cnrs formation entreprises

Direction de l'innovation et des relations avec les entreprises

Phylogénie moléculaire

2018

M. Gouy, V. Daubin et G. Perrière

Laboratoire de Biométrie et Biologie Évolutive

Université Claude Bernard – Lyon 1

https://lbbe.univ-lyon1.fr/

www.cnrs.fr

LBBE
BIOMETRIE ET BIOLOGIE EVOLUTIVE

Alignements et recherche de similarités

Formation CNRS « Phylogénie moléculaire »

Guy Perrière

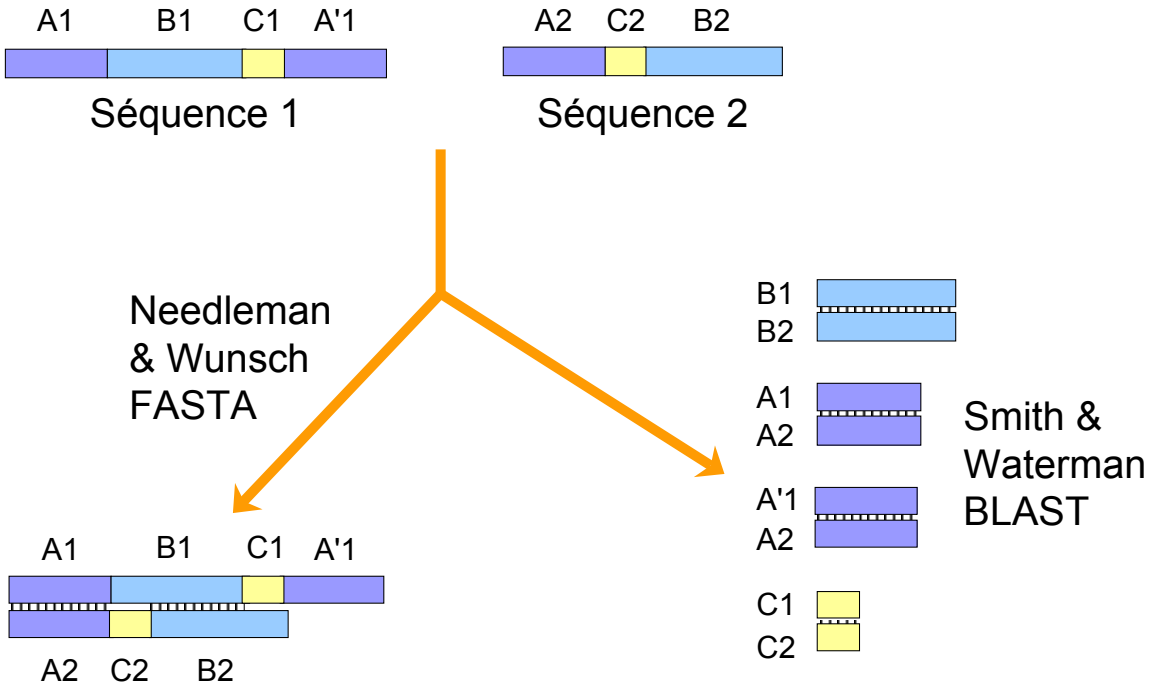
Laboratoire de Biométrie et Biologie Évolutive
UMR CNRS n° 5558
Université Claude Bernard – Lyon 1

20-23 mars 2018

Objectifs poursuivis

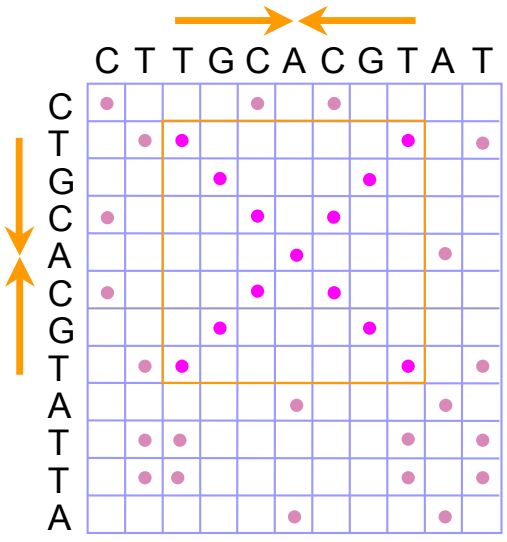
- Identification d'homologues.
- Recherche de contraintes fonctionnelles.
- Prédiction de structure (ARN, protéine).
- Prédiction de fonction.
- Reconstitution des relations évolutives entre séquences (phylogénie).
- Assemblage de lectures (séquençage).

Alignement global et local



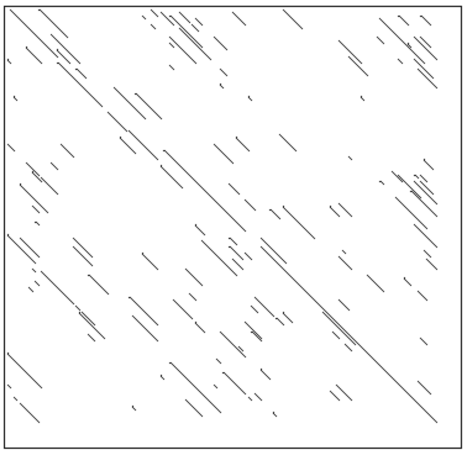
Matrices de points

- Comparaison visuelle de deux séquences :
  - Une diagonale indique une similarité locale.
  - Une croix indique une répétition inverse.
  - Méthode simple et rapide :
    - Algorithme en $O(nm)$.
  - Pas d'alignement ni de score global.

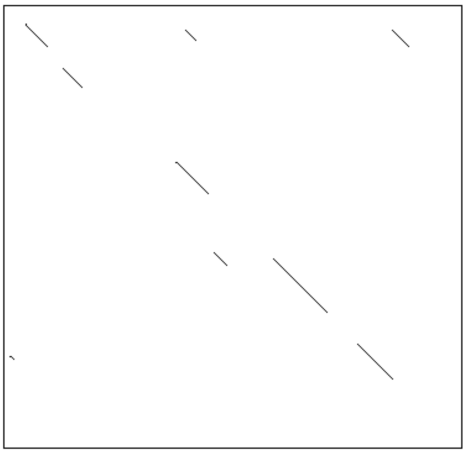


Élimination du bruit de fond

- Filtrage en affichant un point uniquement si plusieurs résidus successifs correspondent :
  - Exemple des hémoglobines α et β humaines :



Identités = 3/10



Identités = 5/10

Représentation

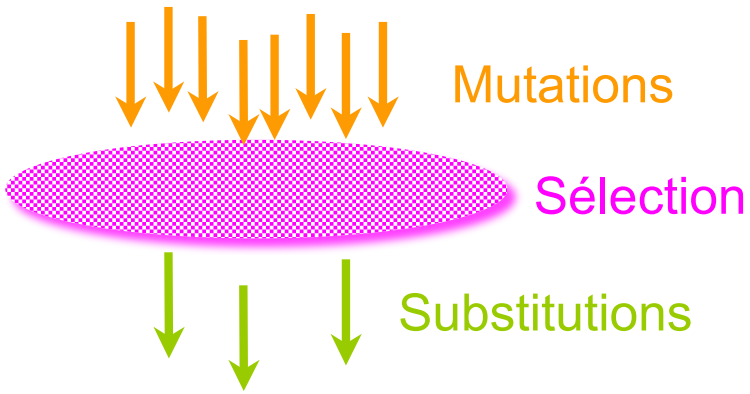
- Les résidus (nucléotides, acides aminés) sont superposés de façon à maximiser les identités entre les séquences :

```
G T T A A G G C G - G G A A A
G T T - - - G C G A G G A C A
* * *       * * *   * * *   *
```

- Il existe deux sortes de différences :
  - Substitutions (*mismatches*).
  - Insertions et délétions (*indels* ou *gaps*).

Mutations et substitutions

- Les différences observées dans un alignement correspondent aux substitutions :
  - Mutations ayant passé le filtre de la sélection :
    - Mutations neutres (*i.e.*, sans effet sur le phénotype) ou avantageuses du point de vue sélectif.



Quel est le bon alignement ?

```
G T T A C G A          G T T A C - G A
G T T - G G A    ou    G T T - - G G A
* * *     * *          * * *       * *
```

```
G T T A C G A
G T T G - G A
* * *     * *
```

- Pour le biologiste, le bon alignement est celui qui représente le scénario évolutif le plus probable :
  - Définition d'une *fonction de score* appropriée.
- Autres choix possibles (*e.g.*, assemblage de lectures).

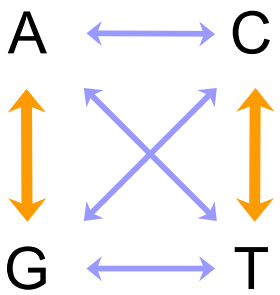
Fonction de score simple

```
G T T A A G G C G - G G A A A
G T T - - - G C G A G G A C A
* * *       * * *   * * *   *
```

$$\text{Score} = \sum \text{Identités} - \sum \text{Différences}$$

$$\left.\begin{array}{ll} \text{Identité} & = +1 \\ \text{Substitution} & = 0 \\ \text{Gap} & = -1 \end{array}\right\} \Longrightarrow \quad \text{Score} = 10 - 4 = 6$$

Modèle d'évolution



$$\mathbb{P}(\text{transition}) > \mathbb{P}(\text{transversion})$$

```
G T T A C G A       G T T A C G A
G T T G - G A   >   G T T - G G A
* * * :   * *       * * *   * *
```

Matrice de substitution ADN

	A	C	G	T
A	1			
C	0	1		
G	0.5	0	1	
T	0	0.5	0	1

$\delta_{AA} = 1.0$
$\delta_{AG} = 0.5$
$\delta_{gap} = -1.0$

```
G T T A C G A          G T T A C G A
G T T G - G A    >     G T T - G G A
* * * :   * *          * * *   * *
```

Score = 4.5 | Score = 4.0

Le cas des acides aminés

- Plus complexe à modéliser que celui des séquences nucléotidiques :
  - Alphabet de vingt lettres au lieu de quatre.
  - Difficulté d'utiliser directement l'information portée par les séquences codantes :
    - Certaines substitutions peuvent avoir plus ou moins d'effet sur la fonction des protéines :

$$\mathbb{P}(\text{AAU}^{\text{Asn}} \rightarrow \text{GAU}^{\text{Asp}}) > \mathbb{P}(\text{AAU}^{\text{Asn}} \rightarrow \text{CAU}^{\text{His}})$$

  - Utilisation de modèles empiriques :
    - Alignement de séquences homologues avec la matrice identité.
    - Construction d'arbres phylogénétiques sur la base de ces alignements.
    - Construction des matrices sur la base du nombre de substitutions observées (ou inférées à partir des arbres).

BLOSUM (*Blocks Substitution Matrices*)

- Matrices fondées sur des alignements locaux de domaines conservés (Henikoff et Henikoff, 1992) :
  - Utilisation de ≈ 2000 domaines provenant de ≈ 500 familles de protéines.
  - Matrice identité.
  - Alignements sans *gaps*.
- Ensemble de quinze matrices créées à partir de domaines comprenant des séquences $\pm$ divergentes :
  - Toutes les paires ayant servi à construire une matrice BLOSUMk ($30 \leq k \leq 100$) ont une identité $\geq$ à $k\%$.
  - Bien adaptées pour l'alignement de séquences très divergentes.

Construction I

- Calcul de la fréquence *observée* de chaque paire d'acide aminé dans l'alignement :
  - Soit n_{ij} le nombre de paires (i, j) observées, $i, j \in \{\text{A, C, D}, \ldots, \text{W}\}$.
  - Soit w la largeur du bloc considéré et a le nombre de séquences alignées.
  - Dans ce cas, le nombre total de paires possibles n est tel que :

$$n = wa(a-1)/2$$

et f_{ij}, la fréquence observée de chaque paire (i, j) est égale à :

$$f_{ij} = n_{ij}/n$$

Construction II

- Calcul de la fréquence *attendue* de chaque paire :
  - Soit π_i la fréquence de l'acide aminé i dans l'alignement.
  - Dans ce cas, les fréquences attendues g_{ij} de chaque paire (i, j) sont égales à :

$$g_{ij} = 2\pi_i\pi_j \quad (i \neq j)$$
$$g_{ii} = \pi_i^2$$

- Calcul des éléments de la matrice de substitution par :

$$\delta_{ij} = 2\log_2(f_{ij}/g_{ij})$$

avec arrondissement à l'entier le plus proche :

- $\delta_{ij} > 0 \Rightarrow$ substitution plus fréquente qu'attendue.
- $\delta_{ij} < 0 \Rightarrow$ substitution moins fréquente qu'attendue.

Exemple simple

$a = 6$, $w = 4$

E	A	E	A
A	A	A	G
A	A	G	G
A	A	E	A
A	A	G	G
A	A	E	G

$\Rightarrow \quad n = 60$

i	π_i
A	14/24
E	4/24
G	6/24

(i, j)	f_{ij}	g_{ij}	δ_{ij}
(A, A)	26/60	196/576	0.70
(A, E)	8/60	112/576	–1.09
(A, G)	10/60	168/576	–1.61
(E, E)	3/60	16/576	1.70
(E, G)	6/60	48/576	0.53
(G, G)	7/60	36/576	1.80

A	1		
E	–1	2	
G	–2	1	2
	A	E	G

BLOSUM62

	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W
C	9																			
S	-1	4																		
T	-1	1	5																	
P	-3	-1	-1	7																
A	0	1	0	-1	4															
G	-3	0	-2	-2	0	6														
N	-3	1	0	-2	-2	0	6													
D	-3	0	-1	-1	-2	-1	1	6												
E	-4	0	-1	-1	-1	-2	0	2	5											
Q	-3	0	-1	-1	-1	-2	0	0	2	5										
H	-3	-1	-2	-2	-2	-2	1	-1	0	0	8									
R	-3	-1	-1	-2	-1	-2	0	-2	0	1	0	5								
K	-3	0	-1	-1	-1	-2	0	-1	1	1	-1	2	5							
M	-1	-1	-1	-2	-1	-3	-2	-3	-2	0	-2	-1	-1	5						
I	-1	-2	-1	-3	-1	-4	-3	-3	-3	-3	-3	-3	-3	1	4					
L	-1	-2	-1	-3	-1	-4	-3	-4	-3	-2	-3	-2	-2	2	2	4				
V	-1	-2	0	-2	0	-3	-3	-3	-2	-2	-3	-3	-2	1	3	1	4			
F	-2	-2	-2	-4	-2	-3	-3	-3	-3	-3	-1	-3	-3	0	0	0	-1	6		
Y	-2	-2	-2	-3	-2	-3	-2	-3	-2	-1	2	-2	-2	-1	-1	-1	-1	3	7	
W	-2	-3	-2	-4	-3	-2	-4	-4	-3	-2	-2	-3	-3	-1	-3	-2	-3	1	2	11

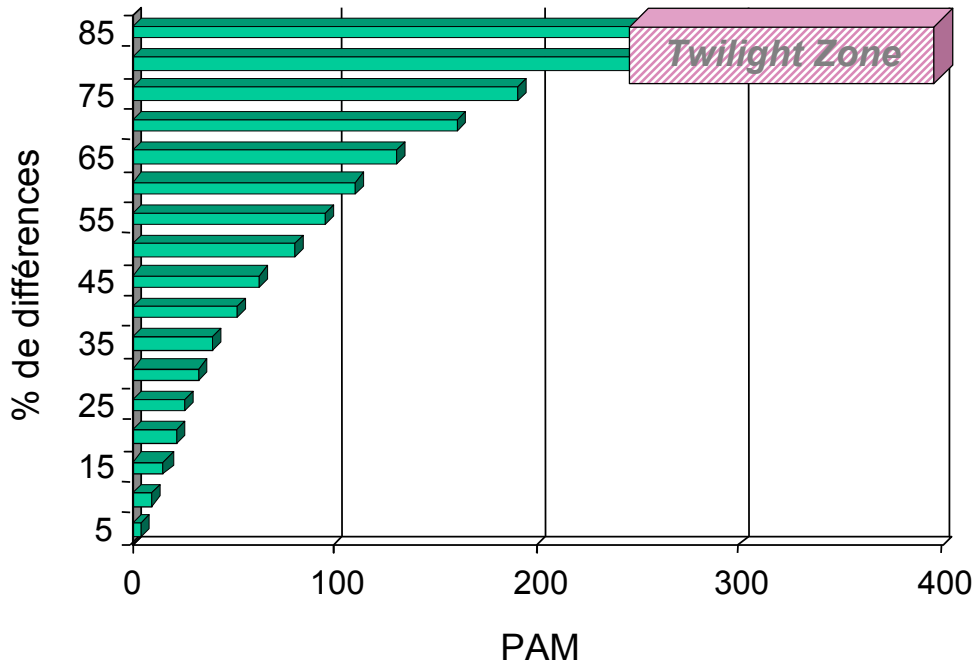
C	sulfhydryle
S	
T	
P	petit hydrophile
A	
G	
N	
D	acide, acide-amide
E	et hydrophile
Q	
H	
R	basique
K	
M	
I	petit hydrophobe
L	
V	
F	
Y	aromatique
W	

Les critères évolutifs recoupent les critères physico-chimiques

PAM (*Point Accepted Mutations*)

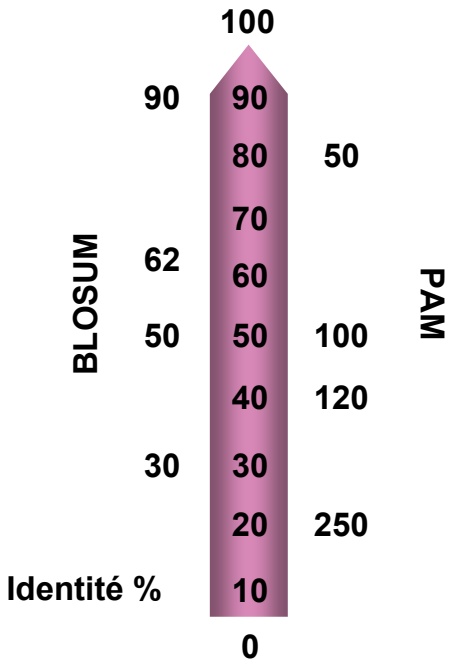
- Matrices construites sous l'hypothèse que les séquences évoluent selon un processus markovien (Dayhoff *et al.*, 1978).
- Alignements globaux de protéines conservées :
  - Utilisation de la matrice identité.
  - Pas de *gaps* dans les alignements.
  - 71 familles de gènes correspondant à 1300 séquences et contenant un total de 1572 substitutions :
    - Identité $\geq 85\%$ entre chaque paire possible au sein d'une famille.
- Matrices JTT (Jones, Taylor et Thornton, 1992) et Gonnet (Gonnet *et al.*, 1992) :
  - Construites avec une méthodologie comparable, mais sur des échantillons bien plus grands.
- Construction : *cf.* cours sur les modèles d'évolution.

Limites de validité



Choix d'une matrice

- Pas de matrice idéale.
- Meilleurs résultats :
  - Matrices construites avec le plus de séquences.
- Critère de choix principal :
  - Degré de similarité entre les séquences.
- Il est recommandé d'expérimenter !



Pénalités pour les *gaps*

- Avec la fonction de score telle que présentée dans la Diapo. 9, la pénalité γ associée à un *gap* de longueur k est égale à :

$$\gamma = k\delta_{\text{gap}}$$

avec δ_{gap}, la pénalité d'un *gap* individuel.

- Amélioration par l'emploi de fonctions *affines* ou *logarithmiques* :

$$\gamma = \delta_{\text{o}} + (k-1)\delta_{\text{e}}$$
$$\gamma = \delta_{\text{o}} + \ln(k-1)\delta_{\text{e}}$$

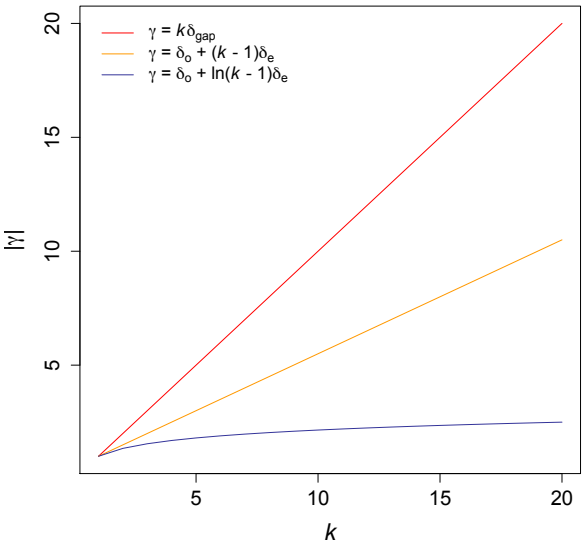
avec δ_{o} la pénalité associée à l'ouverture d'un *gap*, et δ_{e} la pénalité associée à l'extension de ce *gap*.

Influence sur les alignements

- Pénalités affines :
  - Évitement de *gaps* trop rapprochés :

```
TGATATCGCCA      TGATATCGCCA
TGAT---TCCA  >   TGAT-T--CCA
****    ***      **** *  ***
```

  - Alignements plus réalistes du point de vue évolutif.
- Pénalités logarithmiques :
  - Seulement dans le cas où il est nécessaire d'avoir de très longs *gaps*.



Nombre d'alignements

- Le nombre d'alignements possibles entre deux séquences de longueur m et n est égal à (Waterman, 1984) :

$$f(m,n) = f(m-1,n) + f(m-1,n-1) + f(m,n-1)$$

avec $f(0,j) = f(i,0) = 1 \ \forall i,j$.

- Par ailleurs, ce nombre croît de manière exponentielle (Torres *et al.*, 2003) :

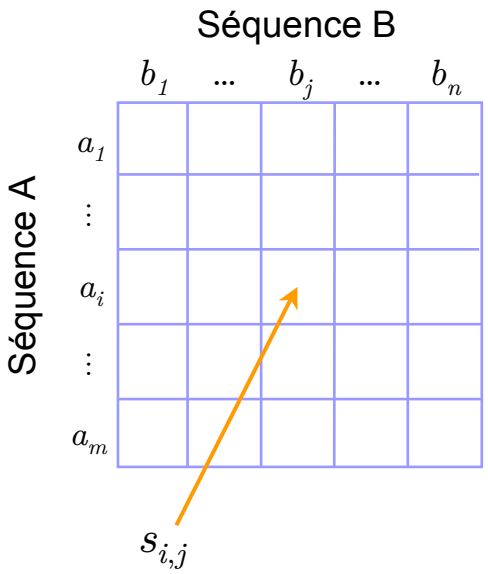
$$f(m,n) = \sum_{k=0}^{\min(m,n)} 2^k \binom{m}{k} \binom{n}{k}$$

Trouver le bon alignement

- Calcul de tous les alignements possibles :
  - Trop long (croissance exponentielle du nombre d'alignements).
  - Peu efficace (recalcul des mêmes valeurs).
- Utilisation d'un algorithme de *programmation dynamique* :
  - Recherche du meilleur alignement possible sur une fraction de la longueur des séquences :
    - Construction progressive de l'alignement.
  - Needleman et Wunsch (1970) : alignement global.
  - Smith et Waterman (1981) : alignement local.

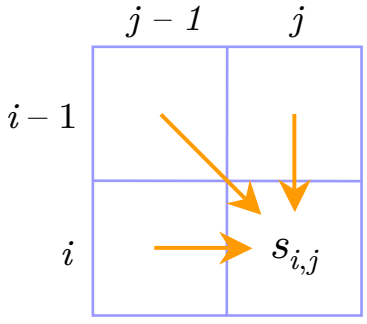
Needleman et Wunsch

- Soit deux séquences A et B de longueurs m et n.
- Soit $\mathbf{S} = (s_{i,j})$, la matrice de chemin associée à ces deux séquences.
- Stockage dans chaque case de $\mathbf{S}$ du score du chemin menant à cette case :
  - Utilisation d'une fonction de score donnée.
  - Le score de l'alignement est la valeur en $s_{m,n}$.



Construction de la matrice

- Soit $s_{i,j}$ la valeur du score optimum dans la case de coordonnées (i, j) :
  - Calcul au moyen des scores dans les trois cases adjacentes $(i-1, j)$, $(i-1, j-1)$ et $(i, j-1)$.
  - Si pas de pénalités affines pour les *gaps* :

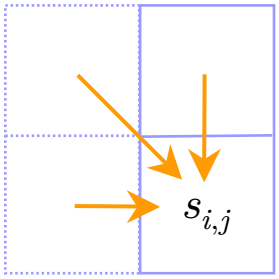


$$s_{i,j} = \max \begin{cases} s_{i-1,j} + \delta(\text{-}) \\ s_{i-1,j-1} + \delta(a_i, b_j) \\ s_{i,j-1} + \delta(\text{-}) \end{cases}$$

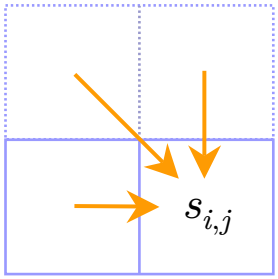
Bords de la matrice

- Les cases situées sur le bord du haut ou le bord gauche ne possèdent pas le total requis de trois cases adjacentes :
  - Ajout d'une ligne et d'une colonne afin d'initialiser la matrice :
    - Le balayage ne se faisant plus qu'avec des indices ≥ 1, on ne rencontre plus de cases nécessitant un traitement particulier.

Bord gauche



Bord du haut



Initialisation de la matrice

- Pénalisation des *gaps* terminaux :

$$\begin{aligned} s_{0,0} &= 0 \\ s_{i,0} &= s_{i-1,0} + \delta(\text{-}) \quad \forall i \in [1, m] \\ s_{0,j} &= s_{0,j-1} + \delta(\text{-}) \quad \forall j \in [1, n] \end{aligned}$$

- Pas de pénalisation des *gaps* terminaux :

$$s_{0,0} = s_{i,0} = s_{0,j} = 0 \quad \forall i \in [1, m],\ \forall j \in [1, n]$$

Option par défaut de beaucoup de programmes d'alignement (*e.g.*, ClustalW).

Exemple de calcul

		A	G	C	T	A
	0	-2	-4	-6	-8	-10
A	-2	-4 +1 -4	-6 -2 -1	-8 -4 -3	-10 -6 -5	-12 -7 -7
T	-4	-1 -2 -6	-3 +1 -3	-5 -1 -1	-7 -2 -3	-9 -5 -4
T	-6	-3 -4 -8	-1 -1 -5	-3 +1 -3	-4 0 -1	-6 -2 -2
A	-8	-5 -5 -10	-3 -3 -7	-1 -1 -5	-2 +1 -3	-4 +1 -1

Identité : +1
Substitution : 0
Gap : -2

```
A  G  C  T  A
A  -  T  T  A
+1 -2 +0 +1 +1
```

$s = +1$

```
A  G  C  T  A
A  T  -  T  A
+1 +0 -2 +1 +1
```

$s = +1$

Pénalités affines pour les gaps

- Soit $u_{i,j}$ et $v_{i,j}$, les scores associés à l'alignement entre les positions i et j et se terminant par un *gap* dans B ou dans A :

$$u_{i,j} = \max \begin{cases} u_{i-1,j} + \delta_e(\text{-}) \\ s_{i-1,j} + \delta_o(\text{-}) + \delta_e(\text{-}) \end{cases}$$

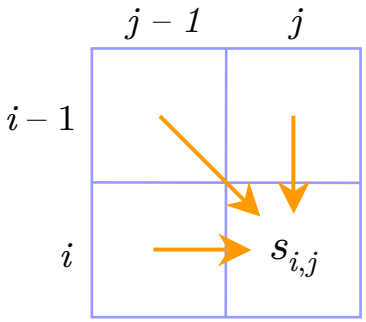
$$v_{i,j} = \max \begin{cases} v_{i,j-1} + \delta_e(\text{-}) \\ s_{i,j-1} + \delta_o(\text{-}) + \delta_e(\text{-}) \end{cases}$$

- Dans ces conditions, le calcul de $s_{i,j}$ est donné par :

$$s_{i,j} = \max \begin{cases} u_{i,j} \\ v_{i,j} \\ s_{i-1,j-1} + \delta(a_i, b_j) \end{cases}$$

Smith et Waterman

- Algorithme dérivé de Needleman et Wunsch :
  - Initialisation des bords à 0.
  - N'importe quelle case de la matrice peut être considérée comme point de départ pour le calcul du score.



$$s_{i,j} = \max \begin{cases} s_{i-1,j} + \delta(\text{-}) \\ s_{i-1,j-1} + \delta(a_i, b_j) \\ s_{i,j-1} + \delta(\text{-}) \\ 0 \end{cases}$$

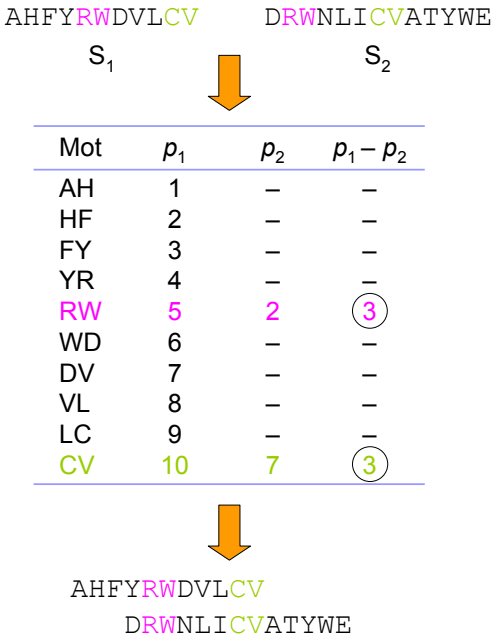
$$s_{i,j} < 0 \Rightarrow s_{i,j} = 0$$

Recherche dans les banques

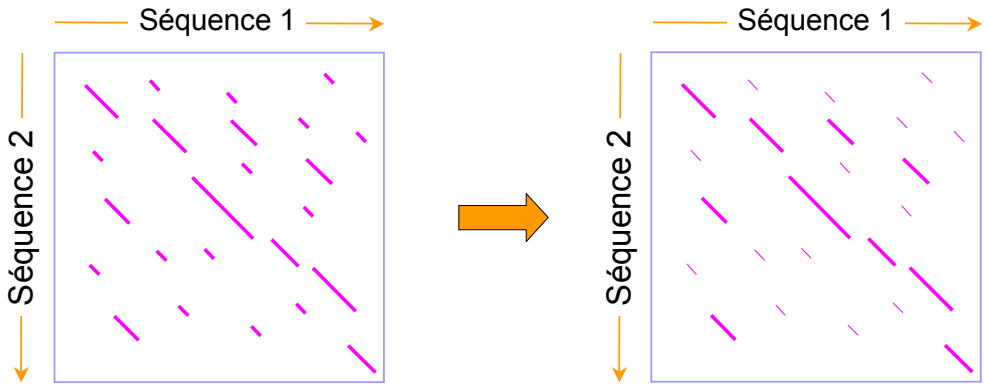
- Algorithme exact (Smith-Waterman) :
  - SSEARCH :
    - Alignements locaux optimaux.
    - Trop lent en pratique et nécessitant beaucoup de mémoire vive.
- Algorithmes fondés sur des heuristiques :
  - FASTA :
    - Fondé sur la recherche de mots *identiques*.
    - Alignement global, ancré sur des régions similaires.
  - BLAST :
    - Fondé sur la recherche de mots identiques (séquences nucléotidiques) ou *similaires* (séquences protéiques).
    - Alignements locaux par extension autour de ces mots.

Recherche des mots

- Mots communs entre la séquence requête et une séquence de la banque :
  - Construction d'une table contenant les positions de chaque mot.
  - Détection des mots communs séparés par un même intervalle.
- Construction d'un alignement local à partir de ces mots communs.

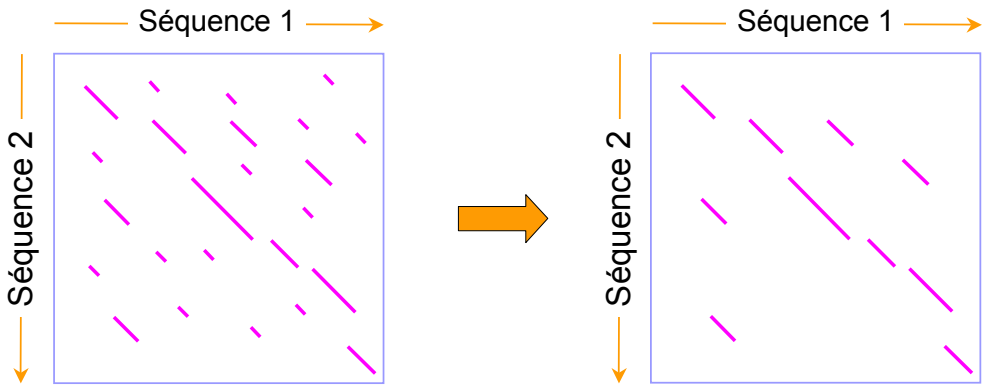


Procédure d'alignement I



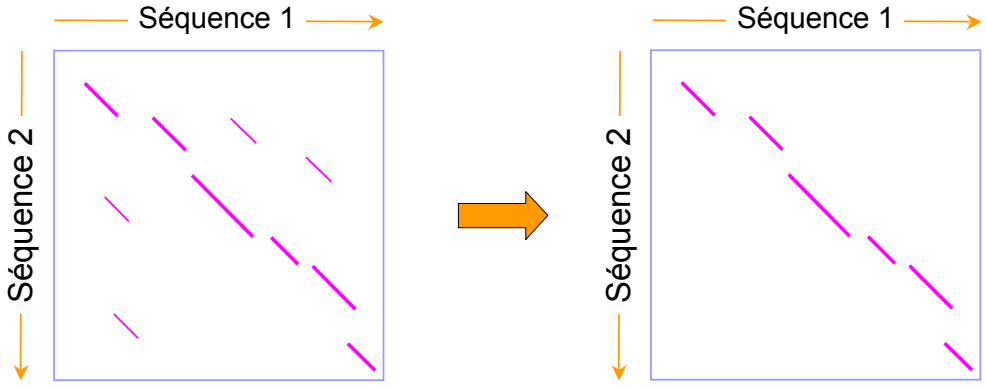
- Calcul de l'ensemble des alignements locaux en partant des mots communs détectés.
- Calcul des scores pour chaque alignement local puis élimination des scores les plus faibles.

Procédure d'alignement I



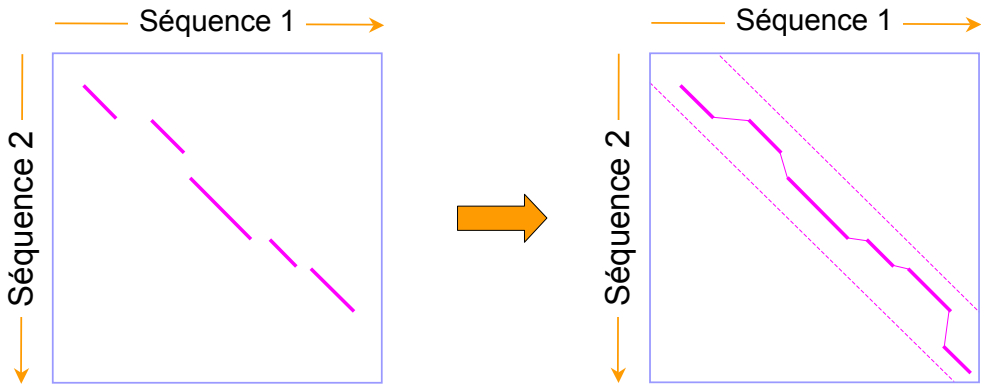
- Calcul de l'ensemble des alignements locaux en partant des mots communs détectés.
- Calcul des scores pour chaque alignement local puis élimination des scores les plus faibles.

Procédure d'alignement II



- Élimination des segments incompatibles avec le segment de score le plus élevé.
- Algorithme de programmation dynamique afin de joindre les segments entre eux, à l'intérieur d'une diagonale.

Procédure d'alignement II



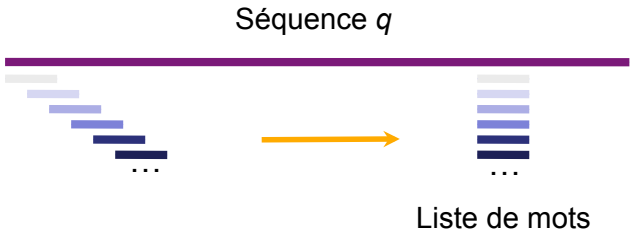
- Élimination des segments incompatibles avec le segment de score le plus élevé.
- Algorithme de programmation dynamique afin de joindre les segments entre eux, à l'intérieur d'une diagonale.

Historique

- Prominence de BLAST depuis son introduction en 1990 :
  - Les deux articles originaux totalisent plus de 94 000 citations !
- Première version, n'autorisant pas les *gaps* dans les alignements (Altschul *et al.*, 1990).
- Nouvelle version (BLAST2), autorisant les *gaps* (Altschul *et al.*, 1997).
- Version améliorée (BLAST+) (Camacho *et al.*, 2009).
- La version actuelle est la 2.7.1 (octobre 2017).

Liste de mots

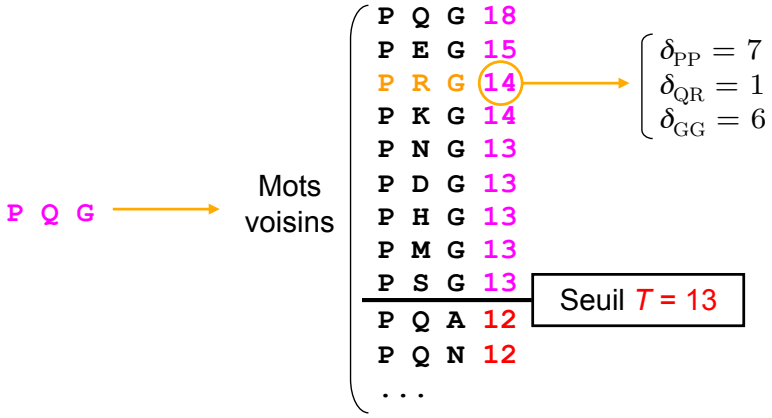
- Construction d'une liste de mots de longueur w à partir d'une séquence q de longueur ℓ :



  Au maximum $\ell - w + 1$ mots différents.
- Séquences nucléotidiques :
  - Mots *exacts*, de longueur $w = 11$ par défaut.
- Séquences protéiques :
  - Mots *similaires*, de longueur $w = 3$ par défaut.

Mots similaires

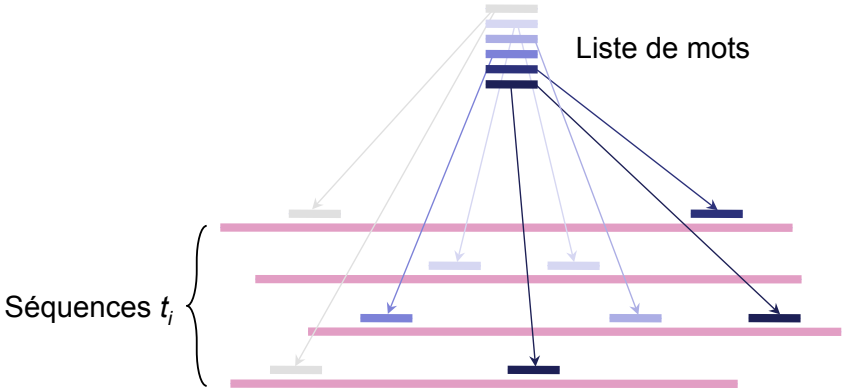
- Pour chaque mot de q, BLAST considère ses voisins les plus proches en utilisant une matrice de substitution :
  - Par défaut, utilisation de BLOSUM62 et seuil fixé à $T = 13$.



  - Dans l'exemple ci-dessus, seulement neuf mots sur les $20^3 = 8000$ possibles présentent une valeur supérieure ou égale au seuil.

Recherche des mots communs

- On recherche les mots de la liste (étendue ou non) construite à partir de q et présents dans les séquences t_i :
  - Constituent ce que l'on appelle les *hits*.

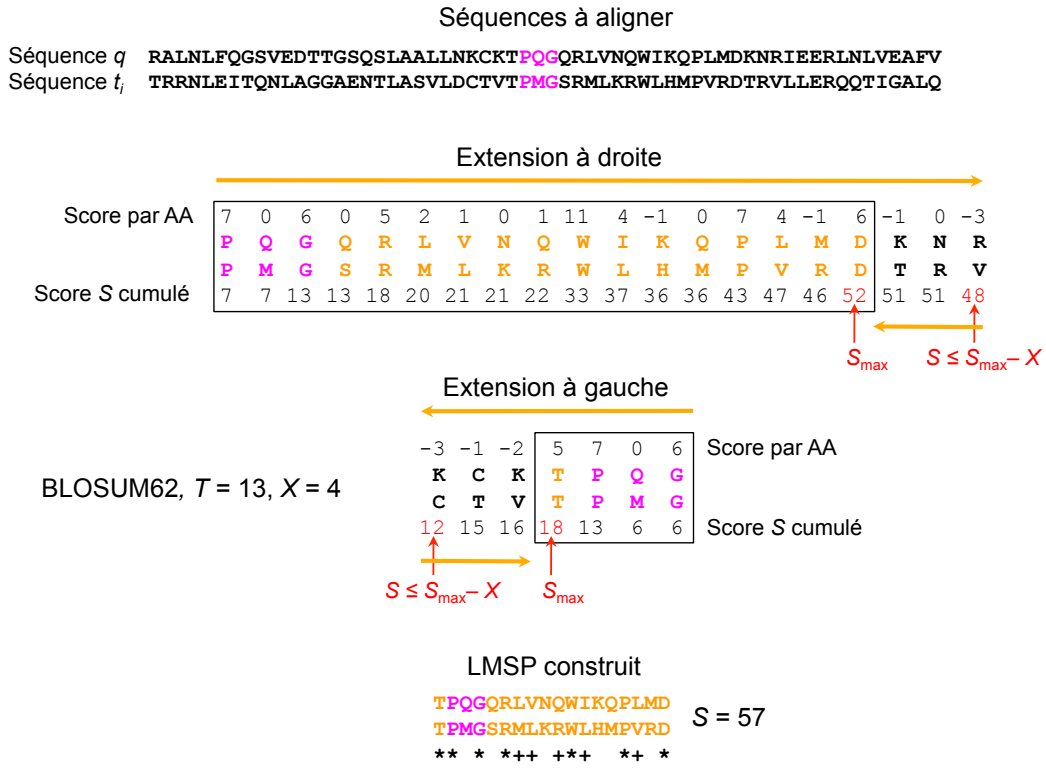


Extension des *hits*

- Pour une séquence t_i donnée, chaque *hit* observé est étendu à gauche et à droite :
  - Arrêt de l'extension lorsque le score S décroit d'une quantité X par rapport à un maximum atteint ($S_{\max}$) :

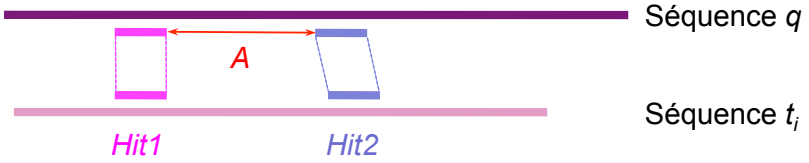
- Chaque *hit* étendu forme un LMSP (*Locally Maximal-scoring Segment Pair*) :
  - Ne sont conservés que les LMSP de score supérieur à un seuil donné, appelés HSP (*High-scoring Segment Pair*).

Arrêt de l'extension



Gestion des *gaps*

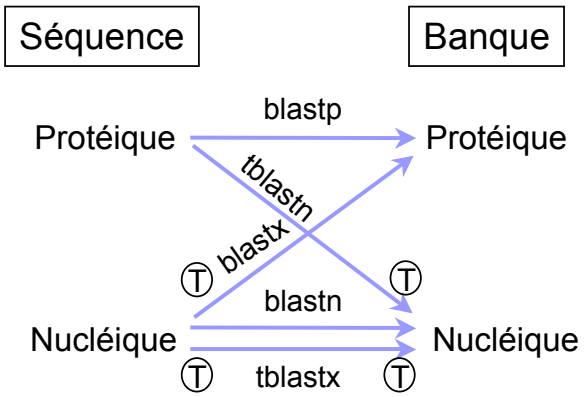
- Disponible à partir de BLAST2 pour les séquences protéiques.
- Introduction entre deux *hits* séparés au maximum par A résidus :



  - Par défaut, seuil fixé à $A = 40$.
- Abaissement de T à 11 pour garder une bonne sensibilité.
- Extension des *hits* comme précédemment avec en plus la possibilité d'introduire des *gaps* :
  - Pénalités par défaut fixées à $\delta_o = -11$ et $\delta_e = -1$.

Programmes

- BLASTP : protéine *vs.* protéine.
- BLASTN : utile pour le non-codant.
- BLASTX : identification de séquences codantes.
- TBLASTN : homologues dans un génome non complètement annoté.



Évaluation statistique

- Porte individuellement sur chaque HSP.
- Similarités détectées :
  - Distinguer les relations significatives des similarités dues au hasard.
- Évaluation de la significativité fondée sur :
  - Le score brut d'alignement observé (S) ou sa valeur normalisée (S').
  - Distribution de ce score.
- Mesure sous la forme :
  - Espérance mathématique (*E-value*).
  - Probabilité (*P-value*).

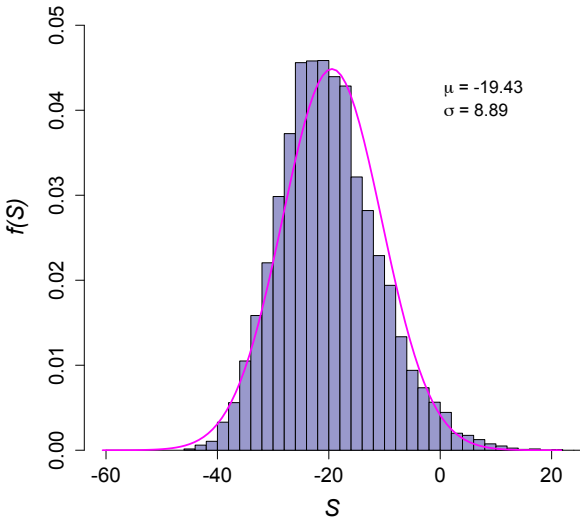
Distribution normale

- La distribution des valeurs de S suit une loi normale :
  - Comparaison avec des séquences *aléatoires*.
  - *Surestimation* de la significativité :

```
TPQGQRLVNQWIKQPLMD
   *+    **     +
HQSGEYRDPQWHYPHAIM
```

$S = 1,\ P < 0.011$

  - Seules les valeurs les plus extrêmes sont d'interêt.

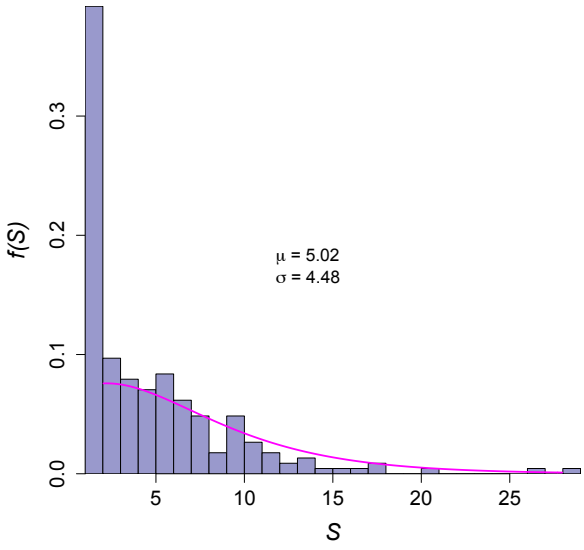


Distribution des valeurs extrêmes

- Loi normale inutilisable en pratique.
- Échantillonnage dans la queue de distribution droite :
  - Distribution des valeurs extrêmes (Gumbel).
  - Score de l'alignement précédent non significatif :

```
TPQGQRLVNQWIKQPLMD
   *+    **     +
HQSGEYRDPQWHYPHAIM
```

$S = 1,\ P = 0.7237$



Estimation de la significativité I

- Tout d'abord le score normalisé S' est donné par :

$$S' = (\lambda S - \ln K)/\ln 2$$

avec K et λ deux paramètres *d'échelle* dépendant de la matrice de substitution utilisée et des pénalités associées aux *gaps*.

- L'espérance mathématique E d'avoir par hasard une HSP dont le score d'alignement serait $\geq$ au score brut observé est égale à :

$$E = Km'n'\mathrm{e}^{-\lambda S}$$

avec m' et n' les tailles *effectives* de la séquence requête et de la banque, calculées à partir des valeurs réelles m et n.

Estimation de la significativité II

- La relation entre E et S' est donc telle que :

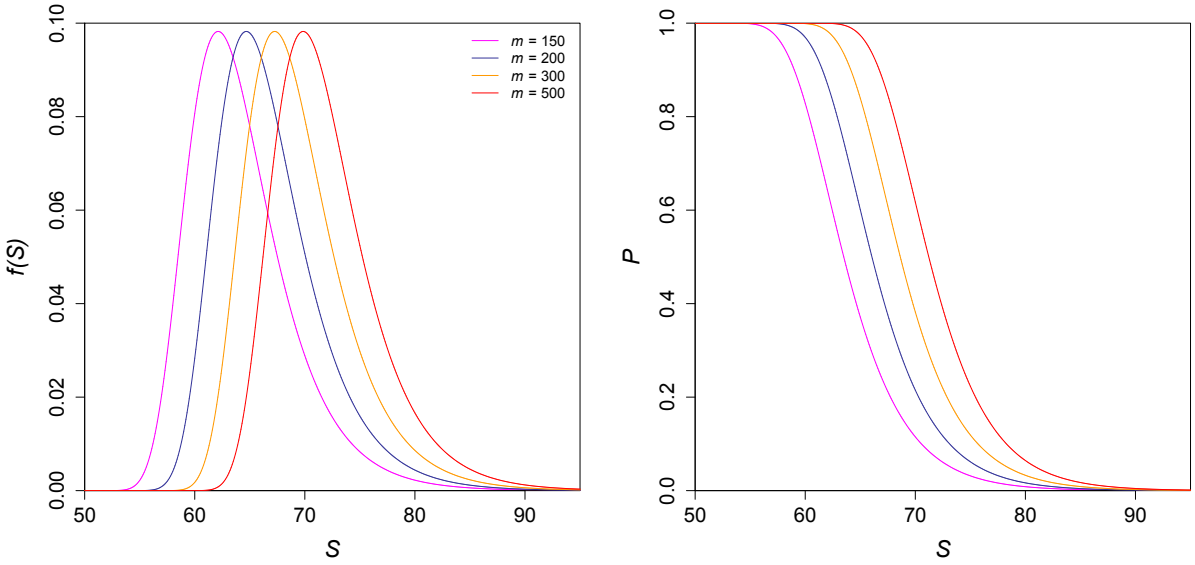
$$E = m'n'2^{-S'}$$

- Enfin, la probabilité P d'avoir par hasard une HSP dont le score d'alignement serait $\geq$ au score brut observé est telle que :

$$P \simeq 1 - \mathrm{e}^{-E}$$

- Utilisation de E plutôt que de P dans les sorties de BLAST :
  - Différences plus directement compréhensibles (*e.g.*, $E = 5$ ou 10 au lieu de $P = 0.9933$ ou 0.99995).
  - A noter que $P \simeq E$ dès que $E \leq 0.01$.

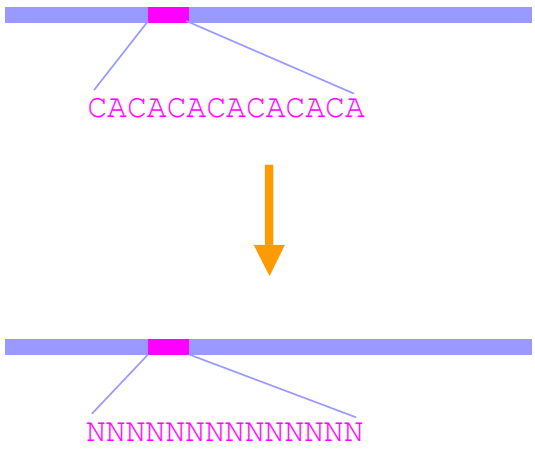
Distribution des valeurs de S



$n = 9418064,\ \delta_o = -11,\ \delta_e = -1$, BLOSUM62

Séquences abondantes

- Immunoglobulines :
  - $>$ 127000 séquences dans GenBank.
- Séquences répétées :
  - 10^6 Alu et 10^5 L1 dans le génome humain.
- Programmes de masquage pour BLAST :
  - DUST, XNU, SEG, RepeatMasker.



Serveurs

- Il existe un grand nombre de serveurs permettant d'effectuer des recherches BLAST mais...
  - Toutes les options ne sont pas toujours accessibles.
  - Peu sont exhaustifs du point de vue des banques de données accessibles.
  - Tous ne permettent pas d'accéder à des banques mises à jour quotidiennement.
  - Les possibilités de filtrage pré- ou post-recherche sont rares et limitées.
  - Généralement pas de lien directs avec d'autres applications (*e.g.*, alignements multiples).

BLAST au NCBI

- Répond à (quasiment) toutes les questions précédentes :
  - Accès aux options.
  - Mises à jour en continu.
  - Filtrage taxonomique pré- et post-recherche.
  - Alignements multiples.
- Est particulièrement rapide.
- Bénéficie d'une interface graphique de visualisation des résultats.
- Est très sollicité !
  - Utilisation en matinée recommandée pour les européens si calculs lourds (*e.g.*, PSI-BLAST).

Quelle approche adopter ?

- Stratégie de recherche (nucléique ou protéique).
- Choix d'un algorithme.
- Banque sur laquelle effectuer la recherche.
- Choix de la matrice de substitution.
- Choix des paramètres (pondération des *gaps*, longueurs des mots, seuils, etc.)
- Traitement du bruit de fond.
- Répétition de la recherche.

Du bon usage de BLAST

- L'annotation par similarité peut conduire à certains abus...

```
MZEORFG     ILNSPDRACNLAKQAFDEAISELDSLGEESYKDSTLIMQLLXDNLTLWTSDTNEDGGDE
BOV1433P    IQNAPEQACLLAKQAFDDAIAELDTLNEDSYKDSTLIMQLLRDNLTLWTSDQQDEEAGE
            * * *:.** *******:** ***.* *:************ ********* .::   *

Score = 87.4 bits (213), Expect = 1e-17
Identities = 41/59 (69%), Positives = 50/59 (84%)

LOCUS       BOV1433P     1696 bp    mRNA            MAM       26-APR-1993
DEFINITION  Bovine brain-specific 14-3-3 protein eta chain mRNA, complete
            cds.

LOCUS       MZEORFG       187 bp    mRNA            PLN       31-MAY-1994
DEFINITION  Zea mays putative brain specific 14-3-3 protein, tau protein
            homolog  mRNA, partial cds.
```

Similarités faibles

```
SéqA    CGRRLILFMLATCGECDTDSSE-...-HICCIKQCDVQDIIRVCC
        ||   |        |||              ||    |         |
SéqB    CGSHLVEALYLVCGERGFFYTP-...-EQCCTSICSLYQLENYCN
          |||    |      |   |              ||   |   |
SéqC    YQSHLLIVLLAITLECFFSDRK-...-KRQWISIFDLQTLRPMTA
```

- Les comparaison par paires présentent des limitations dans le cas de similarités limitées à quelques acides aminés :
  - Paire (A, B) : 25% d'identité.
  - Paire (B, C) : 25% d'identité.
  - Triplet (A, B, C) : $< 5\%$ d'identité.

Exemple : famille de l'insuline

```
                    Chaîne B                                   Chaîne A

Q14641    ELRGCGPRFGKHLLSYCPMPEKTFTTTPGG...[x]58.....SGRHRFDPFCCEVICDDGTSVKLCT
P51460    REKLCGHHFVRALVRVCGGPRWSTEA.......[x]51.....AAATNPARYCCLSGCTQQDLLTLCPY
P04808    VIKLCGRELVRAQIAICGMSTWS..........[x]109.....PYVALFEKCCLIGCTKRSLAKYC
P26732    VHTYCGRHLARTLADLCWEAGVD..........[x]25.........GIVDECCLRPCSVDVLLSYC
P26733    ARTYCGRHLADTLADLCF--GVE..........[x]23.........GVVDECCFRPCTLDVLLSYCG
P26735    SQFYCGDFLARTMSILCWPDMP............[x]25.........GIVDECCYRPCTTDVLKLYCDKQI
P26736    GHIYCGRYLAYKMADLCWRAGFE..........[x]25.........GIADECCLQPCTNDVLLSYC
P15131    VARYCGEKLSNALKLVCRGNYNTMF........[x]58.........GVFDECCRKSCSISELQTYCGRR
P07223    RRGVCGSALADLVDFACSSSNQPAMV.......[x]29.....QGTTNIVCECCMKPCTLSELRQYCP
P25289    PRGICGSNLAGFRAFICSNQNSPSMV.......[x]44.....QRTTNLVCECCFNYCTPDVVRKYCY
P80090    PRGLCGSTLANMVQWLCSTYTTSSKV.......[x]30.....ESRPSIVCECCFNQCTVQELLAYC
P31241    PRGICGSDLADLRAFICSRRNQPAMV.......[x]44.....QRTTNLVCECCYNVCTVDVFYEYCY
P91797    PRGLCGNRLARAHANLCFLLRNTYPDIFPR...[x]86...EVMAEPSLVCDCCYNECSVRKLATYC
P22334    AEYLCGSTLADVLSFVCGNRGYNSQP........[x]31.........GLVEECCYNVCDYSQLESYCNPYS
P01308    NQHLCGSHLVEALYLVCGERGFFYTPKT.....[x]35.........GIVEQCCTSICSLYQLENYCN
P01343    PETLCGAELVDALQFVCGDRGFYF..........[x]12.........GIVDECCFRSCDLRRLEMYCAPLK
P01344    SETLCGGELVDTLQFVCGDRGFYF..........[x]12.........GIVEECCFRSCDLALLETYCATPA
              **         .*                                  **   *    .   *
```

Recherche par profils

- Recherche d'un ensemble d'homologues proches.
- Alignement de ces homologues entre eux.
- Calcul d'une matrice de score position-spécifique (ou *profil*) à partir de l'alignement entier ou d'une région définie :
- Recherche dans la banque en utilisant la matrice au lieu de la séquence :
  - Sélection des *hits* ayant un score supérieur à un seuil.
- Éventuellement, modification itérative de la matrice en incorporant les séquences détectées.

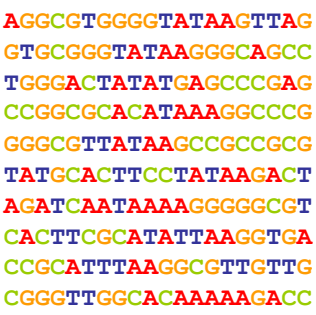
Construction de la matrice

- Soit π_i la fréquence du résidu i dans l'ensemble des séquences utilisées pour construire le profil.
- Soit n le nombre de séquences dans l'alignement et n_{ik} le nombre de résidus i à la position k de l'alignement.
- Positions contenant un ou plusieurs *gaps* :
  - Soit ignorées, soit prises en compte en considérant qu'un *gap* constitue un état de caractère supplémentaire.
- Calcul du score δ_{ik} associé au résidu i à la position k :

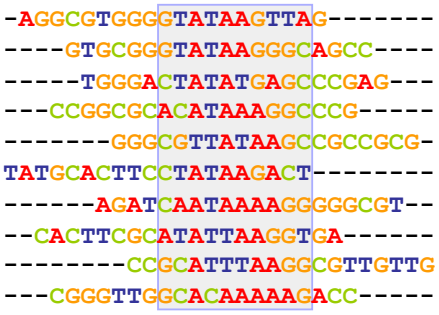
$$\delta_{ik} = 2 \log_2 \left[\frac{n_{ik} + 1}{\pi_i (n + 1)} \right]$$

avec arrondissement à l'entier le plus proche.

Exemple de construction



Famille de séquences



Alignement

A	2	1	9	0	8	7	6	6	1	1
C	4	3	0	1	0	0	0	0	1	4
G	4	1	0	0	0	0	4	3	7	3
T	0	5	1	9	2	3	0	1	1	2

Matrice des fréquences

A	-1	-2	2	-4	2	2	1	1	-2	-2
C	4	3	-1	1	-1	-1	-1	-1	1	4
G	2	-1	-3	-3	-3	-3	2	1	3	1
T	-3	2	-1	4	0	1	-3	-1	-1	0

Matrice de score

Exemples de recherche

Base	Position									
	1	2	3	4	5	6	7	8	9	10
A	-1	-2	2	-4	2	2	1	1	-2	-2
C	4	3	-1	1	-1	-1	-1	-1	1	4
G	2	-1	-3	-3	-3	-3	2	1	3	1
T	-3	2	-1	4	0	1	-3	-1	-1	0

$S_{\max} = 27$

Matrice de pondération

G C A G T A T A A G G G G A...

Fenètre glissante

```
G C A G T A T A A G                S=+2+3+2-3+0+2-3+1-2+1= +3
. C A G T A T A A G G              S=+4-2-3+4+2+1+1+1+3+1=+12
. . A G T A T A A G G G            S=-1-1-1-4+0+2+1+1+3+1= +1
. . . G T A T A A G G G G          S=+2+2+2+4+2+2+2+1+3+1=+21
. . . . T A T A A G G G G A        S=-3-2-1-4+2-3+2+1+3-2= -7
```

Exemples de recherche

Base	Position									
	1	2	3	4	5	6	7	8	9	10
A	-1	-2	2	-4	2	2	1	1	-2	-2
C	4	3	-1	1	-1	-1	-1	-1	1	4
G	2	-1	-3	-3	-3	-3	2	1	3	1
T	-3	2	-1	4	0	1	-3	-1	-1	0

Matrice de pondération

$S_{\max} = 27$

G A A A G G T G A G T C A T . . .

Fenètre glissante

```
G A A A G G T G A G                S=+2-2+2-4-3-3-3+1-2+1=-11
. A A A G G T G A G T              S=-1-2+2-3-3+1+2+1+3+0=  0
. . A A G G T G A G T C            S=-1-2-3-3+0-3+1+1-1+4= -7
. . . A G G T G A G T C A          S=-1-1-3+4-3+2+2-1+1-2= -2
. . . . G G T G A G T C A T        S=+2-1-1-3+2-3-3-1-2+0=-10
```


Test de significativité

- Soit S_{obs}, le meilleur score observé sur une séquence requête :
  - Dans notre exemple $S_{\text{obs}} = 21$.
- Génération de B séquences aléatoires ($B \geq 100$) :
  - Même longueur et même composition que la séquence requête.
  - Soit b, le nombre de ces séquences où l'on observe un score $S \geq S_{\text{obs}}$.
  - Dans ce cas la mesure de la significativité est donnée par :

$$\mathbb{P}(S \geq S_{\text{obs}}) = b/B$$

- Dans notre exemple, pour $B = 1000$, une seule séquence générée aléatoirement avait un score $S \geq 21$:

$$\mathbb{P}(S \geq 21) = 1/1000$$

PSI-BLAST

- *Position-Specific Iterated* BLAST :
  1. Recherche BLAST classique.
  2. Calcul d'un profil au moyen des *hits* significatifs sélectionnés par l'utilisateur (utilisation de la E-value).
  3. Nouvelle recherche en utilisant le profil.
  4. Répétition des étapes 2-3 jusqu'à convergence.

Profils généralisés

- Intègrent des modèles de Markov cachés (*Hidden Markov Models* – HMM) au cours de la construction du profil.
- Programmes disponibles :
  - PFSEARCH (Bücher *et al.*, 1996) :
    - Écrit en Fortran 77.
    - Très lent.
  - HMMER (Eddy, 1998) :
    - Écrit en C.
    - Rapide (comparable à PSI-BLAST).

Alignements multiples

- Généralisation de Needleman et Wunsch ?
  - Pour n séquences de longueur ℓ :
    - Complexité en $O(\ell^n)$.
    - Croissance exponentielle du temps de calcul et de l'espace mémoire.
- Utilisation d'heuristiques :
  - Approximation de l'alignement optimal.

n	t	t
2	0.01 sec	10^{-4} sec
3	1.5 sec	0.015 sec
4	225 sec	2.25 sec
5	9h22 min	5.6 min
6	58 jours	14.06 h
7	24 ans	87.9 jours
8	3612 ans	36 ans

Temps de calcul pour aligner n séquences de longueur $\ell = 150$

Principe de l'alignement progressif

- Construction itérative par ajout progressif de séquences :
  - Établissement d'une *matrice de distances* entre toutes les paires possibles de séquences.
  - Construction d'un *arbre guide* à partir de cette matrice.
  - Ajout de paires et/ou de séquences individuelles dans l'alignement en remontant dans la structure de cet arbre.
- Nombreuses implémentations disponibles :
  - Clustal (Higgins *et al.*, 1988).
  - T-Coffee (Notredame *et al.*, 2000).
  - Mafft (Katoh *et al.*, 2002).
  - Muscle (Edgar, 2004).
  - ProbCons (Do *et al.*, 2005).

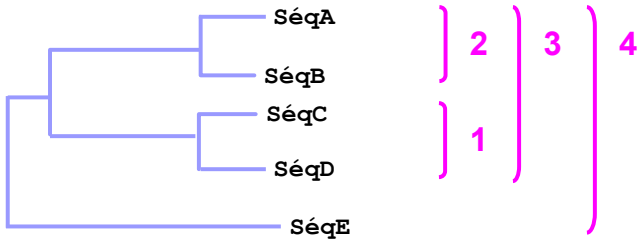
Procédure de Clustal

- Alignement de toutes les paires possibles au moyen d'une heuristique de l'algorithme de Needleman et Wunsch.
- Construction d'une matrice de distance en utilisant les scores d'alignement par paires.
- Construction de l'arbre guide en utilisant cette matrice :
  - La 1ère paire regroupe les deux séquences présentant la distance la plus faible.
  - L'alignement en question est « fixé » et ne changera plus par la suite :
    - Si un *gap* doit être introduit, il le sera à la même position dans les deux séquences.
    - Application du même principe pour toutes les séquences ajoutées par la suite.

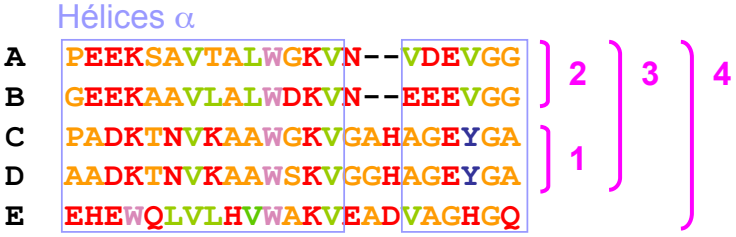
Illustration

SéqB	0.17			
SéqC	0.59	0.60		
SéqD	0.59	0.59	0.13	
SéqE	0.77	0.77	0.75	0.75
	SéqA	**SéqB**	**SéqC**	**SéqD**

Calcul de tous les alignements simples et construction de la matrice de distance



Arbre guide par *Neighbour-joining*



Alignement progressif

Pondérations initiale des *gaps*

- Fonction affine pour les *gaps* dans les alignements par paires.
- Correction des valeurs de δ_o(-) et δ_e(-) en fonction de différents facteurs :
  - Le degré de similarité entre les séquences :

$$\delta_o(\text{-}) \propto \%\ \text{identité(A, B)}$$

  - La longueur des séquences :

$$\delta_o(\text{-}) \propto \log[\min(m, n)]$$

  - La différence de longueur entre les séquences :

$$\delta_e(\text{-}) \propto 1.0 + |\log(n/m)|$$

Pondérations supplémentaires

- Prises en compte au moment du groupement des alignements :
  - Diminution de la pénalité à l'emplacement de *gaps* préexistants.
  - Augmentation de la pénalité au voisinage (8 résidus) de *gaps* préexistants.
  - Réduction de la pénalité au niveau de régions contenant une suite d'acides aminés hydrophiles ($\geq$ 5 résidus) :
    - Lien avec la structure 3D des protéines globulaires.
  - Modification spécifiques en fonction des acides aminés présents (*e.g.*, la pénalité est plus faible avec Gly, Asn, Pro).

Le succès de Clustal

- Fut l'un des tous premiers programmes réellement utilisable disponibles (1ère version en 1988).
- Temps de calcul raisonnable pour des jeux de données de taille importante ($n \leq 500$).
- Fonctionalité de calcul d'arbres phylogénétiques (méthode du *Neighbour-Joining*) avec *bootstrap*.
- Utilisable sur la quasi-totalité des architectures disponibles à l'époque :
  - Windows, Unix/Linux, MacOS, VMS.
- Interface graphique (ClustalX).

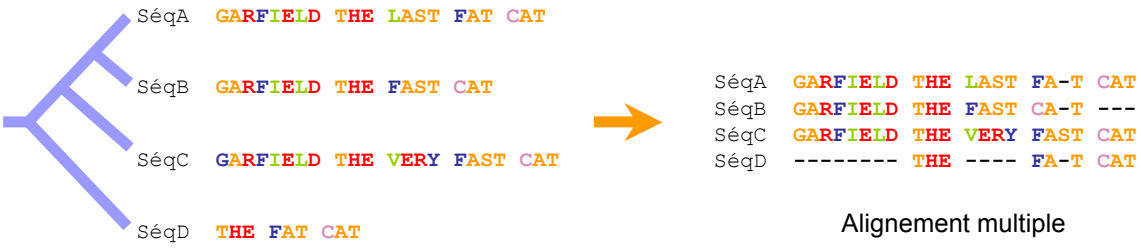
Limitations

- Perte de l'optimalité au moment du regroupement des paires :
  - Existence de minima locaux.
  - Importance de l'ordre dans lequel sont regroupées les séquences.
  - Impossibilité de corriger ces erreurs par la suite.
- Pas de fonction objective :
  - Impossibilité de déterminer la qualité de l'alignement.
- Désormais beaucoup moins performant que ses concurrents directs :
  - Développement d'une nouvelle version appelée ClustalΩ (Sievers *et al.*, 2011).

Importance de l'arbre guide

```
SéqA  GARFIELD THE LAST FAT CAT
SéqB  GARFIELD THE FAST CAT ---

SéqA  GARFIELD THE LAST FA-T CAT
SéqC  GARFIELD THE VERY FAST CAT

SéqA  GARFIELD THE LAST FAT CAT
SéqD  -------- THE ---- FAT CAT

SéqB  GARFIELD THE ---- FAST CAT
SéqC  GARFIELD THE VERY FAST CAT

SéqB  GARFIELD THE FAST CAT
SéqD  -------- THE FA-T CAT

SéqC  GARFIELD THE VERY FAST CAT
SéqD  -------- THE ---- FA-T CAT
```

Alignement par paires



Arbre guide

Alignement multiple

Alignement progressif

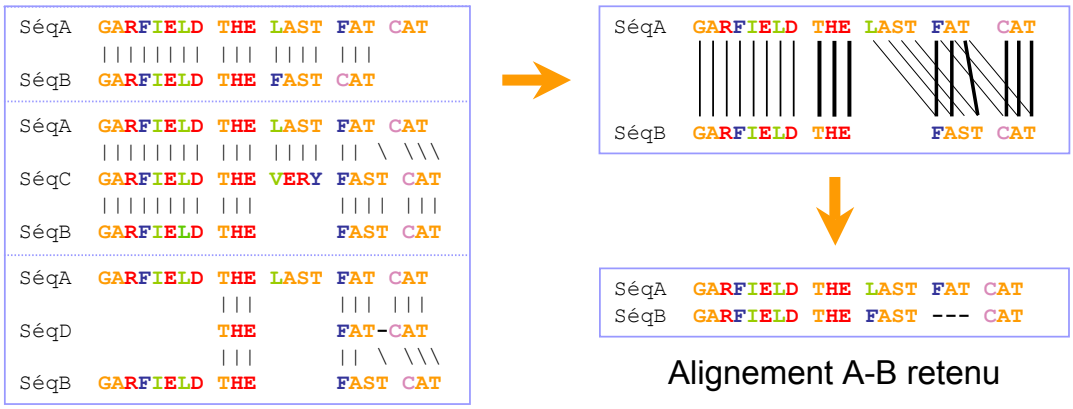
Contraintes de consistance

- Utilisées pour améliorer les méthodes d'alignement progressif.
- Prise en compte de combinaisons consistantes de séquences :
  - Données intrinsèques (alignements par paires et par triplets).
  - Données extrinsèques (*e.g.*, structures 3D).
- Introduites dans différents programmes :
  - T-Coffee (Notredame *et al.*, 2000).
  - ProbCons (Do *et al.*, 2005).
  - ProbAlign (Roshan et Livesay, 2006).

Stratégie de T-Coffee

```
SéqA  GARFIELD THE LAST FAT CAT        SéqB  GARFIELD THE ---- FAST CAT
SéqB  GARFIELD THE FAST CAT ---        SéqC  GARFIELD THE VERY FAST CAT

SéqA  GARFIELD THE LAST FA-T CAT       SéqB  GARFIELD THE FAST CAT
SéqC  GARFIELD THE VERY FAST CAT       SéqD  -------- THE FA-T CAT

SéqA  GARFIELD THE LAST FAT CAT        SéqC  GARFIELD THE VERY FAST CAT
SéqD  -------- THE ---- FAT CAT        SéqD  -------- THE ---- FA-T CAT
```

Alignements par paires de départ



Alignement A-B retenu

Aminoacyl-ARNt synthétases d'*E. coli*

```
SYL_ECOLI   HMGHVRNYTIGDVIARYQRMLGKNVLQPIGWDAFGLPAEGAAVKNNTAPA----------/…/
SYV_ECOLI   HMGHAFQQTIMDTMIRYQRMQGKNTLWQVGTDHAGIATQMVVERKIAAEEGKTRHDYGRE/…/
SYI_ECOLI   HIGHSVNKILKDIIVKSKGLSGYDSPYVPGWDCHGLPIELKVEQEYGKPG----EKFTAA/…/
SYM_ECOLI   HLGHMLEHIQADVWVRYQRMRGHEVNFICADDAHGTPIMLKAQQLGITPE----Q-----/…/
            *:**  :    *   : : :  * :    . *   * .    . :

SYL_ECOLI   LVYTGMSKMSKSKNNGIDPQVMVER-----------------------------------
SYV_ECOLI   -------KMSKSKGNVIDPLDMVDGISLPELLEKRTGNMMQPQLADKIRKRTEKQFPNGI
SYI_ECOLI   -------KMSKSIGNTVSPQDVMNK-----------------------------------
SYM_ECOLI   -------KMSKSRGTFIKASTWLNH-----------------------------------
                  ***** .. :..   ::
```

Alignement de T-Coffee

```
SYL_ECOLI   HMGHVRNYTIGDVIARYQRMLGKNVLQPIGWDAFGLPAEGAAVKNNTAPAP---------/…/
SYV_ECOLI   HMGHAFQQTIMDTMIRYQRMQGKNTLWQVGTDHAGIATQMVVERKIAAEEGKTRHDYGRE/…/
SYI_ECOLI   HIGHSVNKILKDIIVKSKGLSGYDSPYVPGWDCHGLPIELKVEQEYGKPGEK----FTAA/…/
SYM_ECOLI   HLGHMLEHIQADVWVRYQRMRGHEVNFICADDAHGTPIMLKAQQLGITPEQMIG------/…/
            * :*:**  :   *   : : : * :    . *  * .    . :

SYL_ECOLI   MHLLYFRFFHKLMRDAGMVNSDEPAKQLLCQG--MVLADAFYYVGENGERNWVS------
SYV_ECOLI   EGQKMSKSKGNVIDPLDMVDGISLPELLEKRTGNMMQPQLADKIRKRTEKQFPNGIEPHG
SYI_ECOLI   APYRQVLTHGFTVDGQGRKMSKSIGNTVSPQD-------VMNKLGADILRLWVASTDYTG
SYM_ECOLI   ------------------------------------------------------------
```

Alignement de Clustal

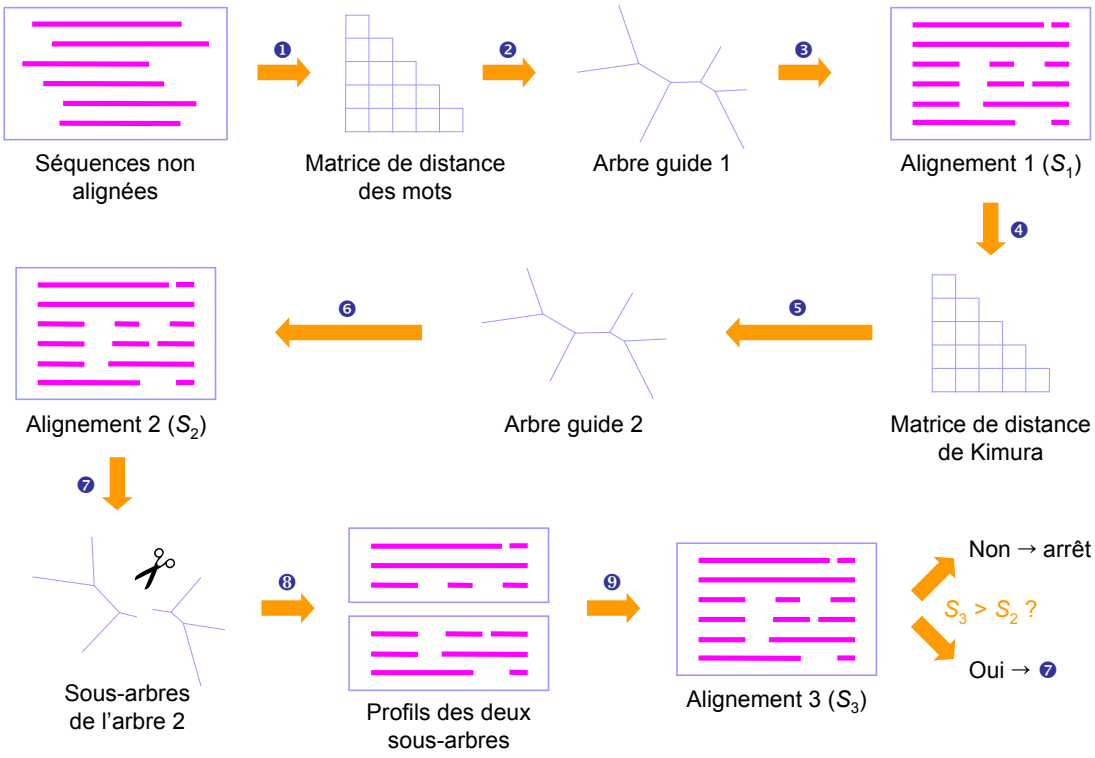
Interêt et limitations

- Résultats très supérieurs à ceux de ClustalW.
- Possibilité d'améliorer les alignements en utilisant des données structurales.
- Intègre une fonction objective d'évaluation de la qualité des alignements.
- Très gourmand en mémoire et en temps de calcul :
  - Jeux de données de taille limitée ($n \leq 100$).

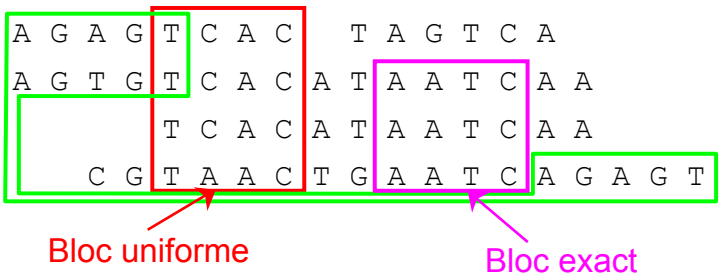
Amélioration par itération

- Premier alignement par la méthode progressive.
- Raffinements successifs de l'alignement de départ jusqu'à convergence.
- Implémentation la plus performante réalisée dans le programme Muscle :
  - Rapide.
  - Peut travailler sur des jeux de données de très grande taille (plusieurs milliers de séquences).
  - Bon résultats en terme de qualité des alignements.

Procédure de Muscle



Alignement par blocs

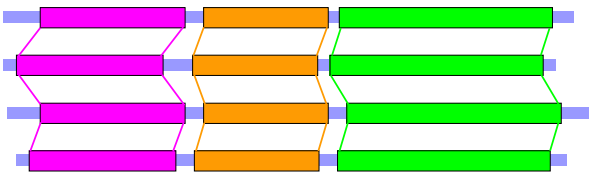


- Recherche des blocs similaires sans *gaps*.
- Sélection de la meilleure combinaison de blocs compatibles entre eux (heuristique).
- Plus lent que les alignements progressifs.
- Implémentation disponible :
  - Dialign2 (Subramanian *et al.*, 2008).

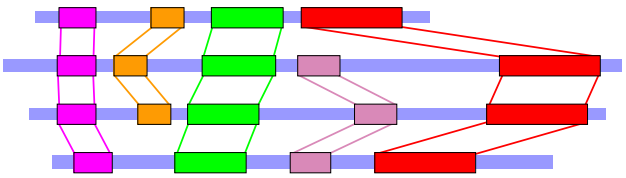
Jeux de données de référence

- Banques créées manuellement, souvent à partir d'alignements structuraux :
  - BAliBASE (Thompson *et al.*, 1999).
  - PALI (Balaji *et al.*, 2001).
  - OXBench (Raghava *et al.*, 2003).
  - PREFAB (Edgar, 2004).
  - SABmark (Van Walle *et al.*, 2005).
  - HomFam (Blackshields *et al.*, 2010).
- Utilisées pour évaluer la qualité des algorithmes disponibles :
  - Certains programmes « passent » mieux sur certaines références que sur d'autres.

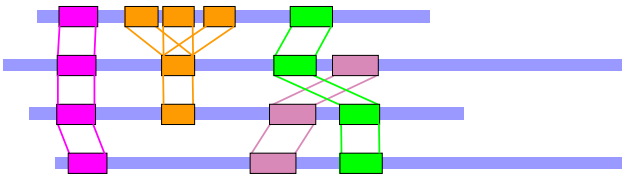
Programmes recommandés



Clustal
Multalin



Dialign2
Muscle
T-Coffee



Match-Box
MEME
PIMA

Cas particuliers

- Alignement de séquences codantes :
  - SeaView (Gouy *et al.*, 2010) :
    - Alignement des séquences protéiques.
    - Calage du nucléique sur l'alignement protéique.
- Alignement ADNc / ADN génomique :
  - SIM4 (Florea *et al.*, 1998).
- Alignement protéine / ADN :
  - GeneWise (Birney *et al.*, 2004).

Quelques conseils

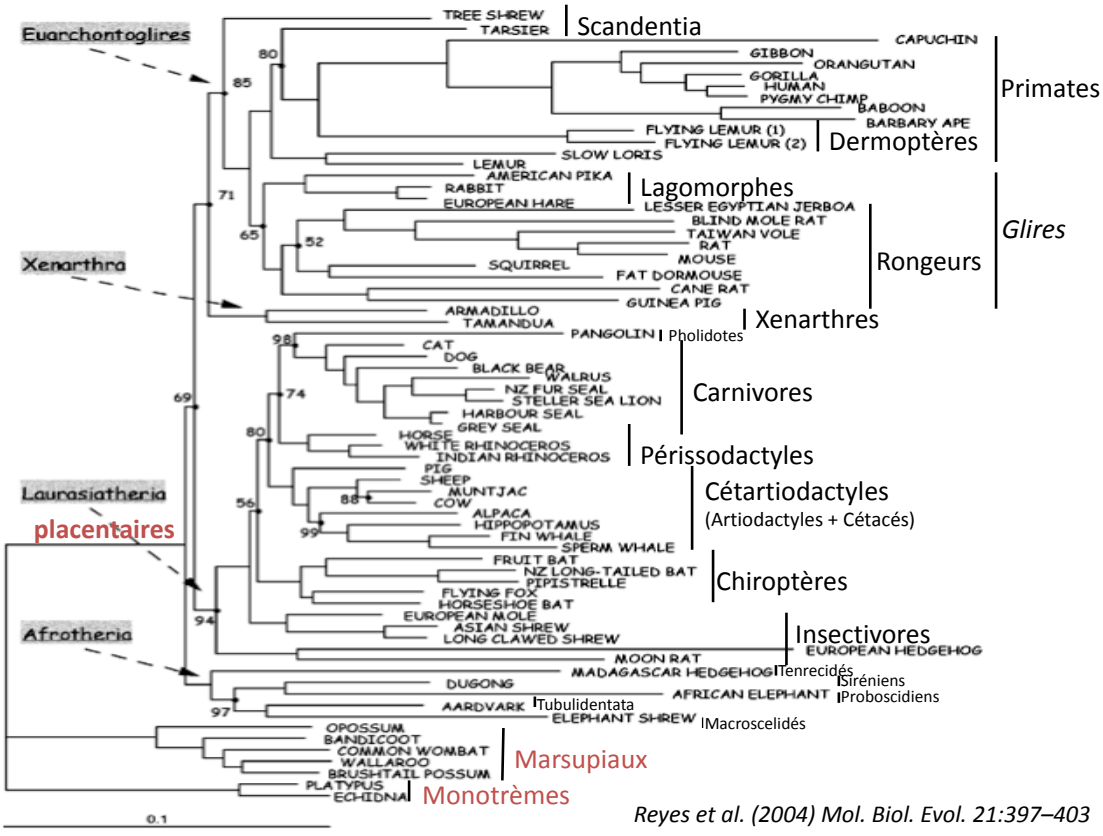
- Considérez les résultats avec recul.
- Pour la phylogénie, n'utilisez que les sites pour lesquels l'hypothèse d'homologie est vraisemblable.
- Essayez d'identifier les régions dont vous pensez qu'elles sont correctement alignées :
  - Un alignement multiple peut (parfois) être amélioré « à la main » à l'aide d'un éditeur :
    - SeaView, STRAP, CINEMA, etc.
  - Il existe des programmes permettant de filtrer automatiquement les régions mal alignées :
    - Gblocks, TrimAl, BMGE, Guidance, etc.

Formation à la phylogénie moléculaire

Lyon – 20-23 mars 2018

Introduction à la phylogénie moléculaire
- aspects biologiques -

Manolo Gouy, CNRS, Lyon



Reyes et al. (2004) Mol. Biol. Evol. 21:397–403

FIG. 1.—Phylogenetic tree of placental mammals reconstructed using the program MrBayes from mitochondrial H-stranded protein-coding genes using ungapped first and second codon positions with the exclusion of Leu synonymous sites. Posterior probabilities (PP) supporting the tree nodes are only reported when less than 100. Marsupialia and Monotremata were used as outgroups. The lengths of the branches are proportional to the number of nucleotide substitutions per site.

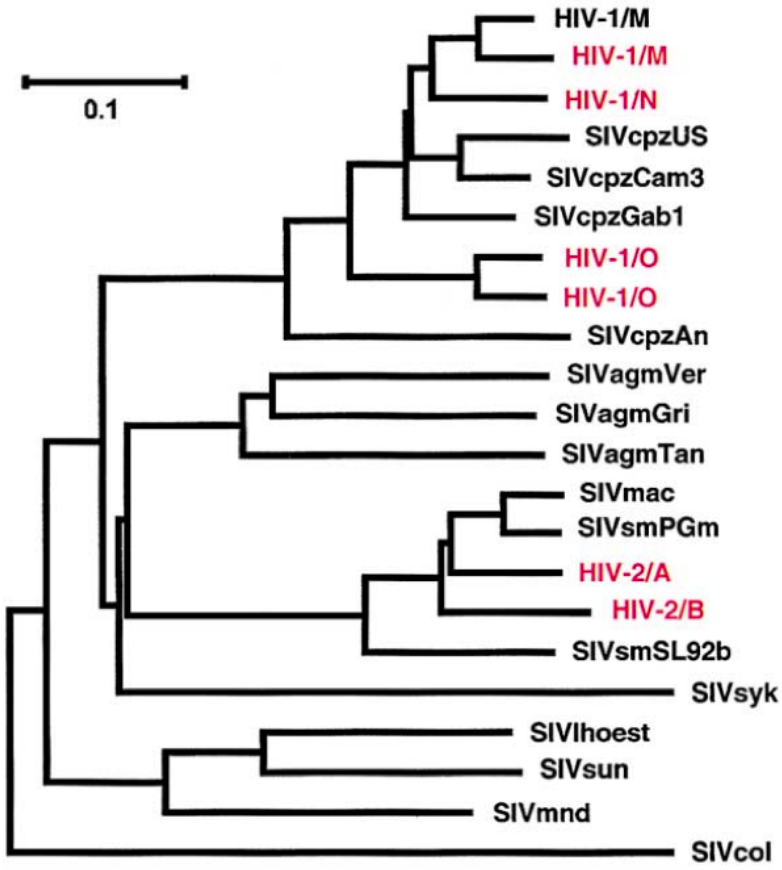


Figure 1. Evolution of AIDS Viruses

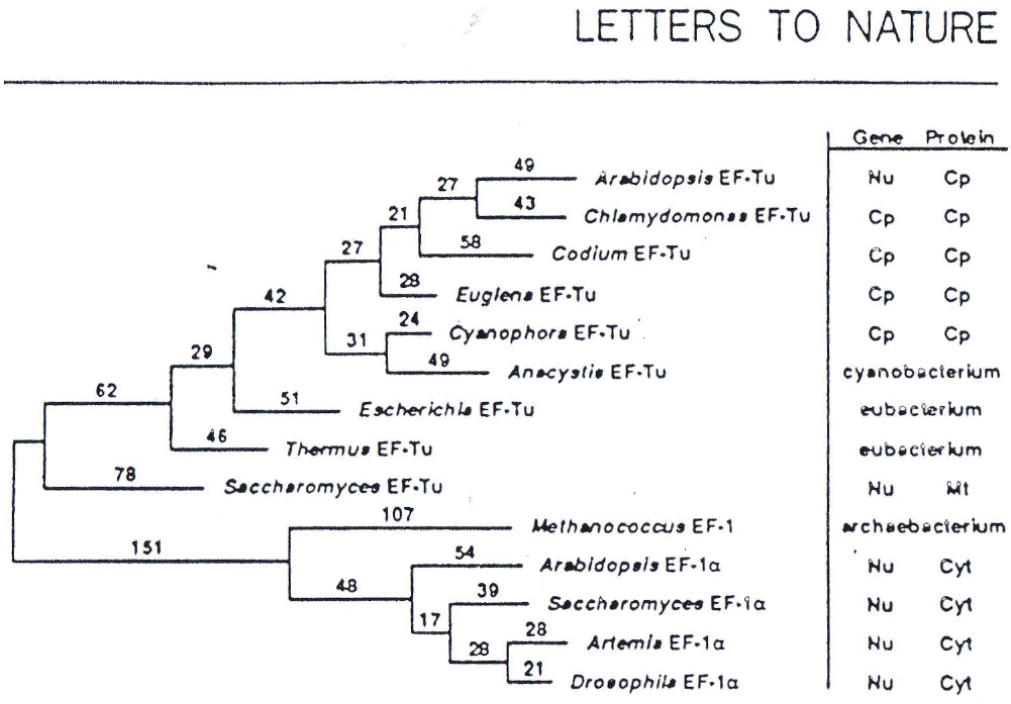
Origine du virus du SIDA

cpz: chimpanzé --> HIV-1
agm: singe vert africain
mac: macaque
sm: *Cercocebus atys* --> HIV-2
syk: *Cercopithecus albogularis*
lhoest: *C. lhoesti*
sun: *C. solatus*
mnd: *Mandrillus sphinx*
col: *Colobus guereza*

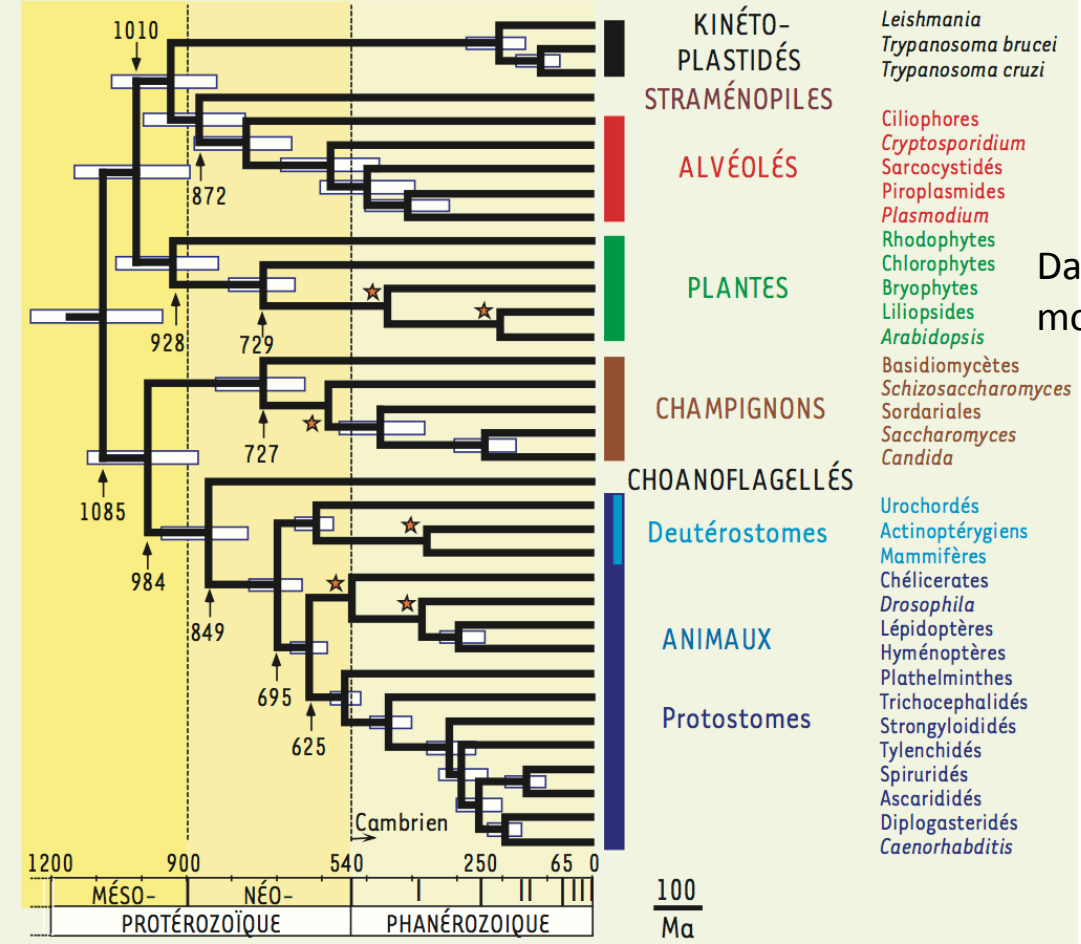
La vitesse d'évolution des molécules utilisées doit être adaptée à l'échelle temporelle du phénomène étudié.

Sharp (2002) Cell 108:305

Evolutionary transfer of the chloroplast tufA gene to the nucleus.



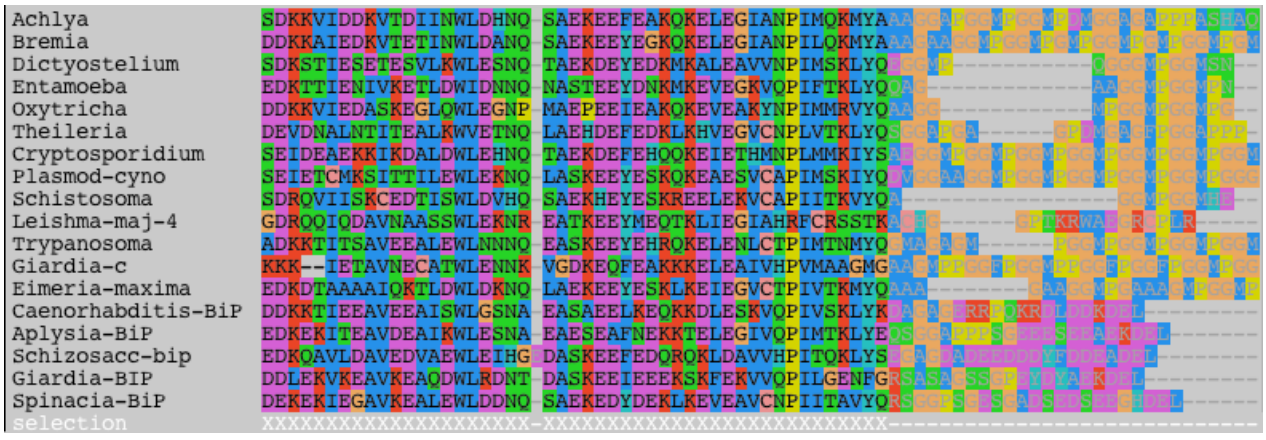
Baldauf & Palmer (1990) *Nature* 344:262.



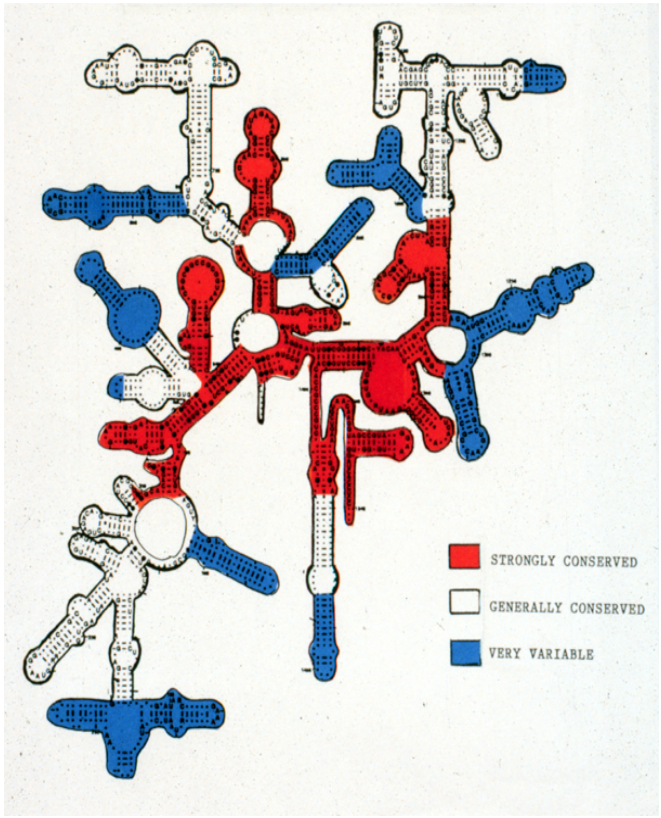
Datation moléculaire

Alignement et Gaps

- La qualité de l'alignement est essentielle : chaque colonne de l'alignement (site) est supposée contenir des résidus homologues (nucléotides, acides aminés) qui dérivent d'un ancêtre commun.
  ==> Les parties non fiables de l'alignement doivent être omises du reste des analyses.
- La plupart des méthodes ne tiennent compte que des substitutions ; les gaps (événements d'insertion/délétion) ne sont pas utilisés.
  ==> les sites contenant des gaps sont ignorés.



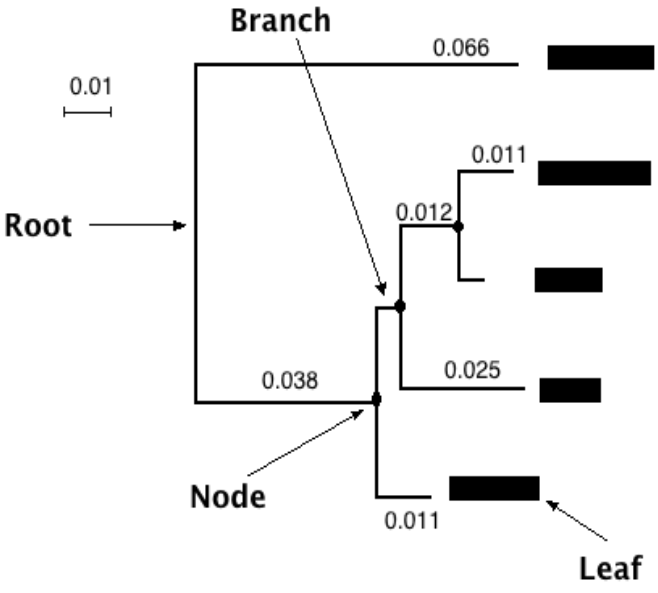
Variation de la vitesse d'évolution entre sites



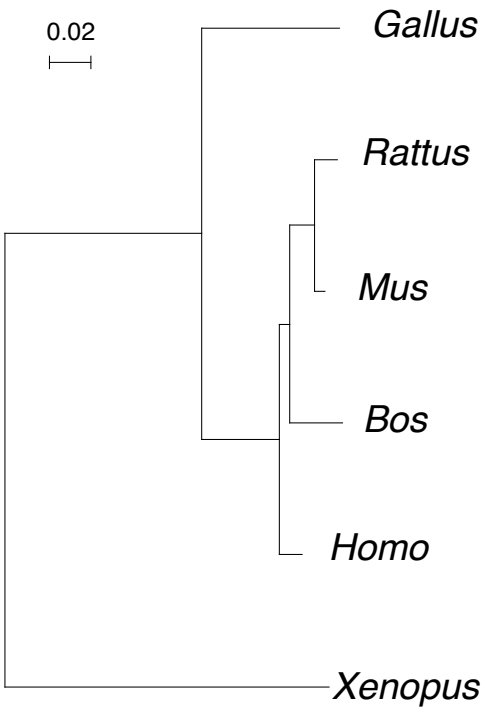
Small subunit ribosomal RNA (18S or 16S)

Arbre Phylogénétique

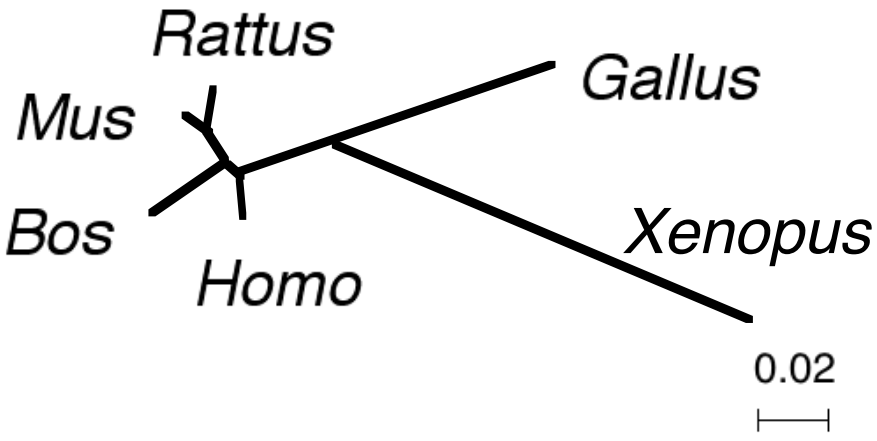
- Branche Interne: entre 2 nœuds. Branche Externe: entre un nœud et une feuille
- Les longueurs des branches horizontales sont proportionnelles aux distances évolutives entre séquences ancestrales (unité = substitution / site).
- Topologie d'arbre = forme de l'arbre = ordre de branchement des nœuds



Arbre raciné

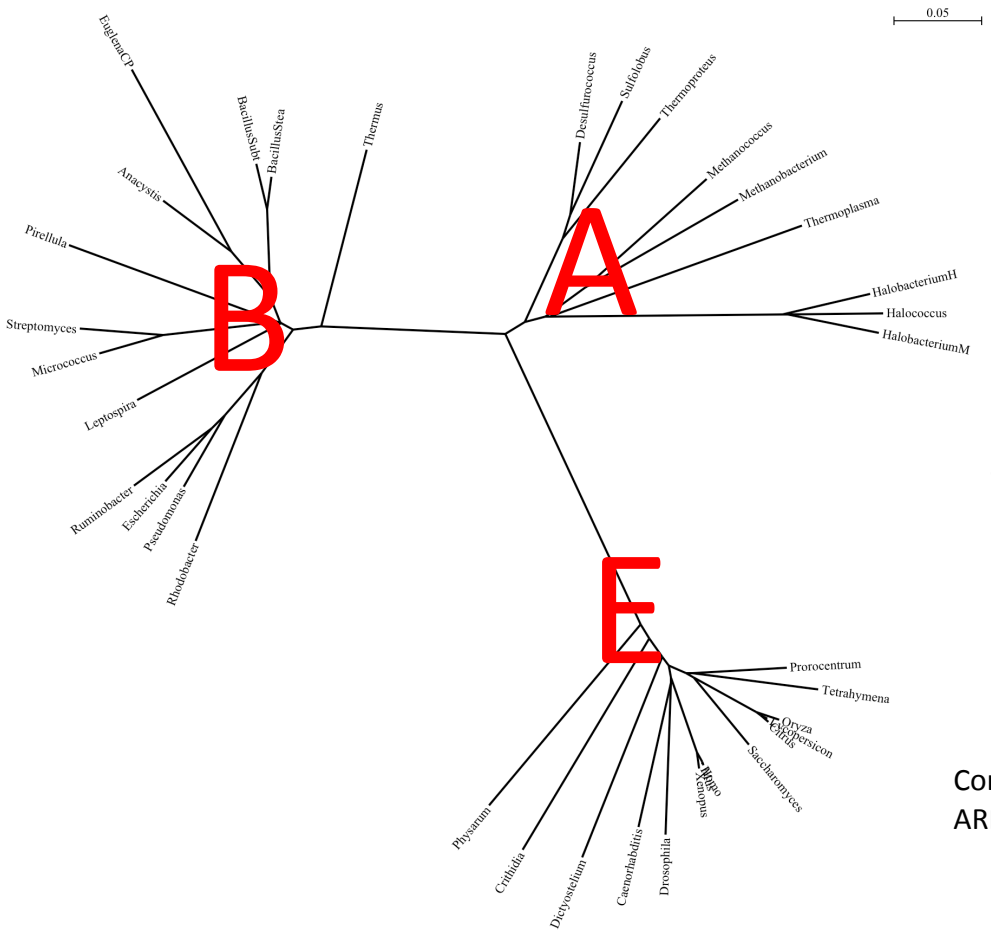


Arbre non raciné



Arbres racinés et non-racinés

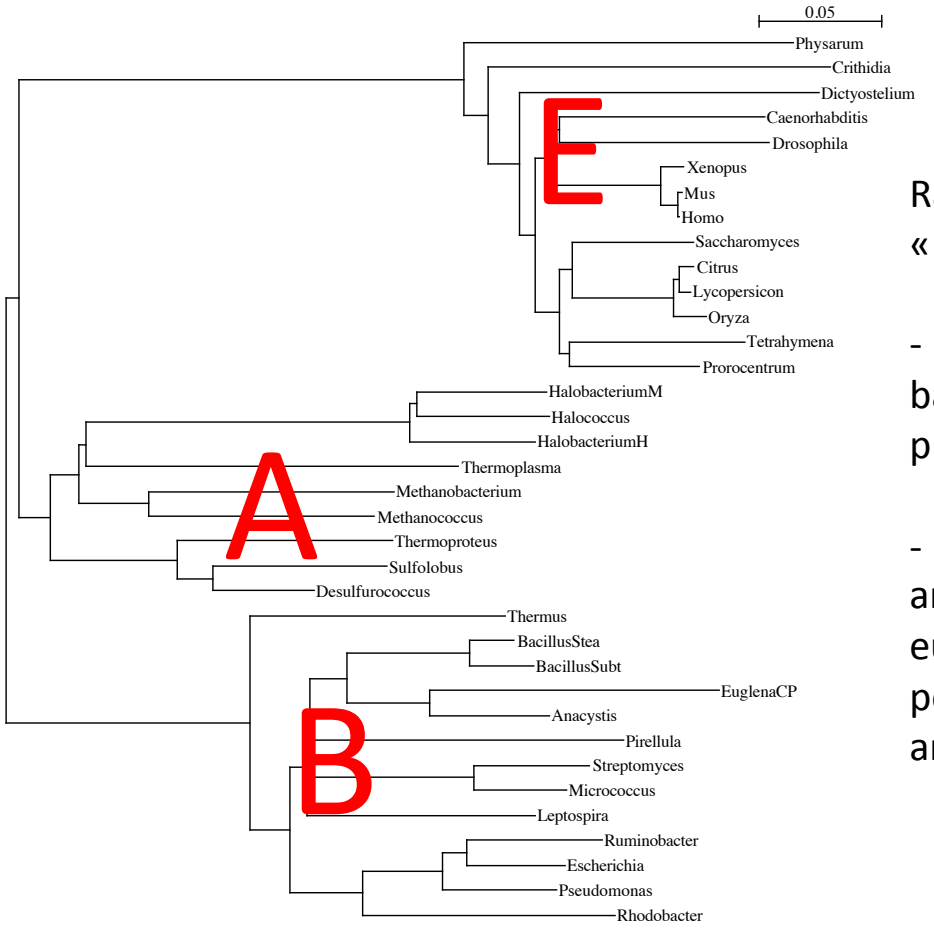
- La plupart des méthodes phylogénétiques produisent des arbres non racinés. La raison est que les méthodes détectent des différences entre séquences, sans avoir le moyen de les orienter temporellement.
- Deux façons d'enraciner un arbre non raciné:
  - Méthode du groupe externe : inclure dans l'analyse un groupe de séquences dont on sait *a priori* qu'elles sont externes au groupe étudié; la racine est sur la branche qui relie le groupe externe aux autres séquences.
  - Faire l'hypothèse de l'horloge moléculaire : toutes les lignées sont supposées évoluer à la même vitesse depuis leur divergence; la racine est au point de l'arbre équidistant de toutes ses feuilles.



Phylogénie universelle non racinée

Trois domaines
Archées
Bactéries
Eucaryotes

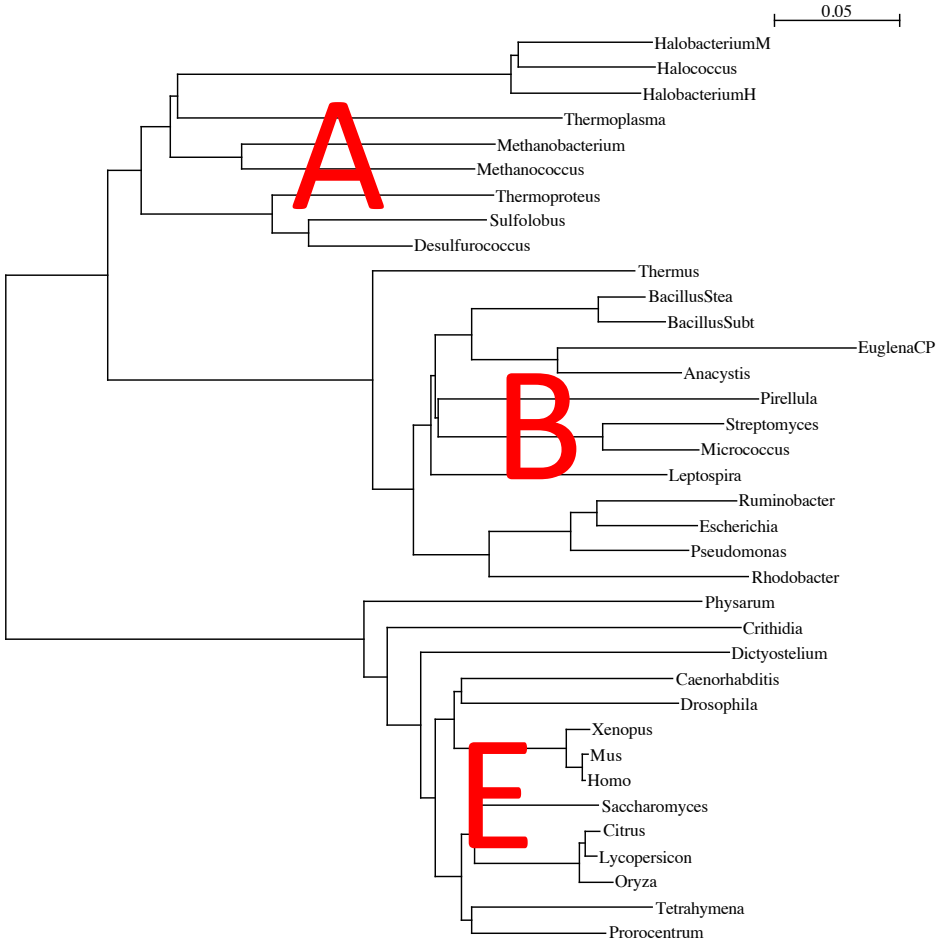
Concaténation des ARNr SSU et LSU



Racinement « standard »:

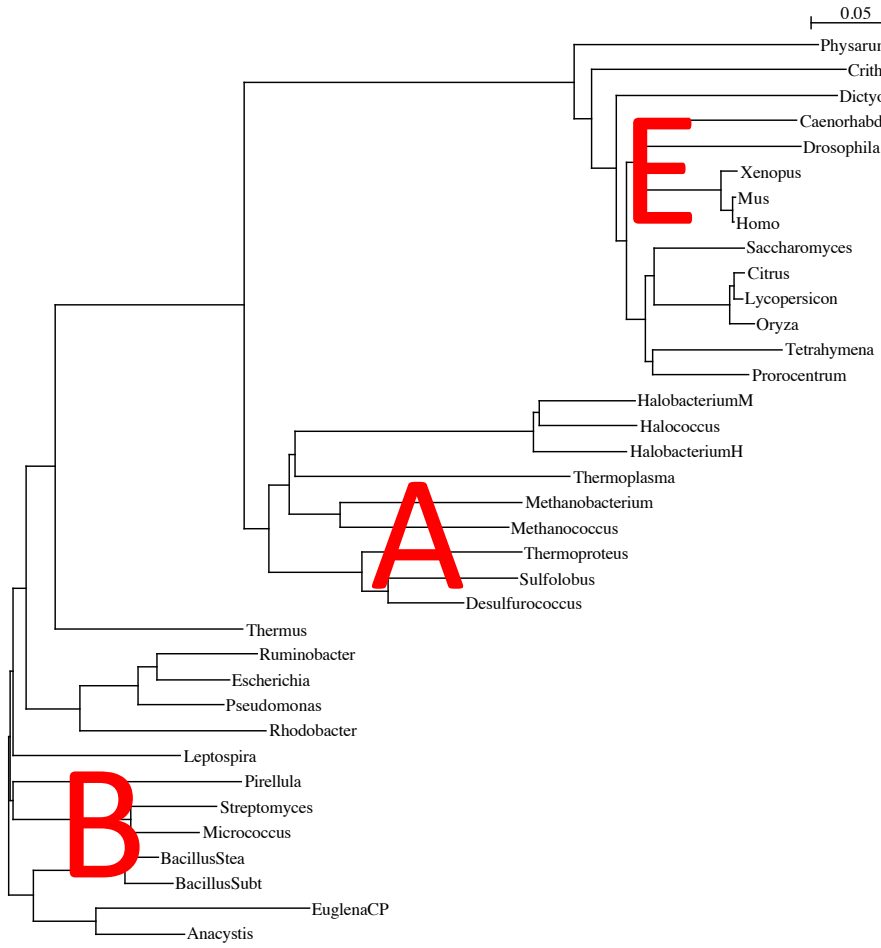
- le domaine bactérien est le premier à diverger

- les domaines archéen et eucaryote possèdent un ancêtre exclusif



Racinement par le centre :

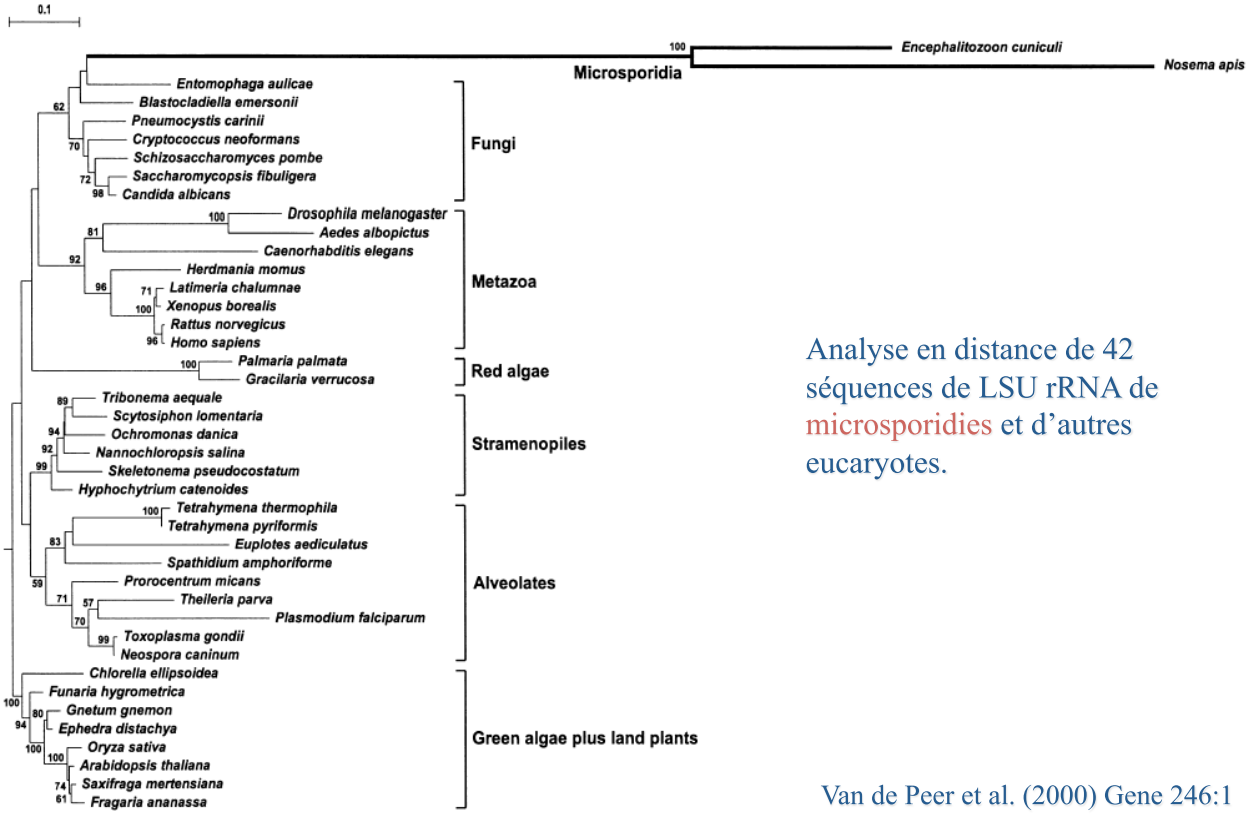
opposition procaryote/ eucaryote



Racinement à l'intérieur du domaine bactérien défendu par certains auteurs :

l'état ancestral est bactérien, les archées et les eucaryotes sont dérivés spécifiquement de certains types de bactéries.

Racinement par le centre: incorrect si fortes différences de vitesse entre lignées



Analyse en distance de 42 séquences de LSU rRNA de microsporidies et d'autres eucaryotes.

Van de Peer et al. (2000) Gene 246:1

Arbre des gènes *vs.* Arbre des espèces

- L'histoire évolutive des gènes reproduit celle des espèces qui les portent, sauf si:
  - Transfert horizontal = transfert de gène entre espèces
  - Duplication génique : orthologie/ paralogie

Orthologie / Paralogie

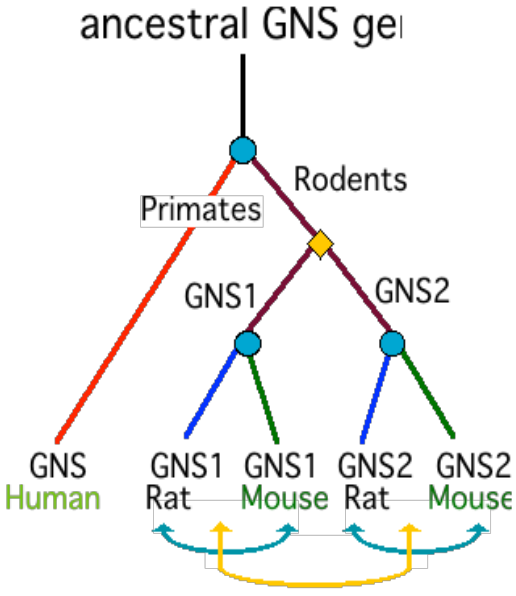
speciatio
duplicatio

Homology : two genes are homologous iff they have a common ancestor.

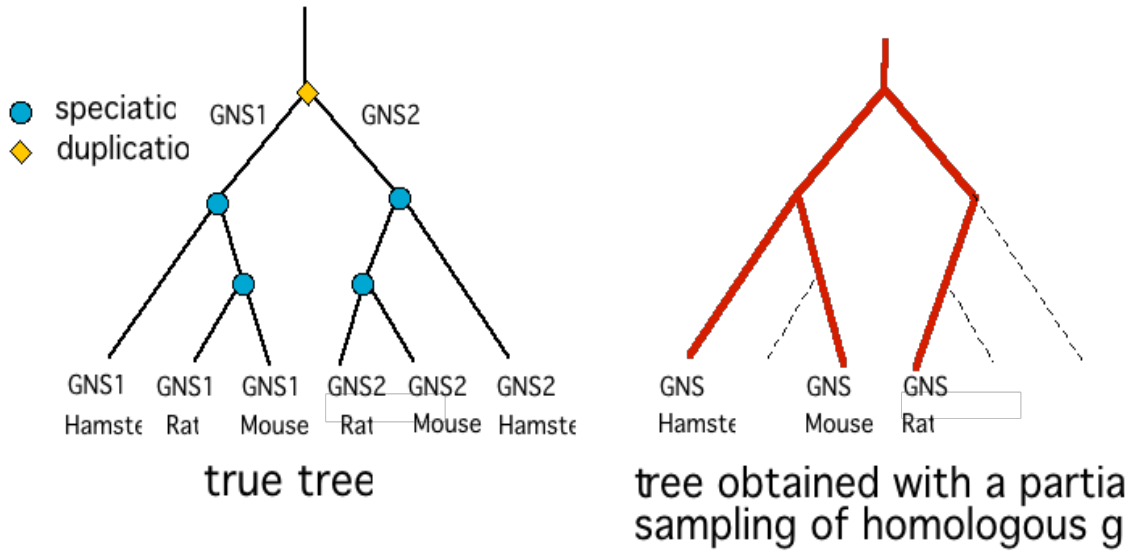
Orthology : two genes are orthologous iff they diverged following a speciation event.

Paralogy : two genes are paralogous iff they diverged following a duplication event.

Orthology ≠ functional equivalence

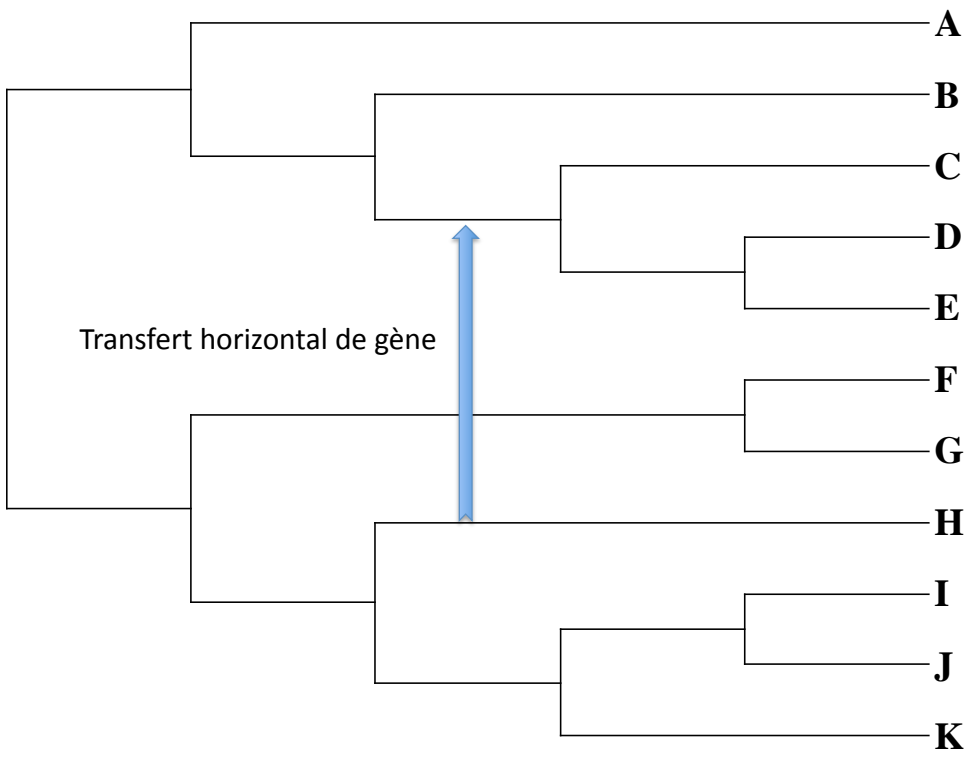


Reconstruction de la phylogénie des espèces: artéfacts dus à la paralogie

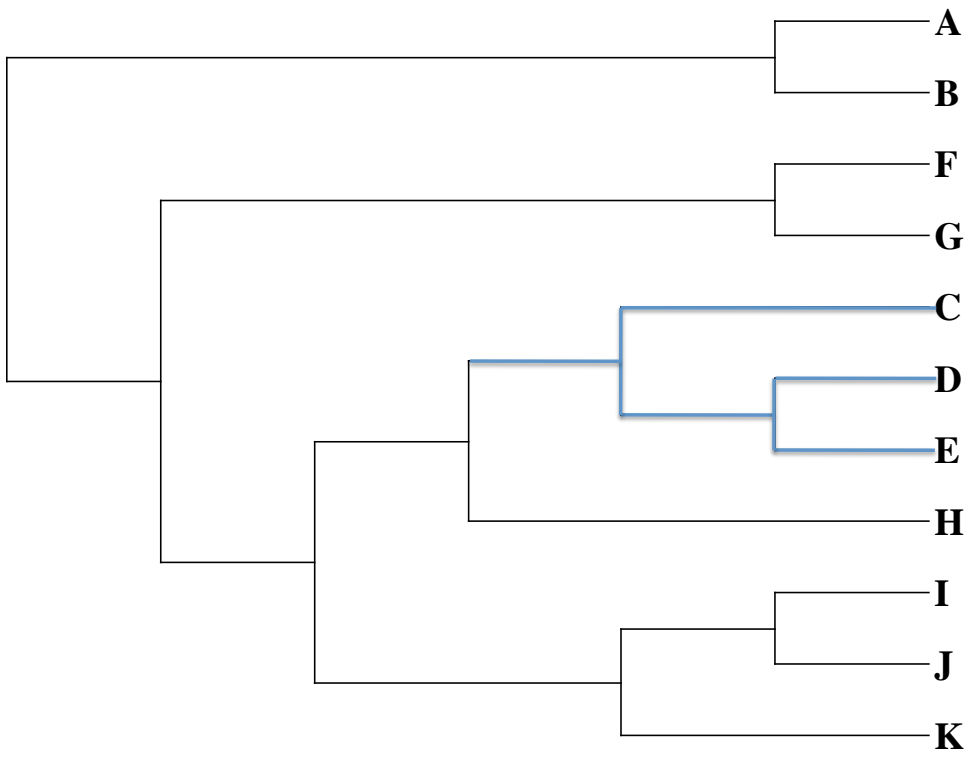


!! Des pertes de gènes peuvent se produire au cours de l'évolution : même avec des séquences génomiques complètes, il peut être difficile de détecter la paralogie !!

Arbre des espèces



Arbre du gène



L’artefact d’attraction des longues branches

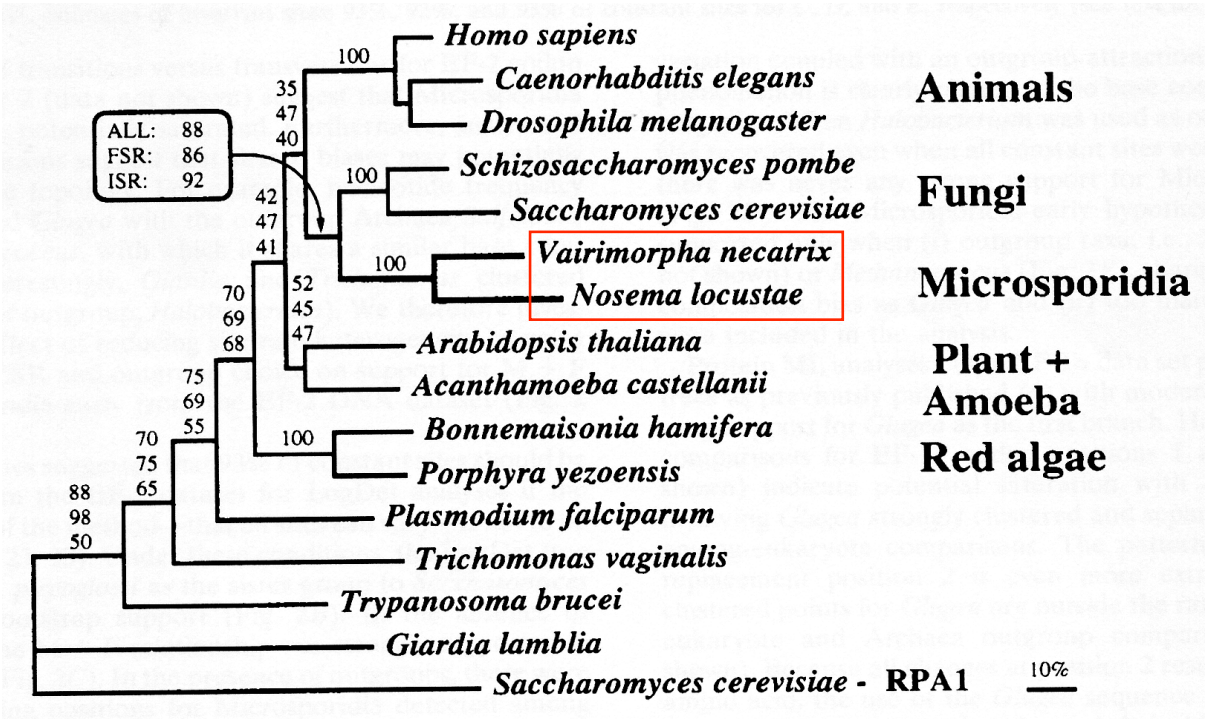
Phylogenetic analysis of eukaryotic small subunit ribosomal RNA

Sans prise en compte de la variation des vitesses entre sites

Cavalier-Smith & Chao (1996) J Mol Evol 43:551

Phylogenetic analysis of RNA polymerase II large subunit

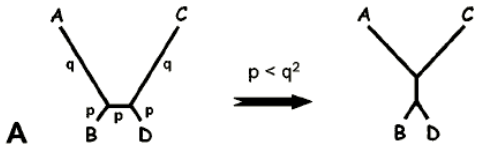
Hirt *et al*. (1999) Proc.Natl.Acad.Sci. USA 96:580

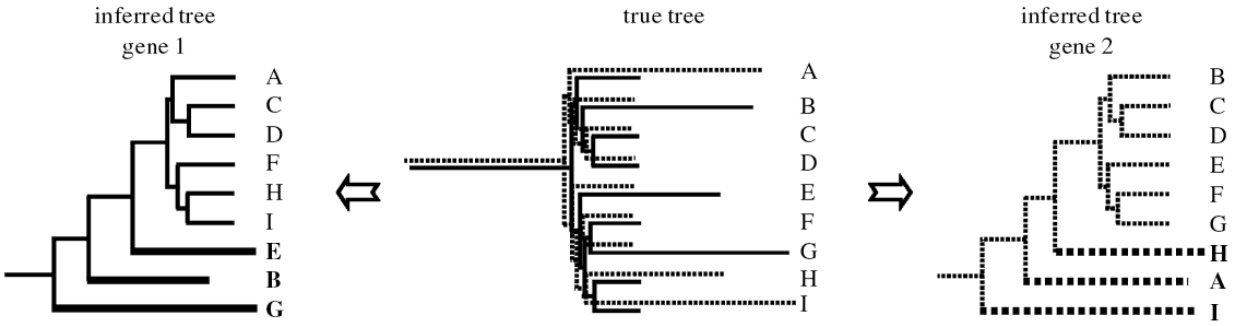


Peut-on réconcilier les ARNr et les ARN polymérases ?

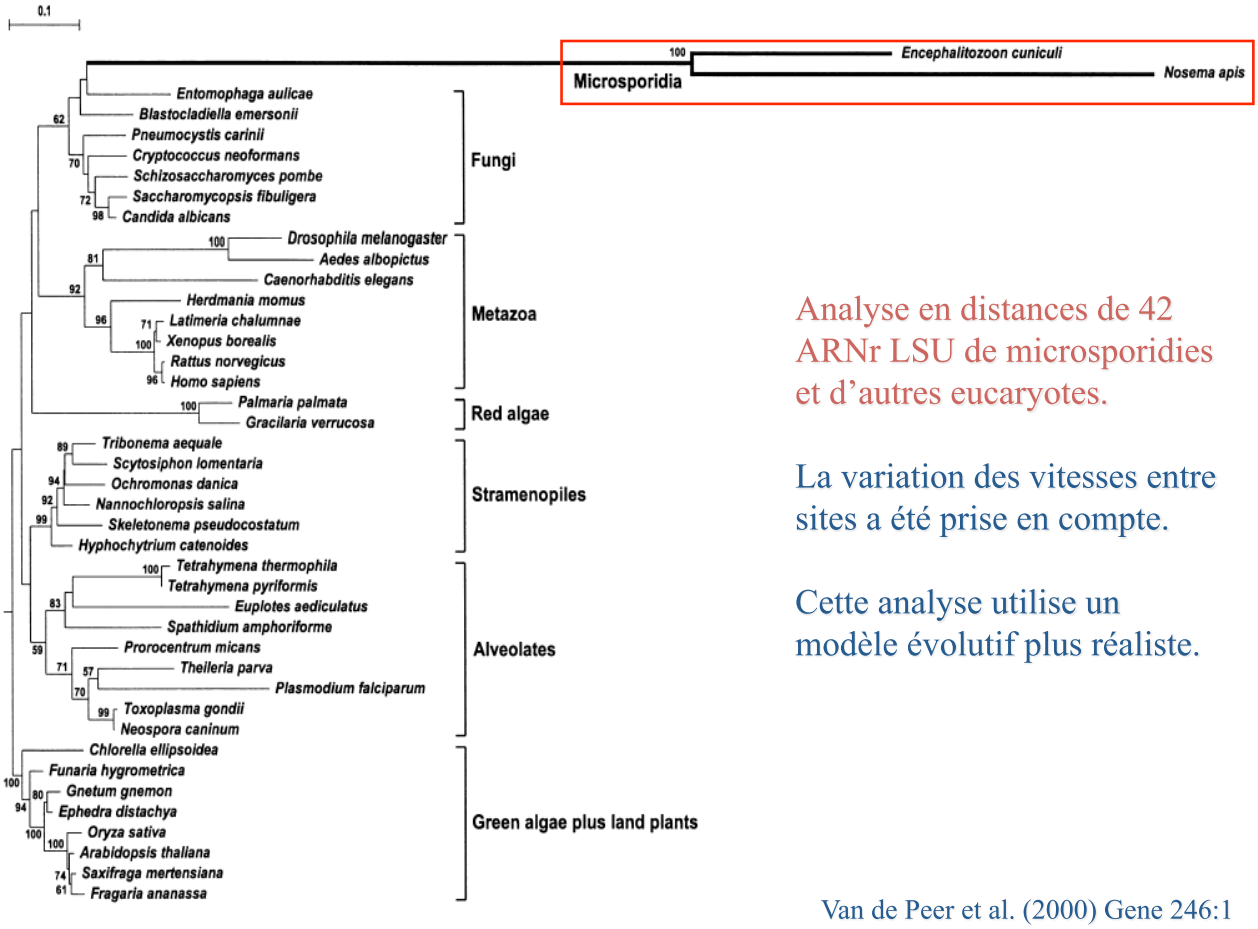
The Long Branch Attraction artifact

[Felsenstein (1978) *Syst Zool* 27:401]





Philippe et al. (2000) Proc. Royal Soc. Lond. B 267:1213.



Analyse en distances de 42 ARNr LSU de microsporidies et d'autres eucaryotes.

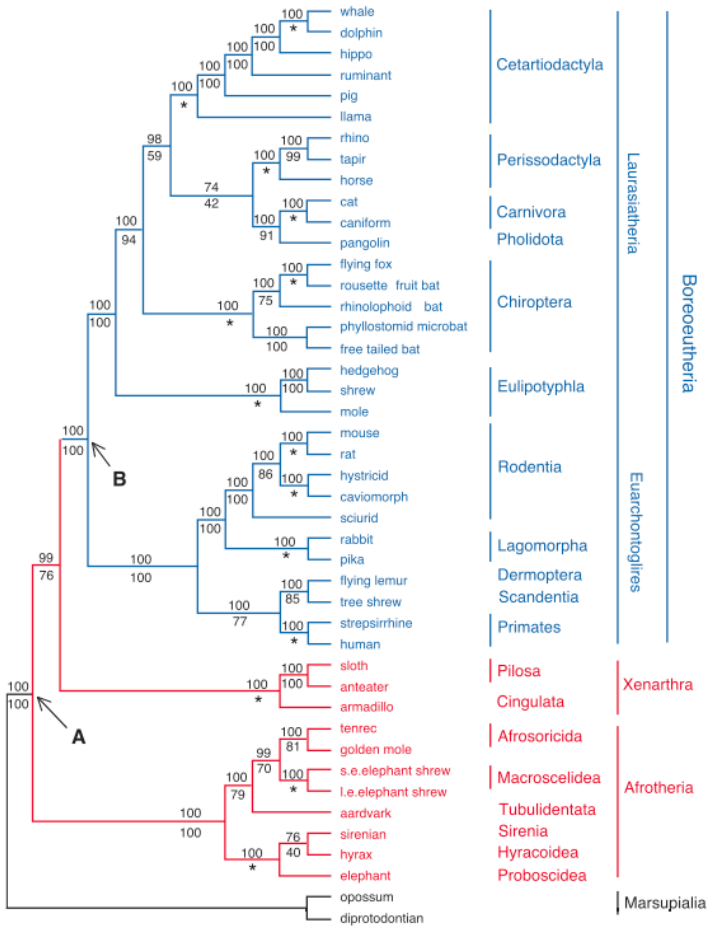
La variation des vitesses entre sites a été prise en compte.

Cette analyse utilise un modèle évolutif plus réaliste.

Van de Peer et al. (2000) Gene 246:1

L'effet du modèle évolutif utilisé:

les modèles plus réalistes sont meilleurs

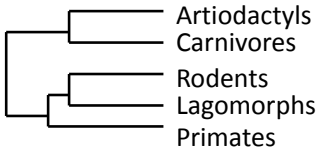


Phylogénie des ordres de mammifères

Resolution of the Early Placental Mammal Radiation Using Bayesian Phylogenetics

William J. Murphy,[1]* Eduardo Eizirik,[1,2]* Stephen J. O'Brien,[1]† Ole Madsen,[3] Mark Scally,[4,5] Christophe J. Douady,[4,5] Emma Teeling,[4,5] Oliver A. Ryder,[6] Michael J. Stanhope,[5,7] Wilfried W. de Jong,[3,8] Mark S. Springer[4]†

True tree :

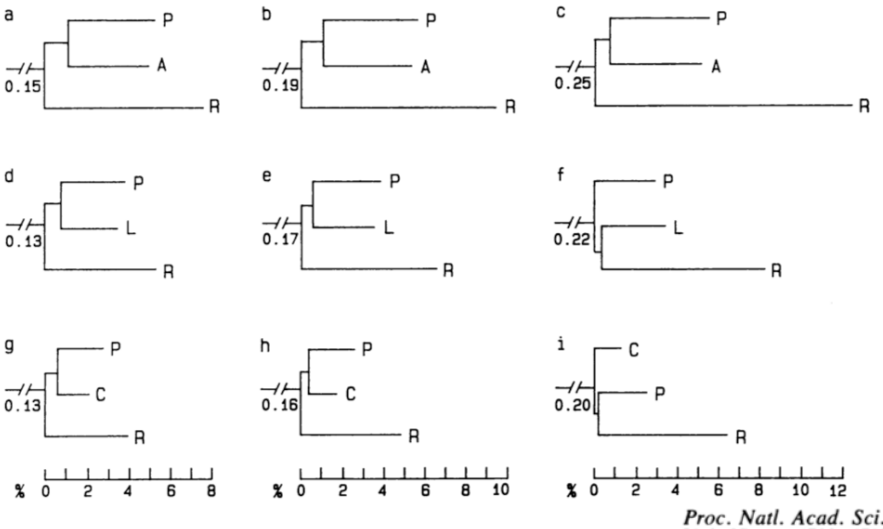


no across-site rate variation

with across-site rate variation

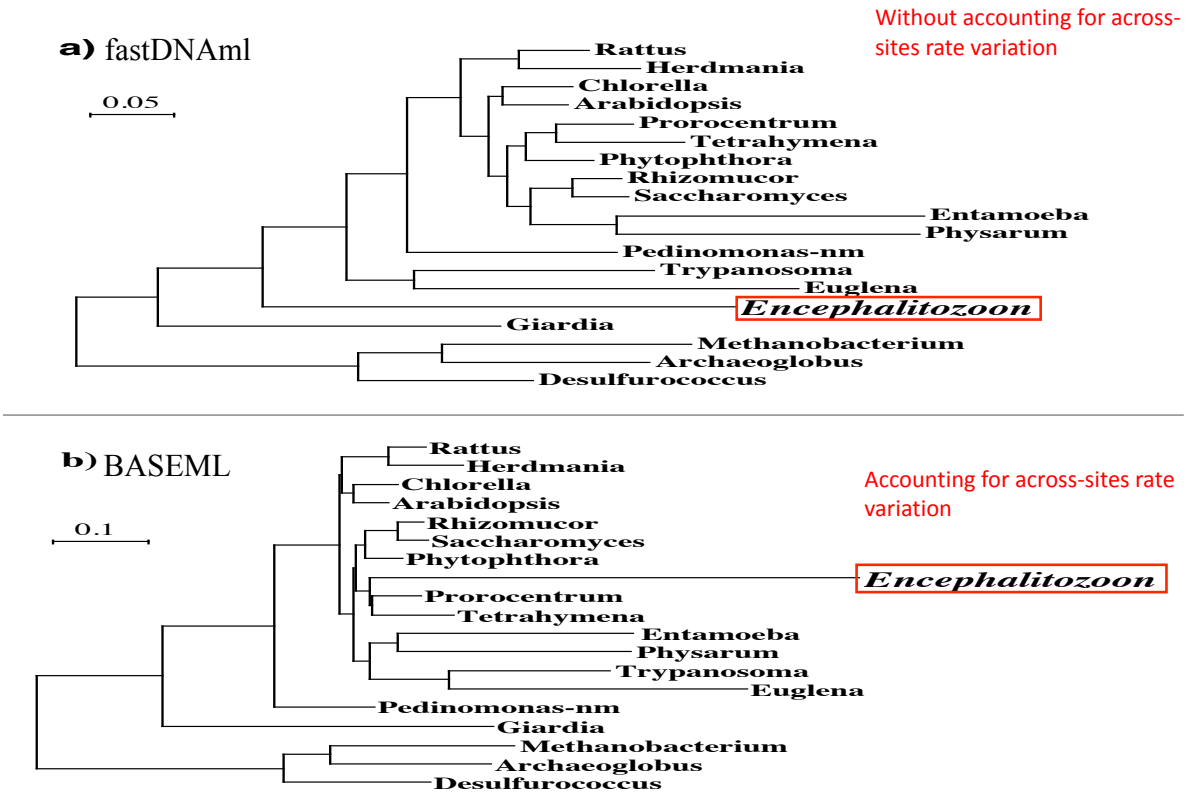
$\alpha = 1$

$\alpha = 0.5$



Proc. Natl. Acad. Sci. USA
Vol. 87, pp. 6703–6707, September 1990

Phylogenetic analysis of LSU rRNA



Without accounting for across-sites rate variation

Accounting for across-sites rate variation

Peyretaillade et coll. (1998) Nucleic Acids Res 26:3513

L'effet de l'échantillonage taxonomique

"Molecular phylogeny of the kingdoms Animalia, Plantae, and Fungi"
Gouy & Li (1989)

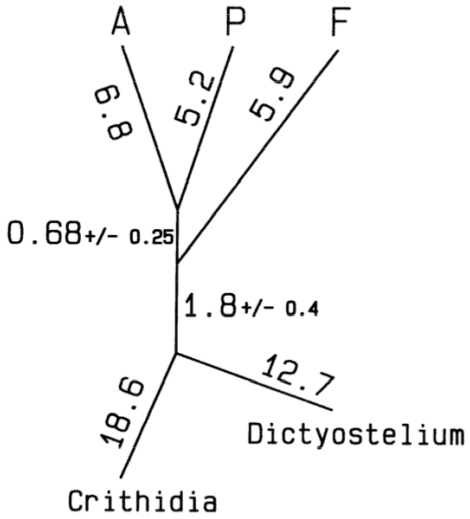


FIG. 2.—Unrooted phylogenetic tree inferred from rRNA sequences. A total of 2,971 sites were analyzed.

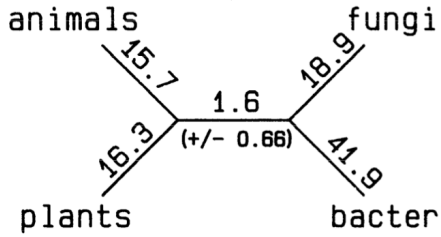
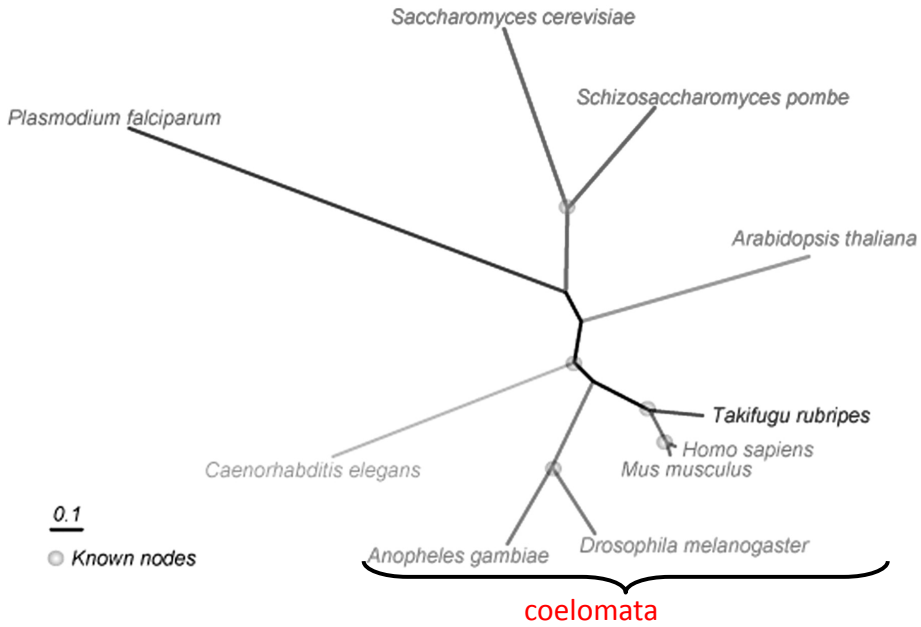


FIG. 4.—Unrooted phylogenetic tree inferred from the pooled protein data set. A total of 1,634 sites were analyzed. Branch lengths are in percent substitutions. The SE of the internal branch length estimate is shown. See table 3 for a description of the data set.

Les temps reculés: à la fois peu de gènes et peu d'espèces

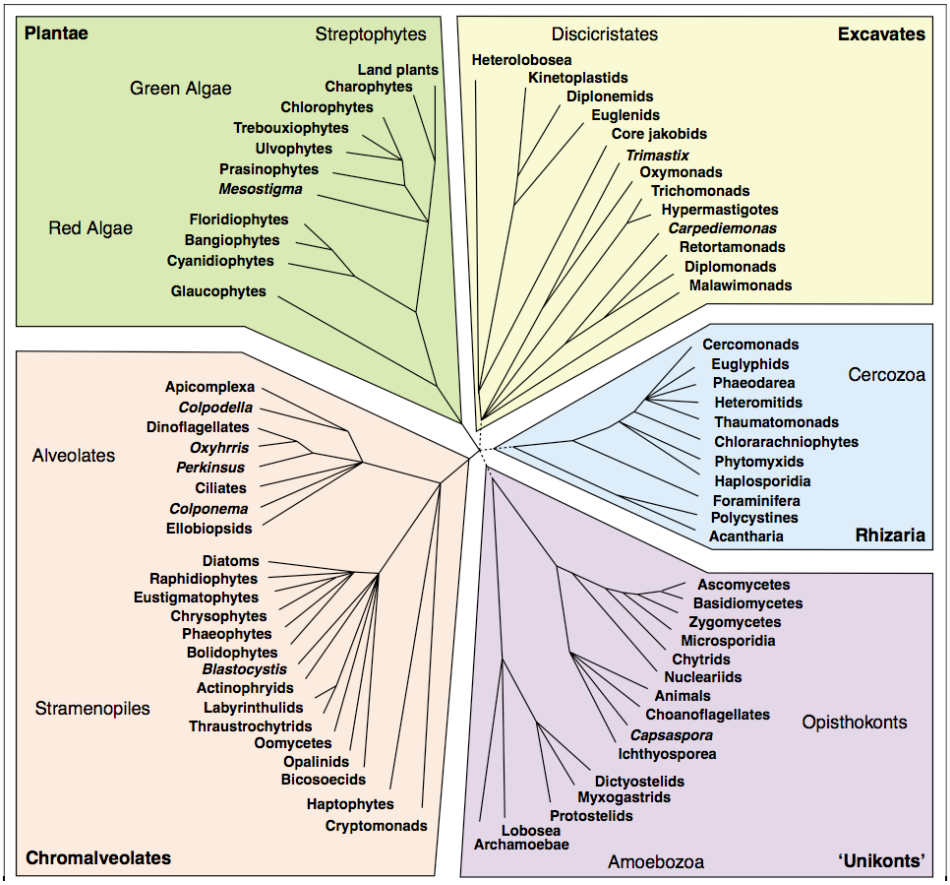
"The Opisthokonta and the Ecdysozoa May Not Be Clades: Stronger Support for the Grouping of Plant and Animal than for Animal and Fungi and Stronger Support for the Coelomata than Ecdysozoa"



Arbres phylogénétiques de 780 gènes de 10 génomes complètement séquencés amalgamés en un unique superarbre.

Philip, G. K. et al. Mol Biol Evol 2005 22:1175-1184

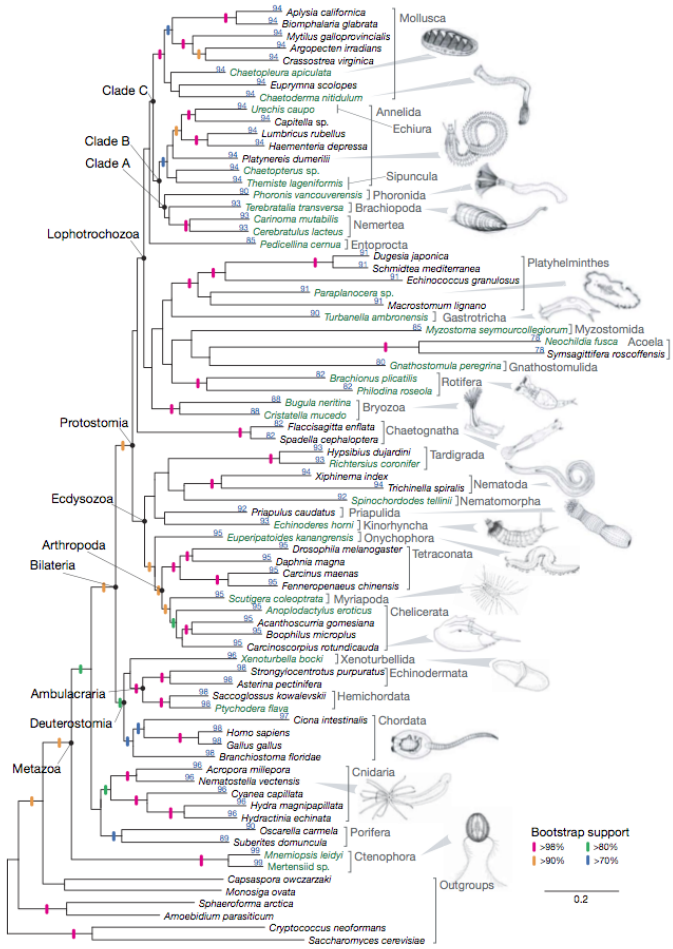
MBE MOLECULAR BIOLOGY AND EVOLUTION



Eukaryotic domain phylogeny.

Emerging consensus for the identification of five super- phyla.

Relationships between them remain very uncertain.



Phylogénie des métazoaires

- Rejet du concept acoelomate, pseudocoelomate, coelomate :

la division pertinente est lophotrochozoa / ecdysozoa

- Bilataria vs. cnidaria, porifera et ctenophora

- Protostomes vs. Deutérostomes

Vol 452|10 April 2008|doi:10.1038/nature06614 nature

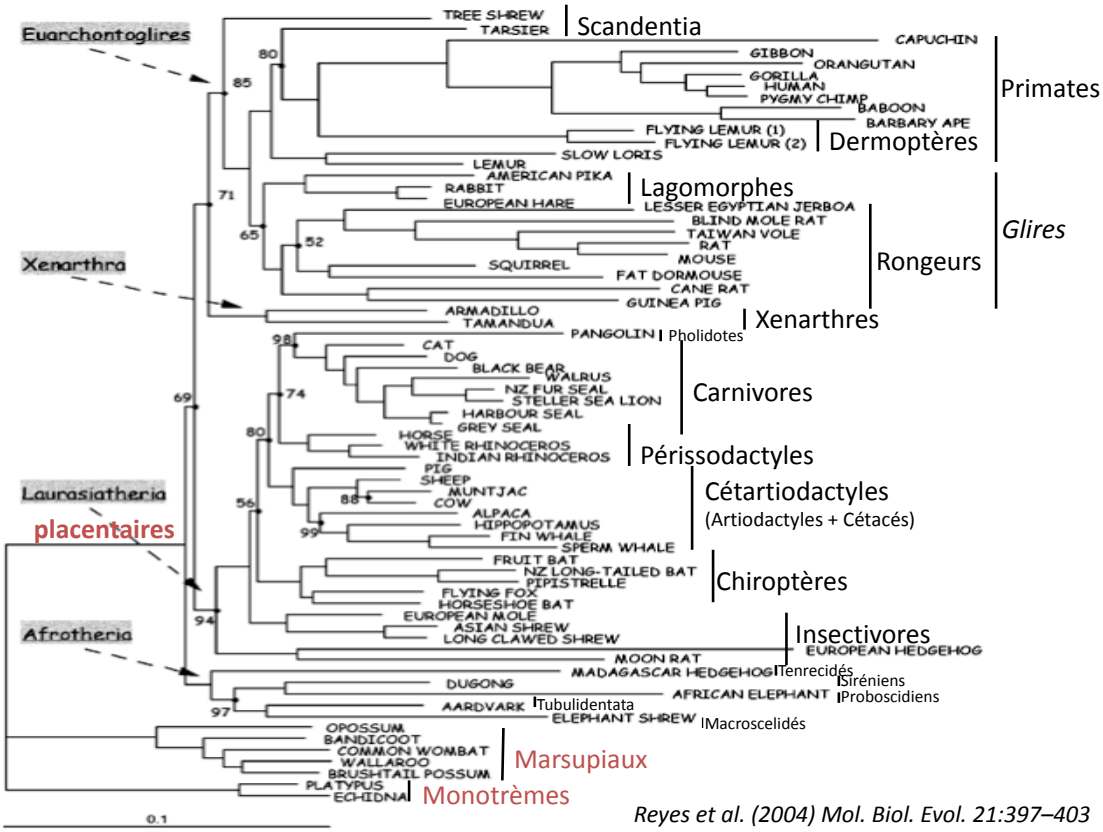
LETTERS

Broad phylogenomic sampling improves resolution of the animal tree of life

Nombre de topologies d'arbres binaires non racinés possibles pour n taxa

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

n	N_{arbres}
3	1
4	3
5	15
6	105
7	945
...	...
10	2,027,025
...	...
20	$\sim 2 \times 10^{20}$



Reyes et al. (2004) Mol. Biol. Evol. 21:397–403

Arbre raciné des 17 ordres de mammifères ⇔ arbre à 18 feuilles non raciné ≈ 2.10^{17} possibilités

Méthodes pour la reconstruction phylogénétique

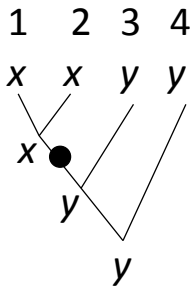
Quatre familles principales de méthodes :

- Parcimonie
- Méthodes de distances
- Maximum de vraisemblance
- Méthodes bayesiennes

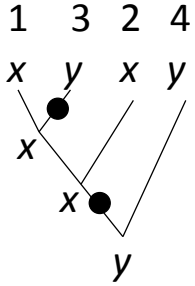
Pourquoi la parcimonie ?

Soit un caractère relevé dans 4 espèces {1, 2, 3, 4} et présentant les états {*x,x,y,y*}. Quelle histoire évolutive a pu conduire à cet état final ?

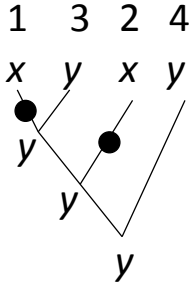
Egalité par ascendance commune: deux espèces possèdent le même état de caractère car elles l'ont hérité sans le transformer de leur dernier ancêtre commun



Présence d'homoplasie: des états identiques sont observés bien qu'ils n'aient pas été hérités, inchangés, du dernier ancêtre.



réversion



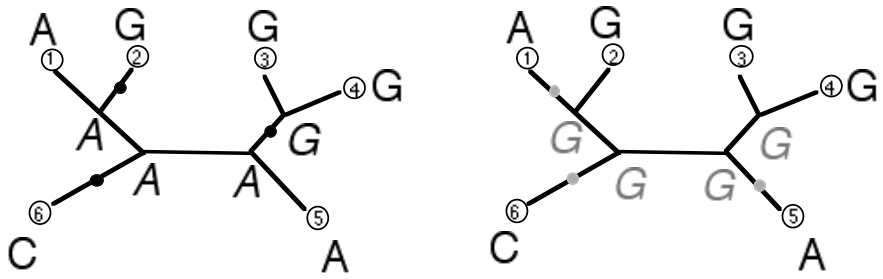
convergence

Les scénarios homoplasiques demandent plus de changements au cours de l'évolution. La parcimonie parie que convergences et réversions sont rares et recherche l'histoire qui demande le moins possible de changements.

Parcimonie (1)

- Etape 1: Pour une topologie d'arbre donnée, et pour un site donné de l'alignement, calculer, à l'aide de l'algorithme de Fitch, le plus petit nombre total de changements dans tout l'arbre.

  Soit *d* ce nombre total de changements.



X: ancestral nucleotide • : substitution event

Exemple: A ce site et pour cette forme d'arbre, au moins 3 substitutions sont nécessaires pour expliquer le pattern de nucléotides présent aux feuilles de l'arbre. Plusieurs scénarios distincts à 3 changements sont possibles.

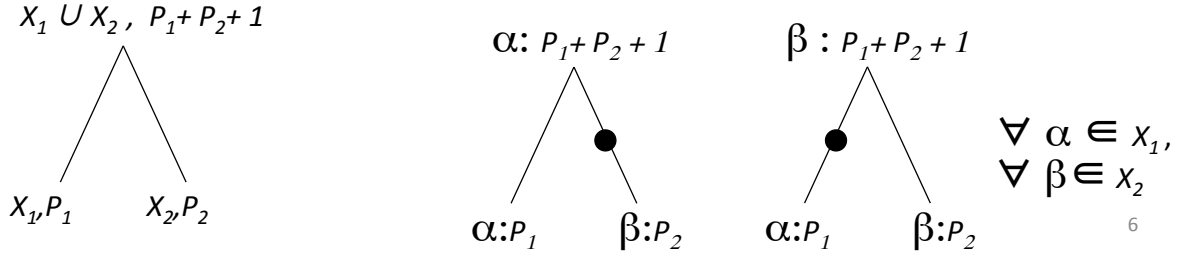
Algorithme de Fitch : calcul du nombre minimal de changements

Raciner arbitrairement l'arbre et calculer récursivement, à chaque nœud, deux objets:
- X: ensemble des résidus tous également possibles à ce nœud
- P: nombre minimal de changements dans le sous-arbre dont ce nœud est racine

---

1er cas: $X_1 \cap X_2$ n'est pas vide

2ème cas: $X_1 \cap X_2$ est vide

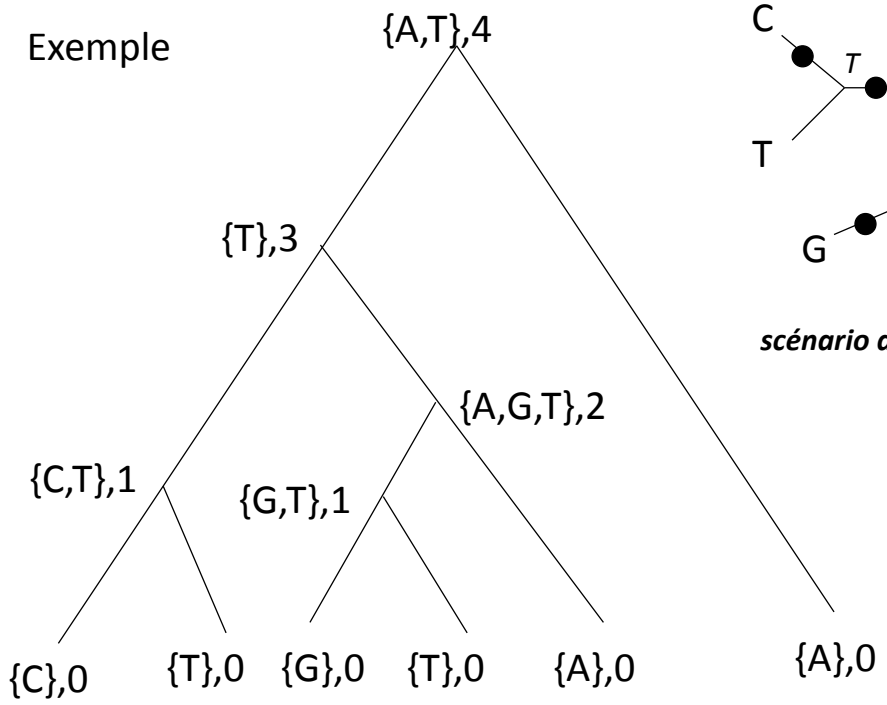


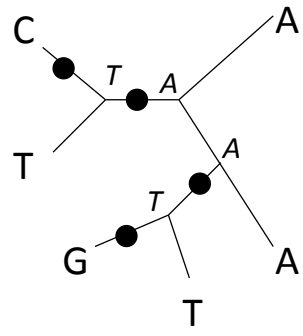
Algorithme de Fitch (suite)

Initialisation du calcul récursif aux feuilles de l'arbre

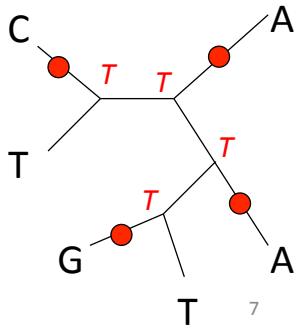
X={résidu présent à cette feuille}, P = 0

Exemple





scénario ancestral non unique !

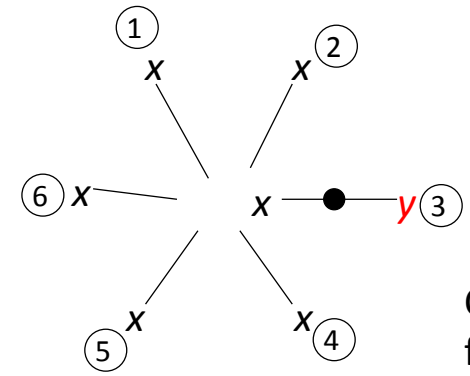


Parcimonie (2)

- Etape 2:
  - calculer *d* (étape 1) pour chaque site de l'alignement.
  - Sommer les valeurs *d* pour tous les sites.
  - Ceci donne la longueur *L* de l'arbre.

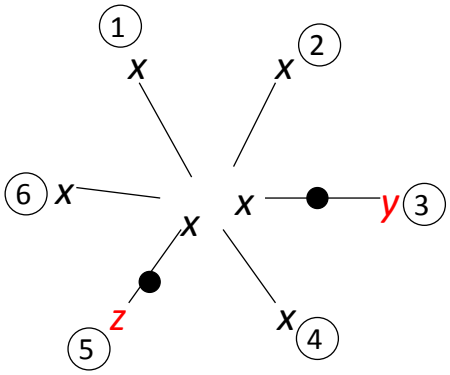
- Etape 3:
  - Calculer la valeur *L* (étape 2) pour toutes les formes d'arbre possibles.
  - Retenir l'arbre le plus court
    = le (ou les) arbre(s) qui nécessite(nt) le plus petit nombre de changements
    = le (ou les) arbre(s) le(s) plus parcimonieux.

Parcimonie : sites informatifs



Quelle que soit la topologie choisie, ce site contribue 1 pas

Ces sites ne contiennent pas d'information favorisant certaines topologies d'arbre:ils sont non-informatifs. Un site est **informatif** si et seulement si au moins 2 états présents chacun au moins 2 fois.



Quelle que soit la topologie choisie, ce site contribue 2 pas

Quelques propriétés de la Parcimonie

- Conduit à des arbres sans racine.
- Algorithme et principe généraux (ADN, protéines, morphologie)
- La position des changements sur chaque branche n'est pas unique => la parcimonie ne permet pas de définir la longueur des branches de façon unique.
- Plusieurs arbres peuvent être également parcimonieux (même longueur, la plus petite de toutes).
- Le nombre d'arbres croit très vite avec le nombre de séquences traitées:

  La recherche de l'arbre le plus court doit être limitée à une fraction de l'ensemble de tous les arbres possibles => On n'a plus de certitude de trouver l'arbre le plus court.

Exemple d'heuristique d'exploration de l'espace des topologies (PHYLIP)

• Définir un ordre, arbitraire, des séquences.

• Débuter avec les 3 premières séquences et l'unique topologie possible; ajouter la séquence suivante dans toutes les positions possibles sur l'arbre courant; retenir la meilleure position.

• Faire des réarrangements locaux: chaque branche interne définit 4 sous-arbres *a, b, c, d* et une topologie entre eux; évaluer les 2 autres topologies alternatives:

$$\begin{matrix} a & & c \\ & > - < & \\ b & & d \end{matrix} \qquad \begin{matrix} a & & b \\ & > - < & \\ c & & d \end{matrix} \qquad \begin{matrix} a & & c \\ & > - < & \\ d & & b \end{matrix}$$

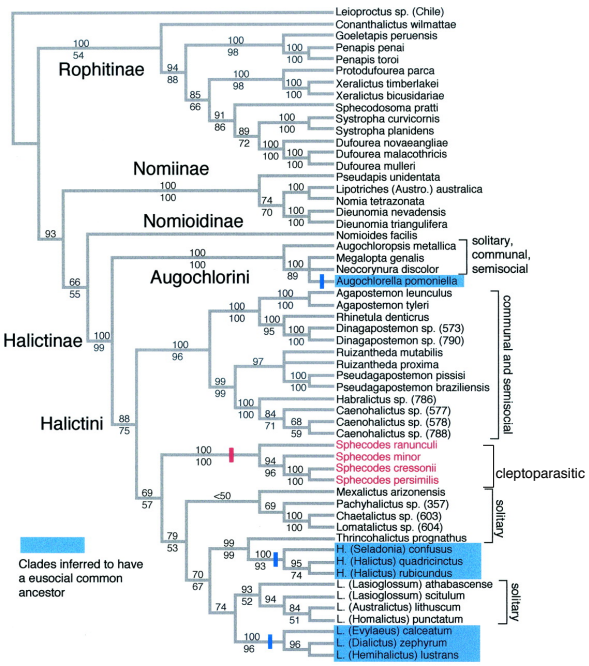
et retenir un arbre alternatif s'il est meilleur.
Cette opération s'appelle « nearest neighbor interchange (NNI) ».

• Recommencer tant qu'il reste des séquences à ajouter.

• Faire des réarrangements globaux: évaluer toutes les positions alternatives de chaque sous-arbre de l'arbre courant; s'arrêter quand aucune alternative n'améliore l'arbre courant. Ces opérations s'appellent « subtree pruning and regrafting (SPR) ».

Ceci transforme un calcul impossible (toutes les topologies) en un calcul assez rapide jusqu'à 20 ou 30 séquences. On répète souvent toute la recherche pour plusieurs ordres initiaux.

Evolution of sociality in a primitively eusocial lineage of bees



Phylogeny of the halictid subfamilies, tribes, and genera. Strict consensus of six trees based on equal weights parsimony analysis of the entire data set of three exons and two introns. Two regions within the introns were excluded because they could not be aligned unambiguously. Gaps coded as a fifth state or according to the methods described in ref. 23 yielded the same six trees. Bootstrap values above the nodes indicate bootstrap support based on the exons introns data set. Bootstrap values below the nodes indicate support based on an analysis of exons only. For the exons introns analysis the data set included 1,541 total aligned sites (619 parsimony-informative sites), the trees were 3,388 steps in length.

Advanced eusocial insects, such as ants, termites, and corbiculate bees, cannot provide insights into the earliest stages of eusocial evolution because eusociality in these taxa evolved long ago (in the Cretaceous) and close solitary relatives are no longer extant. In contrast, primitively eusocial insects, such as halictid bees, provide insights into the early stages of eusocial evolution because eusociality has arisen recently and repeatedly. I show that eusociality has arisen only three times within halictid bees.

Danforth, Bryan N. (2002) Proc. Natl. Acad. Sci. USA 99, 286-290

PNAS

Modèles d'évolution

Formation CNRS « Phylogénie moléculaire »

Guy Perrière

Laboratoire de Biométrie et Biologie Évolutive
UMR CNRS n° 5558
Université Claude Bernard – Lyon 1

20-23 mars 2018

Divergence observée

- Appelée p (ou p-distance), c'est l'estimation la plus simple de la distance entre deux séquences :

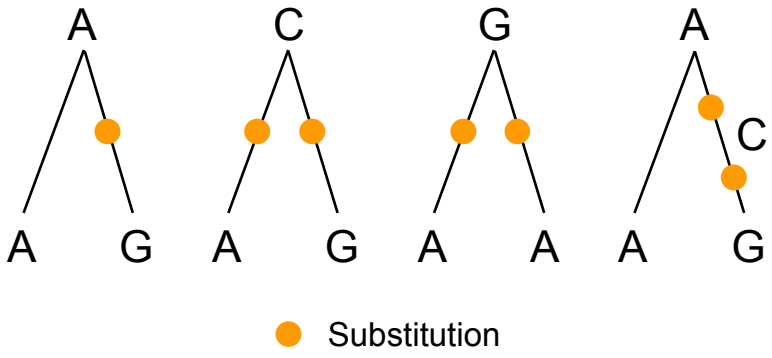
$$p = n/\ell$$

avec n le nombre total de substitutions et ℓ le nombre de sites homologues comparés.

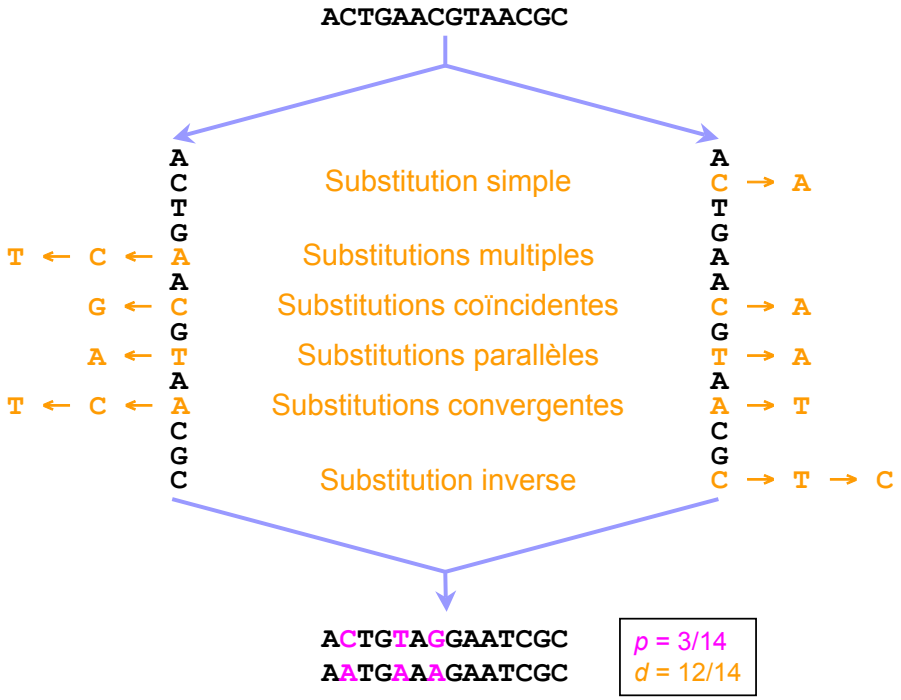
- Variation pour deux séquences de composition homogène :
  - Pour l'ADN : $0 \leq p \leq 0.75$.
  - Pour les protéines : $0 \leq p \leq 0.95$.

Substitutions multiples

- La distance évolutive réelle (d) est généralement supérieure à la divergence observée (p).
- En faisant des hypothèses sur la nature du processus évolutif, il est possible d'estimer d à partir de p.

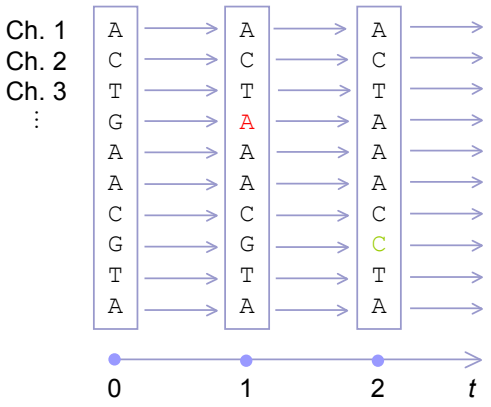


Types de substitutions



Modèles de Markov en phylogénie

- Utilisés pour les séquences nucléotidiques et protéiques.
- Les substitutions se font suivant un *processus de Markov*.
- Impliquent de déterminer des *probabilités de substitution* :
  - 16 valeurs en théorie pour les séquences d'ADN.
  - Moins en pratique :
    - Hypothèses simplificatrices.



Évolution des sites d'une séquence d'ADN selon un processus markovien

Propriétés des modèles

- Temps continu.
- Hypothèses communes à tous les modèles classiques :
  - Homogénéité par branche :
    - Un seul taux global de substitution (λ) tout au long de l'arbre.
  - Homogénéité par sites (ou uniformité) :
    - Tous les sites évoluent suivant le même processus.
  - Stationnarité :
    - La fréquences des nucléotides/acides aminés dans les séquences est la même de la racine aux feuilles de l'arbre.
  - Réversibilité :
    - La quantité de changement d'un nucléotide/acide aminé $i \rightarrow j$ est égale à la quantité de changement $j \rightarrow i$.
- Les modèles les plus récents ne font pas certaines de ces hypothèses.

Nombre de substitutions

- On pose $\Omega = \{\text{A, C, T, G}\}$ l'ensemble des états possibles.
- Soit $\mathbf{N} = (n_{ij})$ $(i, j \in \Omega)$, la matrice contenant le nombre de substitutions $(i \neq j)$ et de conservations $(i = j)$ observées entre deux séquences alignées :

$$\mathbf{N} = \begin{pmatrix} n_{\text{AA}} & n_{\text{AC}} & n_{\text{AT}} & n_{\text{AG}} \\ n_{\text{CA}} & n_{\text{CC}} & n_{\text{CT}} & n_{\text{CG}} \\ n_{\text{TA}} & n_{\text{TC}} & n_{\text{TT}} & n_{\text{TG}} \\ n_{\text{GA}} & n_{\text{GC}} & n_{\text{GT}} & n_{\text{GG}} \end{pmatrix}$$

- Le nombre total de substitutions observées n est tel que :

$$n = \sum_{i \neq j} n_{ij}$$

Fréquence des substitutions

- Soit $\mathbf{F} = (f_{ij})$ $(i, j \in \Omega)$, la matrice contenant les fréquences des substitutions $(i \neq j)$ et des conservations $(i = j)$ observées entre deux séquences alignées :

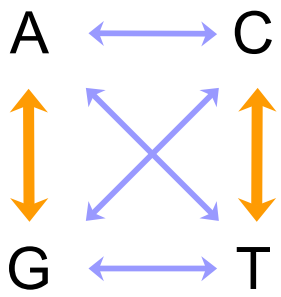
$$\mathbf{F} = \begin{pmatrix} f_{\text{AA}} & f_{\text{AC}} & f_{\text{AT}} & f_{\text{AG}} \\ f_{\text{CA}} & f_{\text{CC}} & f_{\text{CT}} & f_{\text{CG}} \\ f_{\text{TA}} & f_{\text{TC}} & f_{\text{TT}} & f_{\text{TG}} \\ f_{\text{GA}} & f_{\text{GC}} & f_{\text{GT}} & f_{\text{GG}} \end{pmatrix}$$

- Soit ℓ le nombre de sites homologues comparés, dans ce cas :

$$f_{ij} = \frac{n_{ij}}{\ell} \quad \text{et} \quad p = \sum_{i \neq j} f_{ij} = \frac{n}{\ell}$$

Transitions et transversions

- Beaucoup de modèles font la distinction entre les substitutions de type transitions et celles de type transversions :



- Soit r la fréquence des transitions et v celle des transversions, telles que :

$$r = r_{\mathrm{R}} + r_{\mathrm{Y}} = f_{\mathrm{AG}} + f_{\mathrm{GA}} + f_{\mathrm{CT}} + f_{\mathrm{TC}}$$
$$v = f_{\mathrm{AC}} + f_{\mathrm{CA}} + f_{\mathrm{AT}} + f_{\mathrm{TA}} + f_{\mathrm{CG}} + f_{\mathrm{GC}} + f_{\mathrm{GT}} + f_{\mathrm{TG}}$$

Taux de substitution

- Soit q_{ij} $(i \neq j)$ le *taux de substitution instantané* d'un nucléotide i vers un nucléotide j $(i, j \in \Omega)$.
- Dans ce cas le *taux de changement instantané* d'un nucléotide i est défini comme $\lambda_i = \sum_{j \neq i} q_{ij}$.
- L'ensemble des taux de substitutions et des taux de changements peuvent être regroupés dans une matrice $\mathbf{Q} = (q_{ij})$ telle que :

$$\mathbf{Q} = \begin{pmatrix} -\lambda_{\mathrm{A}} & q_{\mathrm{AC}} & q_{\mathrm{AT}} & q_{\mathrm{AG}} \\ q_{\mathrm{CA}} & -\lambda_{\mathrm{C}} & q_{\mathrm{CT}} & q_{\mathrm{CG}} \\ q_{\mathrm{TA}} & q_{\mathrm{TC}} & -\lambda_{\mathrm{T}} & q_{\mathrm{TG}} \\ q_{\mathrm{GA}} & q_{\mathrm{GC}} & q_{\mathrm{GT}} & -\lambda_{\mathrm{G}} \end{pmatrix}$$

Les sommes en ligne de $\mathbf{Q}$ sont égales à 0.

Probabilités de substitution

- La relation entre probabilités de substitutions au cours du temps et taux instantanés est donnée par :

$$\mathbf{P}(t) = \mathrm{e}^{\mathbf{Q}t}$$

avec :

$$\mathbf{P}(t) = \begin{pmatrix} p_{\mathrm{AA}}(t) & p_{\mathrm{AC}}(t) & p_{\mathrm{AT}}(t) & p_{\mathrm{AG}}(t) \\ p_{\mathrm{CA}}(t) & p_{\mathrm{CC}}(t) & p_{\mathrm{CT}}(t) & p_{\mathrm{CG}}(t) \\ p_{\mathrm{TA}}(t) & p_{\mathrm{TC}}(t) & p_{\mathrm{TT}}(t) & p_{\mathrm{TG}}(t) \\ p_{\mathrm{GA}}(t) & p_{\mathrm{GC}}(t) & p_{\mathrm{GT}}(t) & p_{\mathrm{GG}}(t) \end{pmatrix}$$

Les sommes en ligne de $\mathbf{P}(t)$ sont égales à 1.

Stationnarité

- Au bout d'un temps infini, un processus de Markov atteint ce qu'on appelle un *état stationnaire* :
  - Les fréquences des différents états ne changent plus :

$$\lim_{t \to \infty} p_{ij}(t) = \pi_j \ (i, j \in \Omega)$$

avec π_j la *fréquence à l'équilibre* pour le nucléotide j.

- Les modèles classiques considèrent que la stationnarité est atteinte dès la racine de l'arbre :
  - Utilisation des fréquences des bases dans le jeu de données pour estimer les valeurs de π_j.

Réversibilité

- Un processus de Markov est dit *réversible* si, lorsque la stationnarité est atteinte, on a :

$$\pi_i p_{ij}(t) = \pi_j p_{ji}(t), \ \forall i, j \in \Omega$$

  À l'équilibre, la quantité de changement $i \to j$ est égale à la quantité de changement $j \to i$.
- Pas de directionalité dans l'écoulement du temps :
  - Arbres non racinés.

Échangeabilités

- En remplaçant les probabilités de substitution par les taux instantanés, l'équation précédente devient :

$$\pi_i q_{ij} = \pi_j q_{ji}$$

  Soit :

$$\frac{q_{ij}}{\pi_j} = \frac{q_{ji}}{\pi_i} = s_{ij} = s_{ji}$$

  avec $s_{ij} = s_{ji}$ un terme symétrique, appelé paramètre *d'échangeabilité* entre i et j.

Matrices S et Π

- Sous l'hypothèse de réversibilité, l'expression de **Q** peut s'écrire comme étant le produit :

$$\mathbf{Q} = \mathbf{S\Pi} = \begin{pmatrix} \cdot & s_{AC} & s_{AT} & s_{AG} \\ s_{CA} & \cdot & s_{CT} & s_{CG} \\ s_{TA} & s_{TC} & \cdot & s_{TG} \\ s_{GA} & s_{GC} & s_{GT} & \cdot \end{pmatrix} \times \begin{pmatrix} \pi_A & 0 & 0 & 0 \\ 0 & \pi_C & 0 & 0 \\ 0 & 0 & \pi_T & 0 \\ 0 & 0 & 0 & \pi_G \end{pmatrix}$$

avec **S** la matrice des échangeabilités entre nucléotides et $\mathbf{\Pi} = \text{diag}(\pi_i)$ la matrice diagonale contenant les valeurs des fréquences des bases à l'équilibre.

Simplification de l'écriture

- Les échangeabilités étant symétriques, on pose :

$$\begin{aligned} s_{AC} = s_{CA} = \alpha \quad & s_{AT} = s_{TA} = \beta \\ s_{AG} = s_{GA} = \gamma \quad & s_{CT} = s_{TC} = \delta \\ s_{CG} = s_{GC} = \epsilon \quad & s_{TG} = s_{GT} = \eta \end{aligned}$$

- Et le produit matriciel précédent peut s'écrire :

$$\mathbf{Q} = \mathbf{S\Pi} = \begin{pmatrix} \cdot & \alpha & \beta & \gamma \\ \alpha & \cdot & \delta & \epsilon \\ \beta & \delta & \cdot & \eta \\ \gamma & \epsilon & \eta & \cdot \end{pmatrix} \times \begin{pmatrix} \pi_A & 0 & 0 & 0 \\ 0 & \pi_C & 0 & 0 \\ 0 & 0 & \pi_T & 0 \\ 0 & 0 & 0 & \pi_G \end{pmatrix}$$

Expression de Q

- Au moyen du produit matriciel précédent, on en déduit l'expression de $\mathbf{Q}$:

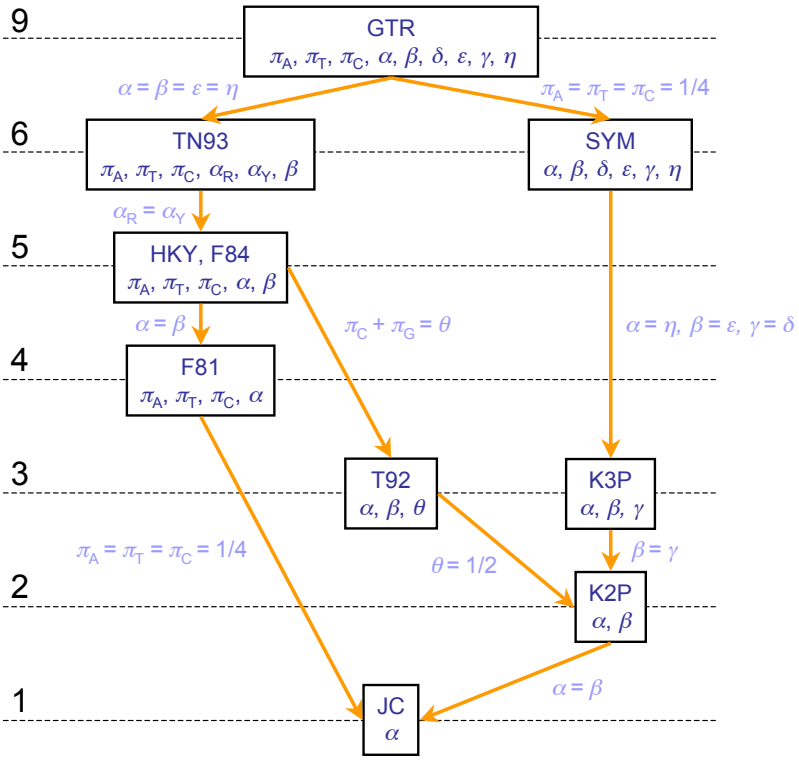
$$\mathbf{Q} = \begin{pmatrix} -\lambda_A & \pi_C\alpha & \pi_T\beta & \pi_G\gamma \\ \pi_A\alpha & -\lambda_C & \pi_T\delta & \pi_G\epsilon \\ \pi_A\beta & \pi_C\delta & -\lambda_T & \pi_G\eta \\ \pi_A\gamma & \pi_C\epsilon & \pi_T\eta & -\lambda_G \end{pmatrix}$$

$$\text{avec} \begin{cases} \lambda_A = \pi_C\alpha + \pi_T\beta + \pi_G\gamma \\ \lambda_C = \pi_A\alpha + \pi_T\delta + \pi_G\epsilon \\ \lambda_T = \pi_A\beta + \pi_C\delta + \pi_G\eta \\ \lambda_G = \pi_A\gamma + \pi_C\epsilon + \pi_T\eta \end{cases}$$

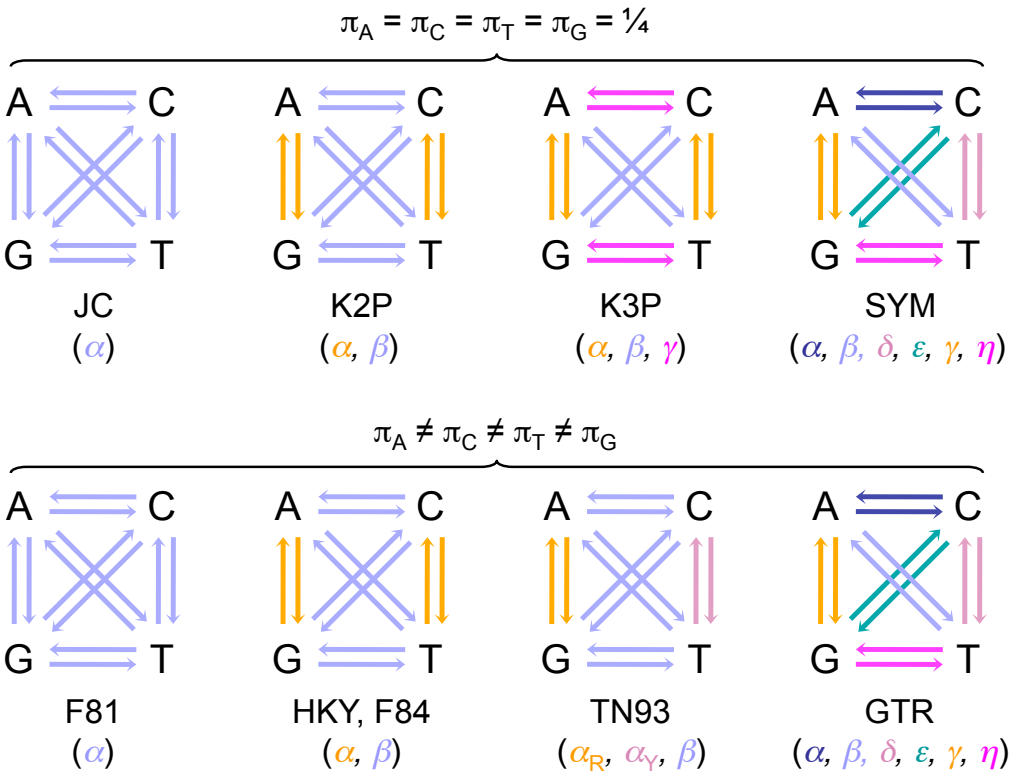
Soit neuf paramètres à estimer :

- Modèle GTR (*Generalised Time Reversible*) ou REV.

Imbrication des modèles



Paramètres des modèles



Calcul de la distance évolutive

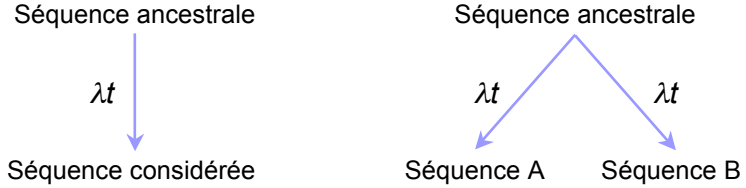
- Soit λ, le *taux global de substitions* dans une séquence. Sous l'hypothèse de réversibilité, ce taux est égal à :

$$\lambda = \sum_i \pi_i \lambda_i, \ \ i \in \Omega$$

avec λ_i le taux de changement instantané d'un nucléotide en n'importe lequel des trois autres.

- Dans ce cas, la distance évolutive entre deux séquences est donnée par la formule :

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



Normalisation

- Par convention, les valeurs des taux instantanés sont normalisées de façon à ce que :

$$\lambda = \sum_i \pi_i \lambda_i = 1$$

- Sous cette contrainte, la distance évolutive entre deux séquences est assimilable au temps écoulé :

$$d = 2\lambda t = 2t$$

Modèle de Jukes et Cantor

- Une seule échangeabilité (α), identique pour chacun des quatre nucléotides.
- Fréquences à l'équilibre : $\pi_A = \pi_C = \pi_T = \pi_G = 1/4$.
- Matrices $\mathbf{Q}$ et $\mathbf{P}(t)$:

$$\mathbf{Q} = \begin{pmatrix} -\lambda & \alpha & \alpha & \alpha \\ \alpha & -\lambda & \alpha & \alpha \\ \alpha & \alpha & -\lambda & \alpha \\ \alpha & \alpha & \alpha & -\lambda \end{pmatrix} \quad \mathbf{P}(t) = \begin{pmatrix} p_0(t) & p_1(t) & p_1(t) & p_1(t) \\ p_1(t) & p_0(t) & p_1(t) & p_1(t) \\ p_1(t) & p_1(t) & p_0(t) & p_1(t) \\ p_1(t) & p_1(t) & p_1(t) & p_0(t) \end{pmatrix}$$

  avec $p_0(t) + 3p_1(t) = 1$.
- Taux global de substitutions : $\lambda = \sum_i \pi_i \lambda_i = 3\alpha$.

Résolution

- Le calcul de $\mathbf{P}(t) = \mathrm{e}^{\mathbf{Q}t}$ permet de déterminer que :

$$p_0(t) = \frac{1}{4} + \frac{3}{4}\mathrm{e}^{-4\alpha t} \quad \text{et} \quad p_1(t) = \frac{1}{4} - \frac{1}{4}\mathrm{e}^{-4\alpha t}$$

- Élimination de t et introduction de d en utilisant la relation $d = 2\lambda t = 6\alpha t$.
- Introduction de la divergence observée entre deux séquences $p = 3p_1(2t)$.
- Formule de Jukes et Cantor pour le calcul de la distance évolutive :

$$d = -\frac{3}{4}\ln\left(1 - \frac{4}{3}p\right)$$

Autres distances I

- Modèle de Kimura à deux paramètres(1980) – K2P :

$$d = -\frac{1}{2}\ln(1 - 2r - v) - \frac{1}{4}\ln(1 - 2v)$$

avec r la fréquence des transitions et v la fréquence des transversions observées entre les deux séquences ($p = r + v$).

- Modèle de Felsenstein (1981) – F81 :

$$d = -a\ln\left(1 - \frac{p}{a}\right)$$

avec $a = 1 - \pi_{\mathrm{A}}^2 - \pi_{\mathrm{C}}^2 - \pi_{\mathrm{T}}^2 - \pi_{\mathrm{G}}^2$.

Autres distances II

- Modèle de Felsenstein (1984) – F84 :

$$d = -2a_1 \ln\left[1 - \frac{r}{2a_1} - \frac{(a_1 - a_2)v}{2a_1 a_3}\right]$$

$$\text{avec} \begin{cases} a_1 = \dfrac{\pi_A \pi_G}{\pi_A + \pi_G} + \dfrac{\pi_C \pi_T}{\pi_C + \pi_T} \\ a_2 = \pi_A \pi_G + \pi_C \pi_T \\ a_3 = (\pi_A + \pi_G)(\pi_C + \pi_T) \end{cases}$$

Autres distances III

- Modèle de Tamura et Nei (1993) – TN93 :

$$d = \frac{2\pi_T \pi_C}{\pi_Y}(a_1 - \pi_R b) + \frac{2\pi_A \pi_G}{\pi_R}(a_2 - \pi_Y b) + 2\pi_Y \pi_R b$$

$$\text{avec} \begin{cases} a_1 = -\ln\left(1 - \dfrac{\pi_Y}{2\pi_T \pi_C} r_Y - \dfrac{1}{2\pi_Y} v\right) \\ a_2 = -\ln\left(1 - \dfrac{\pi_R}{2\pi_A \pi_G} r_R - \dfrac{1}{2\pi_R} v\right) \\ b = -\ln\left(1 - \dfrac{1}{2\pi_R \pi_Y} v\right) \end{cases}$$

Modèle GTR

- Pas de solution analytique au calcul de $\mathbf{P}(t) = \mathrm{e}^{\mathbf{Q}t}$.
- La fréquence des substitutions $i \to j$ observées entre deux séquences au temps t est donnée par :

$$f_{ij}(t) = \pi_i p_{ij}(2t)$$

- Estimation des valeurs de π_i à partir des fréquences des bases dans les deux séquences considérées.
- Estimation des valeurs de $f_{ij}(t)$ en utilisant celles de celles de $\mathbf{F} = (f_{ij})$ (Diapo. 8).
- Calcul des valeurs de $p_{ij}(2t)$ à partir de l'équation précédente.

Utilité des modèles complexes

- Modélisent mieux l'évolution des séquences :
  - Plus proches de la réalité biologique.
- Séquences trop courtes :
  - Erreurs d'échantillonnage (valeurs de $d < 0$).
  - Variance importante.
- Séquences trop divergentes :
  - Méthodes à plus de quatre paramètres fréquemment inapplicables.
- Séquences peu divergentes :
  - Toutes les méthodes donnent des résultats comparables.

Le code génétique

I \ II	T	C	A	G	III
T	TTT Phe F	TCT Ser S	TAT Tyr Y	TGT Cys C	T
	TTC Phe F	TCC Ser S	TAC Tyr Y	TGC Cys C	C
	TTA Leu L	TCA Ser S	TAA Stop	TGA Stop	A
	TTG Leu L	TCG Ser S	TAG Stop	TGG Trp W	G
C	CTT Leu L	CCT Pro P	CAT His H	CGT Arg R	T
	CTC Leu L	CCC Pro P	CAC His H	CGC Arg R	C
	CTA Leu L	CCA Pro P	CAA Gln Q	CGA Arg R	A
	CTG Leu L	CCG Pro P	CAG Gln Q	CGG Arg R	G
A	ATT Ile I	ACT Thr T	AAT Asn N	AGT Ser S	T
	ATC Ile I	ACC Thr T	AAC Asn N	AGC Ser S	C
	ATA Ile I	ACA Thr T	AAA Lys K	AGA Arg R	A
	ATG Met M	ACG Thr T	AAG Lys K	AGG Arg R	G
G	GTT Val V	GCT Ala A	GAT Asp D	GGT Gly G	T
	GTC Val V	GCC Ala A	GAC Asp D	GGC Gly G	C
	GTA Val V	GCA Ala A	GAA Glu E	GGA Gly G	A
	GTG Val V	GCG Ala A	GAG Glu E	GGG Gly G	G

Substitutions synonymes et non synonymes

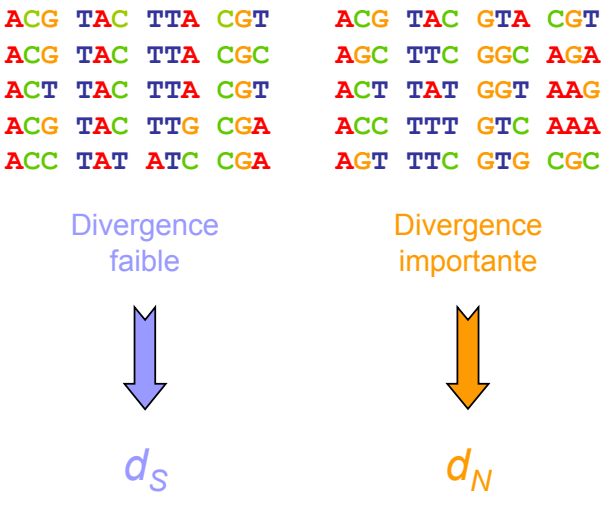
- Exemple d'une transversion C → A dans le codon CTC :
  - En position I :
    $$\mathrm{CTC}^{\mathrm{Leu}} \rightarrow \mathrm{ATC}^{\mathrm{Ile}}$$
    soit une substitution *non synonyme* (ou *non silencieuse*).
  - En position III :
    $$\mathrm{CTC}^{\mathrm{Leu}} \rightarrow \mathrm{CTA}^{\mathrm{Leu}}$$
    soit une substitution *synonyme* (ou *silencieuse*).
- Toutes les substitutions touchant la position II des codons sont non synonymes.

Distances d_N et d_S

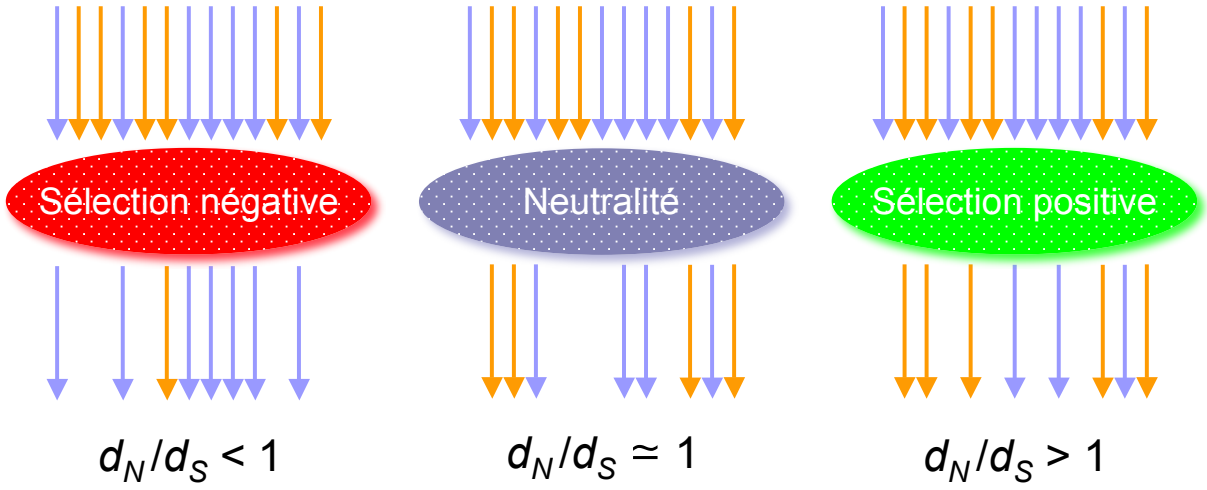
- Dans les gènes protéiques, il existe deux classes de sites ayant des vitesses évolutives différentes :
  - Substitutions non synonymes lentes.
  - Substitutions synonymes rapides.
  - L'hypothèse faite par les modèles d'évolution « classiques » que chaque site évolue en suivant le même processus est fausse.
- Calcul de deux distances évolutives différentes :
  - Distance non synonyme (d_N) :
    - Calcul à partir de p_N = nb. de substitutions non synonymes / nb. de sites non synonymes.
  - Distance synonyme (d_S) :
    - Calcul à partir de p_S = nb. de substitutions synonymes / nb. de sites synonymes.

Utilisation

- On se trouve fréquemment dans l'une ou l'autre de ces deux situations :
  - Séquences évolutivement peu distantes :
    - d_S est informatif, d_N ne l'est pas.
  - Séquences évolutivement très distantes :
    - d_S est saturé, d_N est informatif.



Sélection et neutralité



Substitutions synonymes
Substitutions non synonymes

Méthodes de comptage

- Gojobori et Nei (1986) :
  - Simplification des premières méthodes proposées par Miyata et Yasunaga (1980) et Perler *et al.* (1980).
  - Utilisation du modèle de Jukes et Cantor.
- Ina (1995) :
  - Amélioration de la méthode de Gojobori et Nei en distinguant les transitions des transversions.
  - Utilisation du modèle de Kimura à deux paramètres.
- Li, Wu et Luo (1985) :
  - Utilise la dégénérescence du code génétique et le modèle de Kimura à deux paramètres.

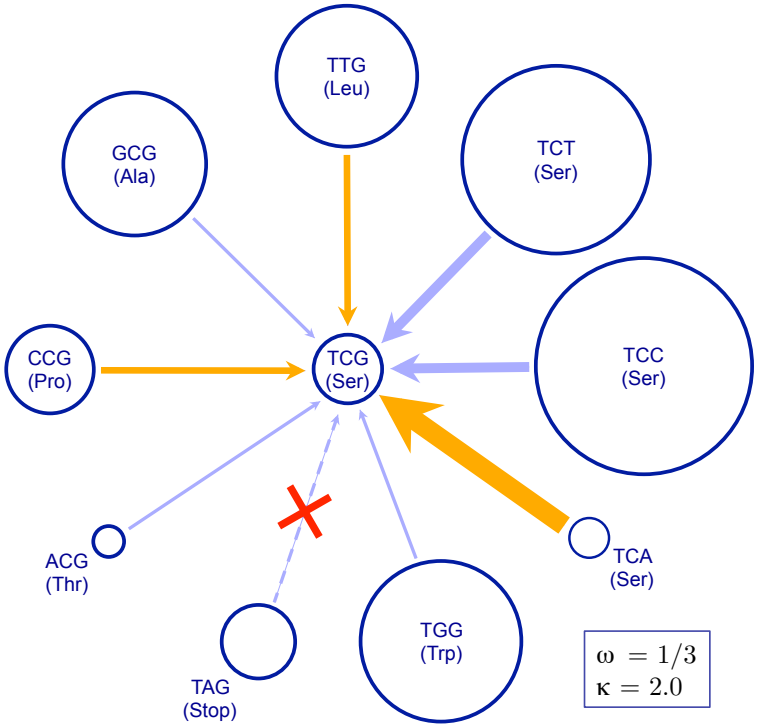
Approches au maximum de vraisemblance

- Publication la même année de deux méthodes similaires :
  - Goldman et Yang (1994) :
    - Simplification ultérieure par Yang et Nielsen (2000).
  - Muse et Gaut (1994).
- Dans les deux cas, définition d'un modèle de substitution spécifique aux codons.

Modèle de Goldman et Yang

- Indépendance des sites à l'intérieur d'un codon.
- À chaque intervalle de temps dt, une seule des trois positions est susceptible de muter.
- À chaque instant t, tout codon est susceptible de se substituer vers l'un de ses voisins :
  - Un voisin est un codon différant du codon d'intérêt par une seule substitution :
    - Position I, II ou III.
    - Codons Stop non considérés.
  - Chaque codon possède au plus neuf voisins.

Exemple du codon TCG$^{\text{Ser}}$



Paramètres du modèle

- Matrice $\mathbf{Q} = (q_{ij})$ des taux instantanés ($i, j \in \{\text{AAA, AAC, AAG, } \ldots, \text{ TTT}\}$, codons Stop exclus) :

$$\mathbf{Q} = \begin{pmatrix} -\lambda_{\text{AAA}} & q_{\text{AAA,AAC}} & \cdots & q_{\text{AAA,TTT}} \\ q_{\text{AAC,AAA}} & -\lambda_{\text{AAC}} & \cdots & q_{\text{AAC,TTT}} \\ \vdots & \vdots & \ddots & \vdots \\ q_{\text{TTT,AAA}} & q_{\text{TTT,AAC}} & \cdots & -\lambda_{\text{TTT}} \end{pmatrix}$$

- Ratio des taux de transitions/transversions κ.
- Ratio des distances non synonymes/synonymes $\omega = d_{\text{N}}/d_{\text{S}}$.

Taux de substitutions instantanés

- Les valeurs de q_{ij} sont telles que :

$$q_{ij} = \begin{cases} 0, & \text{si } i \text{ et } j \text{ diffèrent en plus d'une position} \\ \pi_j, & \text{pour une transversion synonyme} \\ \kappa\pi_j, & \text{pour une transition synonyme} \\ \omega\pi_j, & \text{pour une transversion non synonyme} \\ \omega\kappa\pi_j, & \text{pour une transition non synonyme} \end{cases}$$

- Normalisation de telle façon que le taux moyen de substitutions soit égal à un :

$$\sum_i \pi_i \lambda_i = 1$$

sachant que $\lambda_i = \sum_{j \neq i} q_{ij}$.

Probabilités de transition

- Comme pour les modèles standards, les valeurs des probabilités de transition $p_{ij}(t)$ sont données en résolvant $\mathbf{P}(t) = \mathrm{e}^{\mathbf{Q}t}$.
- À partir des valeurs de $p_{ij}(t)$, calcul des valeurs de $f_{ij}(t)$:

$$f_{ij}(t) = \pi_i p_{ij}(t)$$

soit la probabilité d'observer le codon i de la séquence A aligné avec le codon j de la séquence B.

- Valeurs de π_i :
  - Uniforme ($\pi_i = 1/61, \forall i$).
  - À partir des fréquences des 61 codons dans le jeu de données (F61).
  - À partir des fréquences des nucléotides, toutes positions confondues (F1×4).
  - À partir des fréquences des nucléotides à chacune des trois positions des codons (F3×4).

Modélisation de l'hétérogénéité

- Dans une phylogénie, certaines lignées peuvent être soumises à de la sélection et d'autres non.
- Utilisation de plusieurs modèles afin de pouvoir détecter ces phénomènes :
  - Même valeur de ω pour toutes les branches de l'arbre (homogénéité).
  - Autant de valeurs de ω qu'il existe de branches dans l'arbre (hétérogénéité maximale).
  - Plusieurs intermédiaires entre ces deux extrêmes.
- Comparaison des différents modèles afin de déterminer quel est le scénario le plus vraisemblable.

Séquences protéiques

- Premières séquences biologiques à avoir été utilisées pour construire des phylogénies moléculaires.
- Toujours fréquemment utilisées :
  - Plus conservées que les séquences d'ADN (substitutions synonymes) :
    - Utiles pour des analyses portant sur de longues durées évolutives ou sur des séquences évoluant rapidement.
    - Généralement inutilisables dans le cas d'organismes trop proches.
- Existence de nombreux modèles permettant d'estimer le nombre de substitutions entre deux séquences.

Modèle de Poisson

- Introduit par Zuckerkandl et Pauling (1965).
- Correction la plus simple pour les séquences protéiques :
  - Modélisation par une distribution de Poisson.
- Hypothèses :
  - Tous les sites évoluent indépendemment et selon le même processus.
  - Toutes les substitutions sont équiprobales.
  - Le taux de réversion est négligeable.
- Calcul de la distance au moyen de la formule :

$$d = -\ln(1-p)$$

Modèle GTR pour les protéines ?

- Matrice 20×20 des taux instantanés :

$$\mathbf{Q} = \mathbf{S\Pi} = \begin{pmatrix} -\lambda_{\mathrm{A}} & \pi_{\mathrm{R}} s_{\mathrm{AR}} & \cdots & \pi_{\mathrm{V}} s_{\mathrm{AV}} \\ \pi_{\mathrm{A}} s_{\mathrm{AR}} & -\lambda_{\mathrm{R}} & \cdots & \pi_{\mathrm{V}} s_{\mathrm{RV}} \\ \vdots & \vdots & \ddots & \vdots \\ \pi_{\mathrm{A}} s_{\mathrm{AV}} & \pi_{\mathrm{R}} s_{\mathrm{RV}} & \cdots & -\lambda_{\mathrm{V}} \end{pmatrix}$$

  Soit 190 paramètres d'échangeabilité s_{ij} et 19 fréquences à l'équilibre π_i $(i, j \in \{\mathrm{A}, \mathrm{R}, \mathrm{N}, \ldots, \mathrm{V}\})$.
- Non directement utilisable entre deux séquences :
  - Pas assez de données pour permettre l'estimation d'un si grand nombre de paramètres.
  - Estimation à partir de jeux de données de grande taille, puis fixation des paramètres.

Modèles empiriques

- Pas de définition *formelle* des probabilités de transition.
- Estimation des $p_{ij}(t)$ à partir de la fréquence des substitutions estimées sur des ensembles de séquences alignées :

$$f_{ij}(t) = \pi_i p_{ij}(t)$$

- Valeurs de $f_{ij}(t)$:
  - Inférence par maximum de parcimonie :
    - PAM (*Point Accepted Mutation*, Dayhoff *et al.*, 1978).
    - JTT (Jones, Taylor et Thornton, 1992).
  - Inférence par maximum de vraisemblance :
    - WAG (Whelan et Goldman, 2001).
    - LG (Le et Gascuel, 2008).

Approximation de Kimura

- Calcul rapide d'une distance PAM avec les ordinateurs d'aujourd'hui, mais pas au moment de la conception du modèle.
- Mise en place par Kimura (1983) d'une mesure permettant d'approximer cette distance :

$$d = -\ln(1 - p - 0.2p^2)$$

- Méthode simple et rapide, mais présentant deux inconvénients :
  - Pas de possibilité de prise en compte des fréquences à l'équilibre des séquences étudiées.
  - La précision de l'estimation diminue avec le degré de divergence entre les séquences ($p \leq 0.75$).

Matrices de substitution

- Utilisées par les programmes d'alignement et de recherche de similarités :
  - Différentes des matrices de transition utilisées pour la reconstruction phylogénétique.
- Calcul effectué à partir des matrices de transition $\mathbf{P}(d)$ pour des valeurs *fixées* de d (0.3, 1, 1.5, 2.5, etc.) :
  - Soit $\hat{p}_{ij}(d)$ la probabilité d'une transition $i \to j$ estimée avec $\mathbf{P}(d)$.
  - Chaque élément $\delta_{ij}(d)$ de la matrice de substitution correspondante est défini par :

$$\delta_{ij}(d) = 10 \log \left(\frac{\hat{p}_{ij}(d)}{\pi_j} \right)$$

    avec arrondi à l'entier le plus proche.

Données pour les autres modèles

- Modèle JTT :
  - Utilisation de 16300 séquences totalisant 59190 substitutions.
  - Procédure de construction identique à PAM.
- Modèle WAG :
  - Utilisation de 3905 séquences provenant de 182 familles.
  - Utilisation du maximum de vraisemblance pour estimer les probabilités de transitions :
    - Prise en compte des substitutions multiples.
- Modèle LG :
  - Utilisation de 49637 séquences provenant de 3912 familles.
  - Prise en compte des différences de vitesse d'évolution entre les sites.

Écriture et utilisation

- Dans les publications récentes, indication des valeurs de $\mathbf{S}$ et $\mathbf{\Pi}$ plutôt que de celles de $\mathbf{P}(t)$ ou $\mathbf{Q}$:
  - Fait pour permettre le remplacement facile des valeurs de π_i fournies par le modèle, sachant que :

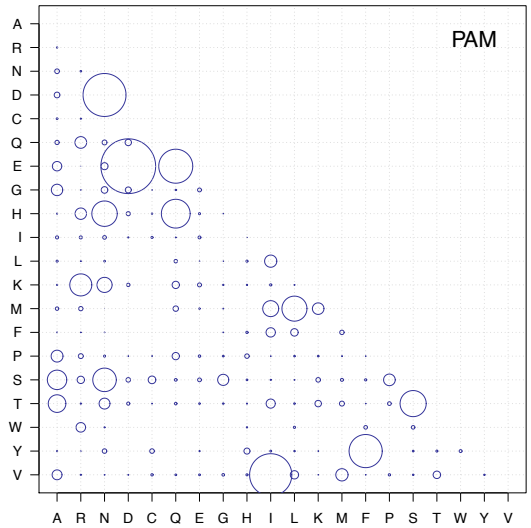
$$\mathbf{Q} = \mathbf{S\Pi}$$

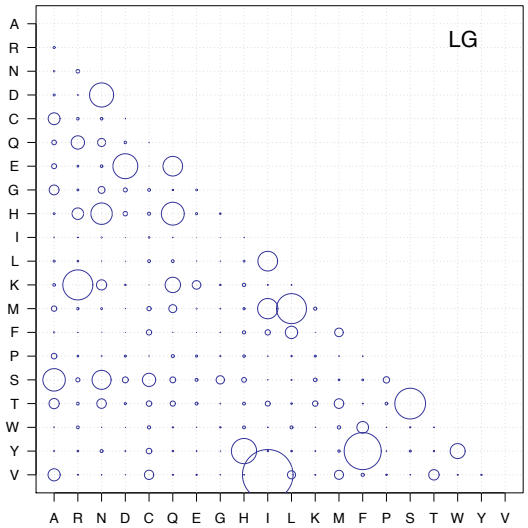
    avec la nécessité habituelle de normaliser les valeurs de q_{ij} de façon à ce que $\sum_i \pi_i \lambda_i = 1$.
  - Déduction des valeurs de $\mathbf{P}(t)$.
  - Calcul des distances évolutives avec la même procédure que celle utilisée pour PAM.

Comparaison des échangeabilités

Sur- ou sous-estimations de certaines valeurs de PAM :

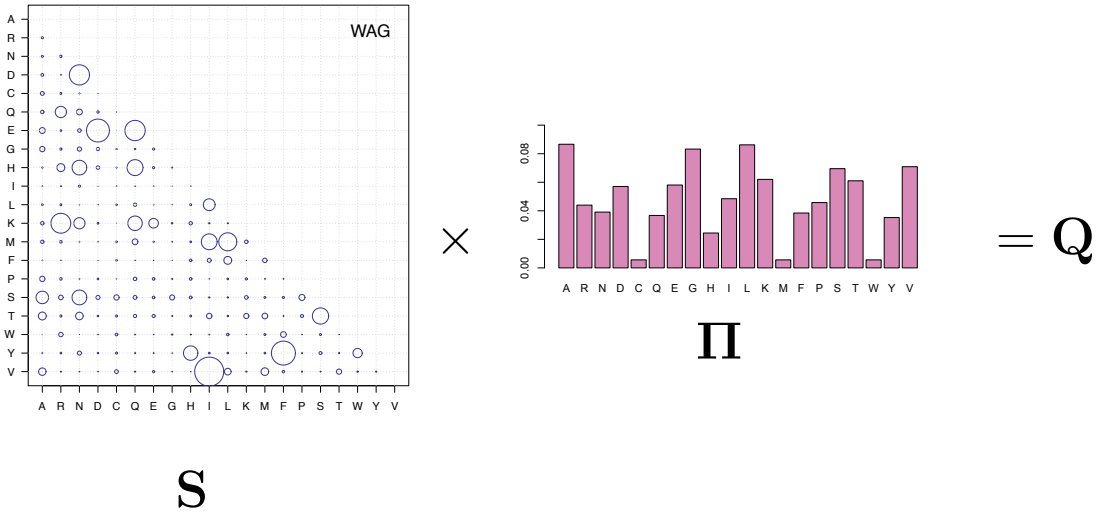
Problème lié à la taille de l'échantillon utilisé



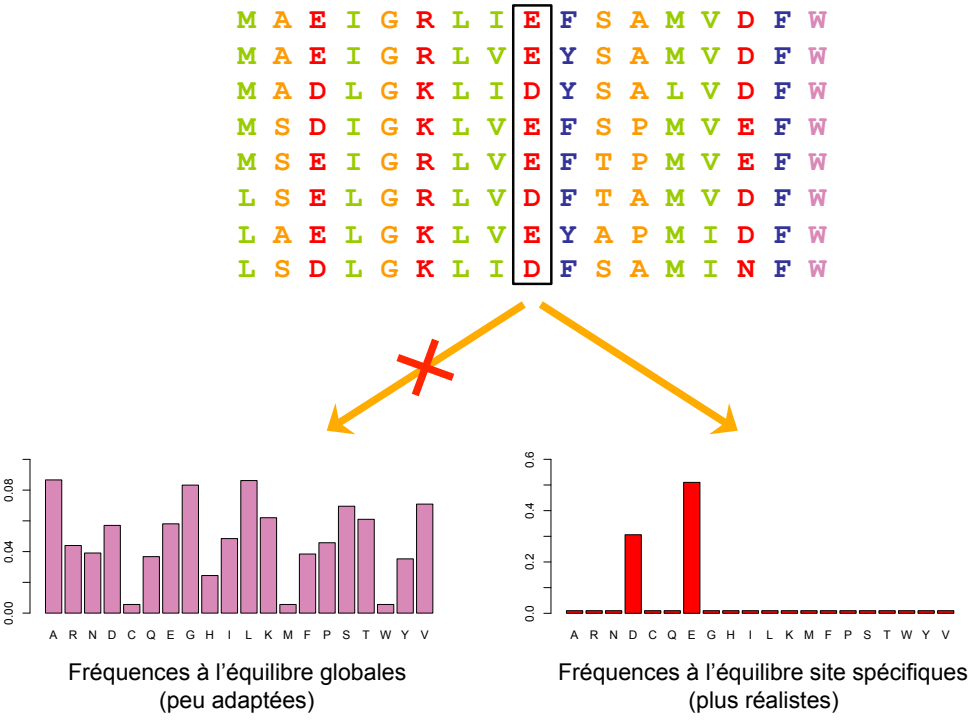


Approche classique

- Échangeabilités estimées à partir d'un jeu de données établi par les concepteurs du modèle.
- Fréquences à l'équilibre provenant du modèle ou obtenues à partir des séquences de l'alignement.



Limites de l'approche classique



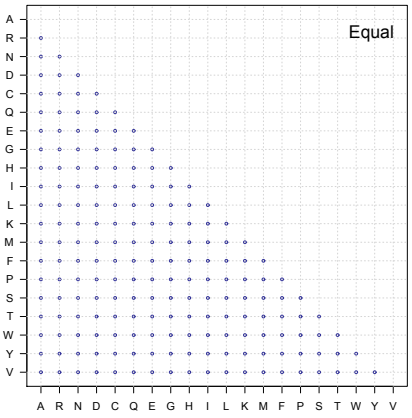
Fréquences à l'équilibre globales (peu adaptées)

Fréquences à l'équilibre site spécifiques (plus réalistes)

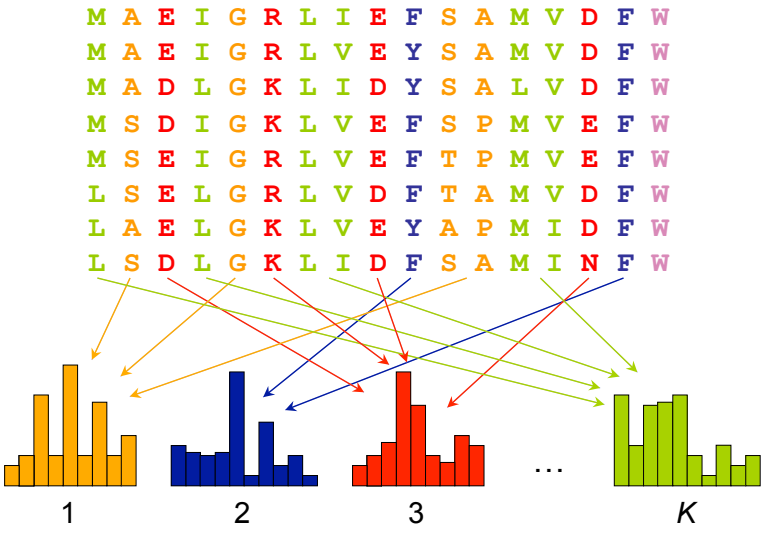
Approche site spécifique

- L'utilisation d'un jeu de valeurs π_i « globales » est non réaliste.
- Il n'est cependant pas possible d'utiliser un jeu par site de l'alignement :
  - Risques de surparamétrisation.
- Développement du modèle CAT (Le *et al.*, 2008) dans lequel il existe des *catégories* de sites :
  - Fréquences à l'équilibre :
    - Un jeu de valeurs de π_i par catégorie.
    - Cinq variantes à 20, 30, 40, 50 et 60 catégories.
  - Échangeabilités :
    - Une valeur unique, à l'image du modèle F81 (CAT-Poissson).
    - Valeurs provenant des modèles classiques (*e.g.*, CAT-JTT).
    - Valeurs estimées sur le jeu de données (CAT-GTR).

CAT-Poisson



Une échangeabilité α



K catégories de valeurs de π_i
(K = 20, 30, 40, 50, 60)

Méthodes de distances

Formation CNRS « Phylogénie moléculaire »

Guy Perrière

Laboratoire de Biométrie et Biologie Évolutive
UMR CNRS n° 5558
Université Claude Bernard – Lyon 1

20-23 mars 2018

Principe général

Alignement de séquences

↓ Mesures de distances évolutives

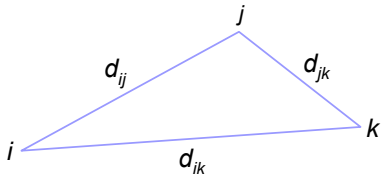
Matrice de distances évolutives entre paires de séquences

↓ Calcul de l'arbre à partir de la matrice

Arbre

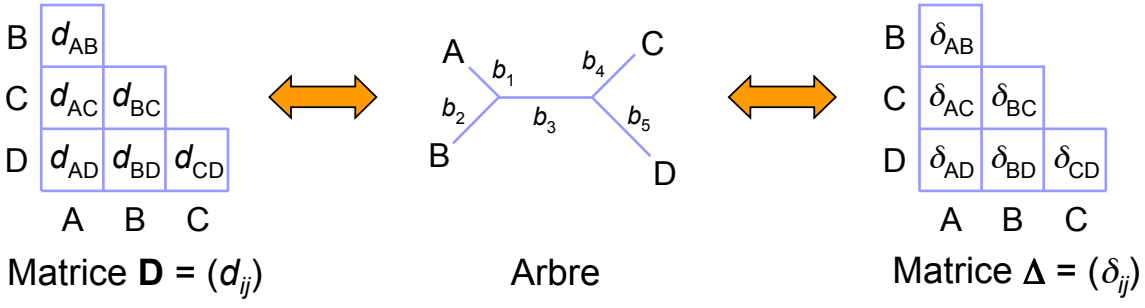
Notion de distance

- En mathématiques, une distance (ou *métrique*) sur un ensemble E est une fonction $d : E \times E \mapsto \mathbb{R}^+$.
- Cette fonction doit satisfaire à trois conditions, ceci $\forall i, j, k \in E$:
  - *Symétrie* – la distance entre deux points est la même, quelle que soit la direction considérée ($d_{ij} = d_{ji}$).
  - *Séparation* – si la distance entre deux points est égale à zéro, alors ces deux points sont confondus ($d_{ij} = 0 \Leftrightarrow i = j$).
  - *Inégalité triangulaire* – le chemin direct entre deux points est le plus court ($d_{ik} \leq d_{ij} + d_{jk}$) :



Distance arborée

- Dans un arbre, la distance δ_{ij} entre deux UTO i et j est donnée par la somme des longueurs de branches les séparant :
  - On parle de distance *arborée* ou *patristique* :
    - Doit vérifier, en plus des trois conditions standard, la *condition des quatre points* ($\delta_{ij} + \delta_{kl} \leq \max(\delta_{ik} + \delta_{jl}, \delta_{il} + \delta_{jk})$).
- Objectif des différentes méthodes de distances :
  - Faire que les valeurs δ_{ij} correspondent le plus fidèlement possible aux valeurs de d_{ij} présentes dans la matrice de départ.



Typologie

- Méthodes nécessitant d'explorer l'ensemble des topologies (optimisation d'un critère) :
  - Moindres carrés (*Least Squares*, LS) :
  - Mimum d'évolution (*Minimum of Evolution*, ME).
- Méthodes construisant un arbre unique :
  - Classification ascendante hiérarchique au lien moyen (*Unweighted Pair-Group Method with Arithmetic means*, UPGMA).
  - *Neighbor Joining* (NJ).

Principe général

- Pour une topologie τ donnée, déterminer quelles sont les valeurs des longueurs de branches minimisant :

$$Q = \sum_{i<j} w_{ij}(d_{ij} - \delta_{ij})^2$$

  avec w_{ij} les valeurs de pondération associées à chaque paire (i, j) :
  - Pondération uniforme ($w_{ij} = 1$).
  - Inverse de la distance ($w_{ij} = 1/d_{ij}$).
  - Inverse du carré de la distance ($w_{ij} = 1/d_{ij}^2$).
- Effectuer ces calculs pour l'ensemble des topologies possibles :
  - Retenir celle pour laquelle Q est minimale.

Avantages et limitations

- Méthode consistante.
- Algorithme de complexité en $O(n^3)$.
- Aussi efficace que le maximum de vraisemblance si les variables suivent une distribution normale :
  - Nécessité d'avoir un grand nombre de sites dans l'alignement.
- Problèmes de dérives numériques :
  - La résolution du problème des moindres carrés nécessite d'effectuer l'inversion d'une matrice pouvant être de grande taille.
  - Utilisation de simplifications ne nécessitant pas d'effectuer cette inversion de matrice :
    - Approximation de Fitch et Margoliash (1967).
    - Simplification de Rzhetsky et Nei (1992).

Approximation de Fitch et Margoliash

- Estimations moins précises que celles obtenues par les moindres carrés proprement dits :
  - Différences observées souvent négligeables.
- Construction en effectuant des groupements par triplets :
  - Correspondance exacte entre distance observée et la distance patristique :
    - Calcul simple des longueurs de branches.
  - Soit d_{AB}, d_{AC} et d_{BC} les valeurs des distances entre trois groupes A, B et C, dans ce cas, il est possible d'écrire que :

$$\begin{cases} d_{AB} = b_A + b_B \\ d_{AC} = b_A + b_C \\ d_{BC} = b_B + b_C \end{cases} \Leftrightarrow \begin{cases} b_A = (d_{AB} + d_{AC} - d_{BC})/2 \\ b_B = (d_{AB} + d_{BC} - d_{AC})/2 \\ b_C = (d_{AC} + d_{BC} - d_{AB})/2 \end{cases}$$

Algorithme I

Pour chacune des $n(n-1)/2$ paires (i,j) possibles, faire :

1. $A \leftarrow i$, $B \leftarrow j$ et regroupement de toutes les autres UTO dans C.
2. Calcul des distances d_{AC} et d_{BC} telles que :

$$d_{AC} = \frac{1}{n_C} \sum_{j \in C} d_{Aj} \quad \text{et} \quad d_{BC} = \frac{1}{n_C} \sum_{j \in C} d_{Bj}$$

   avec $n_C = \text{card}(C)$ le nombre d'éléments présents dans C.
3. Calcul des trois longueurs de branches au moyen de la formule précédente :
   - Soustraction des longueurs déjà calculées le cas échéant.

Algorithme II

4. Regrouper A et B dans un même ensemble $Z = A \cup B$ puis calculer, pour chaque $j \in C$:

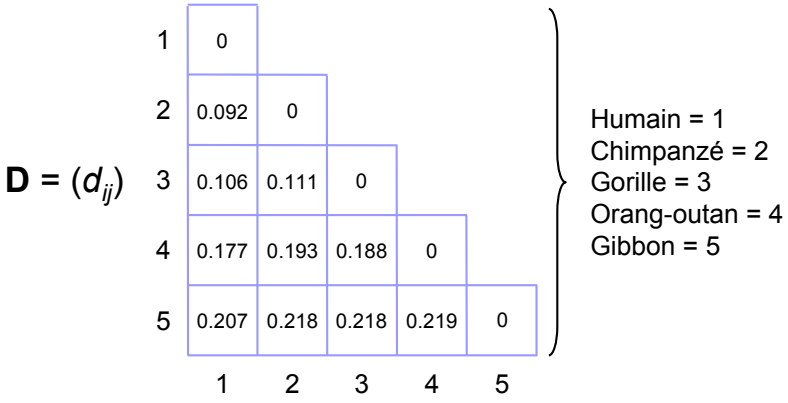
$$d_{Zj} = \frac{1}{n_Z} \sum_{i \in Z} d_{ij}$$

   avec $n_Z = \text{card}(Z)$, le nombre d'éléments présents dans Z. Les valeurs obtenues remplacent celles correspondant à A et à B.
5. Si $\dim(\mathbf{D}) \geq 3$, alors :
   - Réinitialiser A et B avec les UTO ou les groupes d'UTO pour lesquels d_{ij} est minimale et retourner en 2.

   Sinon, aller en 6.
6. Calcul de la valeur de Q.

Jeu de données exemple

- Jeu de données de Brown *et al.* (1982) sur les séquences d'ADN mitochondrial d'Hominoïdes.
- Modèle de Kimura à deux paramètres pour le calcul de la matrice de distances :



Exemple d'utilisation I

- Initialisation en prenant la paire (i, j) telle que d_{ij} soit minimale :

$$A \leftarrow \{1\},\ B \leftarrow \{2\} \text{ et } C \leftarrow \{3, 4, 5\}$$

- Calcul de d_{AB}, d_{AC} et d_{BC} :

$$\begin{cases} d_{AB} = 0.092 \\ d_{AC} = (0.106 + 0.177 + 0.207)/3 = 0.163 \\ d_{BC} = (0.111 + 0.193 + 0.218)/3 = 0.174 \end{cases}$$

- Calcul des longueurs de branches correspondantes :

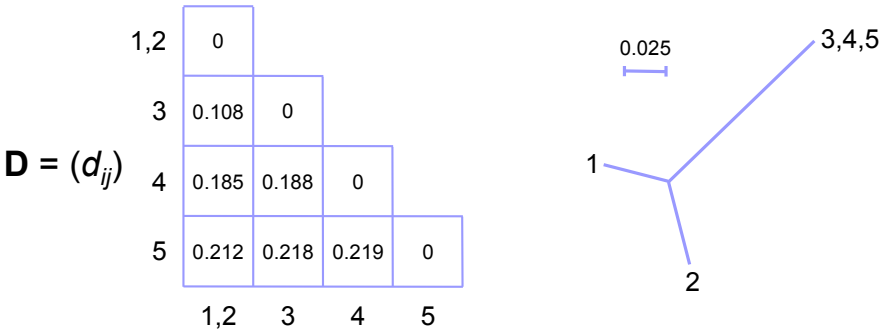
$$\begin{cases} b_A = (0.092 + 0.163 - 0.174)/2 = 0.041 \\ b_B = (0.092 + 0.174 - 0.163)/2 = 0.052 \\ b_C = (0.163 + 0.174 - 0.092)/2 = 0.123 \end{cases}$$

Exemple d'utilisation II

- Calcul des nouvelles distances, avec $Z = A \cup B = \{1, 2\}$:

$$\begin{cases} d_{Z3} = (0.106 + 0.111)/2 = 0.108 \\ d_{Z4} = (0.177 + 0.193)/2 = 0.185 \\ d_{Z5} = (0.207 + 0.218)/2 = 0.212 \end{cases}$$

- Nouvelles valeurs de $\mathbf{D}$ et arbre obtenu :



Exemple d'utilisation III

- Du fait que $\dim(\mathbf{D}) \geq 3$, on relance une itération avec :

$$A \leftarrow \{1, 2\},\ B \leftarrow \{3\} \text{ et } C \leftarrow \{4, 5\}$$

- Calcul de d_{AB}, d_{AC} et d_{BC} :

$$\begin{cases} d_{AB} = 0.108 \\ d_{AC} = (0.185 + 0.212)/2 = 0.199 \\ d_{BC} = (0.188 + 0.218)/2 = 0.203 \end{cases}$$

- Calcul des longueurs de branches correspondantes :

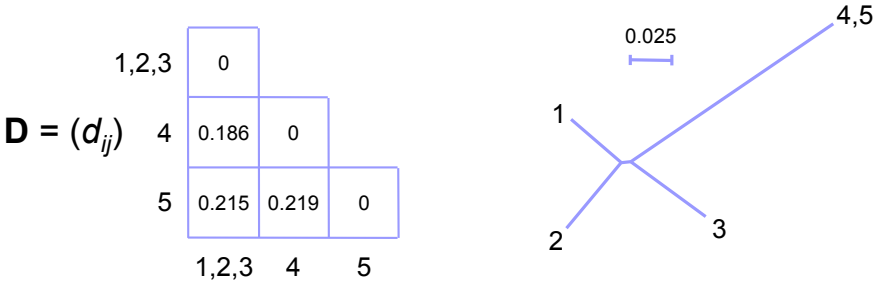
$$\begin{cases} b_A = (0.108 + 0.199 - 0.203)/2 = 0.052 \\ b_B = (0.108 + 0.203 - 0.199)/2 = 0.056 \\ b_C = (0.199 + 0.203 - 0.108)/2 = 0.147 \end{cases}$$

Exemple d'utilisation IV

- Dans le cas de b_A, prise en compte des longueurs de branches existantes conduisant aux éléments de A :
  - La longueur de la branche interne à ajouter est égale à $0.052 - (0.0405 + 0.0515)/2 = 0.006$.
- Calcul des nouvelles distances, avec $Z = A \cup B = \{\{1, 2\}, 3\}$:

$$\begin{cases} d_{Z4} = (0.185 + 0.188)/2 = 0.186 \\ d_{Z5} = (0.212 + 0.218)/2 = 0.215 \end{cases}$$

- Nouvelles valeurs de $\mathbf{D}$ et arbre obtenu :



Exemple d'utilisation V

- Dernière itération avec :

$$A \leftarrow \{1, 2, 3\},\ B \leftarrow \{4\} \text{ et } C \leftarrow \{5\}$$

- Calcul de d_{AB}, d_{AC} et d_{BC} :

$$\begin{cases} d_{AB} = 0.186 \\ d_{AC} = 0.215 \\ d_{BC} = 0.219 \end{cases}$$

- Calcul des longueurs de branches correspondantes :

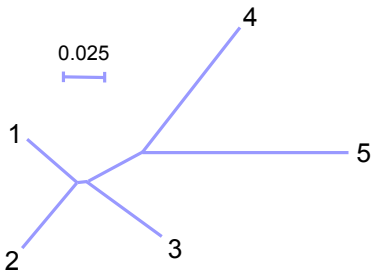
$$\begin{cases} b_A = (0.186 + 0.215 - 0.219)/2 = 0.091 \\ b_B = (0.186 + 0.219 - 0.215)/2 = 0.095 \\ b_C = (0.215 + 0.219 - 0.186)/2 = 0.124 \end{cases}$$

Exemple d'utilisation VI

- Dans le cas de b_A prise en compte des longueurs de branches existantes conduisant aux éléments de A :
  - La longueur de la branche interne à ajouter est égale à $0.091 - (0.0405 + 0.006 + 0.0515 + 0.006 + 0.056)/3 = 0.038$.
- Matrice des distances patristiques et arbre obtenus :

$\Delta = (\delta_{ij})$

	1	2	3	4	5
1	0				
2	0.092	0			
3	0.103	0.113	0		
4	0.179	0.191	0.189	0	
5	0.208	0.220	0.218	0.219	0



Avantages et limitations

- Calcul simultané de la topologie et des longueurs de branches.
- Pas d'exploration de l'ensemble des topologies :
  - Seulement $n(n-1)/2$ itérations (*i.e.*, le nombre de paires possibles entre deux UTO) :
    - Complexité globale de l'algorithme en $O(n^5)$.
  - Pas de garantie que l'arbre obtenu soit effectivement celui des moindres carrés.

Minimum d'évolution

- Méthode très comparable aux moindres carrés (mêmes avantages et inconvénients).
- Pour une topologie τ donnée :
  - Détermination des longueurs de branches par les moindres carrés.
  - Calcul de la longueur de l'arbre S, telle que :

$$S = \sum_{k=1}^{2n-3} b_k$$

- Effectuer ces calculs pour l'ensemble des topologies possibles :
  - Retenir celle pour laquelle S est minimale.

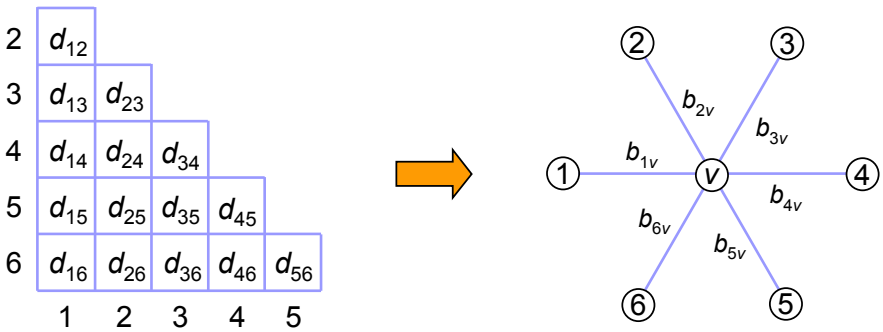
Algorithme I

1. Initialisation à partir d'une topologie en étoile telle que :
   - Branches b_{iv} reliées à un nœud central v.
   - Expression des valeurs de d_{ij} à partir des longueurs de branches :

$$d_{ij} = b_{iv} + b_{jv} \ (i \neq j)$$

   - Longueur de l'arbre déduite :

$$S_0 = \sum_{i=1}^{n} b_{iv} = \frac{1}{n-1} \sum_{i<j} d_{ij}$$

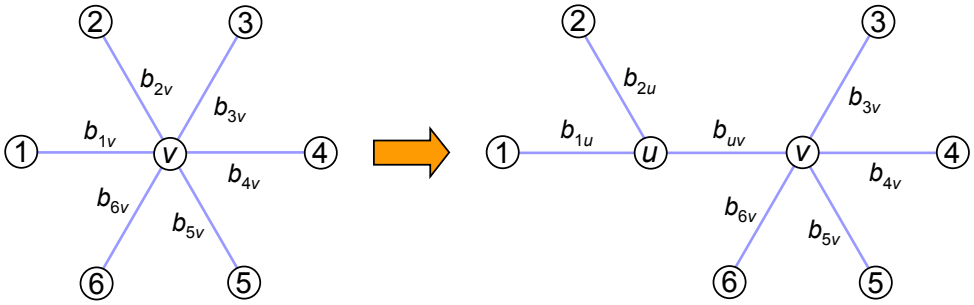


Algorithme II

2. Identification de la paire (i, j) qui, une fois agglomérée, minimise la longueur de l'arbre S_{ij} :
   - Création d'un nœud u connectant i et j.
   - Création d'une branche interne b_{uv} connectant u et v.
   - Dans ce cas, expression de S_{ij} comme :

$$\begin{aligned} S_{ij} &= b_{iu} + b_{ju} + b_{uv} + S_k \\ &= d_{ij} + b_{uv} + S_k \end{aligned}$$

   avec S_k la longueur de l'arbre en étoile contenant les $n-2$ UTO restantes.



Algorithme III

3. Sachant que :

$$S_k = \sum_{k \neq i,j} b_{kv} = \frac{1}{n-3} \sum_{k \neq i,j; k<l} d_{kl}$$

et que :

$$b_{uv} = \frac{1}{2(n-2)} \left[\sum_{k \neq i,j} (d_{ik} + d_{jk}) - (n-2) d_{ij} - 2S_k \right]$$

on en déduit l'expression de S_{ij} :

$$S_{ij} = \frac{1}{2} d_{ij} + \frac{1}{2(n-2)} \sum_{k \neq i,j} (d_{ik} + d_{jk}) + \frac{1}{n-2} \sum_{k \neq i,j; k<l} d_{kl}$$

Algorithme IV

4. Une fois la paire (i, j) identifiée, recalcul des longueurs de branches b_{iu} et b_{ju} au moyen de Fitch-Margoliash :

$$b_{iu} = \frac{1}{2}\left(d_{ij} + \frac{1}{n-2}\sum_{k\neq i,j} d_{ik} - \frac{1}{n-2}\sum_{k\neq i,j} d_{jk}\right)$$

et :

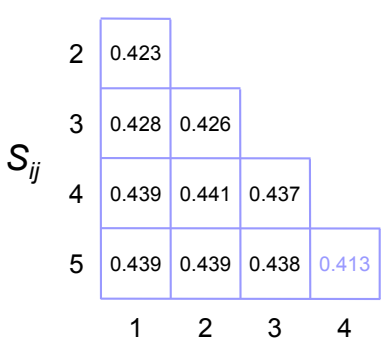
$$b_{ju} = \frac{1}{2}\left(d_{ij} + \frac{1}{n-2}\sum_{k\neq i,j} d_{jk} - \frac{1}{n-2}\sum_{k\neq i,j} d_{ik}\right)$$

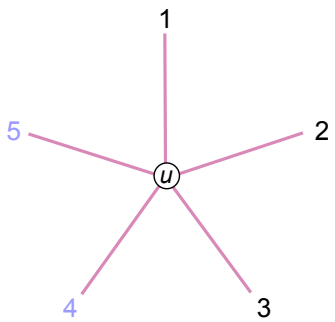
5. Recalcul de la matrice $\mathbf{D}$ en remplaçant les lignes correspondant à i et j par la paire (i, j), telle que :

$$d_{ij,k} = \frac{1}{2}(d_{ik} + d_{jk} - d_{ij})$$

Exemple d'utilisation I

- Initialisation à partir d'une topologie en étoile de longueur $S_0 = 0.432$.
- Calcul de l'ensemble des valeurs de S_{ij} possibles :
  - Identification de la paire $(4, 5)$ comme étant celle minimisant S_{ij} :



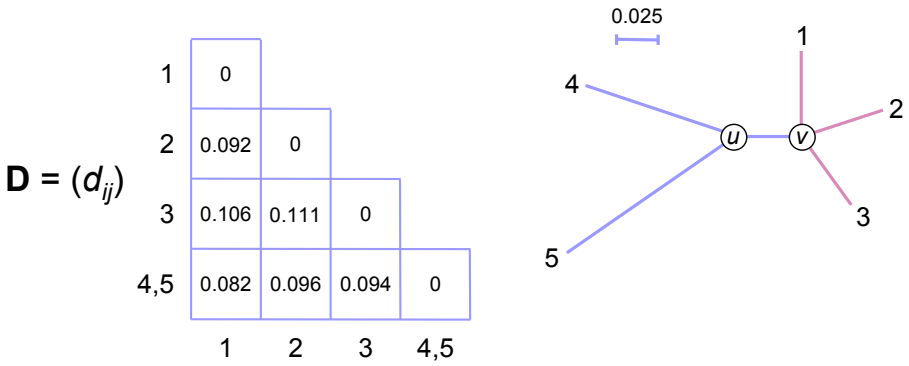


Exemple d'utilisation II

- Calcul des longueurs de branches conduisant à u et calcul de la longueur de la branche interne b_{uv} :

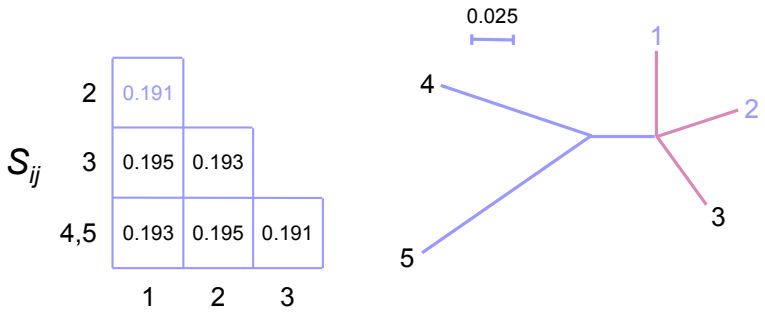
$$b_{4u} = 0.0955, b_{5u} = 0.1238 \text{ et } b_{uv} = 0.0392$$

- Nouvelles valeurs de $\mathbf{D}$ et arbre obtenu :



Exemple d'utilisation III

- Calcul de l'ensemble des nouvelles valeurs de S_{ij} possibles :
  - Identification de la paire $(1, 2)$ comme étant celle minimisant S_{ij} :

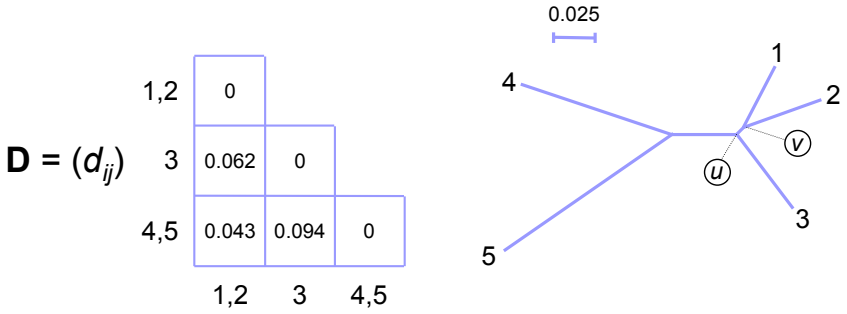


- Calcul des longueurs de branches conduisant à u et calcul de la longueur de la branche interne b_{uv} :

$$b_{1u} = 0.0413, b_{2u} = 0.0505 \text{ et } b_{uv} = 0.006$$

Exemple d'utilisation IV

- Nouvelles valeurs de $\mathbf{D}$ et arbre obtenu :



- Calcul de la longueur de la branche conduisant à $\{3\}$ en utilisant Fitch-Margoliash, soit :

$$(0.0623 + 0.0936 - 0.0432)/2 = 0.0564$$

Avantages et limitations

- Méthode consistante.
- À chaque itération, les longueurs de branches calculées sont une estimation de celles obtenues aux moindres carrés.
- Rapide, même avec des milliers d'UTO :
  - Implémentation originale par Saitou et Nei (1987) avec une complexité en $O(n^5)$.
  - Amélioration de Studier et Keppler (1988) réduisant la complexité en $O(n^3)$.
- L'arbre obtenu est une bonne approximation de l'arbre du minimum d'évolution.

Méthode du Maximum de vraisemblance (1)

(programmes fastDNAml, PAUP*, PROML, PROTML, PhyML)

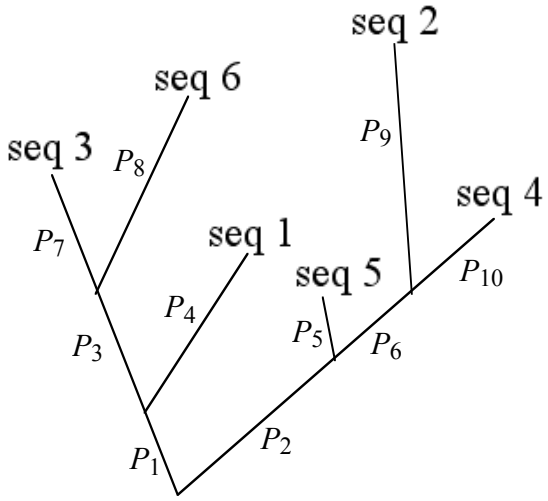
- Hypothèses
  - Le processus de substitution suit un modèle probabiliste dont on connaît l'expression mathématique, mais pas les valeurs numériques.
  - Les sites évoluent indépendamment les uns des autres.
  - Les sites évoluent selon la même loi (on peut aussi modéliser la variation des taux entre sites par une loi gamma).
  - Les taux de substitution ne changent pas au cours du temps le long d'une branche. Ils peuvent varier entre branches.

Méthode du Maximum de vraisemblance(2)

Modèle probabiliste de l'évolution de séquences

Chaque branche est modélisée par un modèle de Markov distinct.

Matrice P_b : probabilités conditionnelles de substitution le long de la branche *b*



En général, on suppose que les matrices P_b ne diffèrent que par leur paramètre de longueur, r_b, et partagent leurs paramètres qualitatifs, θ. Longueur d'une branche :

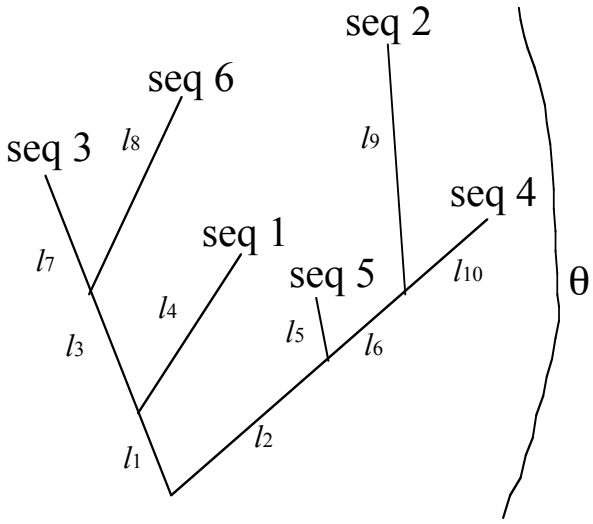
l_b = nbre attendu de subst. sur la branche $b \propto r_b$

Méthode du Maximum de vraisemblance(3)

Modèle probabiliste de l'évolution de séquences

l_b, longueur de la branche b = nbre attendu de substitutions par site le long de la branche

θ, taux relatifs des substitutions (e.g., transition/transversion, biais G+C, fréquences d'équilibre)



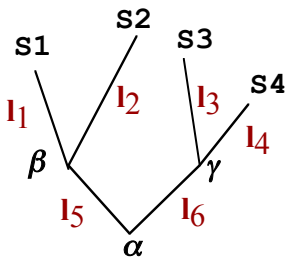
On sait calculer

$P_{branche\ b}(y \text{ en fin} \mid x \text{ en début})$

pour toutes bases x & y, toute branche b, toutes valeurs θ

Algorithme du maximum de vraisemblance (1)

- Etape 1: Pour une forme d'arbre racinée donnée, pour un site donné *y*, et pour un jeu de valeurs des longueurs de branches donné, on calcule la probabilité que le pattern de nucléotides observés à ce site ait évolué le long de cet arbre.



S1, S2, S3, S4: bases observées au site y dans seq. 1, 2, 3, 4
α,β, γ : bases ancestrales inconnues et variables
l1, l2, …, l6: longueurs des branches données

$L(y) = \Sigma_\alpha \Sigma_\beta \Sigma_\gamma\ P_{anc}(\alpha)\ P_{l5}(\alpha,\beta)\ P_{l6}(\alpha,\gamma)\ P_{l1}(\beta,S1)\ P_{l2}(\beta,S2)\ P_{l3}(\gamma,S3)\ P_{l4}(\gamma,S4)$

où $P_{anc}(S7)$ est estimée par les fréquences moyennes des bases dans les séquences.

Algorithme du maximum de vraisemblance(2)

Calcul général de la vraisemblance d'un site

$$L(y) = \sum_{i \in B} P_{anc}(r = i) L^{r,i}(y)$$

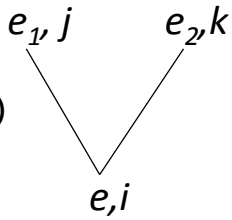
avec y : site; $B = \{A, C, G, T\}$; r : racine; P_{anc} : proba ancestrales des b

$L^{e,i}(y)$: vraisemblance au noeud e de l'arbre conditionnelle à la base i au noeud

Définition récursive de $L^{e,i}(y)$

si e est un noeud interne soient e_1 et e_1 ses 2 descendants

$$L^{e,i}(y) = \sum_{j \in B} \sum_{k \in B} P(e_1 = j \mid e = i) L^{e_1,j}(y) P(e_2 = k \mid e = i) L^{e_2,k}(y)$$



si e est une feuille

$$L^{e,i}(y) = \begin{cases} 1 \text{ si } i \text{ est la base au site } y \text{ de la sequence} \\ 0 \text{ sinon} \end{cases}$$

Algorithme du maximum de vraisemblance(3)

Si le modèle utilisé est réversible, homogène (les paramètres qualitatifs ne varient pas entre branches), et stationnaire (à l'équilibre des fréquences de bases),

alors

la **vraisemblance est indépendante de la position de la racine** dans l'arbre.

Algorithme du maximum de vraisemblance(4)

- Etape 2: calculer la probabilité que les séquences entières aient évolué :

$$L = \prod_{\text{sites } y} L(y)$$

C'est la vraisemblance du modèle. En pratique on calcule $log(L) = \Sigma log(L(y))$

- Etape 3: calculer les longueurs des branches $l1, l2, \ldots, l6$ et les valeurs du paramètre θ qui correspondent à la valeur maximale de L.
- Etape 4: calculer la vraisemblance de tous les arbres possibles. Retenir l'arbre associé à la plus haute vraisemblance.

Maximum de vraisemblance : propriétés

- 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.

Modélisation de la variation du taux d'évolution entre sites

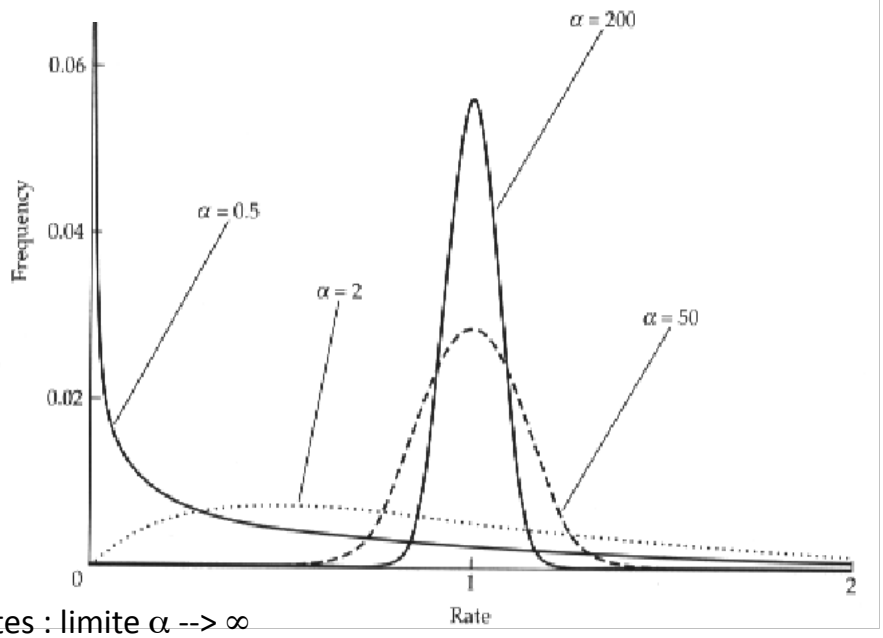
La distribution gamma n'a pas de justification biologique, uniquement commodité mathématique.

Densité *f(r)* de la distribution gamma:

$$f(r) = \frac{1}{\Gamma(\alpha)\beta^{\alpha}} r^{\alpha-1} e^{-r/\beta}$$

α: paramètre de forme
β: paramètre d'échelle

moyenne: $\alpha\beta$
variance: $\alpha\beta^2$



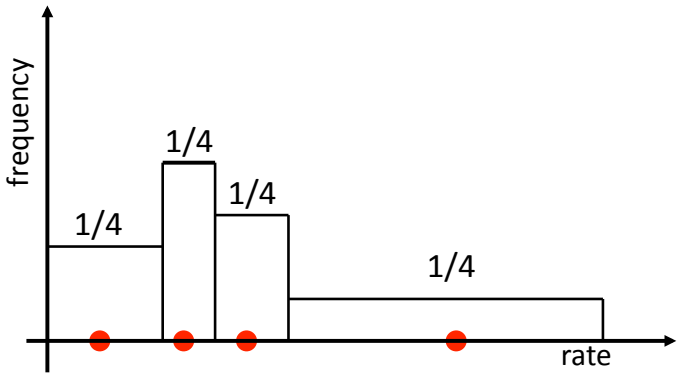
En phylogénie, utilisée pour modéliser la distribution des taux d'évolution entre sites avec $\beta=1/\alpha$ pour avoir
moyenne = 1
variance = $1/\alpha$

Pas de variation entre sites : limite α --> ∞

Modélisation de la variation du taux d'évolution entre sites (2)

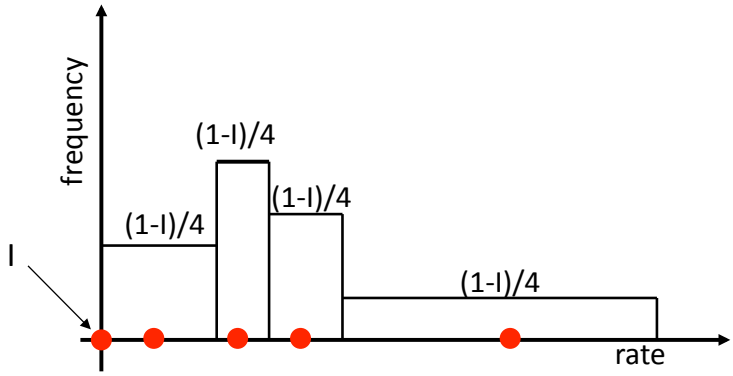
La distribution gamma est souvent discrétisée pour faciliter les calculs.

Exemple de discrétisation en 4 classes de poids égaux:



Modélisation de la variation du taux d'évolution entre sites (3)

On ajoute aussi souvent une autre classe de sites: les sites invariables
C'est le modèle G + I.
La fraction de sites invariables est estimée à partir des séquences.



Maximum de vraisemblance : vitesse d'évolution variable entre sites

On peut ajouter l'hypothèse que les vitesses d'évolution des sites varient selon la distribution gamma f_α de paramètre de forme α et de moyenne 1. La vraisemblance du site *y* devient

$$L(y) = \int_0^\infty f_\alpha(u) L(y,u) du$$

où *L(y,u)* est la vraisemblance calculée comme plus haut en multipliant par *u* tous les paramètres de longueur du modèle probabiliste.

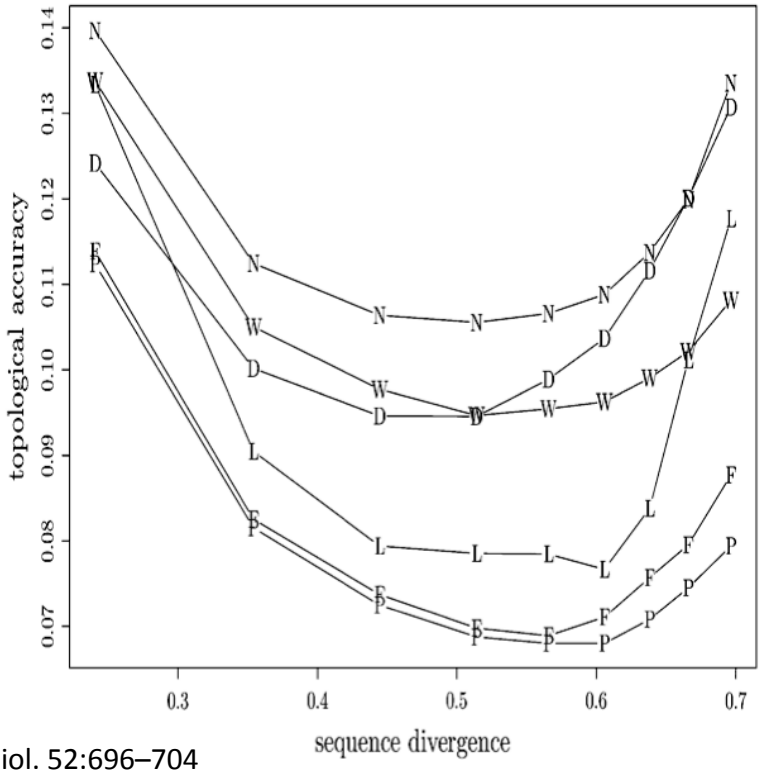
Pour obtenir quelque chose de calculable, on discrétise la distribution gamma en k quartiles représentés par leur moyenne w_j:

$$L(y) \approx \sum_{j=1}^{k} \frac{1}{k} L(y, w_j)$$

Comparaison des performances des méthodes par expériences de simulation de séquences et d'arbres

P, PHYML
F, fastDNAml
L, NJML
D, DNAPARS
N, NJ

5000 arbres aléatoires
40 taxons, 500 bases
pas d'horloge moléculaire
Niveau de divergence variable
K2P, $\alpha = 2$



Guindon & Gascuel (2003) Syst. Biol. 52:696–704

Evaluation et tests des phylogénies

Formation CNRS « Phylogénie moléculaire »

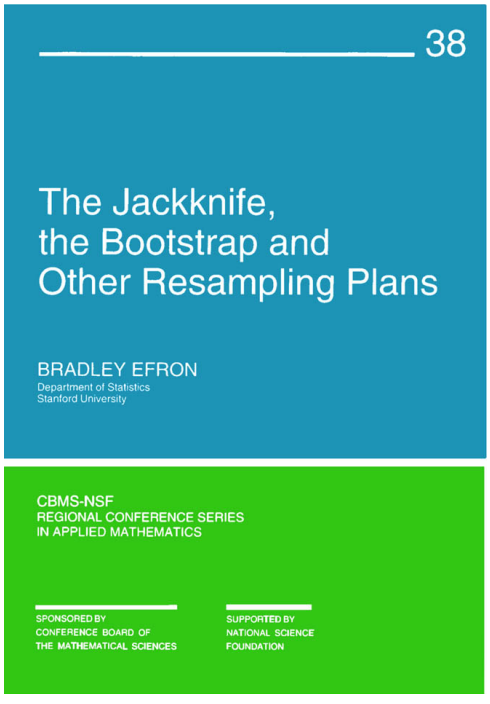
Guy Perrière

Laboratoire de Biométrie et Biologie Évolutive
UMR CNRS n° 5558
Université Claude Bernard – Lyon 1

20-23 mars 2018

Le *bootstrap*

- Bases mathématiques établies par Efron (1979) :
  - Construction d'intervalles de confiance.
  - Mesure de la précision d'une estimation.
- Adaptation à la phylogénie par Felsenstein (1985) :
  - Méthode aujourd'hui la plus couramment utilisée.

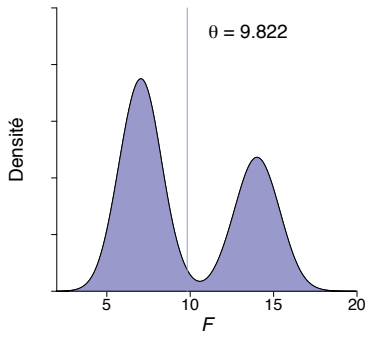


Principe général

- Soit un échantillon $\mathbf{x} = (x_1, x_2, \ldots, x_\ell)$ de ℓ observations tirées d'une distribution $\mathcal{F}$, de paramètre θ inconnu :
  - Soit $\hat{\mathcal{F}}$ la distribution observée dans cet échantillon :
    - Estimation de θ à partir de $\hat{\mathcal{F}}$.
- Mesure de l'intervalle de confiance de l'estimation précédente au moyen du *bootstrap* :
  - Tirage de B échantillons $\mathbf{x}^* = (x_1^*, x_2^*, \ldots, x_\ell^*)$ à partir de $\hat{\mathcal{F}}$.
  - Chaque $\mathbf{x}^*$ est construit par ℓ tirages avec remise dans $\mathbf{x}$ et constitue ce que l'on appelle un *réplicat de bootstrap*.
  - $I(\theta)$ à 95% obtenu en retirant les 2.5% de valeurs les plus hautes et les 2.5% de valeurs les plus basses.
  - Nécessité que B et ℓ soient grands et que les observations de $\mathbf{x}$ soient i.i.d.

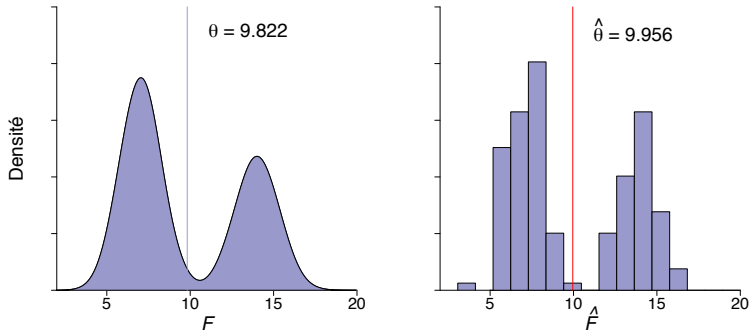
Moyenne d'une distribution

- Construction d'une distribution $\mathcal{F}$ par le mélange de deux lois normales :
  - $\mathcal{N}(7, 1)$, pour 60% des effectifs et $\mathcal{N}(14, 1)$, pour 40% des effectifs :
    - Moyenne de la distribution : $\theta = 9.822$.



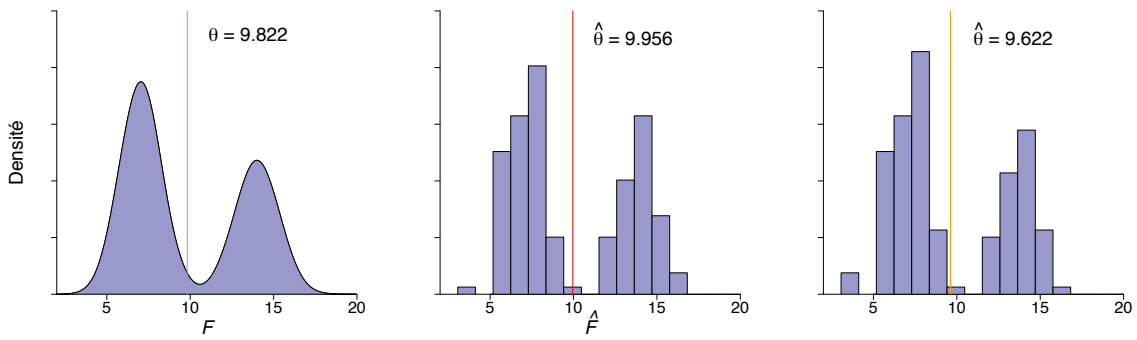
Moyenne d'une distribution

- Construction d'une distribution $\mathcal{F}$ par le mélange de deux lois normales :
  - $\mathcal{N}(7,1)$, pour 60% des effectifs et $\mathcal{N}(14,1)$, pour 40% des effectifs :
    - Moyenne de la distribution : $\theta = 9.822$.
- Tirage de $\ell = 150$ individus dans $\mathcal{F}$ pour construire $\hat{\mathcal{F}}$:
  - Moyenne estimée : $\hat{\theta} = 9.956$.



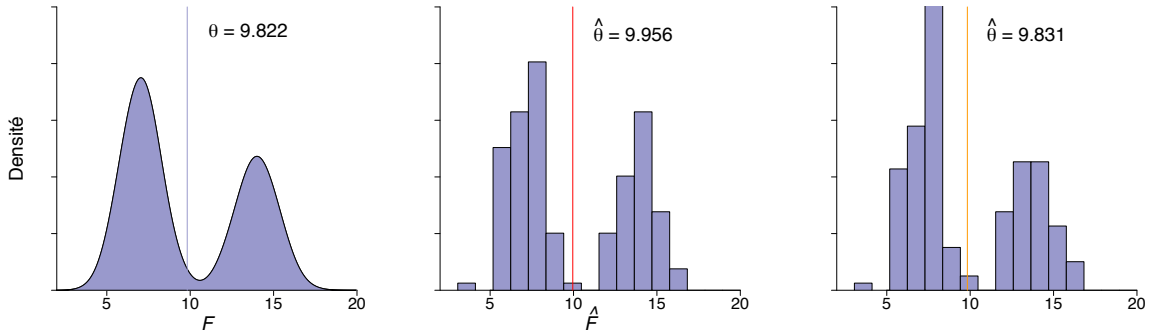
Moyenne d'une distribution

- Construction d'une distribution $\mathcal{F}$ par le mélange de deux lois normales :
  - $\mathcal{N}(7,1)$, pour 60% des effectifs et $\mathcal{N}(14,1)$, pour 40% des effectifs :
    - Moyenne de la distribution : $\theta = 9.822$.
- Tirage de $\ell = 150$ individus dans $\mathcal{F}$ pour construire $\hat{\mathcal{F}}$:
  - Moyenne estimée : $\hat{\theta} = 9.956$.
  - Mesure de la validité de cette estimation par *bootstrap* :



Moyenne d'une distribution

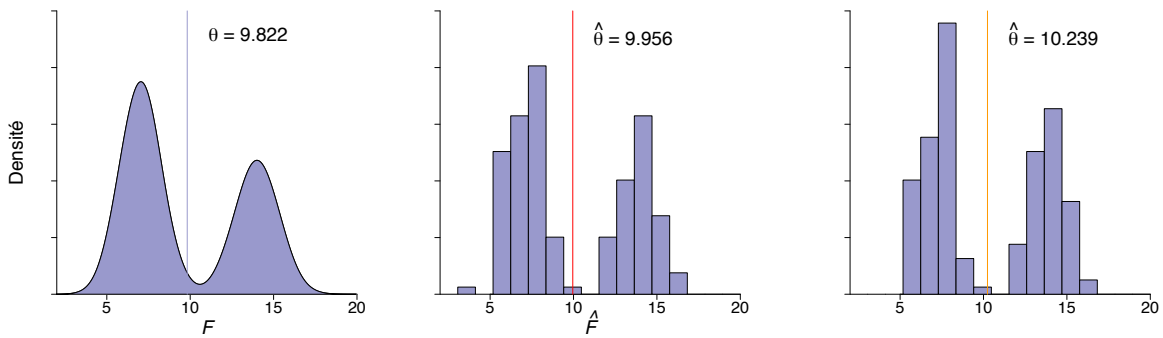
- Construction d'une distribution $\mathcal{F}$ par le mélange de deux lois normales :
  - $\mathcal{N}(7,1)$, pour 60% des effectifs et $\mathcal{N}(14,1)$, pour 40% des effectifs :
    - Moyenne de la distribution : $\theta = 9.822$.
- Tirage de $\ell = 150$ individus dans $\mathcal{F}$ pour construire $\hat{\mathcal{F}}$:
  - Moyenne estimée : $\hat{\theta} = 9.956$.
  - Mesure de la validité de cette estimation par *bootstrap* :



Bootstrap

Moyenne d'une distribution

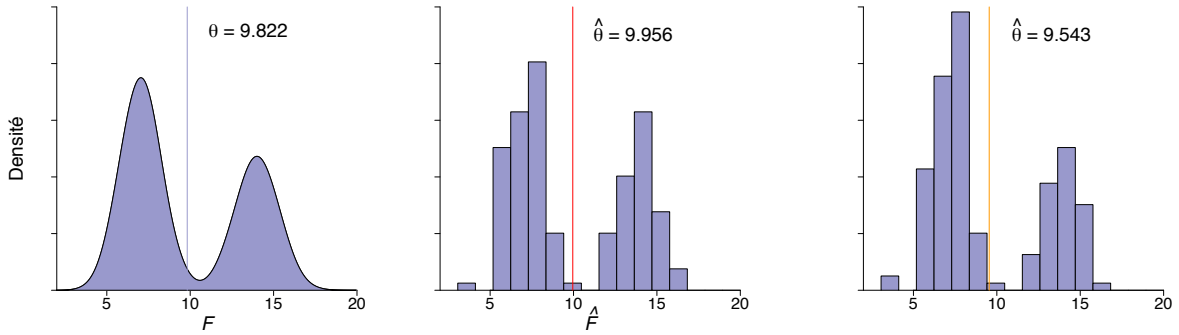
- Construction d'une distribution $\mathcal{F}$ par le mélange de deux lois normales :
  - $\mathcal{N}(7,1)$, pour 60% des effectifs et $\mathcal{N}(14,1)$, pour 40% des effectifs :
    - Moyenne de la distribution : $\theta = 9.822$.
- Tirage de $\ell = 150$ individus dans $\mathcal{F}$ pour construire $\hat{\mathcal{F}}$:
  - Moyenne estimée : $\hat{\theta} = 9.956$.
  - Mesure de la validité de cette estimation par *bootstrap* :



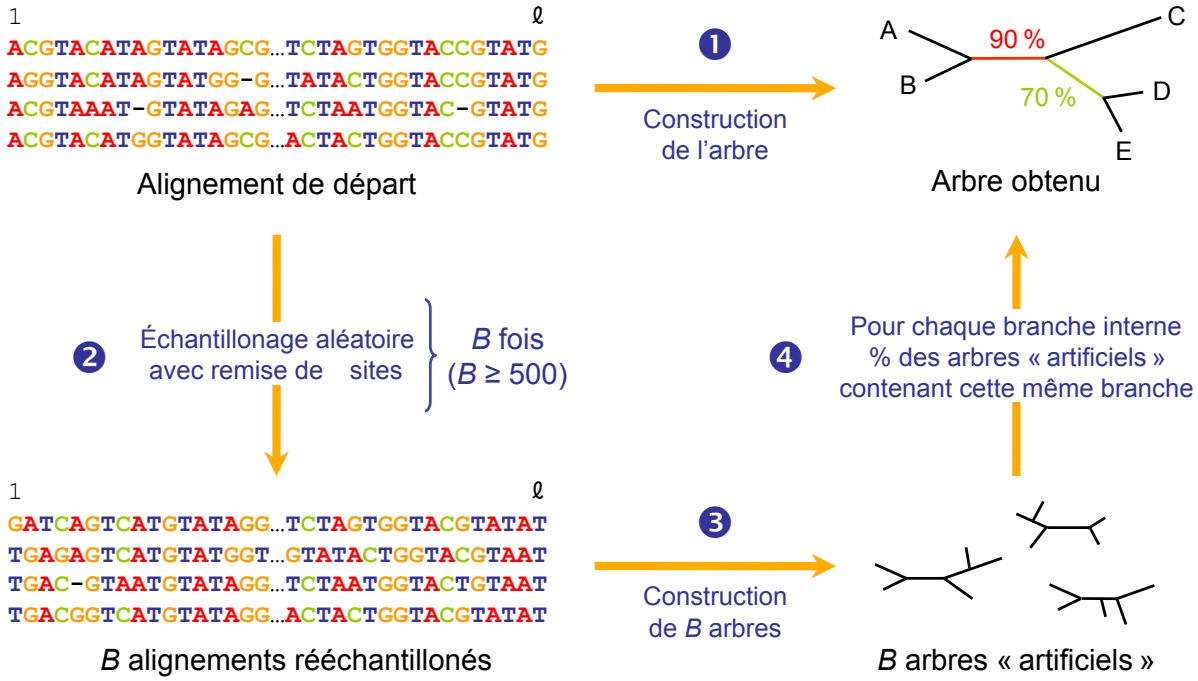
Guy Perrière (BBE) Tests 20-23 mars 2018 4 / 25

Moyenne d'une distribution

- Construction d'une distribution $\mathcal{F}$ par le mélange de deux lois normales :
  - $\mathcal{N}(7,1)$, pour 60% des effectifs et $\mathcal{N}(14,1)$, pour 40% des effectifs :
    - Moyenne de la distribution : $\theta = 9.822$.
- Tirage de $\ell = 150$ individus dans $\mathcal{F}$ pour construire $\hat{\mathcal{F}}$:
  - Moyenne estimée : $\hat{\theta} = 9.956$.
  - Mesure de la validité de cette estimation par *bootstrap* :



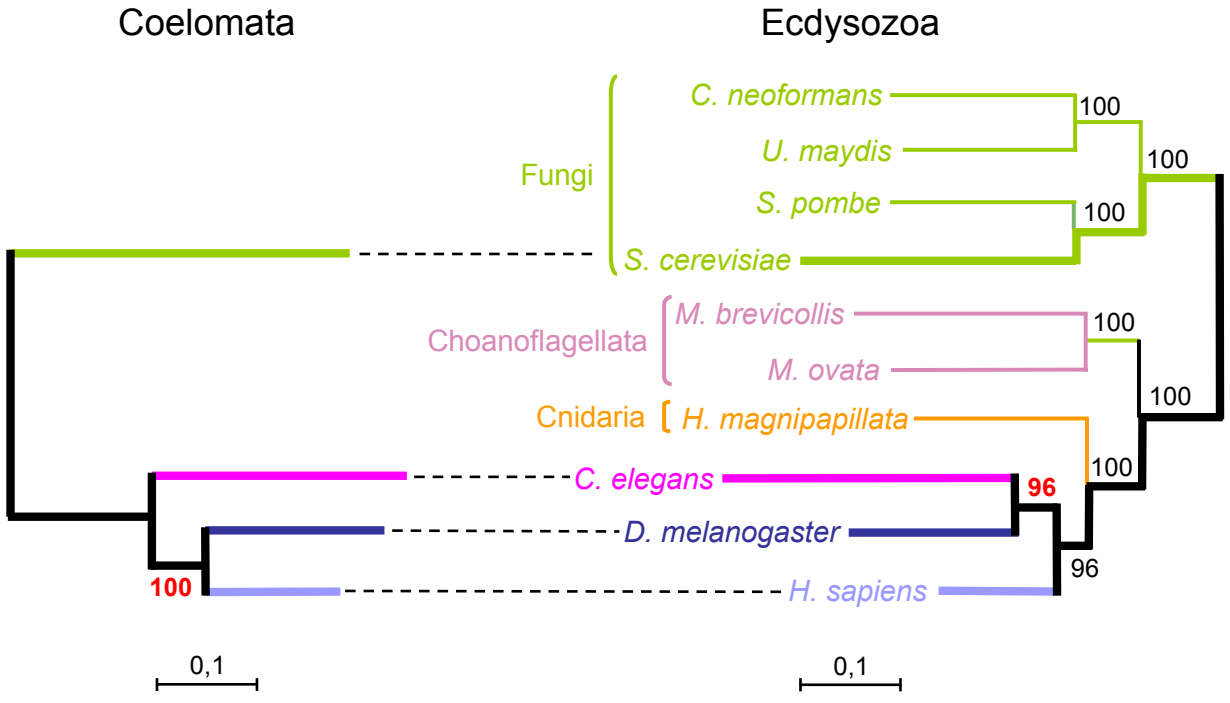
Application à la phylogénie



Limitations et usage

- Ne permet pas de déterminer si un arbre est vrai ou faux :
  - Un arbre faux peut avoir des branches soutenues par de fortes valeurs de *bootstrap*.
- Non-indépendance des observations (sites) :
  - Surestimation des scores faibles et sous-estimation des scores forts.
- En théorie, seuil en fonction d'un risque d'erreur fixé *a priori* :
  - En pratique, valeurs fluctuantes suivant les utilisateurs.
  - Seuils communément admis :
    - 100% : robustesse maximale.
    - 95-99% : très fort soutien par les données.
    - 90-94% : fort soutien par les données.
    - 80-89% : soutien modéré par les données.
    - $< 80\%$: pas de soutien.

Un exemple classique



Approximate Likelihood Ratio Test (aLRT)

- Alternative à l'utilisation du *bootstrap*, très coûteux en temps de calcul dans le cas du maximum de vraisemblance.
- Calcul de la statistique :
  - Soit τ_1 la topologie présentant la vraisemblance maximale $L(\tau_1)$.
  - Soit τ_2 la topologie présentant la *deuxième* vraisemblance maximale $L(\tau_2)$:
    - Obtention par réarrangement NNI autour de la branche d'intérêt b_k.
    - Fixation des autres paramètres $(\mathbf{b}, \boldsymbol{\vartheta}, \alpha)$.
  - Le rapport des vraisemblances est donné par :

$$\Lambda_k = 2 \ln \left[\frac{L(\tau_1)}{L(\tau_2)} \right] = 2 \left[\ln L(\tau_1) - \ln L(\tau_2) \right]$$

- Calcul du test :

$$\Lambda_k \sim \frac{1}{2} \left[\chi^2(0) + \chi^2(1) \right]$$

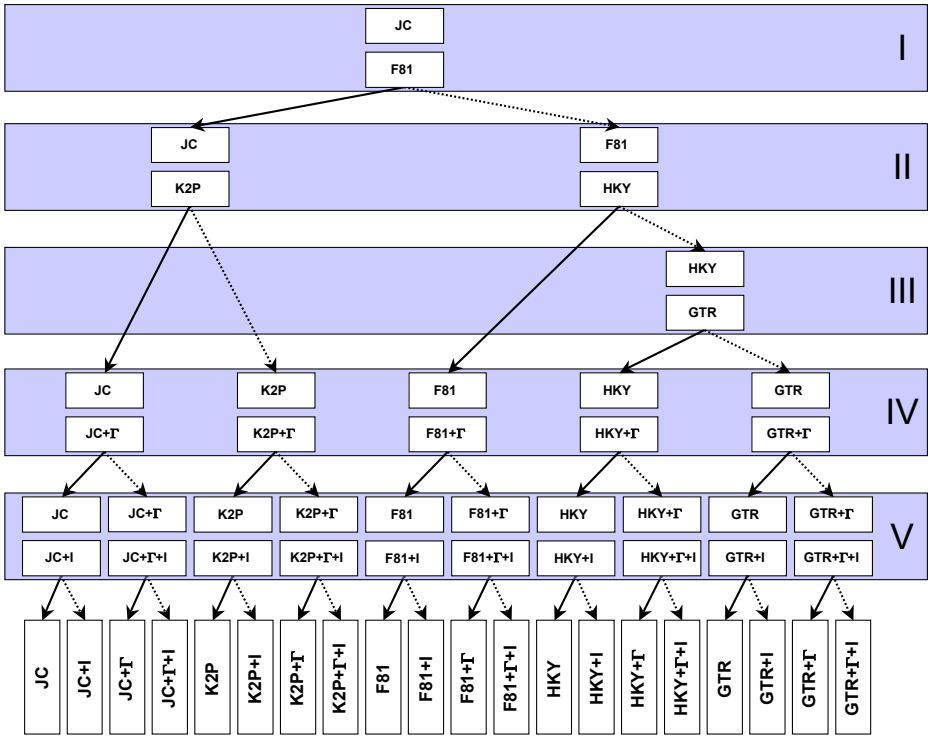
Likelihood Ratio Test (LRT)

- Soient M_0 et M_1 deux modèles caractérisés par leurs vecteurs de paramètres $\boldsymbol{\vartheta}_0$ et $\boldsymbol{\vartheta}_1$ tels que $k_0 = \dim(\boldsymbol{\vartheta}_0)$ et $k_1 = \dim(\boldsymbol{\vartheta}_1)$:
  - M_0 doit être *imbriqué* dans M_1 ($k_0 < k_1$).
- Le rapport des vraisemblances est donné par :

$$\Lambda = 2 \ln \left[\frac{L(\boldsymbol{\vartheta}_1)}{L(\boldsymbol{\vartheta}_0)} \right] = 2 [\ln L(\boldsymbol{\vartheta}_1) - \ln L(\boldsymbol{\vartheta}_0)]$$

  avec $L(\boldsymbol{\vartheta}_0)$ et $L(\boldsymbol{\vartheta}_1)$ les vraisemblances associés à M_0 et M_1.
- Pour le calcul du test proprement dit, on considère que $\Lambda \sim \chi^2(k_1 - k_0)$.

Arbre de décision du LRT



Akaike Information Criterion (AIC)

- Test AIC standard :

$$\mathrm{AIC} = -2 \ln L(\boldsymbol{\vartheta}) + 2k$$

avec $k = \dim(\boldsymbol{\vartheta})$ le nombre de paramètres du modèle.

- Test AICc, incluant une correction par la taille de l'échantillon :

$$\mathrm{AICc} = \mathrm{AIC} + \frac{2k(k+1)}{\ell - k - 1}$$

avec ℓ la longueur de l'alignement.

- Dans les deux cas, sélection du modèle présentant la plus faible valeur au test.

Bayesian Information Criterion (BIC)

- Test BIC standard :

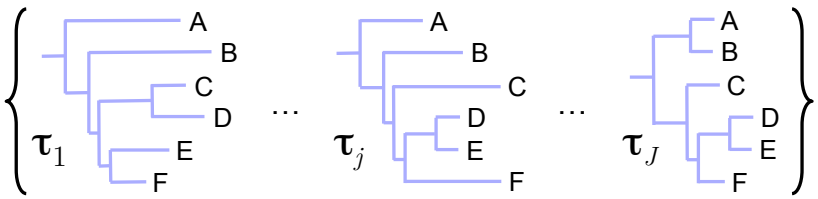
$$\mathrm{BIC} = -2 \ln L(\boldsymbol{\vartheta}) + k \ln \ell$$

- Comme dans le cas de l'AIC, sélection du modèle présentant la plus faible valeur au test.
- Approximation du test de comparaison de modèles utilisant les Facteurs de Bayes (*cf.* cours sur l'inférence bayésienne) :

$$2 \ln \mathrm{BF}_{10} \approx \mathrm{BIC}_1 - \mathrm{BIC}_0$$

Nécessité

- Différents jeux de données peuvent retourner différents arbres.
- Différentes méthodes peuvent retourner différents arbres.
- Une même méthode peut retourner différents arbres.
- Les différences observées sont-elles significatives ?



Utilisation de tests de vraisemblance

Tests courants

- Kishino et Hasegawa (KH – Kishino et Hasegawa, 1989).
- Shimodaira et Hasegawa (SH – Shimodaira et Hasegawa, 1999).
- *Expected Likelihood Weight* (ELW – Strimmer et Rambaut, 2001).
- *Approximately Unbiased* (AU – Shimodaira, 2002).

Test de Kishino et Hasegawa

- Soit S un alignement de séquences de longueur ℓ et $L(\boldsymbol{\theta}_1)$ et $L(\boldsymbol{\theta}_2)$ les vraisemblances de deux arbres obtenus à partir de S.
- On pose $Y_1 = \ln L(\boldsymbol{\theta}_1)$ et $Y_2 = \ln L(\boldsymbol{\theta}_2)$ et $\Delta = Y_1 - Y_2$.
- Le test KH consiste à tester si Δ est significativement différent de zéro, ce qui revient à la formulation :

$$H_0 : \mathbb{E}(\Delta) = 0$$
$$H_1 : \mathbb{E}(\Delta) \neq 0$$

- Le problème est que la distribution de Δ n'est pas connue :
  - Estimation de de la variance de Δ au moyen de différentes méthodes.

Approche classique (I)

- Soit $y_1^{(i)} = \ln L^{(i)}(\boldsymbol{\theta}_1)$ et $y_2^{(i)} = \ln L^{(i)}(\boldsymbol{\theta}_2)$, dans ce cas les valeurs de Y_1 et Y_2 sont telles que :

$$Y_1 = \sum_{i=1}^{\ell} y_1^{(i)} \quad \text{et} \quad Y_2 = \sum_{i=1}^{\ell} y_2^{(i)}$$

- Soit $\delta^{(i)} = y_1^{(i)} - y_2^{(i)}$, la différence des valeurs de vraisemblance par site, dans ce cas :

$$\Delta = Y_1 - Y_2 = \sum_{i=1}^{\ell} \delta^{(i)}$$

Approche classique (II)

- La moyenne des différences des valeurs de vraisemblances est donc égale à :

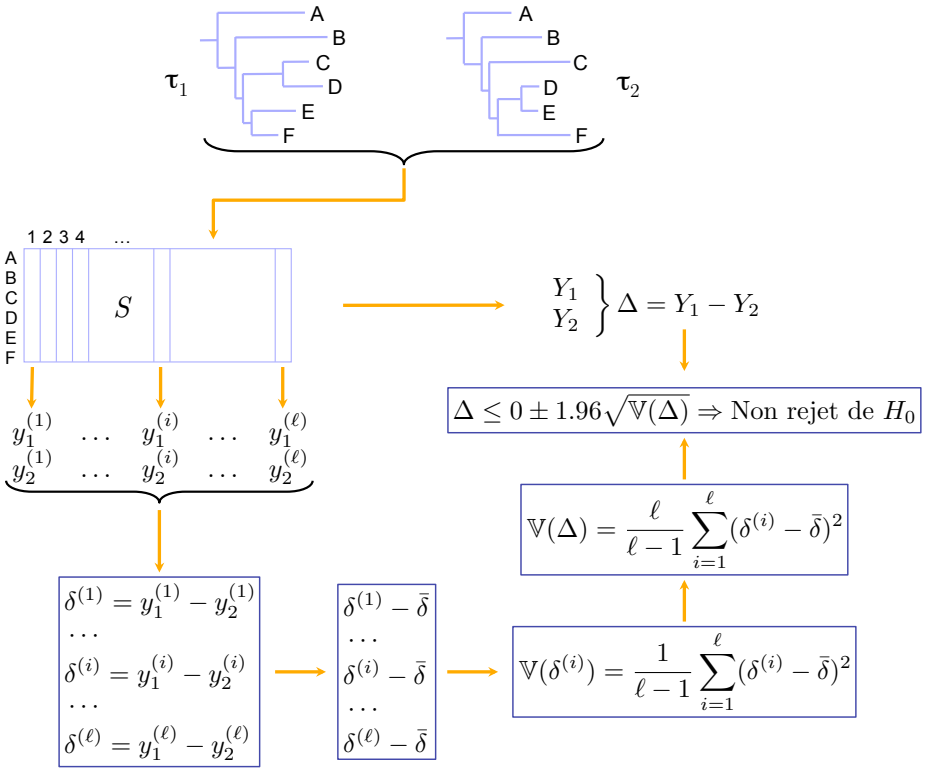
$$\bar{\delta} = \frac{1}{\ell} \sum_{i=1}^{\ell} \delta^{(i)} = \frac{\Delta}{\ell}$$

- Estimation de la variance de Δ par :

$$\mathbb{V}(\Delta) = \mathbb{V}\left(\delta^{(i)}\right) = \frac{1}{\ell - 1} \sum_{i=1}^{\ell} \left(\delta^{(i)} - \bar{\delta}\right)^2$$

- Utilisation de cette estimation pour réaliser un test bilatéral sous l'hypothèse que $\Delta \sim \mathcal{N}(0, \mathbb{V}(\Delta))$.

Schéma général



Approche par *bootstrap* (I)

- Réalisation de B rééchantillonages des sites de S par une approche de type *bootstrap*.
- Calcul, pour chaque réplicat k ($1 \leq k \leq B$), des vraisemblances *approchées* $Y'_{1(k)}$ et $Y'_{2(k)}$ associées aux topologies τ_1 et τ_2 :
  - Utilisation des valeurs de vraisemblances par sites provenant de S pour effectuer ce calcul.
- Calcul pour chaque réplicat de $\Delta'_{(k)} = Y'_{1(k)} - Y'_{2(k)}$.
- La moyenne des valeurs de $\Delta'_{(k)}$ est telle que :

$$\bar{\Delta}' = \frac{1}{B}\sum_{k=1}^{B}\Delta'_{(k)}$$

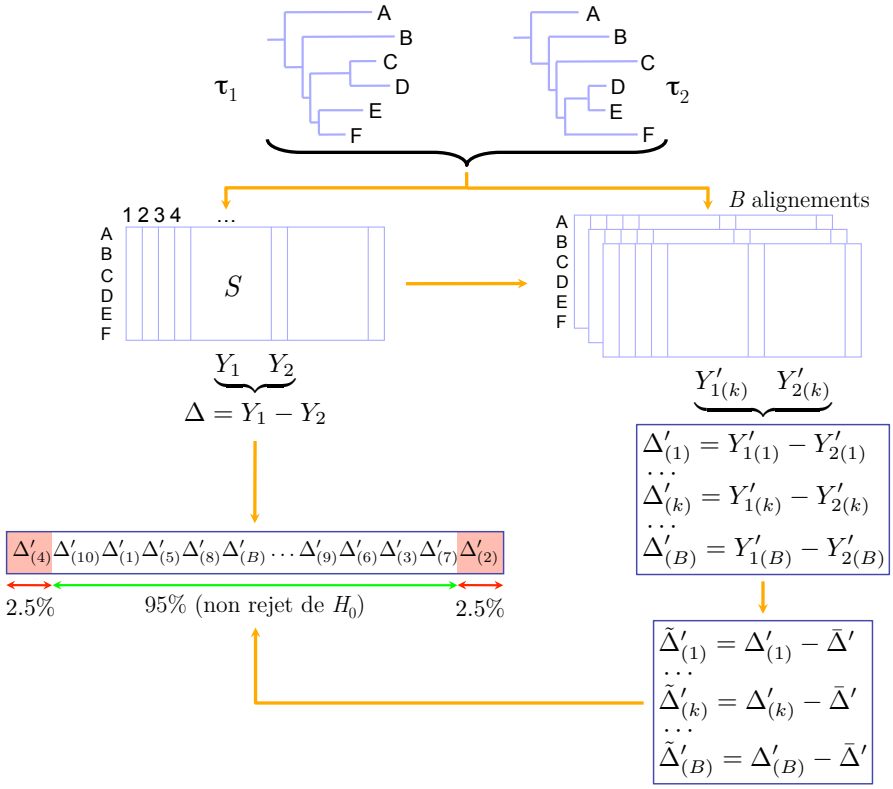
Approche par *bootstrap* (II)

- Calcul des valeurs de $\Delta'_{(k)}$ centrées par la moyenne :

$$\tilde{\Delta}'_{(k)} = \Delta'_{(k)} - \bar{\Delta}'$$

- Estimation de la variance de Δ par celle de $\tilde{\Delta}'_{(k)}$.
- Utilisation de cette variance pour réaliser un test bilatéral sous l'hypothèse que $\Delta \sim \mathcal{N}\left(0, \mathbb{V}\left(\tilde{\Delta}'_{(k)}\right)\right)$.
- Une autre possibilité est la comparaison directe de Δ avec la distribution des $\tilde{\Delta}'_{(k)}$.

Schéma général

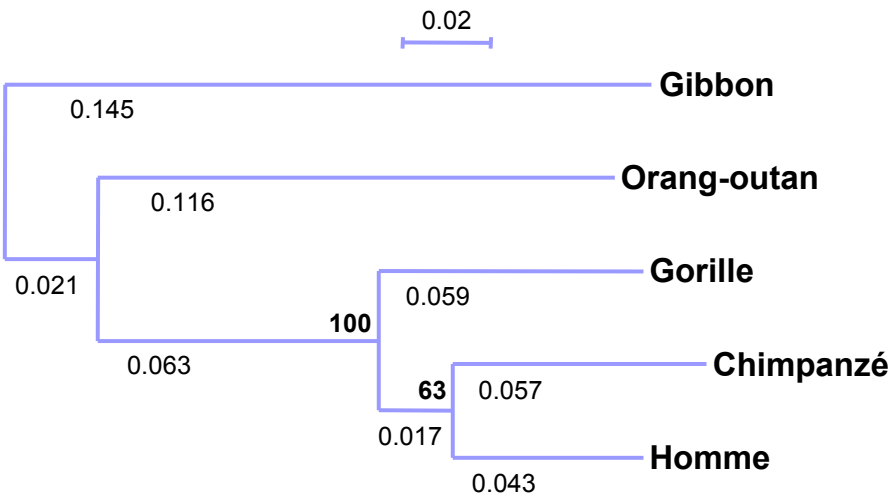


Limitations

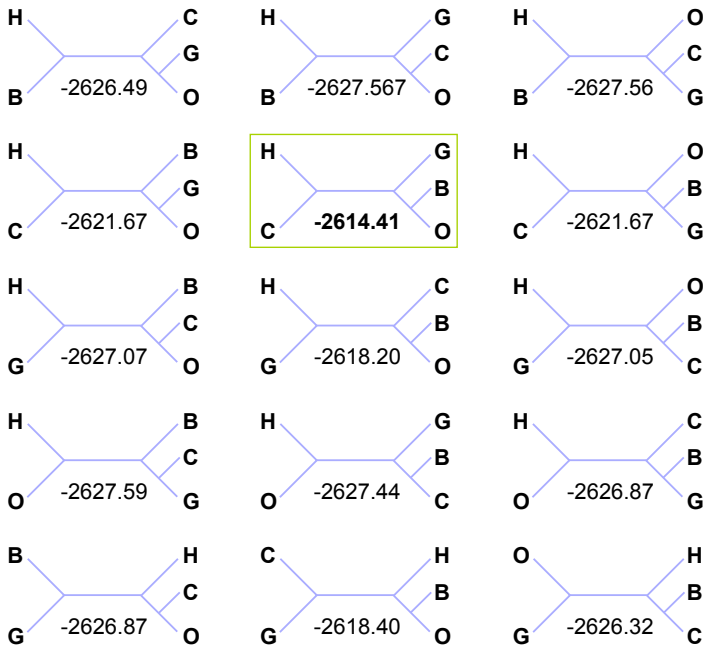
- Test limité à la comparaison de deux topologies :
  - Pas de correction pour les tests multiples.
- Les arbres testés doivent êtres choisis *indépendamment* des données utilisées pour réaliser le test :
  - Indispensable pour justifier l'hypothèse nulle sous laquelle $\mathbb{E}(\Delta) = 0$.
  - Le choix ne peut donc pas se faire sur la base de la vraisemblance.
- A malheureusement été fréquemment utilisé en violation de ces deux conditions !
- Les autres méthodes (SH, AU, ELW) utilisent un principe similaire mais corrigent ces défauts.

Phylogénie des Hominoïdes

- Sélection du modèle HKY+Γ après un test BIC.
- Racinement avec la séquence du Gibbon.
- 500 réplicats de *bootstrap*.



Vraisemblances des topologies



B = Gibbon, H = Homme, C = Chimpanzé, G = Gorille, O = Orang-outan

Comparaison des topologies

j	τ_j	Y_j	Δ	KH	SH	ELW	AU
1	((H,B),(G,O),C)	−2626.486	12.074	0.0050	0.0150	0.0013	0.0620
2	((H,B),(C,O),G)	−2627.563	13.150	0.0150	0.0190	0.0019	0.0100
3	((H,B),(C,G),O)	−2627.563	13.150	0.0150	0.0190	0.0019	0.0068
4	((H,C),(G,O),B)	−2621.668	7.256	0.0490	0.1560	0.0270	0.0414
5	((H,C),(B,O),G)	−2614.413	0.000	0.8390	1.0000	0.7224	0.9490
6	((H,C),(B,G),O)	−2621.668	7.256	0.0500	0.1570	0.0270	0.0399
7	((H,G),(C,O),B)	−2627.071	12.659	0.0220	0.0270	0.0040	0.0449
8	((H,G),(B,O),C)	−2618.205	3.793	0.1610	0.4250	0.1187	0.2531
9	((H,G),(B,C),O)	−2627.051	12.639	0.0220	0.0260	0.0043	0.0512
10	((H,O),(C,G),B)	−2627.590	13.177	0.0130	0.0160	0.0017	0.0193
11	((H,O),(B,C),G)	−2627.441	13.029	0.0170	0.0210	0.0025	0.0516
12	((H,O),(B,G),C)	−2626.874	12.461	0.0080	0.0140	0.0010	0.0174
13	((B,G),(C,O),H)	−2626.874	12.461	0.0080	0.0140	0.0010	0.0150
14	((C,G),(B,O),H)	−2618.401	3.989	0.1470	0.4090	0.0833	0.0536
15	((O,G),(B,C),H)	−2626.316	11.904	0.0070	0.0160	0.0019	0.0763

SH, ELW, AU : tests multiples ; KH : test simple entre τ_5 et chacune des topologies τ_j

Inférence bayésienne

Formation CNRS « Phylogénie moléculaire »

Guy Perrière

Laboratoire de Biométrie et Biologie Évolutive
UMR CNRS n° 5558
Université Claude Bernard – Lyon 1

20-23 mars 2018

Historique

- Théorème de Bayes établi au XVIII^{e} siècle :
  - Utilisation courante en probabilités.
- Introduction récente en phylogénie moléculaire :
  - Yang et Rannala (1996).
- Détermination analytique des probabilités postérieures fréquemment impossible :
  - Utilisation d'approximations numériques.

LII. *An Eſſay towards ſolving a Problem in the Doctrine of Chances. By the late Rev. Mr.* Bayes, *F. R. S. communicated by Mr.* Price, *in a Letter to* John Canton, *A. M. F. R. S.*

Dear Sir,

Read Dec. 23, 1763. I Now ſend you an eſſay which I have found among the papers of our deceaſed friend Mr. Bayes, and which, in my opinion, has great merit, and well deſerves to be preſerved. Experimental philoſophy, you will find, is nearly intereſted in the ſubject of it; and on this account there ſeems to be particular reaſon for thinking that a communication of it to the Royal Society cannot be improper.

He had, you know, the honour of being a member of that illuſtrious Society, and was much eſteemed by many in it as a very able mathematician. In an introduction which he has writ to this Eſſay, he ſays, that his deſign at firſt in thinking on the ſubject of it was, to find out a method by which we might judge concerning the probability that an event has to happen, in given circumſtances, upon ſuppoſition that we know nothing concerning it but that, under the ſame circum-

Théorème de Bayes

- Une définition classique des probabilités conditionnelles est que :

$$\mathbb{P}(A \cap B) = \mathbb{P}(B)\mathbb{P}(A|B) = \mathbb{P}(A)\mathbb{P}(B|A)$$

- En divisant les deux termes de l'équation précédente par $\mathbb{P}(B)$ on obtient la formulation la plus simple du théorème de Bayes, soit :

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(A)\mathbb{P}(B|A)}{\mathbb{P}(B)}$$

avec :

  - $\mathbb{P}(A|B)$, la probabilité *a posteriori* (ou postérieure) de A sachant B.
  - $\mathbb{P}(A)$, la probabilité *a priori* de A.
  - $\mathbb{P}(B|A)$, la *vraisemblance* de A.
  - $\mathbb{P}(B)$, la probabilité *marginale* de B ou *constante de normalisation*.

Généralisation

- Étant donné que :

$$\mathbb{P}(B) = \mathbb{P}(A \cap B) + \mathbb{P}(\bar{A} \cap B) = \mathbb{P}(A)\mathbb{P}(B|A) + \mathbb{P}(\bar{A})\mathbb{P}(B|\bar{A})$$

on en déduit :

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(A)\mathbb{P}(B|A)}{\mathbb{P}(A)\mathbb{P}(B|A) + \mathbb{P}(\bar{A})\mathbb{P}(B|\bar{A})}$$

- Ce qui peut se généraliser sous la forme :

$$\mathbb{P}(A_i|B) = \frac{\mathbb{P}(A_i)\mathbb{P}(B|A_i)}{\sum_j \mathbb{P}(A_j)\mathbb{P}(B|A_j)}$$

pour tout élément du s.c.e. $\{A_i\}$, avec i un des éléments de l'ensemble des valeurs possibles de j.

Un exemple classique

- Quelle est la probabilité d'avoir des *faux positifs* lors d'un test de diagnostic ?
- Soit un test de dépistage d'une maladie quelconque :
  - Si un patient a contracté la maladie, le test est positif dans 99% des cas.
  - Si un patient est sain, le test est négatif dans 95% des cas.
  - On estime que la fréquence de la maladie dans la population est de 1‰.
- Quelle est la probabilité qu'un individu testé positif soit effectivement atteint ?

Résolution

- Dans cet exemple, la probabilité *a priori* est égale à la fréquence de la maladie dans la population, soit $\mathbb{P}(A) = 0.001$:
  - On en déduit $\mathbb{P}(\bar{A}) = 1 - \mathbb{P}(A) = 0.999$.
- Par ailleurs, la probabilité que le test soit positif si le patient est malade est $\mathbb{P}(B|A) = 0.99$.
- Enfin, la probabilité que le test soit négatif si le patient est sain est $\mathbb{P}(\bar{B}|\bar{A}) = 0.95$:
  - On en déduit $\mathbb{P}(B|\bar{A}) = 1 - \mathbb{P}(\bar{B}|\bar{A}) = 0.05$.
- On en déduit la probabilité $\mathbb{P}(A|B)$ qu'un individu soit malade si le test est positif :

$$\mathbb{P}(A|B) = \frac{0.001 \times 0.99}{0.001 \times 0.99 + 0.999 \times 0.05} \simeq 0.019$$

Remarques sur le résultat

- Bien que le test précédent soit apparemment précis, la probabilité d'avoir des faux positifs est très importante (98.1%) :
  - Problème lié au fait que la probabilité *a priori* est faible.
  - Cas fréquent pour les tests de diagnostic :
    - Utilisation de plusieurs tests réalisés de façon séquentielle.
- Dans cet exemple, détermination de l'*a priori* à partir de la fréquence de la pathologie dans la population :
  - L'utilisation du théorème de Bayes ne souffre pas de discussion.
- Dans de nombreux cas, les probabilités *a priori* ne peuvent pas être facilement estimées :
  - Utilisation de valeurs représentant l'appréciation *subjective* de la personne effectuant l'analyse.

Notation en statistiques

- En statistiques, le s.c.e. $\{A_i\}$ correspond à un ensemble d'hypothèses, alors que B correspond aux données observées.
- Dans ce cas, écriture du théorème de Bayes sous la forme :

$$\mathbb{P}(H_i|D) = \frac{\mathbb{P}(H_i)\mathbb{P}(D|H_i)}{\sum_j \mathbb{P}(H_j)\mathbb{P}(D|H_j)}$$

  avec $\mathbb{P}(H_i|D)$, la probabilité conditionnelle d'une hypothèse H_i sous les données D.
- Les différentes hypothèses pouvant correspondre à différentes valeurs pour un paramètre θ, avec $H_1 : \theta = \theta_1$, $H_2 : \theta = \theta_2$, etc.
- Dans le cas où le modèle utilisé comprend plus d'un paramètre, θ correspond alors au vecteur $\boldsymbol{\theta}$ des dits paramètres.

Données continues

- Expression sous la forme de fonctions de densités quand les hypothèses concernent des paramètres *continus* :

$$f(\boldsymbol{\theta}|\mathbf{x}) = \frac{f(\boldsymbol{\theta})f(\mathbf{x}|\boldsymbol{\theta})}{f(\mathbf{x})} = \frac{f(\boldsymbol{\theta})f(\mathbf{x}|\boldsymbol{\theta})}{\int f(\mathbf{x}|\boldsymbol{\theta})f(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta}}$$

- La constante de normalisation $f(\mathbf{x})$ est obtenue en intégrant la vraisemblance sur la distribution *a priori* de $\boldsymbol{\theta}$:
  - Permet d'avoir $\int f(\boldsymbol{\theta}|\mathbf{x}) = 1$.
  - Si $\boldsymbol{\theta}$ correspond à un vecteur comprenant de nombreux paramètres :
    - Pas de solution analytique au calcul de cette intégrale.
    - Calcul de la probabilité postérieure au moyen d'approximations numériques telles que les *Chaînes de Markov avec technique de Monte-Carlo* (MCMC).

Interprétation des résultats

- Le résultat d'une analyse statistique bayésienne est représenté par la distribution des probabilités postérieures.
- Utilisation de valeurs ponctuelles pour faciliter l'interprétation :
  - Moyenne :

$$\mathbb{E}(\boldsymbol{\theta}|\mathbf{x}) = \int \boldsymbol{\theta} f(\boldsymbol{\theta}|\mathbf{x})\mathrm{d}\boldsymbol{\theta}$$

  - Médiane.
  - Maximum *a posteriori* :
    - Conceptuellement similaire au maximum de vraisemblance.
- Détermination d'un intervalle de *crédibilité* $[a, b]$ au seuil α tel que :

$$\int_a^b f(\boldsymbol{\theta}|\mathbf{x})\mathrm{d}\boldsymbol{\theta} = 1 - \alpha$$

Distributions *a priori*

- Conjuguées :
  - Un *a priori* est dit conjugué si $f(\boldsymbol{\theta})$ et $f(\boldsymbol{\theta}|\mathbf{x})$ appartiennent à la même famille de distributions.
  - Permettent de simplifier les calculs (pas de résolution d'intégrales complexes).
- Non informatives ou vagues :
  - $f(\boldsymbol{\theta})$ est non informative si son impact sur $f(\boldsymbol{\theta}|\mathbf{x})$ est faible :
    - Prédominance de la vraisemblance.
  - Utilisées quand aucune information préalable n'est disponible sur les variations du paramètre.
- Informatives :
  - $f(\boldsymbol{\theta})$ est informative si son impact sur $f(\boldsymbol{\theta}|\mathbf{x})$ est fort.
  - Cas de l'analyse bayésienne séquentielle :
    - *A posteriori* d'une étude précédente utilisé comme *a priori* pour l'étude courante.

Critiques de l'*a priori*

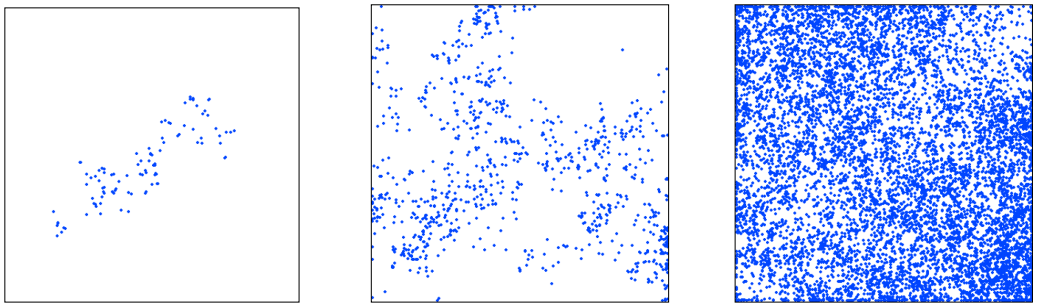
- Depuis le XVIII^{e} siècle, les critiques du bayésien portent essentiellement sur l'*a priori*.
- Résultats différents en fonction d'un *a priori* donné :
  - Rejet par les statisticiens « classiques » de la notion de probabilité subjective.
- Existence d'une école « objective » prônant l'utilisation d'*a priori* les moins informatifs possibles :
  - Distributions uniformes.
  - Loi *a priori* de Jeffreys (1961).
  - Loi de référence de Bernardo (1979).

Principe des MCMC

- En analyse bayésienne, impossibilité de déterminer la constante de normalisation si le nombre de paramètres est élevé :
  - Impossibilité de calculer directement la probabilité postérieure.
- Utilisation d'une chaîne de Markov suivant une marche guidée dans l'espace multidimensionnel des paramètres :
  - À la stationnarité, convergence vers les valeurs attendues des probabilités postérieures.

Analogie du randonneur

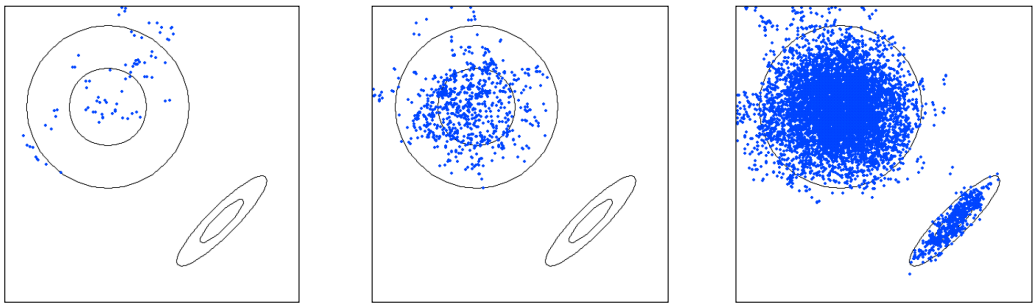
- Soit un randonneur se déplaçant sur une surface plane délimitée en faisant des pas de longueur variable :
  - Amplitude maximale fixée au préalable.
  - Chaque pas est effectué en choisissant aléatoirement une direction quelconque.
  - Rebond si un pas conduit à l'extérieur.
- Au bout d'un certain temps, exploration de l'intégralité de la surface :



100 pas 1000 pas 10000 pas

Exploration de reliefs

- Introduction de deux règles supplémentaires :
  - Si la direction prise par le randonneur le conduit vers une position plus élevée, il le fait toujours.
  - Si au contraire cette direction est descendante, possibilité de choix :
    - Calcul de $r = h^*/h$, avec h^* la hauteur atteinte en cas de descente et h la hauteur actuelle.
    - Tirage de $u \sim \mathcal{U}(0,1)$.
    - Si $u < r$, le randonneur descend, sinon il reste où il est.
- Visite préférentielle des points situés en altitude :



100 pas | 1000 pas | 10000 pas

Problèmes rencontrés

- Nécessité d'éliminer les premiers pas – qui constituent ce que l'on appelle communément la *zone d'approche* ou *burn-in* :
  - Démarrage du trajet en un point sélectionné aléatoirement, point pouvant être situé à une distance importante des reliefs.
- Évitement des maxima locaux :
  - Nécessité d'avoir un nombre de pas suffisamment élevé :
    - Pas toujours suffisant si les pics sont éloignés les uns des autres.
  - Lancement de plusieurs chaînes ayant des points de départ différents :
    - Poursuite de l'exploration jusqu'à convergence des résultats entre les différentes chaînes.

Algorithme de Metropolis-Hastings

1. Soit $\boldsymbol{\theta}_i$, le vecteur des paramètres caractérisant l'état de la chaîne de Markov au temps i.
2. Soit $\boldsymbol{\theta}^*$ le vecteur des paramètres caractérisant un état *candidat* pour constituer le maillon suivant de la chaîne.
3. Calcul de la *probabilité d'acceptation* r, telle que :

$$r = \min\left[1, \frac{f(\boldsymbol{\theta}^*|\mathbf{x})}{f(\boldsymbol{\theta}_i|\mathbf{x})}\right] = \min\left[1, \frac{f(\boldsymbol{\theta}^*)f(\mathbf{x}|\boldsymbol{\theta}^*)}{f(\boldsymbol{\theta}_i)f(\mathbf{x}|\boldsymbol{\theta}_i)}\right]$$

4. Si $r = 1$, alors $\boldsymbol{\theta}_{i+1} = \boldsymbol{\theta}^*$.
5. Si $r < 1$, tirage de $u \sim \mathcal{U}(0,1)$:
   - Si $u < r$ alors $\boldsymbol{\theta}_{i+1} = \boldsymbol{\theta}^*$, sinon $\boldsymbol{\theta}_{i+1} = \boldsymbol{\theta}_i$.
6. Retour à l'étape 1.

Caractéristiques

- Le calcul de r n'implique pas de connaître $f(\mathbf{x})$.
- Initialisation avec un ensemble de paramètres $\boldsymbol{\theta}$ choisis aléatoirement.
- La construction de $\boldsymbol{\theta}^*$ se fait en faisant varier de façon aléatoire les paramètres :
  - Utilisation d'algorithmes générant ce que l'on appelle des *propositions* :
    - Distributions uniformes de type $\mathcal{U}(-w/2, w/2)$, avec w l'amplitude maximale autorisée pour la variation des paramètres.
    - Distributions normales de type $\mathcal{N}(\mu, \sigma^2)$.
- La séquence des états visités forme une chaîne de Markov :
  - Estimation de la probabilité postérieure par la fréquence à laquelle les états sont visités une fois la stationnarité atteinte.

Fréquence d'acceptation

- Proportion du nombre de propositions acceptées dans la chaîne.
- Ne doit être ni trop grande ni trop petite.
- Valeurs optimales :
  - $\approx 50\%$ si $\boldsymbol{\theta}$ ne comprend qu'un seul paramètre.
  - $\approx 26\%$ si $\boldsymbol{\theta}$ comprend plusieurs paramètres.
- Valeurs recommandées :
  - 20-70% si $\boldsymbol{\theta}$ ne comprend qu'un seul paramètre.
  - 15-40% si $\boldsymbol{\theta}$ comprend plusieurs paramètres.

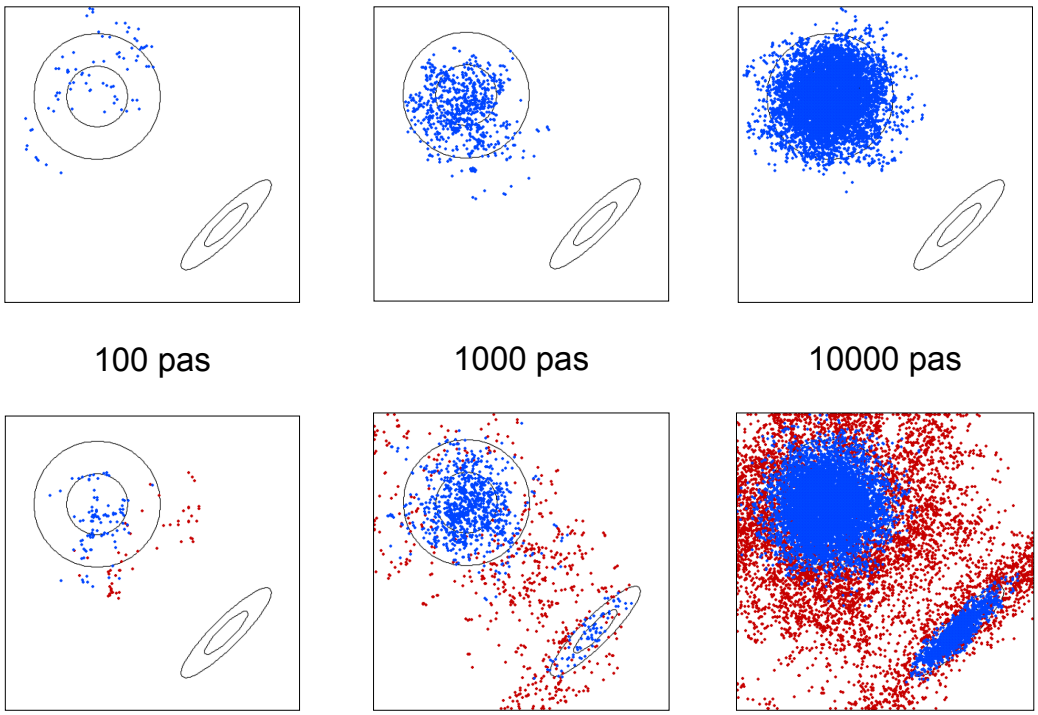
Couplage de Metropolis des MCMC

- Piégeage possible de la chaîne en cas de maximum local.
- Utilisation de plusieurs chaînes au lieu d'une :
  - Couplage de Metropolis des MCMC (MCMCMC ou MC^3).
  - Parmi toutes les chaînes lancées seules les chaînes dites « froides » (faible amplitude des pas) ont besoin de converger :
    - Utilisation de chaînes « chaudes » pour permettre une exploration plus vaste de l'espace des paramètres.
  - Tests à intervalles réguliers pour faire passer une chaîne froide dans une région explorée par une des chaînes chaudes :

$$r = \min\left[1, \frac{\pi_i(\boldsymbol{\theta}_j)\pi_j(\boldsymbol{\theta}_i)}{\pi_i(\boldsymbol{\theta}_i)\pi_j(\boldsymbol{\theta}_j)}\right]$$

    où i et j correspondent aux états de deux chaînes de Markov pour lesquelles la possibilité d'échange est testée.

Application au problème du randonneur



Détermination de la convergence

- Quand faut-il interrompre une MCMC ?
  - A-t-on atteint la distribution stationnaire de la chaîne ?
- Outils disponibles :
  - Inspection visuelle du graphe montrant les déplacements dans l'espace des paramètres.
  - Étude de la variation des valeurs de vraisemblance :
    - Pas de tendances particulières attendues à la stationnarité.
  - Mesure de l'autocorrélation des valeurs successives des paramètres :
    - Absence d'autocorrélation si convergence.
  - Tests statistiques :
    - Test de Gelman et Rubin (1992), ou *Potential Scale Reduction Factor* (PSRF) dans MrBayes.

Probabilité *a priori*

- Estimation par approche bayésienne de la distance évolutive entre deux séquences d'ADN sous le modèle de Jukes et Cantor.
- Calcul de la probabilité *a priori* :
  - Choix d'une distribution exponentielle :

$$f(d) = \frac{1}{\mu} \mathrm{e}^{-d/\mu}$$

    avec μ la moyenne de cette distribution et d la distance évolutive :
    - La probabilité d'obtenir des distances importantes tend rapidement vers 0.
  - D'autres choix sont possibles :
    - Distribution uniforme.

Vraisemblance

- Le calcul de la distance évolutive entre deux séquences au moyen du modèle de Jukes et Cantor est donnée par la formule :

$$d = -\frac{3}{4} \ln \left(1 - \frac{4}{3} p\right) \Leftrightarrow p = \frac{3}{4} - \frac{3}{4} \mathrm{e}^{-4d/3}$$

- Soit ℓ le nombre de sites dans l'alignement et n le nombre de sites pour lesquels il y a une substitution entre les deux séquences :
  - Dans ce cas, la fonction de vraisemblance pour d est donnée par la distribution binomiale $\mathcal{B}(\ell, p)$ telle que :

$$\begin{aligned} L(d) = f(p|d) &= \binom{\ell}{n} p^n (1-p)^{\ell-n} \\ &= \frac{\ell!}{n!(\ell-n)!} \left(\frac{3}{4} - \frac{3}{4}\mathrm{e}^{-4d/3}\right)^n \left(\frac{1}{4} + \frac{3}{4}\mathrm{e}^{-4d/3}\right)^{\ell-n} \end{aligned}$$

Probabilité postérieure

- Probabilité postérieure, sans la constante de normalisation :

$$f(d|p) \propto f(d)f(p|d)$$
$$\propto \frac{1}{\mu}\mathrm{e}^{-d/\mu}\left(\frac{3}{4}-\frac{3}{4}\mathrm{e}^{-4d/3}\right)^{n}\left(\frac{1}{4}+\frac{3}{4}\mathrm{e}^{-4d/3}\right)^{\ell-n}$$

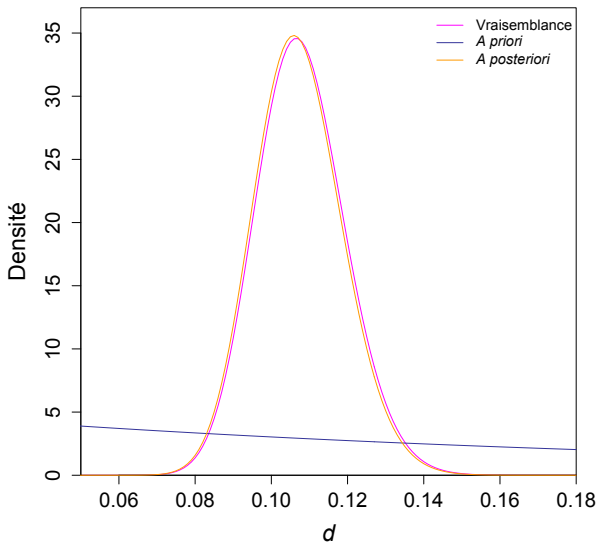
  - Le coefficient binomial étant lui aussi une constante, il peut être omis de cette expression.
- Valeur de la constante de normalisation donnée par :

$$f(p) = \int_0^\infty f(d)f(p|d)\mathrm{d}d$$

  - Solution analytique ou intégration numérique.

Application numérique

- Paire Homme-Gorille du jeu de données de Brown *et al.* (1982) :
  - $\ell = 896$
  - $n = 89$
- Moyenne de la distribution *a priori* fixée à $\mu = 0.2$.
- Estimation au maximum de vraisemblance :
  - $d \simeq 0.1066$
- Estimation bayésienne via la moyenne :
  - $\mathbb{E}(d|p) \simeq 0.1072$



Approximation par MCMC

- Calcul de la probabilité d'acceptation :

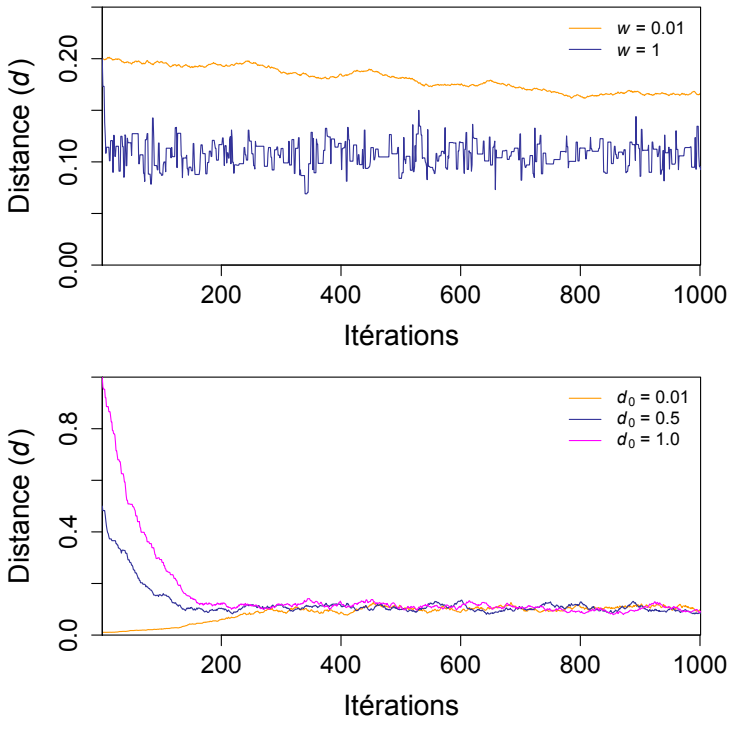
$$r = \min\left[1, \frac{f(d^*)f(p|d^*)}{f(d_i)f(p|d_i)}\right]$$

- Choix des propositions pour d :
  - Distribution uniforme, centrée sur la valeur actuelle et ayant une largeur égale à w :

$$d^* = |d_i + u|, \text{ avec } u \sim \mathcal{U}(-w/2, w/2)$$

- Choix de différentes valeurs pour l'amplitude (w) et la distance (d_0) utilisées pour initialiser la chaîne de Markov :
  - Valeurs variables pour w (0.01 et 1) et valeur fixe pour d_0 (0.2).
  - Valeur fixe pour w (0.1) et valeurs variables pour d_0 (0.01, 0.5 et 1).

Convergence des chaînes



Estimations de la distance

- Paramètres choisis : $\mu = 0.2$, $w = 0.1$ et $d_0 = 0.5$.
- Élimination de la zone d'approche (400 premières itérations).
- Échantillonnage de 1000 itérations prélevées à intervalles réguliers dans une chaîne :
  - Utilisation de la moyenne des valeurs pour l'estimation.
- Estimations obtenues après :
  - 1400 itérations : $\mathbb{E}(d|p) = 0.1073 \pm 5.18 \times 10^{-4}$
  - 10000 itérations : $\mathbb{E}(d|p) = 0.1072 \pm 7.03 \times 10^{-4}$
  - 100000 itérations : $\mathbb{E}(d|p) = 0.1071 \pm 7.23 \times 10^{-4}$

  avec, dans chaque cas, un intervalle de crédibilité à 95%.
- Variations stochastiques autour de la valeur obtenue par calcul direct.

Notations pour la phylogénie

- En phylogénie moléculaire, les données sont représentées par un ensemble de séquences alignées S.
- Par ailleurs, le vecteur des paramètres est $\boldsymbol{\theta} = (\tau, \mathbf{b}, \boldsymbol{\vartheta}, \alpha)$, avec :
  - τ la topologie de l'arbre.
  - $\mathbf{b}$ le vecteur des longueurs de branches.
  - $\boldsymbol{\vartheta}$ le vecteur des paramètres du modèle d'évolution utilisé.
  - α le paramètre de forme de la loi Gamma, le cas échéant.
- Le formule permettant de déterminer la probabilité postérieure est donc égale à :

$$f(\tau, \mathbf{b}, \boldsymbol{\vartheta}, \alpha|S) = \frac{f(\tau, \mathbf{b}, \boldsymbol{\vartheta}, \alpha) f(S|\tau, \mathbf{b}, \boldsymbol{\vartheta}, \alpha)}{f(S)}$$

avec :

$$f(S) = \sum_{\tau} \int_{\mathbf{b}} \int_{\boldsymbol{\vartheta}} \int_{\alpha} f(S|\tau, \mathbf{b}, \boldsymbol{\vartheta}, \alpha) f(\mathbf{b}) f(\boldsymbol{\vartheta}) f(\alpha) \mathrm{d}\mathbf{b} \mathrm{d}\boldsymbol{\vartheta} \mathrm{d}\alpha$$

Choix possibles pour les *a priori*

- Topologies :
  - Distribution uniforme $\mathcal{U}(N)$.
- Longueurs des branches :
  - Distribution uniforme $\mathcal{U}(0, 10)$.
  - Distribution exponentielle $\mathcal{E}(0.1)$.
- Paramètres du modèle d'évolution :
  - Distributions de Dirichlet plates $\mathcal{D}(1, 1, 1, 1)$ pour les échangeabilités et les fréquences à l'équilibre.
- Paramètre α de la loi Gamma :
  - Distribution exponentielle $\mathcal{E}(1)$.

Facteur de Bayes

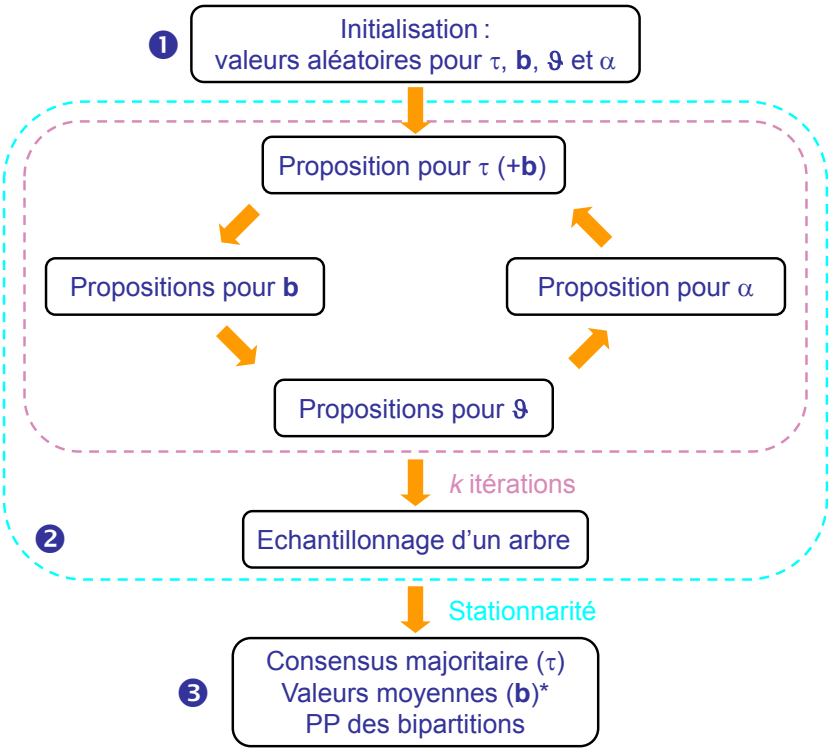
- Défini comme étant le rapport des *vraisemblances marginales* associées aux modèles M_0 et M_1 comparés, soit :

$$\mathrm{BF}_{10} = \frac{f(\mathbf{x}|M_1)}{f(\mathbf{x}|M_0)} = \frac{\int f(\boldsymbol{\vartheta}_1|M_1) f(\mathbf{x}|\boldsymbol{\vartheta}_1, M_1) \mathrm{d}\boldsymbol{\vartheta}_1}{\int f(\boldsymbol{\vartheta}_0|M_0) f(\mathbf{x}|\boldsymbol{\vartheta}_0, M_0) \mathrm{d}\boldsymbol{\vartheta}_0}$$

- Si $H_0 = M_0$, alors interprétation en utilisant l'échelle de Kass et Raftery (1995) :

log(BF)	BF	Évidence
< 0	< 1	Négative
$0 - 0.5$	$1 - 3.2$	Faible
$0.5 - 1$	$3.2 - 10$	Substancielle
$1 - 2$	$10 - 100$	Forte
> 2	> 100	Décisive

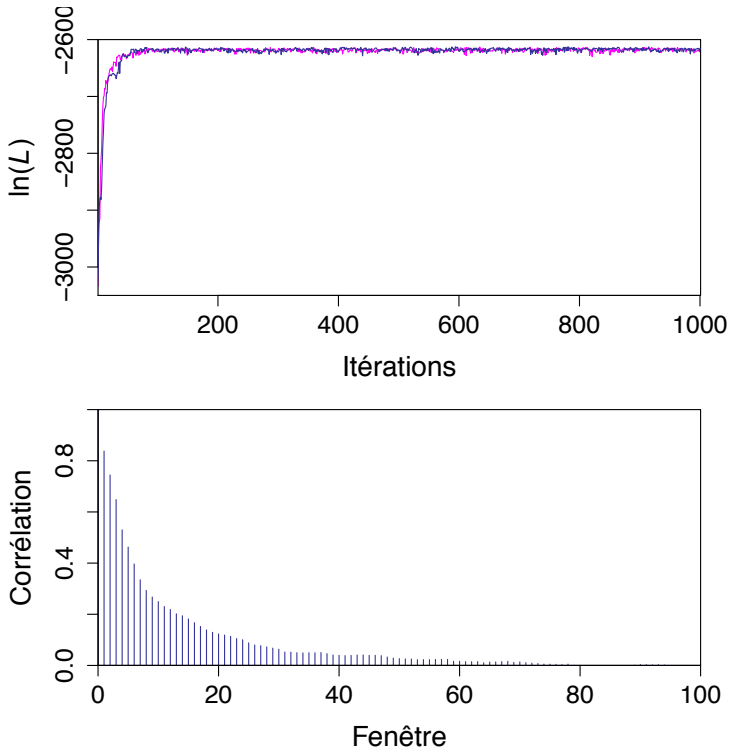
Procédure générale



Phylogénie des Hominoïdes

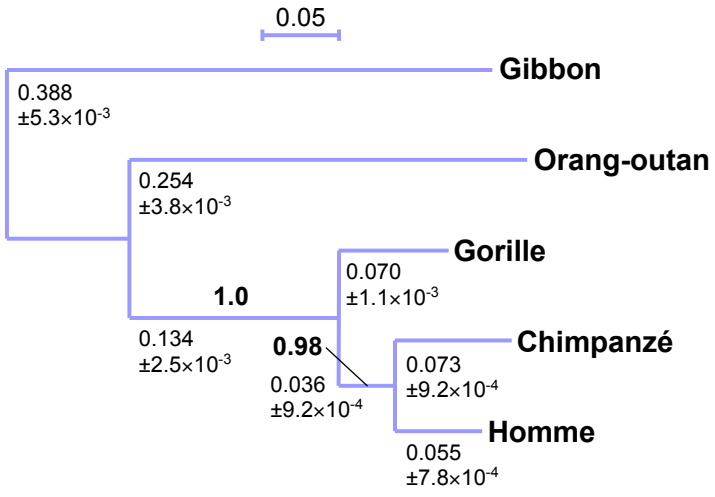
- Sélection du modèle HKY+Γ après un test BIC.
- Utilisation de MrBayes pour reconstruire la phylogénie :
  - Valeurs par défaut des probabilités *a priori*.
  - Deux chaînes froides partant de points de départ différents.
  - Trois chaînes chaudes lancées en parallèle de chaque chaîne froide.
  - Test de Gelman et Rubin pour déterminer si convergence.
  - Arrêt après 10000 itérations et fréquence d'échantillonnage de 1/10 :
    - Jeu de données de petite taille.

Convergence des chaînes



Arbre obtenu

- Construction par consensus majoritaire à 50% sur les itérations échantillonnées hors *burn-in*.
- Racinement avec la séquence du Gibbon.
- Longueurs des branches avec intervalles de crédibilité à 95%.



Avantages et limitations

- Méthode la mieux justifiée du point de vue théorique (si vous êtes bayésien).
- Meilleur comportement que le maximum de vraisemblance avec des modèles comprenant de nombreux paramètres.
- Temps de calcul biens plus longs :
  - Avec les MC^3, de nombreuses chaînes sont lancées en parallèle.
  - Nécessité d'atteindre la distribution stationnaire pour les chaînes froides.
- Pas de nécessité d'effectuer du rééchantillonnage de type *bootstrap* :
  - Utilisation des valeurs de probabilités postérieures des clades :
    - Valeurs directement interprétables en termes de probabilités.

Bootstrap et probabilités postérieures

- Construction de six phylogénies (Douady *et al.*, 2003) :
  - Vraisemblance et bayésien.
- Comparaison entre valeurs de *bootstrap* (BP) et :
  - Probabilités postérieures (PP) des clades.
  - *Bootstrap* des probabilités postérieures (BPP).
- Valeurs des PP systématiquement plus élevées que celles des BP.