypothèses :
  - tous les sites évoluent indépendamment et suivant un même processus
  - toutes les substitutions sont équiprobables
  - le taux de réversion est négligeable

$$d = -Log(1-p)$$

*p* = *p*-distance = distance observée

# 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.

| Modèle | Référence |
|---|---|
| PAM | Dayhoff 1978 |
| BLOSUM | Henikoff 1992 |
| JTT (réactualisation de la PAM) | Jones 1992 |
| WAG & LG | Whelan 2001, Le et Gascuel |
| Spécifiques (organelles etc..) | |

# Calcul des distances entre deux séquences protéiques

Modèle basés sur le maximum de vraisemblance

Problème avec le modèle PAM ou JTT : utilisation de séquences très similaires pour estimer les taux de substitutions. Pour séquences distantes inférés.

Utilisation du maximum de vraisemblance pour estimer les taux de substitutions des acides aminés.

Premières tentatives, modèles adaptés :

- aux séquences mitochondriales de vertébrés (mtREV).
- aux séquences mitochondriales de mammifères (mtMAM).
- aux séquences chloroplastiques (cpREV).

Modèle plus général proposé par Whelan et Goldman : modèle WAG

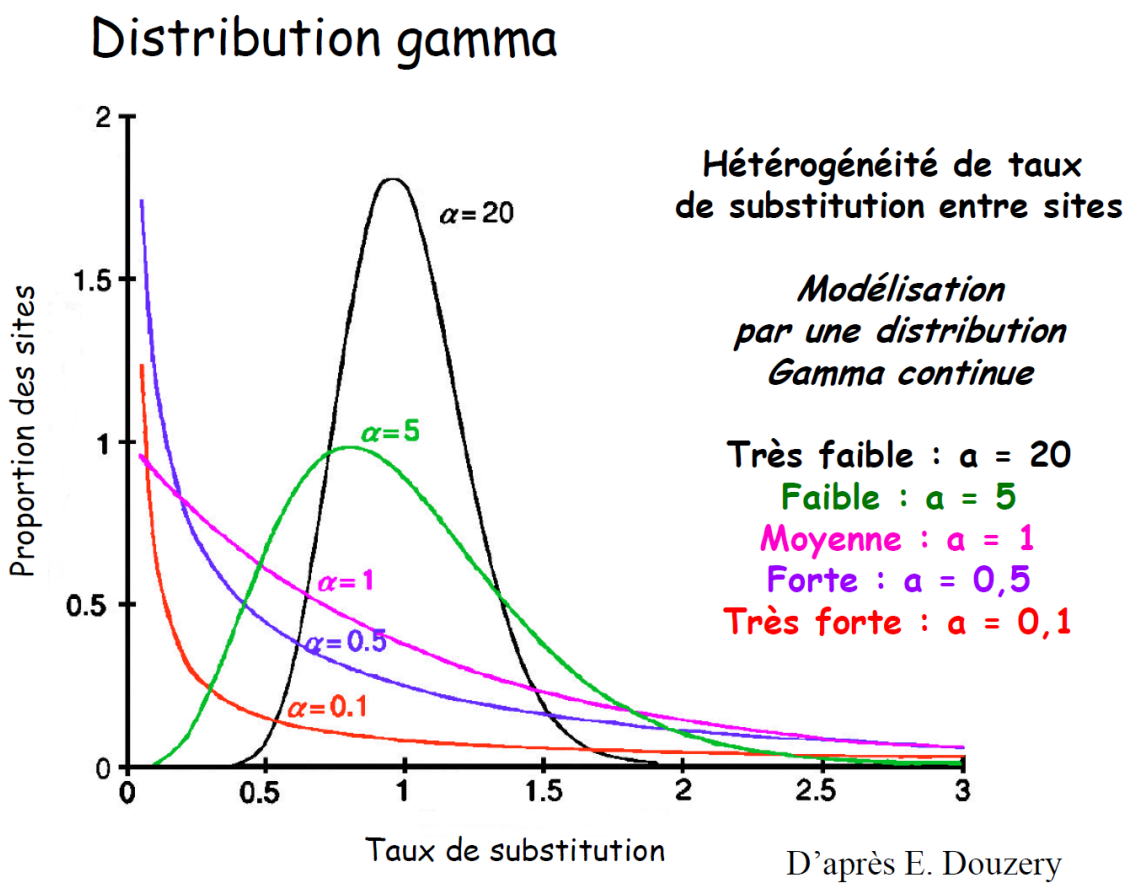
- utilise 182 familles de protéines homologues (3905 séquences).
- utilisation du modèle WAG permet d'obtenir des arbres dont la vraisemblance est significativement supérieure à ceux obtenus avec les modèles PAM ou JTT
- faiblesse du modèle : hypothèse d'uniformité (même vitesse d'évolution pour tous les sites)

Modèle proposé par Le et Gascuel : modèle LG extension du modèle WAG

- prise en compte de différentes vitesses d'évolution pour les sites
- construit à partir de 3912 familles comprenant au total 49637 séquences

# Correction des distances pour différentes vitesses d'évolution

Hypothèse des différents modèles évolutifs présentés : tous les sites évoluent à la même vitesse, or les contraintes fonctionnelles engendrent des taux d'évolution (*r*) différents selon les sites. Il a été démontré que ce taux *r* est modélisable par une loi Gamma (séquences nucléiques ou protéiques). Choix de la distribution Gamma : pas de justification biologique mais commodité mathématique car la forme de la distribution ne dépend que d'un seul paramètre $\alpha$.




D'après E. Douzery

Si $\alpha > 1 \rightarrow$ forme de cloche.
Plus $\alpha$ est grand, plus la variance de *r* diminue traduisant une faible hétérogénéité des taux de substitutions par rapport à la moyenne.

Si $\alpha \leq 1 \rightarrow$ forme de L.
Nombre important de sites avec un r proche de 0 (sites quasiment invariants).
Donc forte hétérogénéité dans les taux d'évolution.

$\alpha$ est estimé à partir des données.
Distribution Gamma est discrétisée (nombre de catégories pour *r* variant de 4 à 8).

# Correction des distances pour différentes vitesses d'évolution

La plupart des modèles vus précédemment peuvent intégrer dans leur calcul de la distance une correction par la loi Gamma.

Exemple séquences nucléiques : le modèle de Jukes et Cantor (JC89) qui s'identifie par JC89+Γ

Modèle JC89

$$d = -\frac{3}{4} Log\left(1 - \frac{4}{3} p_{dist}\right)$$

Modèle JC89+Γ

$$d = \frac{3}{4}\alpha\left[\left(1 - \frac{4}{3} p_{dist}\right)^{-1/\alpha} - 1\right]$$

Exemple séquences protéiques : le modèle de Poisson (Poisson+Γ)

Modèle Poisson

$$d = -Log(1 - p)$$

Modèle Poisson+Γ

$$d = \alpha\left[(1 - p)^{-1/\alpha} - 1\right]$$

# Correction des distances pour différentes vitesses d'évolution

Exemple de l'estimation du taux de substitution par site : chaîne $\alpha$ hémoglobine

| | P-distance | PC-distance | PC + Gamma-distance |
|---|---|---|---|
| Human/cow | 0.121 | 0.129 | 0.134 |
| Human/kangaroo | 0.186 | 0.205 | 0.216 |
| Human/carp | 0.486 | 0.665 | 0.789 |

PC = correction de poisson

# Choix des modèles évolutifs

# 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.

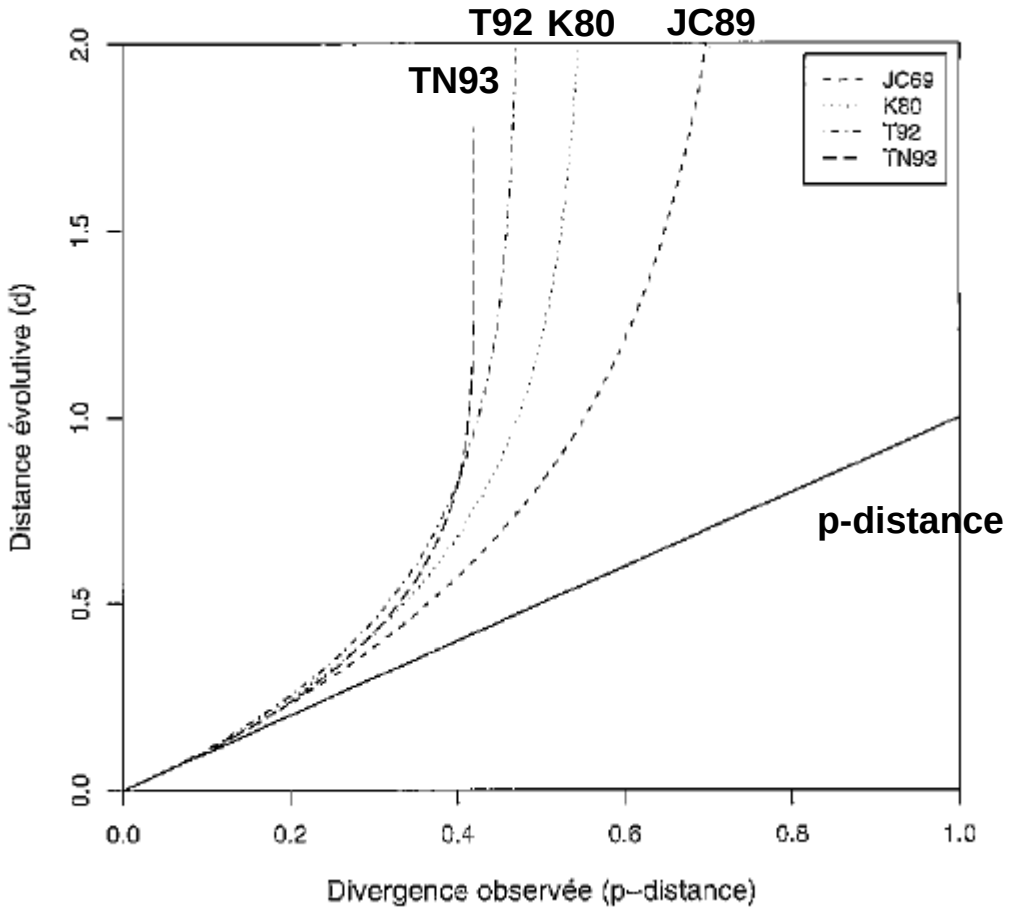
<u>Quelques règles simples :</u>

- construction d'une phylogénie à partir de gènes protéiques :
  - 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 (cf. figure suivante) donc si séquences très divergentes, fréquemment impossible de calculer les distances.
  - ✓ si même résultat avec deux modèles, utiliser le plus simple car la variance de la distance augmente avec le nombre de paramètres.
  - ✓ application de la correction Gamma que si nombre de sites utilisés important car nécessite d'estimer un paramètre supplémentaire (la forme $\alpha$ de la distribution).

# Choix d'un modèle évolutif

## Comparaison des différents modèles évolutifs de séquences d'ADN



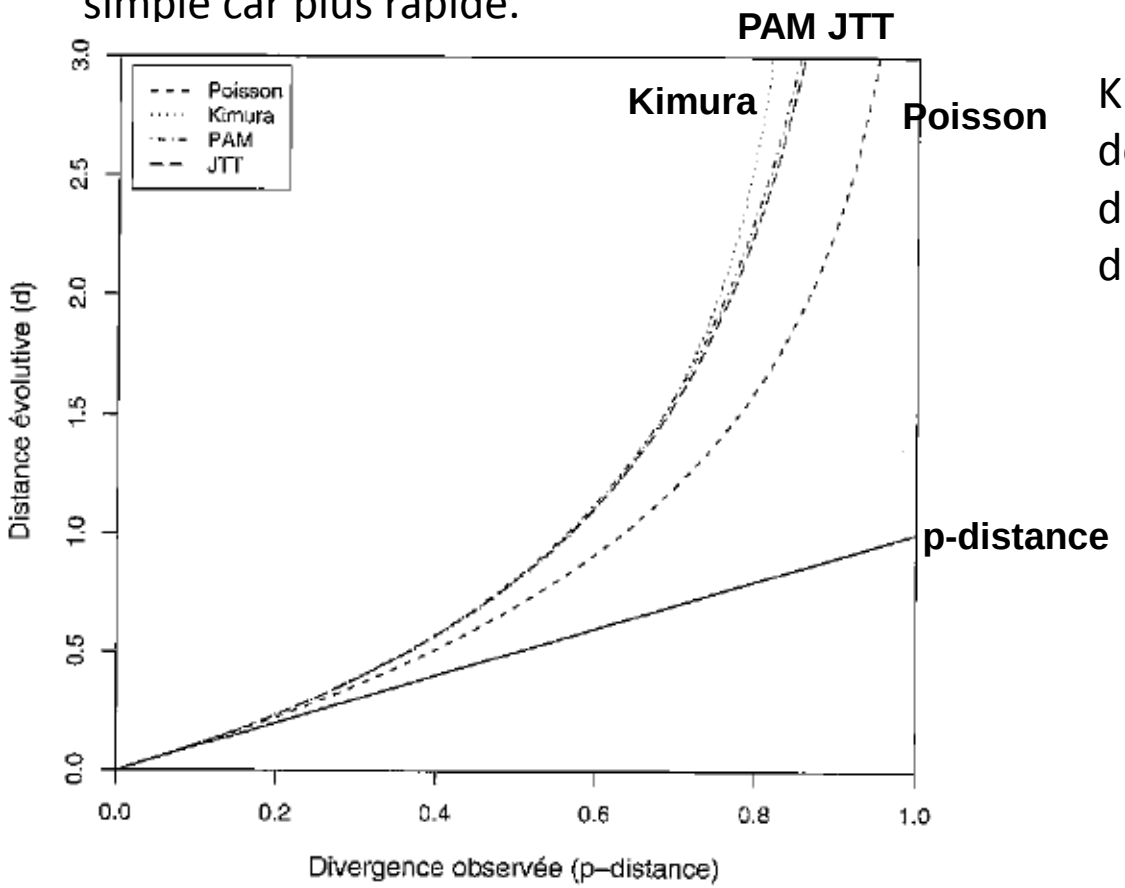

*Extrait de Perrière et Brochier-Armanet (2010) Concepts et méthodes en phylogénie moléculaire.*

Paramètres fixés :
taux *GC* ($\theta$) = 0.3
$\pi_A$ = 0.3, $\pi_T$ = 0.4, $\pi_C$ = 0.2 et $\pi_G$ = 0.1
$\kappa$ = transition/transversion = 4

- d ≤ 0.1 tous les modèles : même résultat (on peut utiliser modèle simple)
- 0.5 > d > 0.1 on peut utiliser JC89 ou K80, K80 préférable si séquence k >5
- 1 > d > 0.5 utilisation de modèles avec nombre de paramètres plus important. Les plus simples sous-estiment le nombre de substitutions (distance évolutive) .
- d > 1 pour beaucoup de paires de séquences, prudence sur la fiabilité de l'arbre. Eliminer les sites saturés manuellement ou avec méthodes appropriées.

# Choix d'un modèle évolutif

- Utilisation de séquences protéiques
  - 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.




Kimura : modèle simplifié donnant une estimation de la distance PAM en fonction de la p-distance $p$ :

$$d = -Log\left(1 - p - 0.2p^2\right)$$

# Choix d'un modèle évolutif

Grand nombre de modèles d'évolution dont certains très complexes et intégrant un grand nombre de paramètres.

Problème : la précision de l'estimation des paramètres peut être mauvaise notamment quand peu de données (nombre de séquences et/ou de sites).

- Primordial de choisir le modèle qui est le plus en adéquation avec les données.

- Etape indispensable à toute analyse phylogénétique rigoureuse.

Les tests de vraisemblance sont des méthodes bien adaptées qui permettent non seulement de déterminer les hypothèses qui expliquent le mieux le jeu de données mais aussi de comparer des hypothèses.

- Test du rapport de vraisemblance appelé LRT pour Likelihood Ratio Test
- Akaike Information Criterion (AIC).

# Likelihood Ratio Test

Nécessite que les modèles que l'on veut tester soit imbriqués (du plus simple au plus complexe)

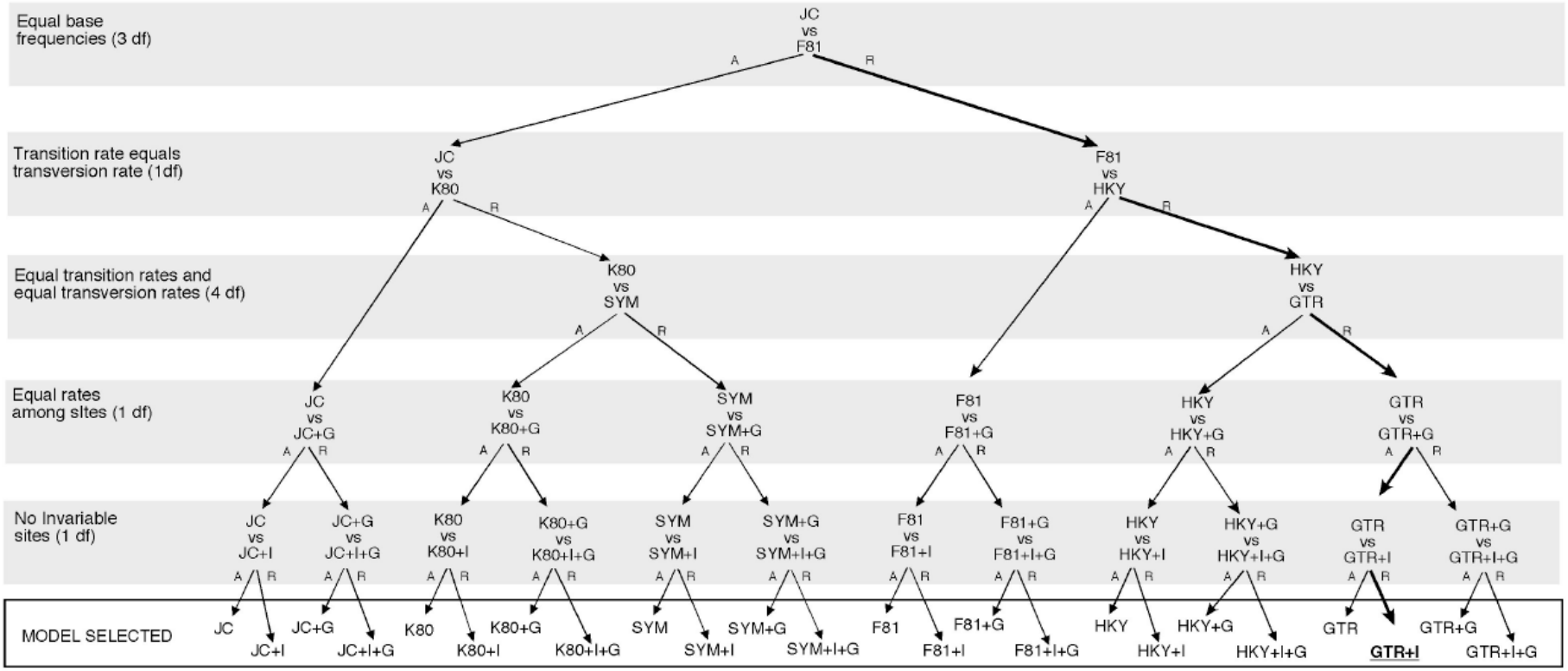



Figure 17. Example of a particular forward hierarchy of likelihood ratio tests for 24 models. At any level the null hypothesis (model on top) is either accepted (A) or rejected (R). In this example the model selected is GTR+I.

*(Extrait du manuel de JModelTest)*

# Likelihood Ratio Test

Ce test est utilisé quand l'on désire comparer deux arbres qui ont la même topologie mais qui ont été obtenus avec des modèles d'évolution différents. On compare deux modèles :

- ✓ Le modèle $M_0$ (le plus simple, *i.e.* qui a le plus petit nombre de paramètres $k_0$) qui correspondra à l'hypothèse nulle H0.
- ✓ Le modèle $M_1$ (le plus complexe, $k_1$ paramètres, $k_1 > k_0$) qui correspondra à l'hypothèse alternative H1.

Le rapport de vraisemblance est donné par : $$\Delta = 2\ln\left[\frac{L(\Theta_1)}{L(\Theta_0)}\right] = 2\left[\ln L(\Theta_1) - \ln L(\Theta_0)\right]$$

$\Delta$ suit une loi du $\chi^2$ à $k_1 - k_0$ d.d.l., soit le nombre de paramètre du modèle $M_1$ à contraindre pour se ramener au modèle $M_0$. Le modèle nul sera rejeté si $\Delta$ est supérieur au niveau de confiance fixé par l'utilisateur

Critiques majeures de ce test :

- la sélection des modèles testés dépend du parcours de l'arbre hiérarchique. Par exemple, si le modèle le plus adapté est le F81+I+G, il ne pourra pas être testé si à l'étape précédente on a rejeté le modèle F81 au profit du modèle HKY. Pour palier à ce problème, on peut faire des tests dynamiques.

# Akaike Information Criterion (AIC)

C'est un estimateur qui correspond à la minimisation de la distance attendue entre un modèle vrai et son estimation. Les modèles correspondant aux valeurs minimales de l'AIC sont considérés comme les plus appropriés pour la reconstruction. Une même topologie de référence doit être utilisée pour tester les différents modèles. L'AIC permet de tester des modèles sans que ceux-ci soient imbriqués.

$$AIC = -2\ln L(\Theta) + 2k$$

*k* = nombre de paramètres libres du modèle

L'AIC apparaît biaisé pour les modèles riches en paramètres comparativement au LRT.

Si la taille *n* du jeu de données est petite comparée au nombre de paramètres *k* du modèle ($n/k < 40$) l'utilisation de l'AIC corrigé AICc est recommandée.

$$AIC_c = AIC + \frac{2k(k+1)}{n-k-1}$$

Deux logiciels : `JModelTest` pour les séquences d'acides nucléiques et `ProtTest`

# Choix de la méthode

# Choix de la méthode

Il existe plusieurs méthodes pour inférer des reconstructions phylogénétiques.
Quatre familles principales:

| Données | Méthode | Principes |
|---|---|---|
| Matrice de distances | Neighbor-Joining (NJ) | à partir des distances entre chaque paire de séquences, recherche l'arbre qui représente au mieux les distances évolutives entre les données. |
| Etat des caractères | Maximum de Parcimonie | à partir d'un ensemble de caractères choisis, recherche l'arbre qui minimise le nombre de changements permettant d'expliquer les données. |
| | Maximum de vraisemblance (ML) | à partir des probabilités de l'apparition des transformation d'un état de caractères en un autre, recherche l'arbre le plus vraisemblable en fonction du modèle évolutif considéré. |
| | Inférence bayésienne | |

# Méthode de distance

- ✓ Introduites en phylogénie dans les années 1960

- ✓ Deux grands types de méthodes :
  - ➢ celles basées sur des algorithmes de clustering (UPGMA (n'est plus utilisé aujourd'hui dans le cas des séquences ).
  - ➢ celles basées sur des critères d'optimisation (moindre carré, Neighbor-Joining).

Méthode de distances actuellement la plus utilisée : la Neighbor-Joining (NJ) et ses variantes.

L'objectif des méthodes de distance : distances d'arbre (ou patristiques) représentent au mieux les distances présentes dans la matrice de distance.

# Méthodes de distance : la NJ

Saitou et Nei (1987) *Mol. Biol. Evol.*, 4, 406-25

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 longueur des branches (arbre de longueur minimum).

NJ : algorithme qui à chaque étape sélectionne la paire de taxon qui une fois agglomérée produit l'arbre minimum. Produit un arbre non enraciné.

Principe

Alignement de séquences

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

Choix d'un modèle évolutif
Calcul d'une matrice de distances

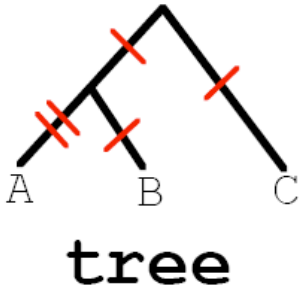
Matrice des distances évolutives entre paires de séquences

| | 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)



# Méthodes de distance : la NJ

Variante : la BIONJ (Gascuel, 1997). La BIONJ apporte de 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. Applicable 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**
- ✓ Souvent utilisée pour chercher des arbres qui vont servir de point de départ pour des méthodes plus coûteuses en temps calcul comme la méthode du maximum de vraisemblance
- ✓ Elle peut être appliquée sur n'importe quel type de distances évolutives.
- ✓ Peut conduire à des distances négatives notamment pour les branches terminales mais l'application de la contrainte de non-négativité permet de s'affranchir du problème.
- ✓ 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.

# Méthode du maximum de Parcimonie

Principe : l'estimation la plus plausible d'un arbre évolutif est celle qui fait appel à la quantité minimale d'évolution *i.e.*, celui qui implique le moins d'évènements évolutifs pour expliquer les données

A utiliser plutôt sur des données de présence/absence (données morphologiques, présence/absence de gènes…

# Exemple d'une reconstruction cladistique

Extrait de P. Tassy, Pour la SCIENCE Dossier L'Evolution (Janvier 1997, page 74)
Exemple à partir de données morphologiques

Distribution de cinq caractères crâniens chez les proboscidiens et les siréniens

- 1) remplacement dentaire dit "horizontal"
- 2) orbite antérieure
- 3) forme particulière de l'os tympanique
- 4) configuration du trou auditif externe
- 5) fosses nasales reculées au-dessus ou en arrière des orbites

Les cases contenant un 1 indiquent l'état transformé du caractère.

| Caractères / Espèces | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| Autres Mammifères | 0 | 0 | 0 | 0 | 0 |
| Lamantin | 1 | 1 | 1 | 0 | 0 |
| Dugong | 0 | 1 | 1 | 0 | 0 |
| Moeritherium | 0 | 1 | 0 | 1 | 0 |
| Phomia | 0 | 0 | 0 | 1 | 1 |
| Eléphants | 1 | 1 | 0 | 1 | 1 |

Représentants des siréniens :
- Lamantin
- Dugong

Représentants des proboscidiens :
- Moeritherium (fossile 40 millions d'années)
- Phomia (fossile 30 millions d'années)
- Eléphant

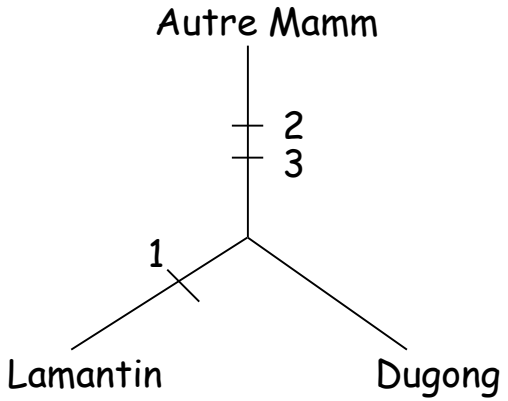
Groupe externe :
- Autres mammifères

# Exemple d'une reconstruction cladistique

| Caractères / Espèces | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| Autres Mammifères | 0 | 0 | 0 | 0 | 0 |
| Lamantin | 1 | 1 | 1 | 0 | 0 |
| Dugong | 0 | 1 | 1 | 0 | 0 |
| Moeritherium | 0 | 1 | 0 | 1 | 0 |
| Phomia | 0 | 0 | 0 | 1 | 1 |
| Eléphants | 1 | 1 | 0 | 1 | 1 |

A partir de ce tableau on va déterminer les relations de parenté par une méthode de parcimonie.

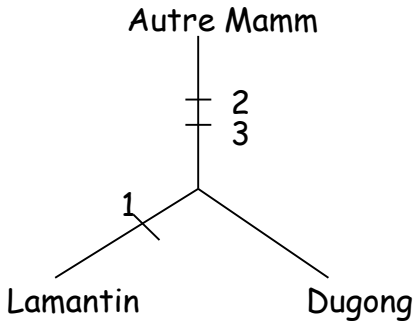
**1ère étape** : On construit un arbre avec les 3 premières espèces et on reporte sur les branches le numéro du caractère transformé.

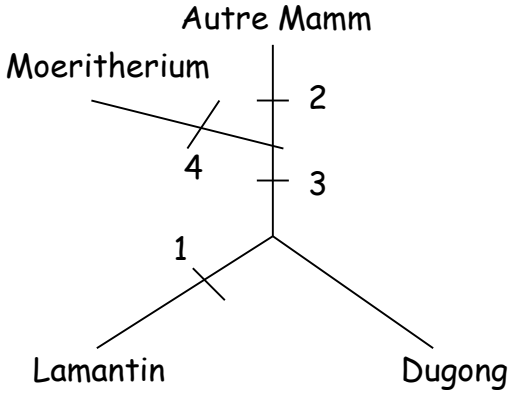


# Exemple d'une reconstruction cladistique

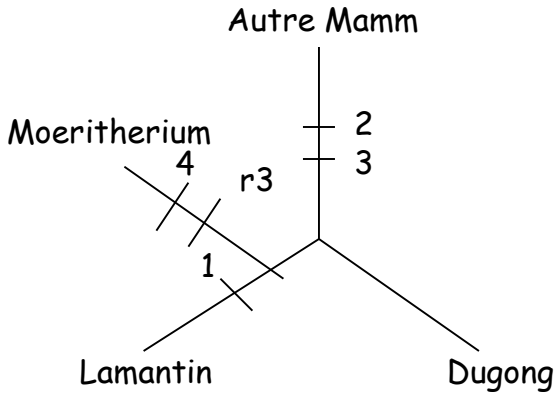
**2ème étape** : on rajoute la 4ème espèce, 3 possibilités sur chacune des 3 branches.

| Espèces \ Caractères | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| Autres Mammifères | 0 | 0 | 0 | 0 | 0 |
| Lamantin | 1 | 1 | 1 | 0 | 0 |
| Dugong | 0 | 1 | 1 | 0 | 0 |
| Moeritherium | 0 | 1 | 0 | 1 | 0 |
| Phomia | 0 | 0 | 0 | 1 | 1 |
| Eléphants | 1 | 1 | 0 | 1 | 1 |




Arbre 1 : 4 changements

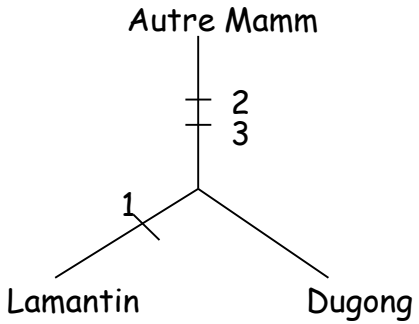


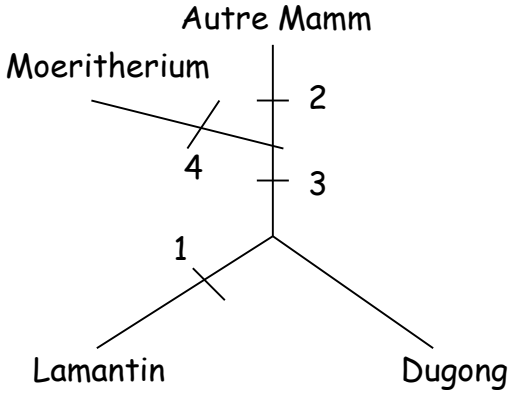

Arbre 2 : 5 changements

# Exemple d'une reconstruction cladistique

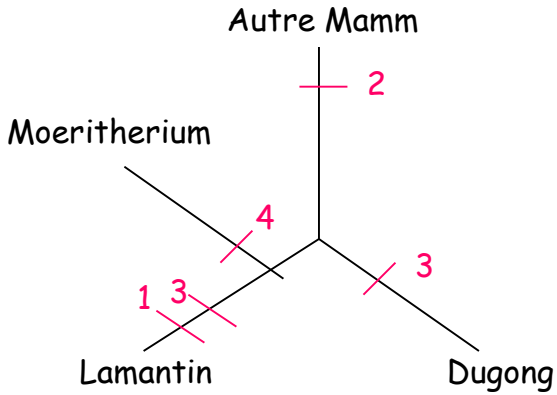
**2ème étape** : on rajoute la 4ème espèce, 3 possibilités sur chacune des 3 branches.

| Espèces \ Caractères | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| Autres Mammifères | 0 | 0 | 0 | 0 | 0 |
| Lamantin | 1 | 1 | 1 | 0 | 0 |
| Dugong | 0 | 1 | 1 | 0 | 0 |
| Moeritherium | 0 | 1 | 0 | 1 | 0 |
| Phomia | 0 | 0 | 0 | 1 | 1 |
| Eléphants | 1 | 1 | 0 | 1 | 1 |

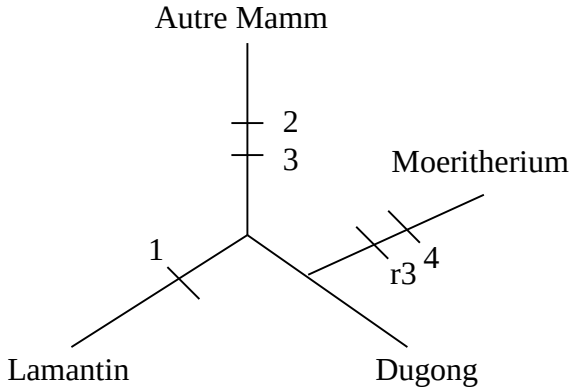



Arbre 1 : 4 changements




Arbre 2 : 5 changements

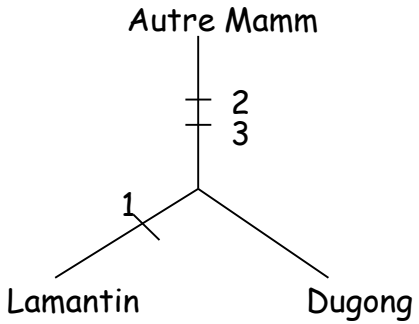



Arbre 3: 5 changements

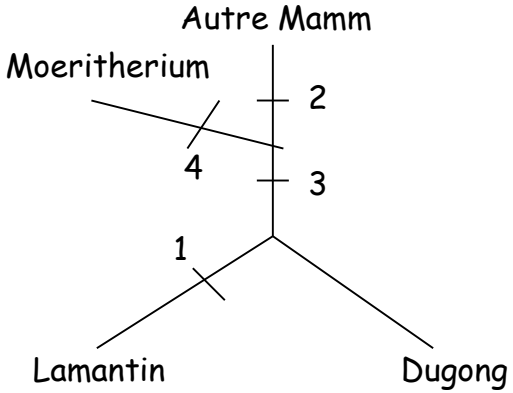
# Exemple d'une reconstruction cladistique

**2ème étape** : on rajoute la 4ème espèce, 3 possibilités sur chacune des 3 branches.

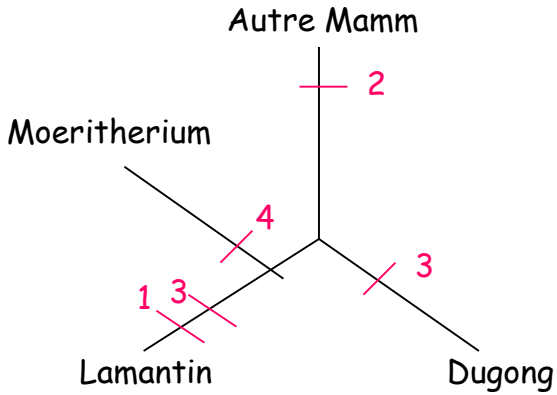
| Espèces \ Caractères | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| Autres Mammifères | 0 | 0 | 0 | 0 | 0 |
| Lamantin | 1 | 1 | 1 | 0 | 0 |
| Dugong | 0 | 1 | 1 | 0 | 0 |
| Moeritherium | 0 | 1 | 0 | 1 | 0 |
| Phomia | 0 | 0 | 0 | 1 | 1 |
| Eléphants | 1 | 1 | 0 | 1 | 1 |



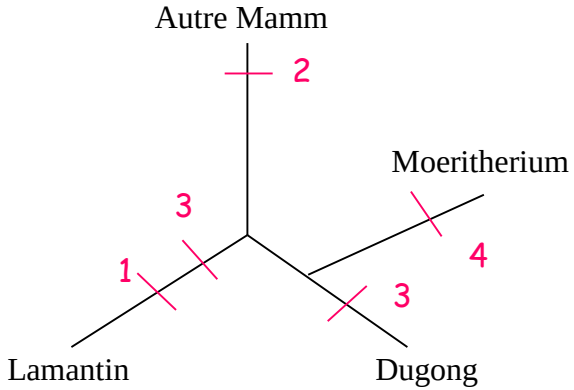

2 explications avec même nb de changements pour la même topologie




Arbre 1 : 4 changements




Arbre 2 : 5 changements

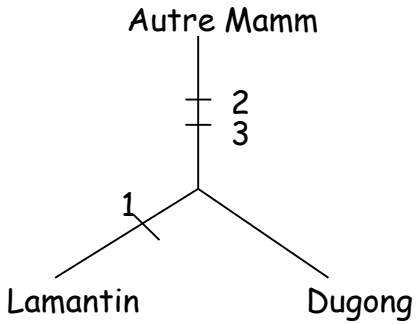



Arbre 3: 5 changements

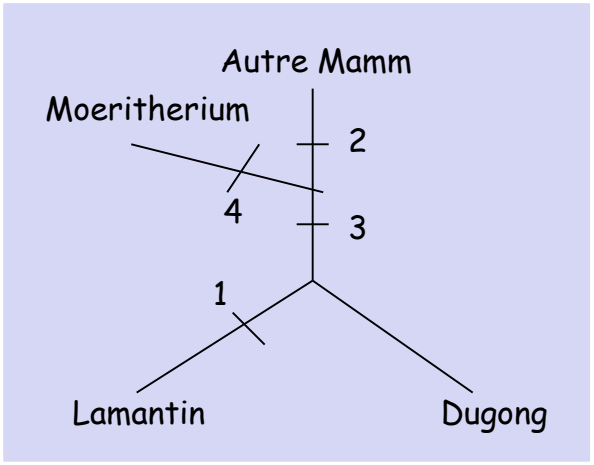
# Exemple d'une reconstruction cladistique

**2ème étape** : on rajoute la 4ème espèce, 3 possibilités sur chacune des 3 branches.

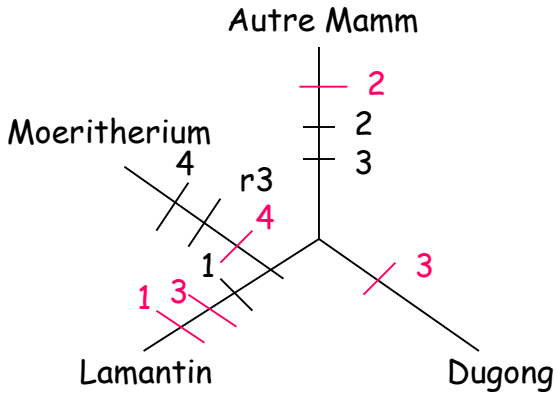
| Espèces \ Caractères | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| Autres Mammifères | 0 | 0 | 0 | 0 | 0 |
| Lamantin | 1 | 1 | 1 | 0 | 0 |
| Dugong | 0 | 1 | 1 | 0 | 0 |
| Moeritherium | 0 | 1 | 0 | 1 | 0 |
| Phomia | 0 | 0 | 0 | 1 | 1 |
| Eléphants | 1 | 1 | 0 | 1 | 1 |



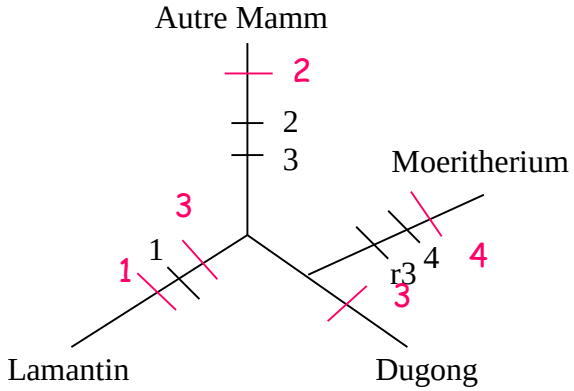

2 explications avec même nb de changements pour la même topologie




Arbre 1 : 4 changements




Arbre 2 : 5 changements



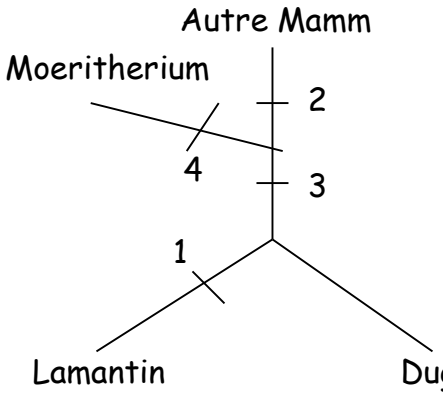

Arbre 3: 5 changements

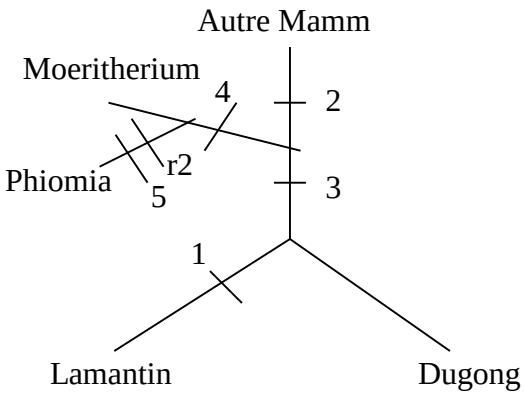
**Rouge : caractère 3 convergence; Noire : caractère 3 Réversion**

# Exemple d'une reconstruction cladistique

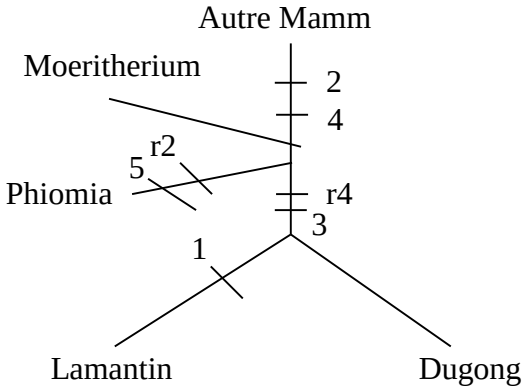
**3ème étape** : on rajoute la 5ème espèce, 5 possibilités sur chacune des 5 branches.

| Caractères / Espèces | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| Autres Mammifères | 0 | 0 | 0 | 0 | 0 |
| Lamantin | 1 | 1 | 1 | 0 | 0 |
| Dugong | 0 | 1 | 1 | 0 | 0 |
| Moeritherium | 0 | 1 | 0 | 1 | 0 |
| Phomia | 0 | 0 | 0 | 1 | 1 |
| Eléphants | 1 | 1 | 0 | 1 | 1 |




Arbre 1 : 6 changements

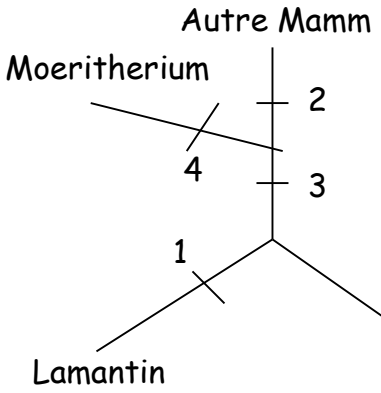


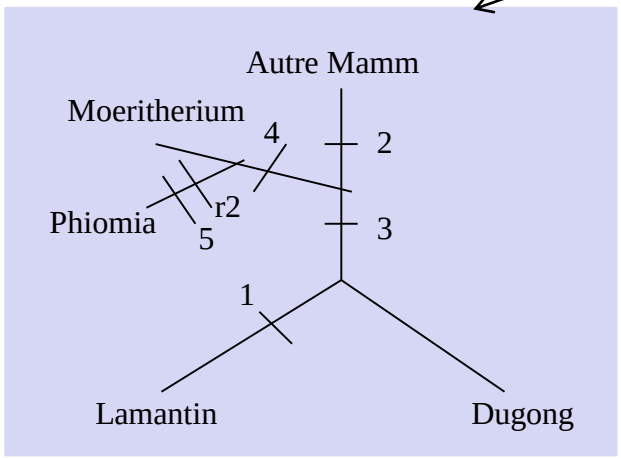

Arbre 2 : 7 changements

# Exemple d'une reconstruction cladistique

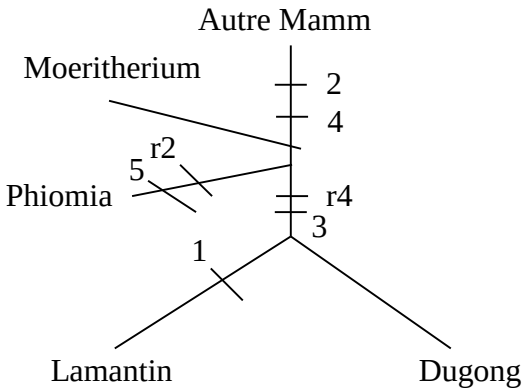
**3ème étape** : on rajoute la 5ème espèce, 5 possibilités sur chacune des 5 branches.

| Caractères / Espèces | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| Autres Mammifères | 0 | 0 | 0 | 0 | 0 |
| Lamantin | 1 | 1 | 1 | 0 | 0 |
| Dugong | 0 | 1 | 1 | 0 | 0 |
| Moeritherium | 0 | 1 | 0 | 1 | 0 |
| Phomia | 0 | 0 | 0 | 1 | 1 |
| Eléphants | 1 | 1 | 0 | 1 | 1 |




Arbre 1 : 6 changements



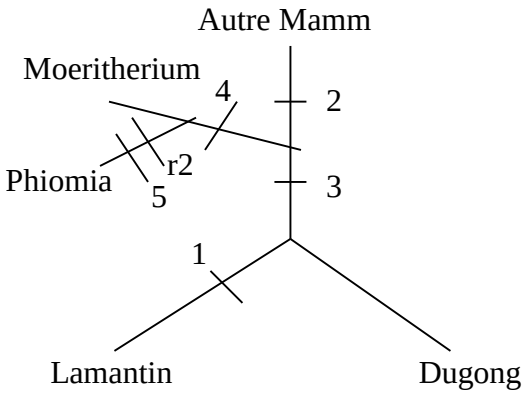

Arbre 2 : 7 changements

Le premier arbre est le plus parcimonieux (vous pouvez tester les autres topologies) et sera utilisé pour ajouter la dernière espèce.

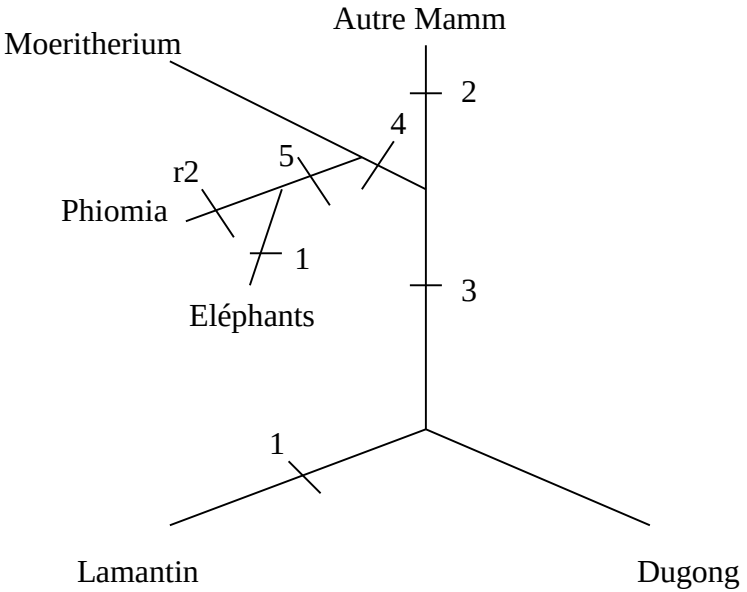
# Exemple d'une reconstruction cladistique

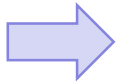
**4ème étape** : on rajoute la 6ème espèce, 7 possibilités sur chacune des 7 branches.

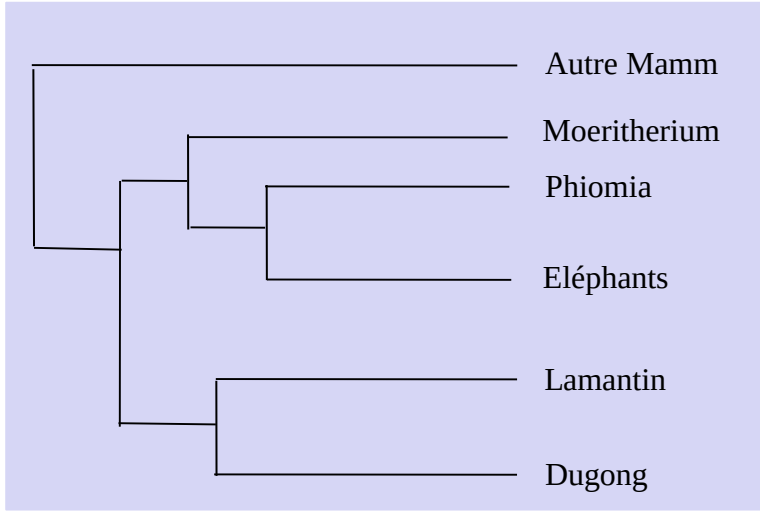
| Caractères / Espèces | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| Autres Mammifères | 0 | 0 | 0 | 0 | 0 |
| Lamantin | 1 | 1 | 1 | 0 | 0 |
| Dugong | 0 | 1 | 1 | 0 | 0 |
| Moeritherium | 0 | 1 | 0 | 1 | 0 |
| Phomia | 0 | 0 | 0 | 1 | 1 |
| Eléphants | 1 | 1 | 0 | 1 | 1 |




Arbre le plus parcimonieux : 7 changements





# Méthode du maximum de Parcimonie

Dans le cadre de données séquences :

Hypothèses :

- Les séquences ont évolué à partir d'une séquence ancestrale commune au travers d'un processus de mutation-sélection.
- Les différents sites évoluent indépendamment.
- Les lignées se différencient les unes des autres de façon autonome.
- La vitesse d'évolution est lente et constante au cours du temps

Cette approche :

- ✓ ne prend en compte que les sites informatifs, *i. e.*, les sites qui permettent de discriminer une topologie par rapport aux autres. Sont donc exclus les sites invariants et les sites variables impliquant le même nombre de substitutions quelle que soit la topologie.
- ✓ ne fait pas de correction pour les substitutions multiples.
- ✓ ne donne aucune information sur la longueur des branches
- ✓ ne fait aucune distinction entre les changements évolutifs ce qui est irréaliste.

# Méthode du maximum de Parcimonie

Conclusion :

- 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 le construction de l'arbre. On n'obtiendra pas forcément le même arbre si on change l'ordre des séquences. Pour pallier à 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 consistante pouvant conduire sous certaines conditions à des résultats erronés (démontré par Felsenstein (1978, Syst. Zool, 27, 401-10)).

# Méthode du maximum de vraisemblance

Les objets étudiés en phylogénie moléculaire, les séquences, sont le résultat d'une histoire évolutive qui nous est inconnue mais que l'on peut essayer de reconstruire sachant que cette histoire intègre plusieurs composantes :

- les relations de parenté entre les séquences représentées par la topologie *t* de l'arbre.
- la quantité d'évolution qui s'est écoulée entre chacune des lignées étudiées et qui est représentée par l'ensemble des longueurs des branches $b_i$ .
- le processus qui gouverne l'évolution de ces séquences, le modèle évolutif considéré composé lui-même d'un certain nombre de paramètres $\theta$.

Les valeurs des différents paramètres ne sont que très rarement connues.
On va donc devoir estimer, en fonction des données actuelles, *i.e.*, les séquences, cet ensemble $\Theta$ de paramètres (la topologie *t*, les longueurs de branches $b_i$ et les paramètres $\theta$ du modèle évolutif).

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 de $\Theta$ qui maximisent la probabilité d'observer les séquences actuelles.

# Méthode du maximum de vraisemblance

Hypothèses

- Le processus de substitution suit un modèle probabiliste dont on connaît l'expression mathématique, mais pas les valeurs numériques.
- Les sites évoluent indépendamment les uns des autres (restrictive).
- Les sites évoluent selon la même loi (on peut affaiblir cette hypothèse).
- Les taux de substitution ne changent pas au cours du temps le long d'une branche. Ils peuvent varier entre branches, c'est-à-dire, que l'évolution des séquences est indépendante d'une lignée à l'autre.

Deux applications du maximum de vraisemblance en phylogénie :

- Estimer la vraisemblance d'un ensemble d'hypothèses.
- Rechercher parmi tous les ensembles $\Theta$ de valeurs de paramètres possibles celui qui possède la vraisemblance la plus élevée. Comme on a vu que la topologie *t* faisait partie de ces paramètres, cela permet de rechercher l'arbre qui possède la plus forte vraisemblance étant donné la valeur des autres paramètres.

# Méthode du maximum de vraisemblance

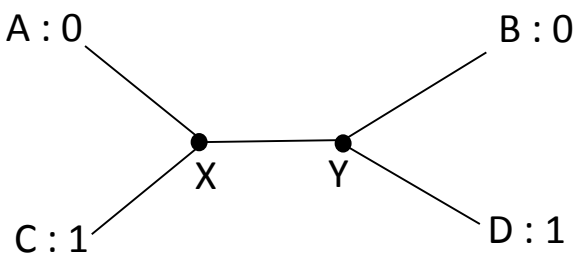
Etant donné un modèle d'évolution  trouver l'arbre maximisant la probabilité d'obtenir les séquences actuelles.

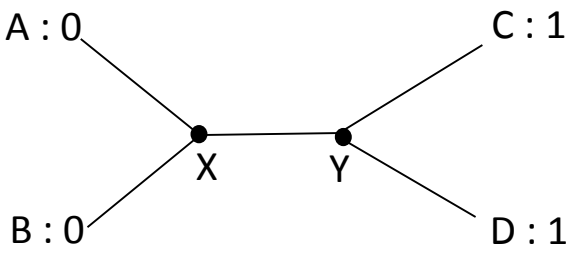
Considérons un exemple simple :

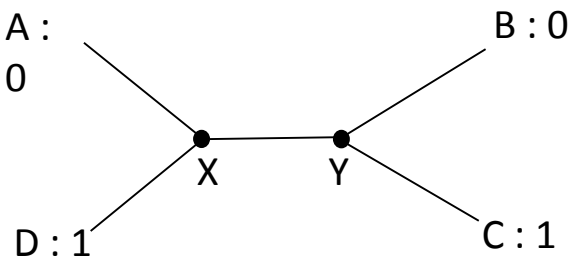
- 4 OTU (A, B, C, D) et un caractère qui peut avoir deux états 0 ou 1.
- On observe l'état de caractère 0 pour A et B et l'état 1 pour C et D.
- On a le modèle d'évolution suivant :

| | 0 | 1 |
|---|---|---|
| 0 | 0.8 (4/5) | 0.2 (1/5) |
| 1 | 0.2 (1/5) | 0.8 (4/5) |

Il existe trois  arbres possibles avec une arête interne :

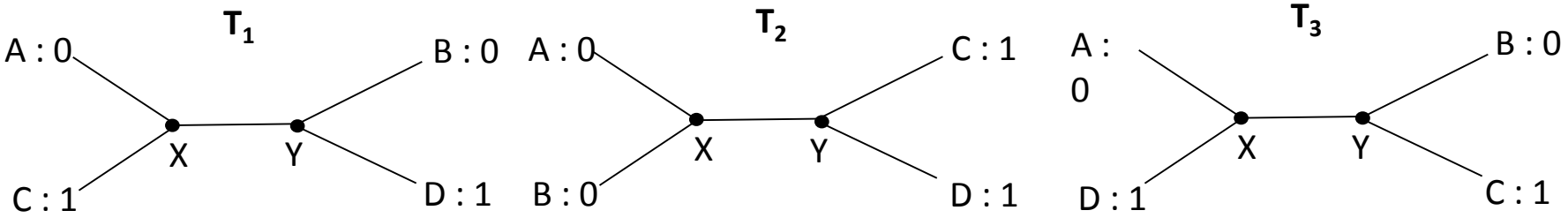


# Méthode du maximum de vraisemblance

On va calculer la vraisemblance L(T) (Likelihood) de chaque topologie :




Les topologies $T_1$ et $T_3$ sont équivalentes, donc on calcule uniquement L(T1).

| X | Y | $L(T_1)$ | $L(T_2)$ |
|---|---|---|---|
| 0 | 0 | $(4/5)^3(1/5)^2$ | $(4/5)^3(1/5)^2$ |
| 0 | 1 | $(4/5)^2(1/5)^3$ | $(4/5)^4(1/5)^1$ |
| 1 | 0 | $(4/5)^2(1/5)^3$ | $(4/5)^0(1/5)^5$ |
| 1 | 1 | $(4/5)^3(1/5)^2$ | $(4/5)^3(1/5)^2$ |
| Somme | | 32/625=0.0512 | 77/625=0.1232 |

**L'arbre $T_2$ est plus vraisemblable que l'arbre $T_1$**

# Méthode du maximum de vraisemblance

Le cas d'une topologie calculée sur des séquences nucléiques :

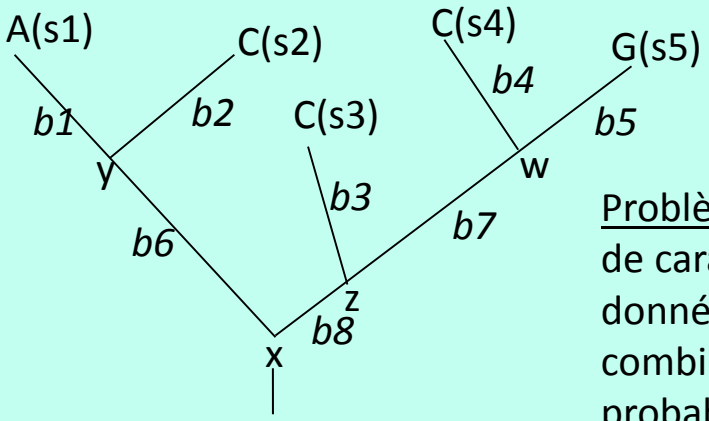
Première étape :

- calculer la vraisemblance à un site quelconque, c'est-à-dire la probabilité que les hypothèses soient à l'origine des états de caractères observés à ce site.

Soit un ensemble de données *D* composé par :

un arbre enraciné donné obtenu sur 5 séquences nucléiques
un site donné *i*
un jeu de longueurs des branches

Hypothèse T: les bases observées à ce site *i* ont évolué le long de cet arbre.
Vraisemblance de T par rapport à *D(i)* : probabilité que les données correspondent à l'hypothèse




$$L(T) = Prob(D(i)|T)$$

Problème : pour calculer cette probabilité, il faut connaître les états de caractères présents aux nœuds internes et à la racine. Or ces données sont inconnues. Il faut donc évaluer, pour chaque combinaisons d'états de caractères présents aux nœuds internes, la probabilité qu'ils aient conduits aux bases actuelles observées (feuilles). Par exemple, quelle est la probabilité d'observer A(s1), C(s2), C(s3), C(s4) et G(s5) sachant que x = A, y = A, z = A et w = A.

# Méthode du maximum de vraisemblance

Si nous prenons maintenant en compte les *m* sites alignés, comme nous avons comme hypothèse que les sites évoluent de façon indépendante, la vraisemblance de la topologie par rapport aux données D est donnée par le produit des vraisemblances calculées pour chaque site :

$$L = Prob(D|T) = \prod_{i=1}^{m} Prob(D_{(i)}|T)$$

La valeur de *L* correspond à la probabilité que les séquences aient évoluées d'après l'arbre T.

# Méthode du maximum de vraisemblance

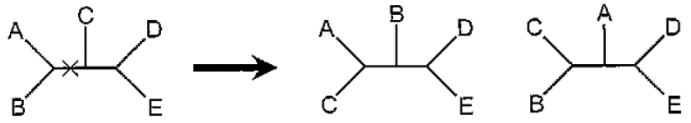
- C'est la méthode la mieux justifiée au plan théorique.
- Des expériences de simulation de séquences ont montré que cette méthode est supérieure aux autres dans la plupart des cas.
- Mais c'est une méthode très lourde en calculs.
- Il est presque toujours impossible d'évaluer tous les arbres possibles 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.

- Logiciels : PhyML (http://atgc.lirmm.fr/phyml/)
  RaXML (http://sco.h-its.org/exelixis/web/software/raxml/index.html)
  PHYLIP (dnaml, protml)

# Techniques de réajustement

Techniques de réajustement les plus courantes :

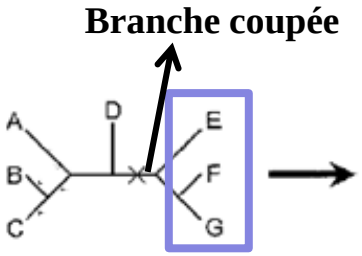
❖ Nearest Neighbor Interchange (NNI) : examiner les arbres qui se trouve à une distance topologique de 2.

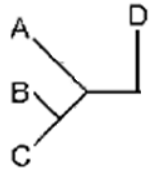



*Exemple extrait de Perrière et Brochier-Armanet (2010) Concepts et méthodes en phylogénie moléculaire*

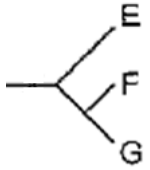
Réarrangement portant sur la branche interne marquée par x.
Seulement deux topologies à une distance topologique de 2

❖ Subtree Pruning et Regrafting (SPR) : l'arbre est coupée au niveau d'une branche ce qui conduit à deux sous-arbres : l'arbre élagué (pruned tree) et l'arbre résiduel. La méthode consiste à rebrancher l'arbre élagué sur chacune des branches de l'arbre résiduel et à calculer pour chaque topologie le critère d'optimisation.




Arbre résiduel

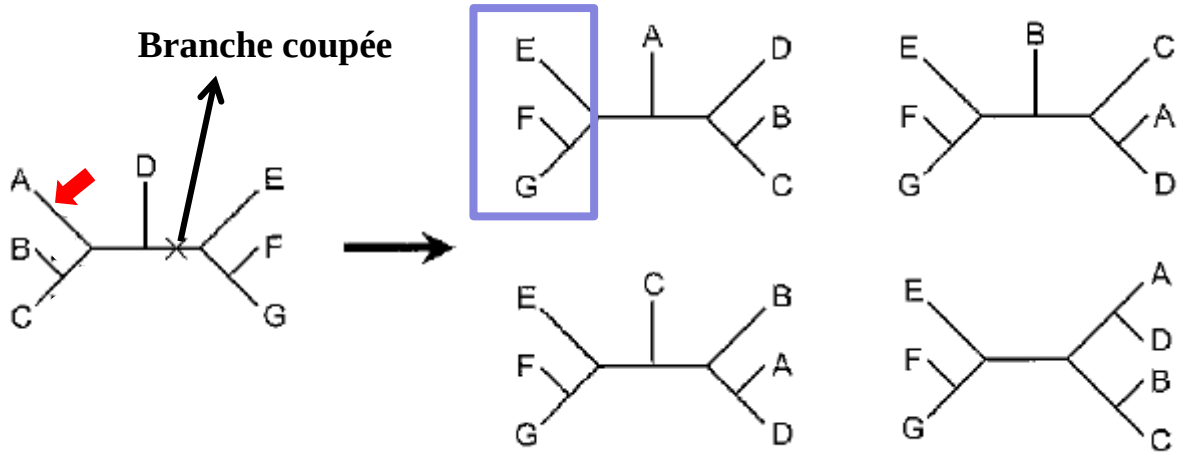



Arbre élagué

*Exemple extrait de Perrière et Brochier-Armanet (2010) Concepts et méthodes en phylogénie moléculaire*

# Techniques de réajustement

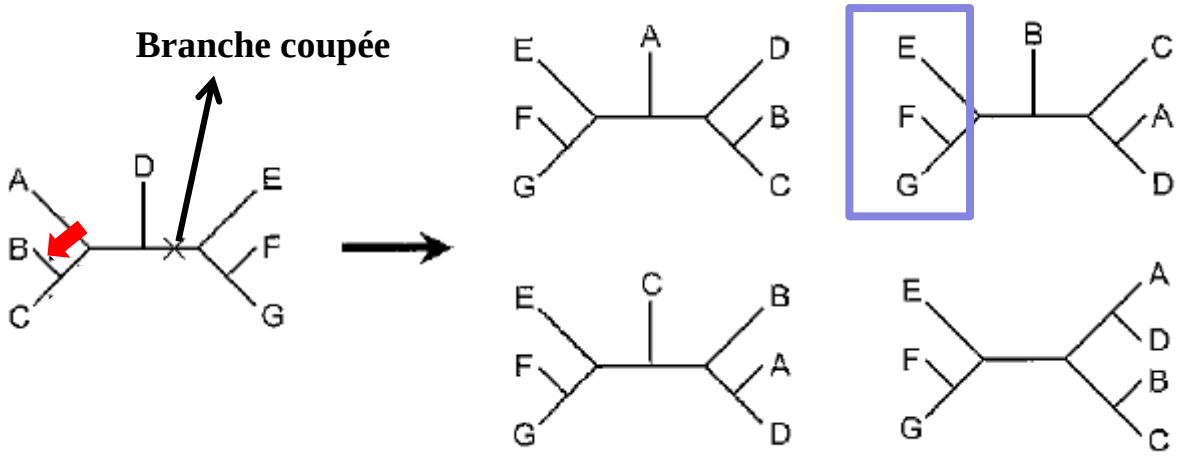
❖ Subtree Pruning et Regrafting (SPR) : l'arbre est coupée au niveau d'une branche ce qui conduit à deux sous-arbres : l'arbre élagué (pruned tree) et l'arbre résiduel. La méthode consiste à rebrancher l'arbre élagué sur chacune des branches de l'arbre résiduel et à calculer pour chaque topologie le critère d'optimisation.




*Exemple extrait de Perrière et Brochier-Armanet (2010) Concepts et méthodes en phylogénie moléculaire*

# Techniques de réajustement

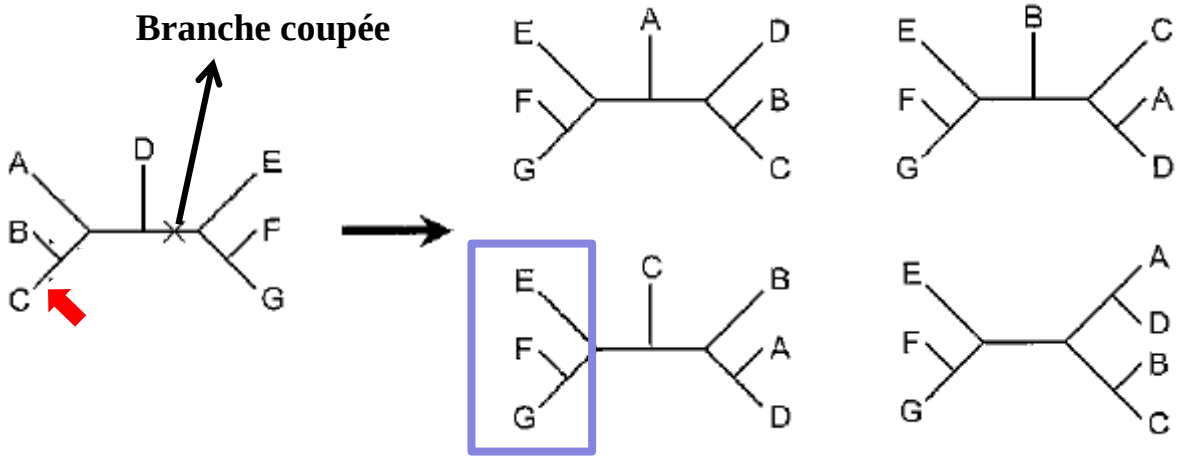
❖ Subtree Pruning et Regrafting (SPR) : l'arbre est coupée au niveau d'une branche ce qui conduit à deux sous-arbres : l'arbre élagué (pruned tree) et l'arbre résiduel. La méthode consiste à rebrancher l'arbre élagué sur chacune des branches de l'arbre résiduel et à calculer pour chaque topologie le critère d'optimisation.




*Exemple extrait de Perrière et Brochier-Armanet (2010) Concepts et méthodes en phylogénie moléculaire*

# Techniques de réajustement

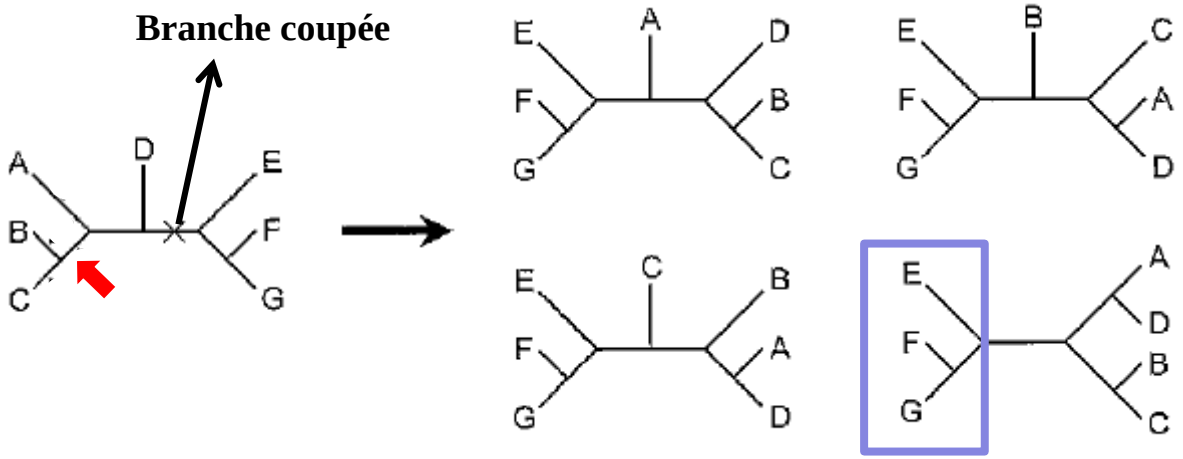
❖ Subtree Pruning et Regrafting (SPR) : l'arbre est coupée au niveau d'une branche ce qui conduit à deux sous-arbres : l'arbre élagué (pruned tree) et l'arbre résiduel. La méthode consiste à rebrancher l'arbre élagué sur chacune des branches de l'arbre résiduel et à calculer pour chaque topologie le critère d'optimisation.




*Exemple extrait de Perrière et Brochier-Armanet (2010) Concepts et méthodes en phylogénie moléculaire*

# Techniques de réajustement

❖ Subtree Pruning et Regrafting (SPR) : l'arbre est coupée au niveau d'une branche ce qui conduit à deux sous-arbres : l'arbre élagué (pruned tree) et l'arbre résiduel. La méthode consiste à rebrancher l'arbre élagué sur chacune des branches de l'arbre résiduel et à calculer pour chaque topologie le critère d'optimisation.

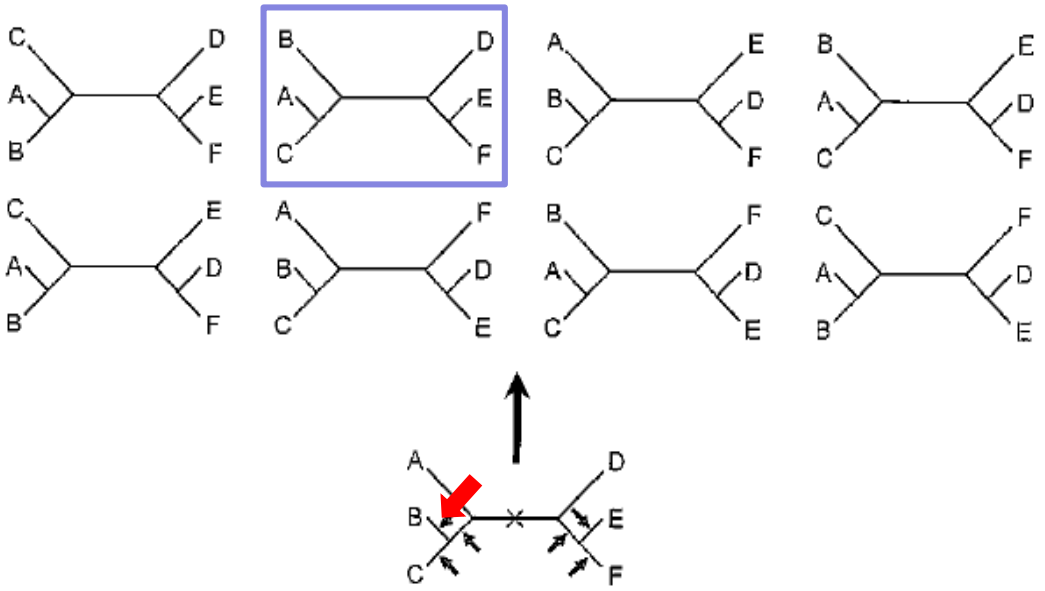



*Exemple extrait de Perrière et Brochier-Armanet (2010) Concepts et méthodes en phylogénie moléculaire*

# Techniques de réajustement

Techniques de réajustement les plus courantes :

❖ Tree Bisection and Reconnection (TBR) : Variante de la SPR. Dans ce cas les deux sous-arbres sont considérés comme indépendants. Toutes les topologies correspondant à toutes les connections possibles entre chacune des branches des deux sous-arbres sont évaluées. Répétée pour l'ensemble des branches internes




*Exemple extrait de Perrière et Brochier-Armanet (2010) Concepts et méthodes en phylogénie moléculaire*

# Robustesse des topologies

Les arbres phylogénétiques construits avec une méthode quelconque ne sont qu'une estimation de l'histoire évolutive des séquences. Il est donc important de disposer de méthodes permettant d'évaluer statistiquement cet estimateur qu'est l'arbre.
Les méthodes les plus couramment utilisées seront décrites.

# Robustesse des topologies

Bootstrap et Jackknife :

Deux méthodes basées sur des techniques de ré-échantillonage.
Méthodes empiriques permettant d'inférer la variabilité des paramètres quand les modèles sont trop complexes pour pouvoir en calculer la variance.
Introduites en phylogénie par Felsenstein (1985) .
Elles sont basées toutes les deux sur l'hypothèse que l'évolution des sites est indépendante.

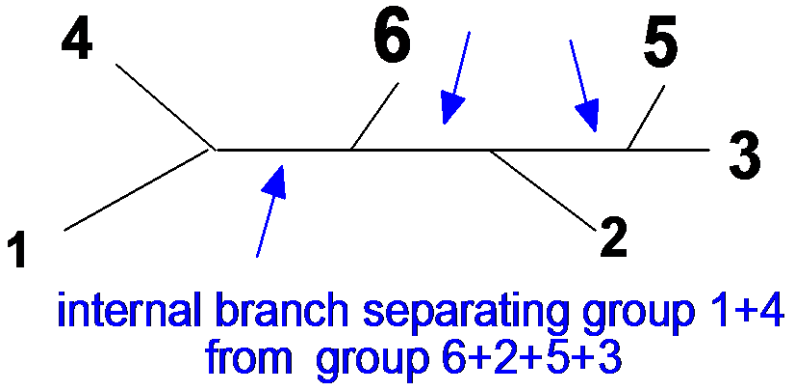
➡ Si un arbre est robuste, c'est-à-dire fortement soutenu par les données, alors sa variance sera faible.

➡ Si un arbre est peu robuste alors il présentera une grande variabilité

Le bootstrap non paramétrique (appelé couramment bootstrap) est la méthode la plus couramment utilisée pour mesurer les incertitudes sur les arbres.
Peut être utilisée en combinaison de n'importe quelle méthode de reconstruction.

# Robustesse des topologies : le bootstrap

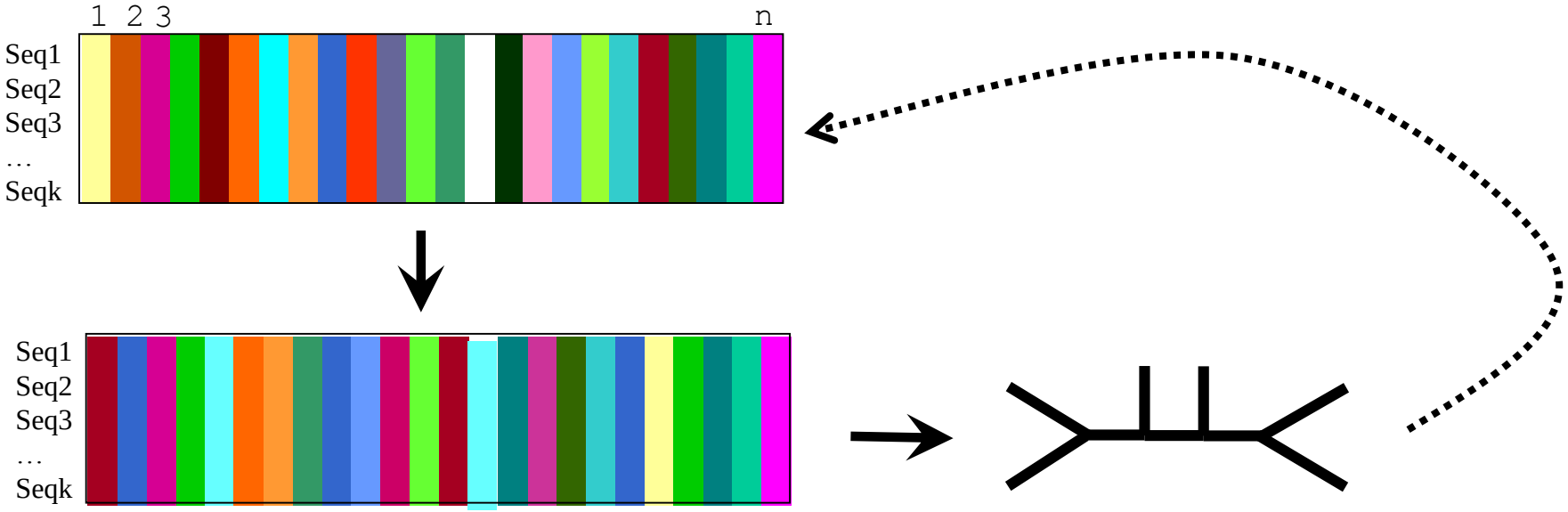
- L'information phylogénétique contenue dans un arbre non raciné réside entièrement dans ses branches internes.




- 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.

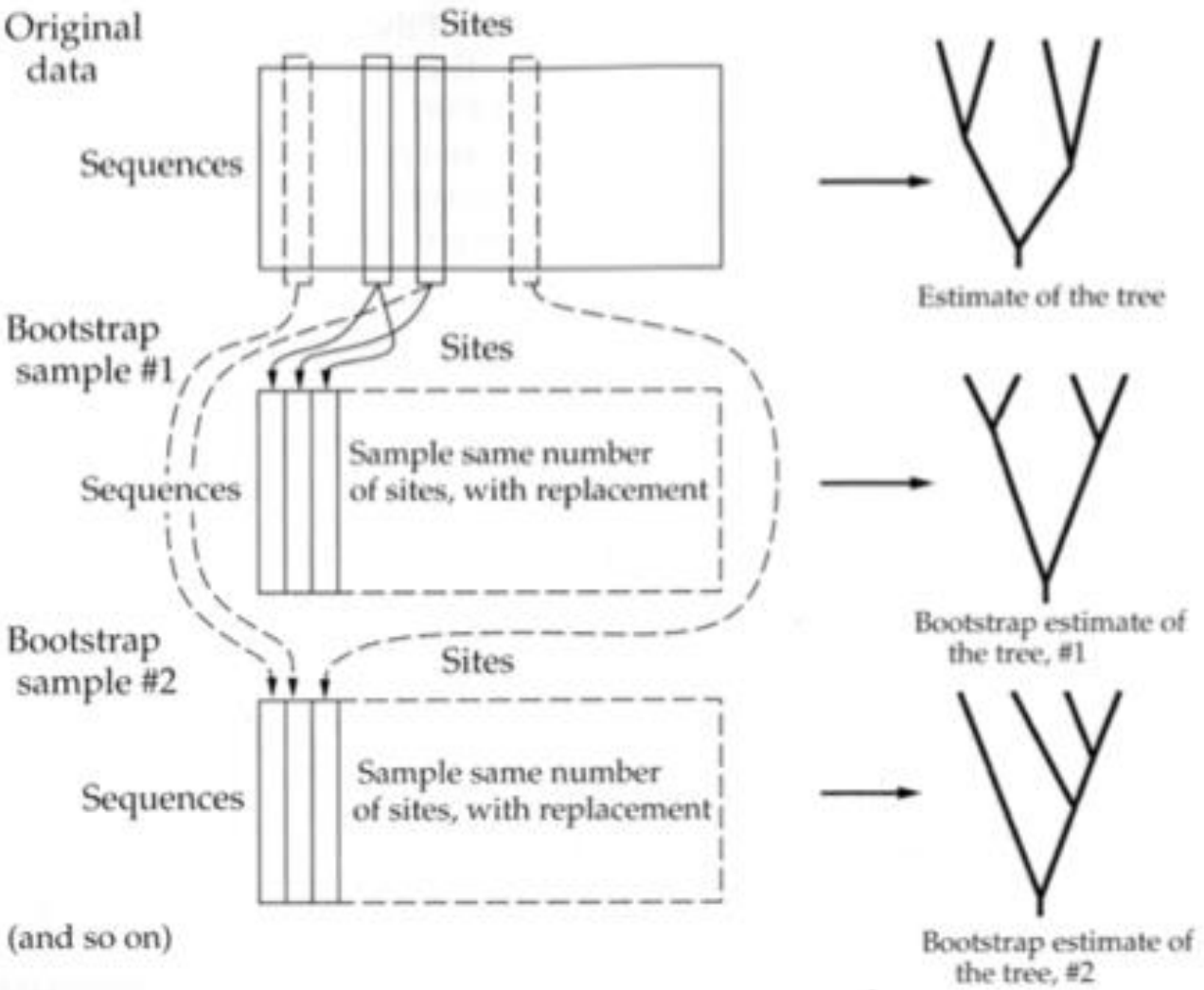
# Robustesse des topologies : le bootstrap

- 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)

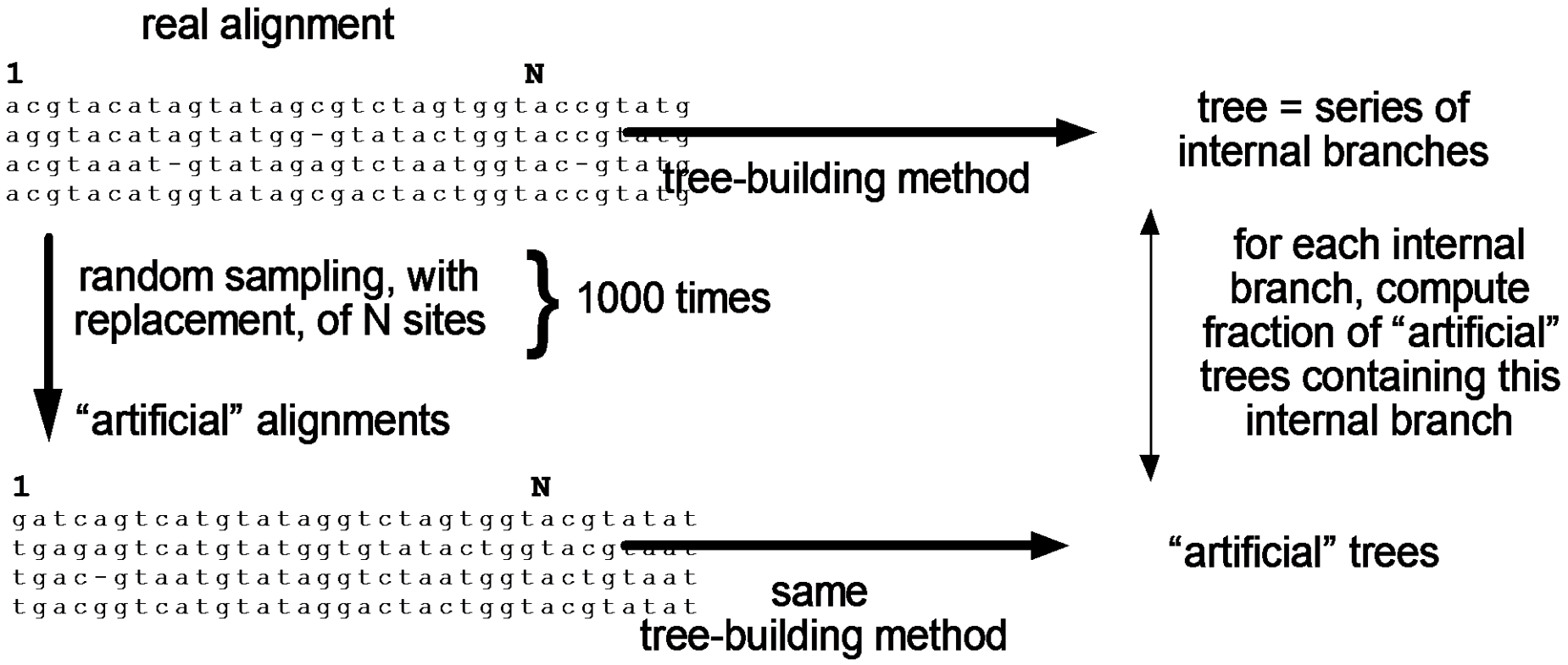



Pour chaque arbre reconstruit on compte le nombre de fois que l'on observe la branche interne. Le soutien de la branche est exprimé en pourcentage de réplications. Si la branche est observée dans tous les arbres, la valeur du bootstrap est égale à 100.

# Robustesse des topologies : le bootstrap

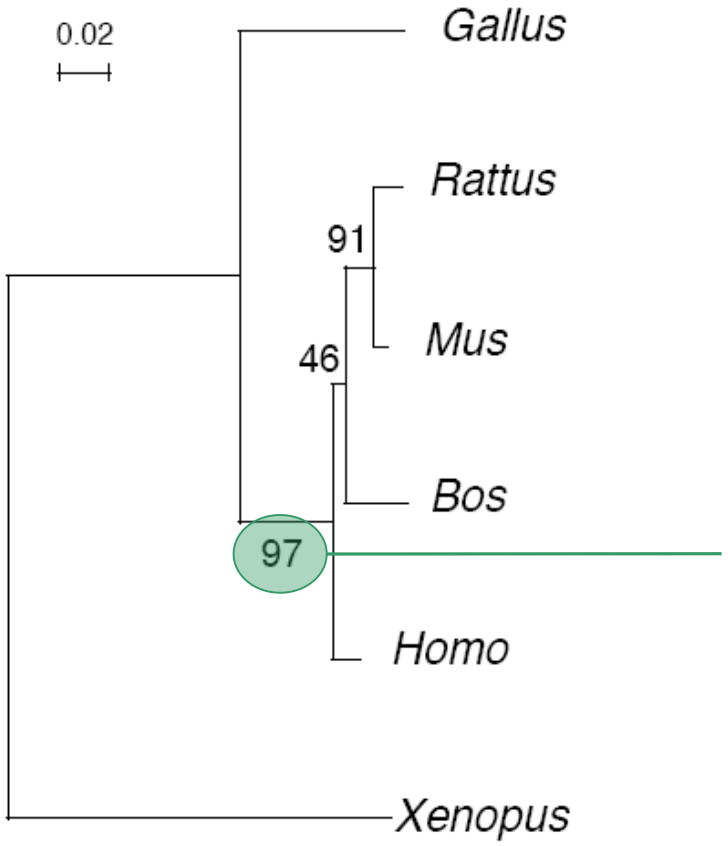


# Robustesse des topologies : le bootstrap



# Robustesse des topologies : 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.



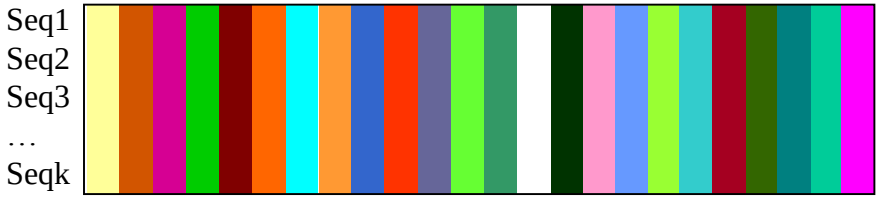

**Estimation** statistique de la **confiance** à accorder à une branche

97% des 1000 arbres contenaient cette branche

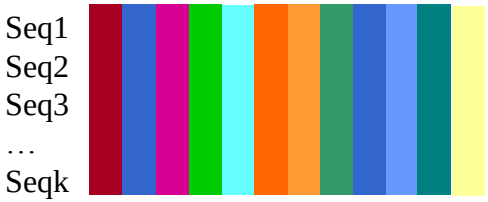
# Robustesse des topologies : le jackknife

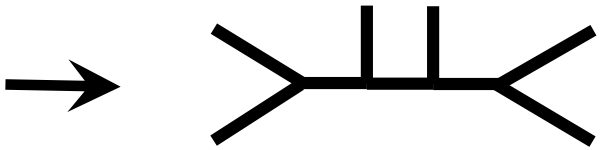
**Principe :** Tirage aléatoire sans remise afin de construire les échantillons. En pratique, approche la plus utilisée, construire des échantillons de taille *n*/2, c'est-à-dire contenant la moitié des sites de l'alignement.

Alignement de départ




Alignement après tirage de *n*/2 sites



- Comme pour le bootstrap, une fois les *X* réplicats construits, les arbres correspondant à chaque réplicat sont calculés.
- La mesure de la robustesse de l'arbre se fait également en calculant le pourcentage de fois où chaque branche interne de l'arbre de départ est retrouvée dans les arbres issus du ré-échantillonage.
- Donc bootstrap et jackknife utilisent des techniques très similaires.

# 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.
Ce n'est pas parce que le boostrap est de 95% que cela signifie que votre clade a 95% d'être « vrai ».

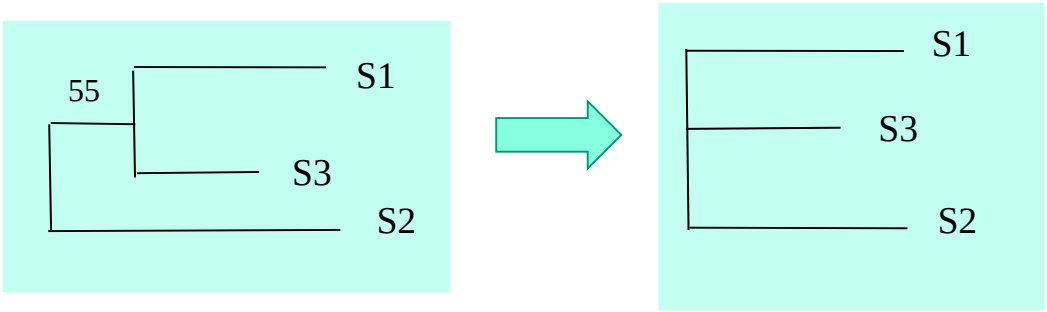
Quel seuil ? Interprétation plutôt qualitative, plus la valeur du bootstrap est élevée, plus vous pouvez avoir confiance dans la branche
Si on applique les critères standards utilisés en statistique, il ne faudrait considérer comme valide que les branches ayant un support de bootstrap ≥ 95%. Des travaux 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.

La procédure de bootstrap n'aide pas à déterminer si la méthode de construction d'arbre est bonne. Un arbre faux peut avoir un score de bootstrap de 100 % pour chacune de ses branches !

# Le bootstrap : interprétation

Remplacer par des multifurcations.

Quand faibles valeurs de bootstrap, possible de remplacer les branches incriminées par des multifurcations indiquant que les données ne permettent pas de résoudre sans ambiguïté l'ordre d'émergence des différentes lignées.



# L'attraction des longues branches

- Toutes les méthodes sont sujettes à l'artéfact dit d'attraction des branches longues, mais la parcimonie y est plus sensible.

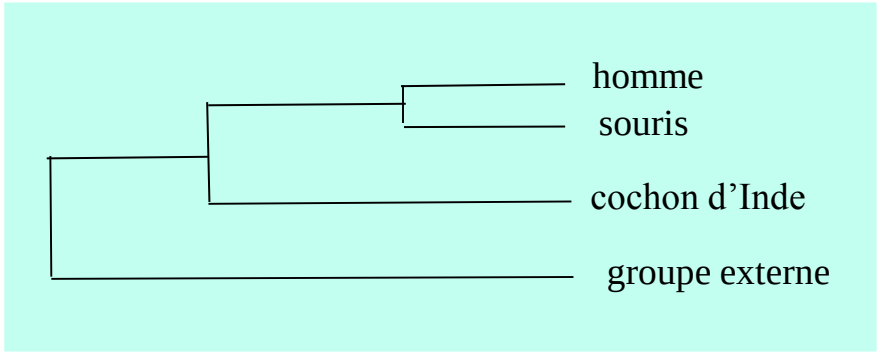
- Cet artéfact provient des inégalités du taux d'évolution des caractères entre les lignées analysées.

- Les espèces qui évoluent plus vite que les autres pour les caractères utilisés se traduisent dans un arbre par une branche propre plus longue.

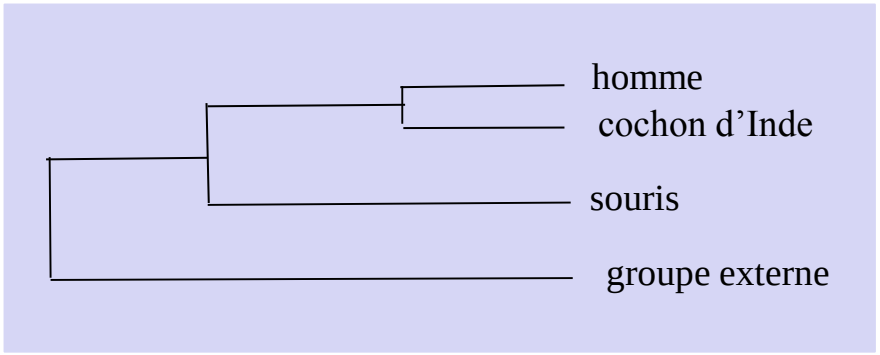
- On a pu montrer théoriquement et expérimentalement qu'au-delà d'un certain écart de vitesse d'évolution entre les espèces, les espèces qui évoluent plus vite ont plus de chance d'avoir des états de caractères communs que par ascendance commune, et que le nombre de caractères communs ainsi acquis devenait supérieur aux caractères qui auraient dû les séparer.

- Par conséquent, elles sont regroupées ensemble dans l'arbre indépendamment des parentés.

# L'attraction des longues branches



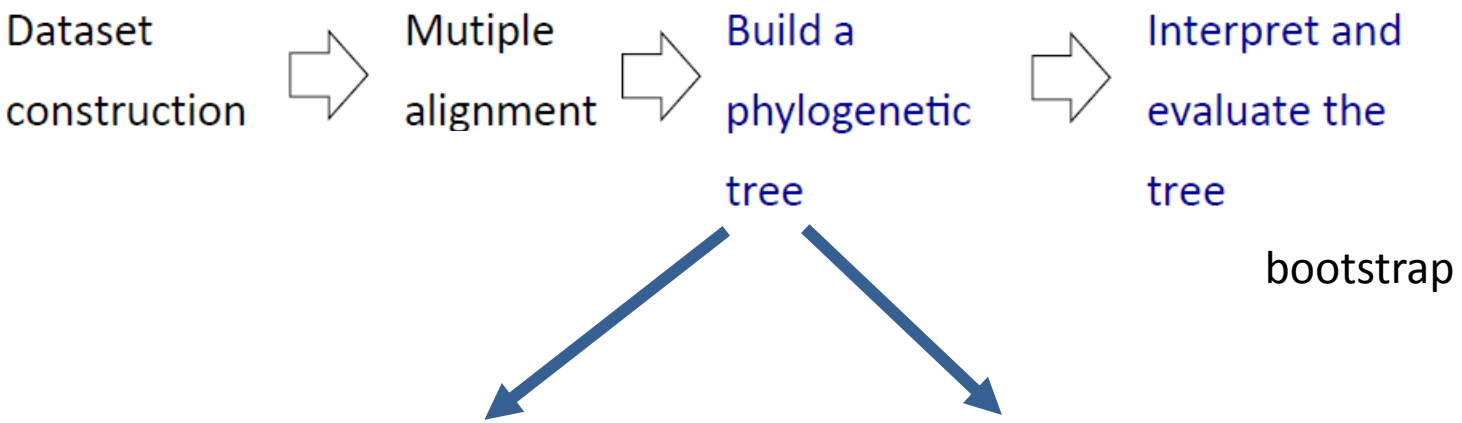

1ère étude sur 15 gènes qui ont évolués plus vite chez le cochon d'Inde que chez la souris.




2ème étude sur l'ADN mitochondrial qui a évolué plus vite chez la souris que chez le cochon d'Inde.

Le groupe externe utilisé pour enraciner l'arbre présente une longue branche, ce qui en général le cas surtout si celui-ci est distant du point de vue évolutif. Dans ce cas, les longues branches du groupe d'étude sont attirées par celle du groupe externe.
Conseil : pour le groupe externe ne pas considérer une seule espèce mais plusieurs présentant des distances évolutives étalonnées pour « casser » les longues branches.

# Conclusion

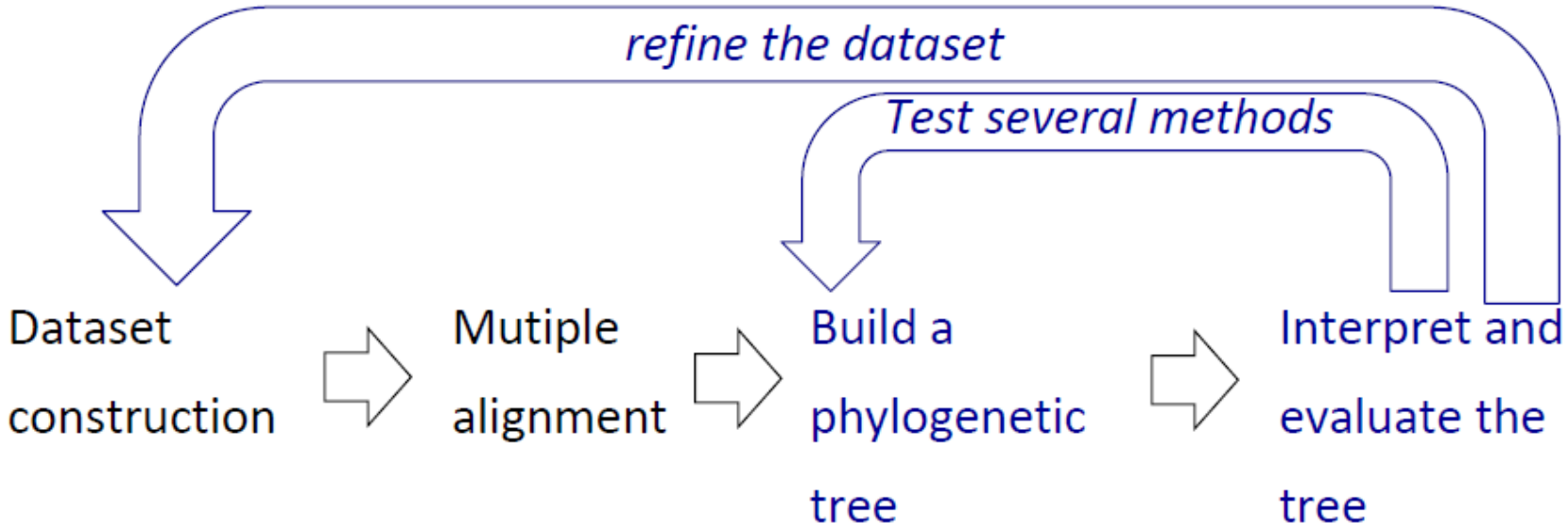



Choix du modèle évolutif

Choix de la méthode de construction

- Distance
- parcimonie
- Maximum de vraisemblance
- bayésienne

# Conclusion




- ✓ Si des séquences ont des longues branches ou perturbent l'arbre, les supprimer et refaire la procédure
- ✓ Fortement conseillé d'utiliser deux méthodes de reconstruction d'arbre différentes. Si les deux méthodes donnent le même (ou très semblable) arbre, cela renforce la confiance dans la topologie obtenue