Analyse de séquences et phylogénie moléculaire -

Maximum de parcimonie

Ecole doctorale E2M2 - 2015-2016

(http://www.frangun.org)

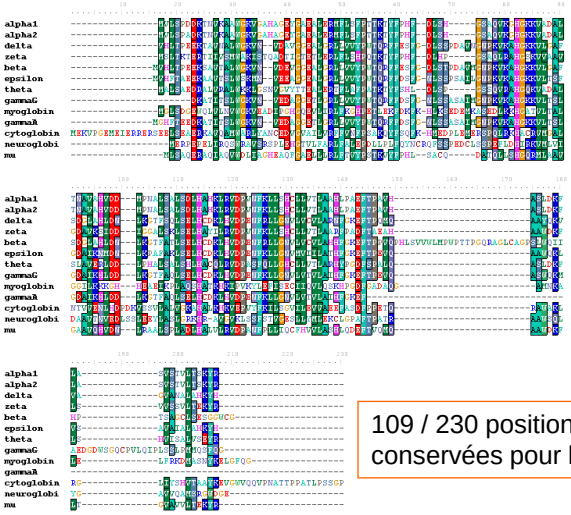
Céline Brochier (celine.brochier-armanet@univ-lyon1.fr)
Guy Perrière (guy.perriere@univ-lyon1.fr)

Méthodes de reconstruction phylogénétique

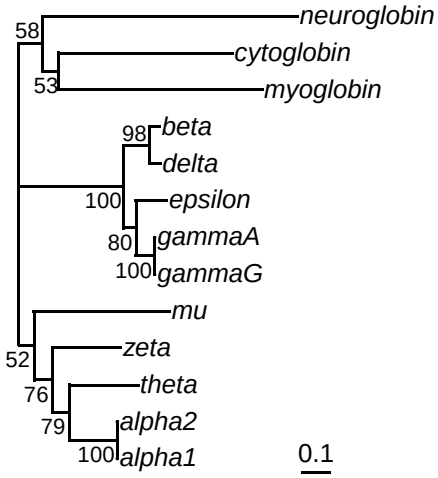
- Quatre grandes familles de méthodes
  - Parcimonie
  - Méthodes de distance
  - Maximum de vraisemblance
  - Méthodes bayésiennes

Données utilisées en phylogénie moléculaire

- Point de départ = alignement de séquences homologues
- Arrivée = arbre décrivant les liens évolutifs entre les séquences de l'alignement



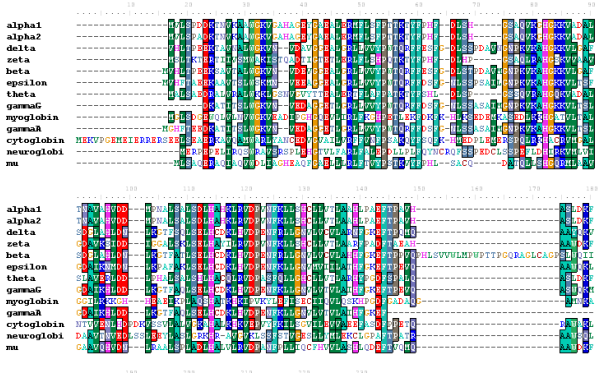
109 / 230 positions conservées pour l'analyse



(Alignement des 13 globines humaines réalisé avec clustalW (http://www.frangun.org/HSglobin_A.fasta), arbre construit avec Seaview (BioNJ, 100 réplicats de bootstrap))

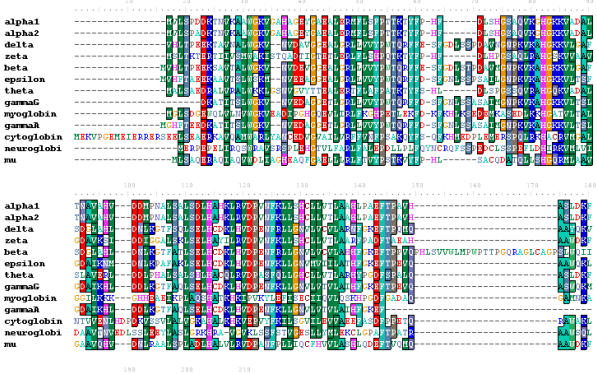
Alignements et gaps

- Chaque colonne de l'alignement représente une position (ou site) composée de résidus homologues, cad dérivant d'un même site ancêtre
- La qualité des alignements est essentielle
  - ⇒ Les régions où l'alignement est ambigu doivent être retirées (automatiquement ou manuellement) avant l'analyse phylogénique
- La plupart des méthodes de reconstruction ne prend en compte que les substitutions et non les événements d'insertions/délétions
  - ⇒ Les sites contenant des gaps sont ignorés



230

(ClustalW)



218

(Muscle)

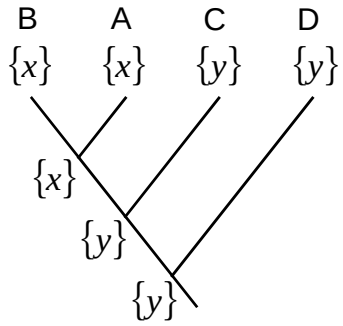
La parcimonie

- Fondement: rasoir d'Occam
  - « Les multiples ne doivent pas être utilisés sans nécessité. » (pluralitas non est ponenda sine necessitate) ou sous une forme plus moderne « les hypothèses les plus simples sont les plus vraisemblables »

Le critère de parcimonie

- Soit un caractère relevé dans 4 espèces {A,B,C,D} (dont on connaît la phylogénie) et présentant les états de caractères $\{x, x, y, y\}$

⇒ Quelle histoire a pu conduire à cet état final?

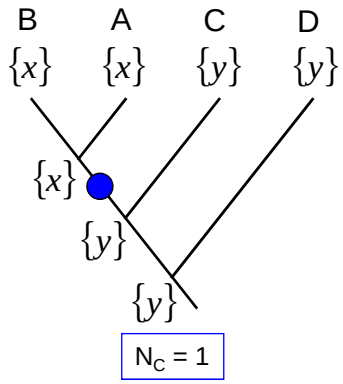


Le critère de parcimonie

- Soit un caractère relevé dans 4 espèces {A,B,C,D} (dont on connaît la phylogénie) et présentant les états de caractères $\{x, x, y, y\}$

⇒ Quelle histoire a pu conduire à cet état final?

Substitution y => x
Substitution x => y



$N_C = 1$

Similarité par ascendance commune

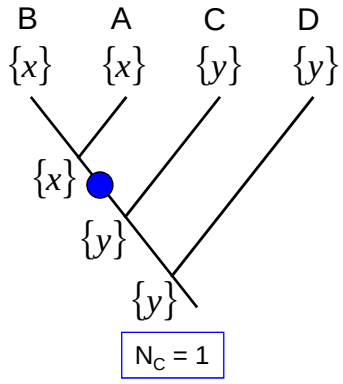
Analyse de séquences et phylogénie moléculaire (Céline Brochier-Armanet 2015-2016)

Le critère de parcimonie

- Soit un caractère relevé dans 4 espèces {A,B,C,D} (dont on connaît la phylogénie) et présentant les états de caractères $\{x, x, y, y\}$

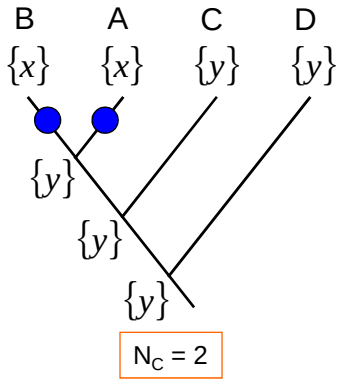
⇒ Quelle histoire a pu conduire à cet état final?

Substitution y => x
Substitution x => y



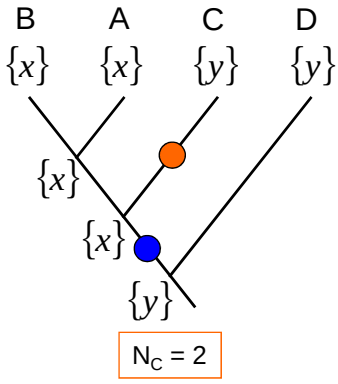
$N_C = 1$

Similarité par ascendance commune



$N_C = 2$

Similarité par convergence



$N_C = 2$

Similarité par réversion

Les scénarios homoplasiques demandent plus de changements évolutifs. L'emploi du critère de parcimonie en phylogénie moléculaire n'est justifié que si les convergences et les réversions sont rares.

Le maximum de parcimonie

- Principe: rechercher parmi l'espace des arbres définissant les liens entre n séquences la topologie qui minimise le nombre de changements évolutifs

⇒ Quelle est la topologie qui implique le moins de changements d'état de caractères pour rendre compte des différences observées entre les UTO étudiées

- Procédure:
  - 1) pour une topologie T fixée et pour un site donné de l'alignement, calculer (N_C) le nombre de changements évolutifs nécessaires pour expliquer les états de caractères observés
  - 2) calculer (N_C) pour chaque site de l'alignement => L , la longueur de l'arbre
  - 3) calculer L pour toutes les topologies T possibles => retenir l'arbre le plus parcimonieux (cad l'arbre le plus court)

Analyse de séquences et phylogénie moléculaire (Céline Brochier-Armanet 2015-2016)

Parcimonie: Etape 1

- Pour une topologie T fixée et pour un site donné de l'alignement, calculer (N_C) le nombre de changements évolutifs nécessaires pour expliquer les états de caractères observés

Algorithme de Fitch: calcul du nombre minimal de changements évolutifs

- Soit une topologie T fixée et racinée de manière arbitraire, soit V l'ensemble de ses nœuds
  - Pour tout $p \in V$ on définit:
    - C_p , le nombre minimal de changements dans le sous-arbre dont p est la racine
    - S_p , l'état de p , cad l'ensemble des résidus en p compatibles avec C_p changements évolutifs dans le sous-arbre raciné par p .
- Soit q et r les deux nœuds fils de p

Algorithme 1 Nombre de substitutions avec Fitch

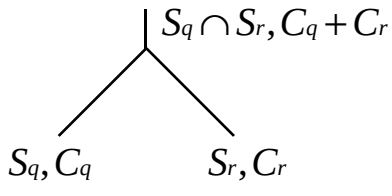
si $S_q \cap S_r \neq \emptyset$ **alors**
  $S_p \leftarrow S_q \cap S_r$
  $c_p \leftarrow c_q + c_r$
sinon
  $S_p \leftarrow S_q \cup S_r$
  $c_p \leftarrow c_q + c_r + 1$
fin si

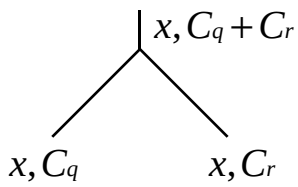
(Perrière & Brochier-Armanet, (2010) Concepts et méthodes en phylogénie moléculaire, Springer)

Analyse de séquences et phylogénie moléculaire (Céline Brochier-Armanet 2015-2016)

Algorithme de Fitch: Illustration

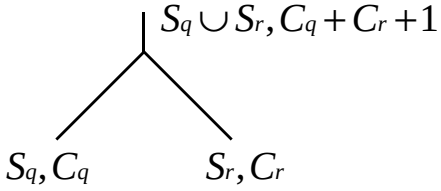
Cas n°1: $S_q \cap S_r \neq 0$

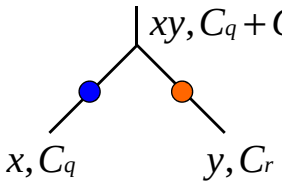




$\forall x \in S_q \cap S_r$

Cas n°2: $S_q \cap S_r = 0$





$\forall x \in S_q$

$\forall y \in S_r$

Le +1 reflète **le** changement évolutif qui a nécessairement eu lieu dans l'une des deux branches filles
soit x => y
Soit y => x

Algorithme de Fitch: Application

Algorithme 1 Nombre de substitutions avec Fitch

si $S_q \cap S_r \neq \emptyset$ **alors**
  $S_p \leftarrow S_q \cap S_r$
  $c_p \leftarrow c_q + c_r$
sinon
  $S_p \leftarrow S_q \cup S_r$
  $c_p \leftarrow c_q + c_r + 1$
fin si

Initialisation du calcul récursif aux feuilles de l'arbre
-P = {x} = résidu présent à cette feuille
-C_p = 0

La racine est placée de manière arbitraire et n'a aucune influence sur le nombre de changements évolutifs inférés

Les états de caractères inférés aux nœuds ne représentent pas des caractères ancestraux, ni tous les états de caractères possibles !

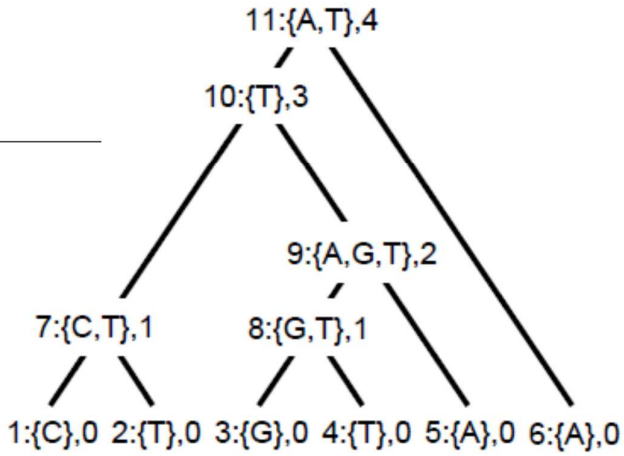
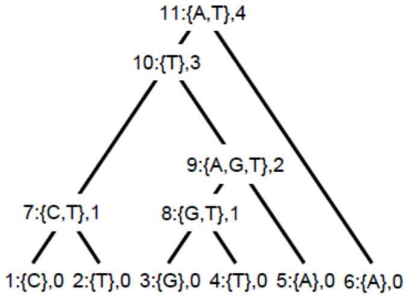


Figure 3.1 – Calcul du nombre de changements pour un site avec l'algorithme de Fitch. Les nœuds sont numérotés de 1 à 11 et les ensembles générés ainsi que le nombre de substitutions inférées sont indiqués. Dans cet exemple, le nombre minimum de changements est égal à 4.

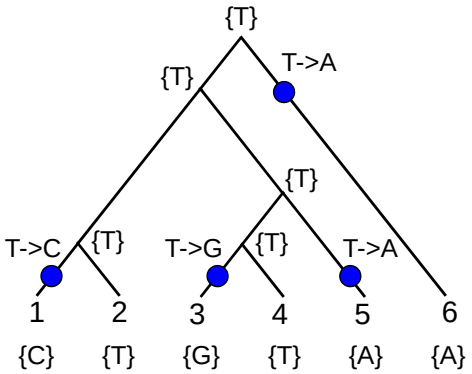
$N_C = 4$

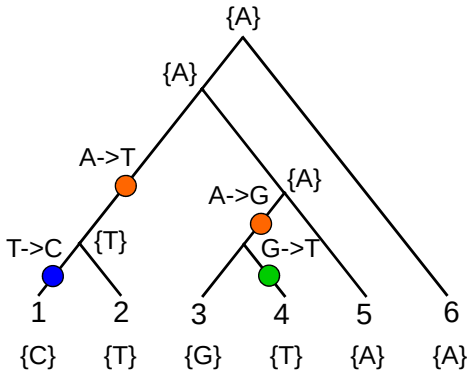
(Perrière & Brochier-Armanet, (2010) Concepts et méthodes en phylogénie moléculaire, Springer)

Des scénarios multiples



Il existe plusieurs scénarios impliquant N_C = 4 changements évolutifs





Parcimonie: Etapes 2 et 3

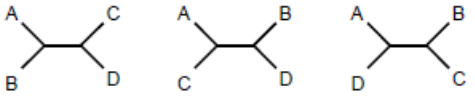
- Etape 2:
  - Calculer N_C pour chaque site de l'alignement
  - Sommer tous les valeurs de N_C pour l'ensemble des sites
  - Calculer L , la longueur totale de l'arbre
- Etape 3:
  - Répéter l'étape 2 pour chaque topologie T composant l'espace des arbres possibles à n feuilles
  - Retenir l'arbre de longueur L minimale ⇔ arbre le plus parcimonieux

Tous les sites ne sont pas équivalents

- Tous les sites ne contiennent pas une information permettant de discriminer les topologies
- Les sites constants (1 seul état de caractère)
  - Ne sont pas informatifs
- Sites variables (au moins 2 états de caractères)
  - Informatifs: présentent au moins deux états de caractères chacun partagés par au moins deux séquences
  - Non informatifs: tous les autres

Tous les sites ne sont pas équivalents

- Soit A, B, C et D quatre séquences d'ADN homologues alignées, il existe
  - 3 topologies non racinées possibles
  - 4 états de caractères {A,T,C,G}
  - 4^4 = 256 motifs possibles à une position
  - ⇒ 36 sont informatifs, 12 => {*x*,*x*,*y*,*y*}, 12 => {*x*,*y*,*x*,*y*} et 12=> {*x*,*y*,*y*,*x*} (avec $x \neq y$ et $x,y \in$ {A,T,C,G})
  - ⇒ Le nb de pas de chaque topo. dépend de la fréquence de ces 36 motifs.



ABCD			
AAAA	0	0	0
AAAC	1	1	1
AAAG	1	1	1
AAAT	1	1	1
AACA	1	1	1
AACC	**1**	**2**	**2**
AACG	2	2	2
AACT	2	2	2
AAGA	1	1	1
AAGC	2	2	2
AAGG	**1**	**2**	**2**
AAGT	2	2	2
AATA	1	1	1
AATC	2	2	2
AATG	2	2	2
AATT	**1**	**2**	**2**
ACAA	1	1	1
...	...	...	...
TTTT	0	0	0

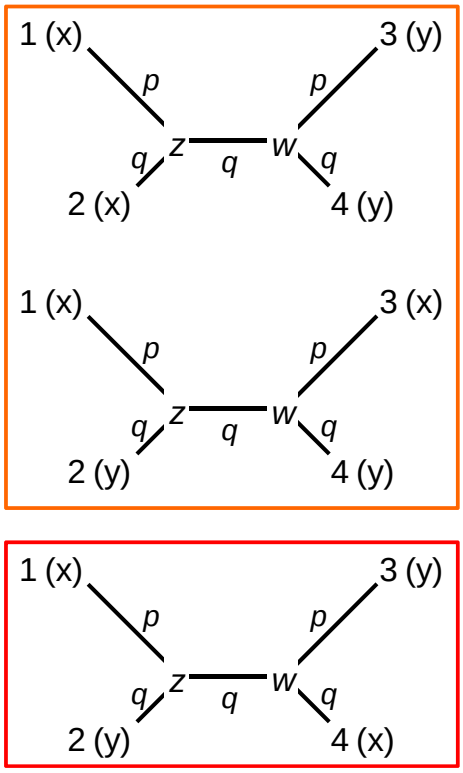
Figure 3.2 – Sous-ensemble des 256 motifs possibles pour un site dans le cas d'un arbre à quatre UTO (A, B, C et D) et mesure du nombre de substitutions inférées pour chacune des trois topologies non racinées possibles. En gras figurent les cas où ce nombre est différent entre plusieurs topologies.

Parcimonie et consistance

- Un estimateur est dit consistant si il converge vers la vraie valeur du paramètre avec une prob. = 1 quand L -> +∞.
- Soit T une topologie NR à 4 feuilles, avec des longueurs de branches et un modèle d'évolution sous-jacent (avec ici 2 prob. de p et q substitution).
  - ⇒ {xxyy} = motif en accord avec T
  - ⇒ Calcul de P(xxyy), P(xyxy) et P(xyyx) en fonction p et q.

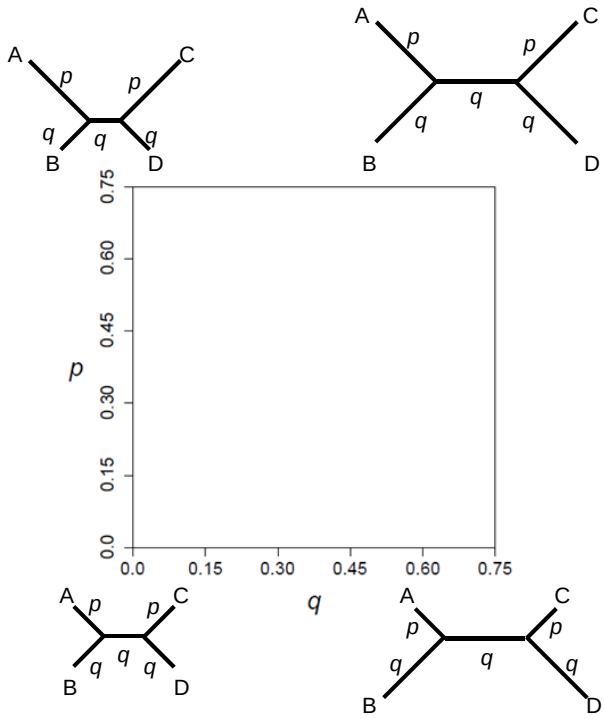
Condition de consistance => P(xxyy) > P(xyxy)

$$p < \frac{-18q + 24q^2 + \sqrt{243q - 567q^2 + 648q^3 - 288q^4}}{9 - 24q + 32q^2}$$



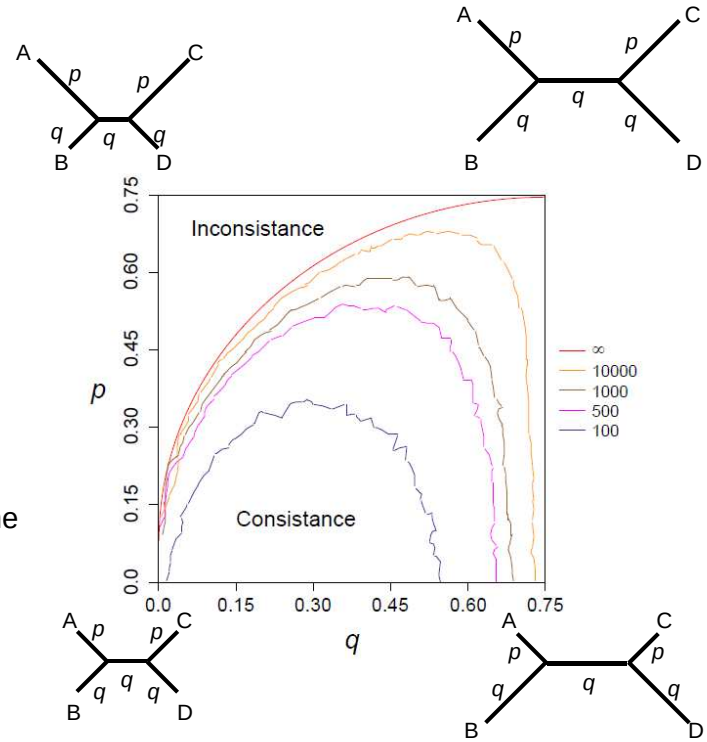
Parcimonie et inconsistance

- Mise en évidence initiale par Felsenstein (1978).
- Généralisation par Huelsenbeck (1995).
- Simulation à partir d'un arbre à quatre UTO :
- Topologie non racinée ((A,B),(C,D)).
- Deux longueurs de branches, p et q ∈ [0; 0:75].
- Longueurs des séquences L ∈ [100;+∞[

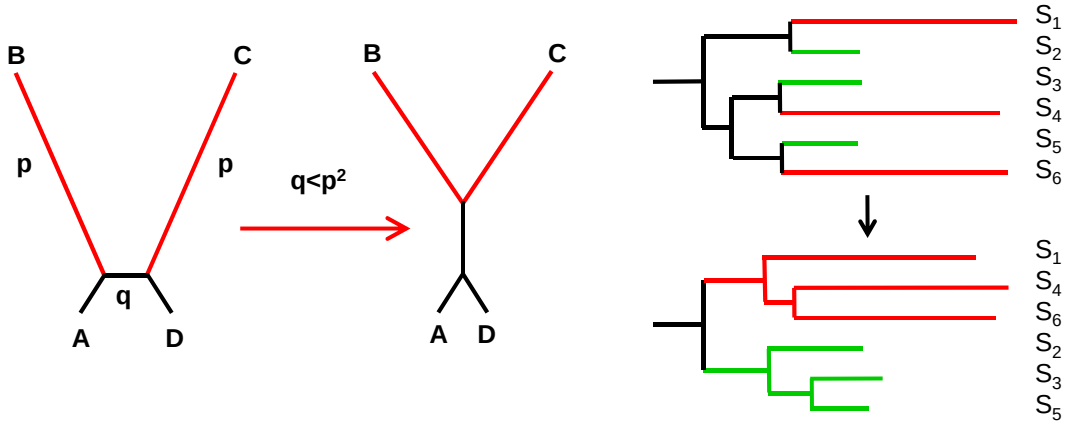


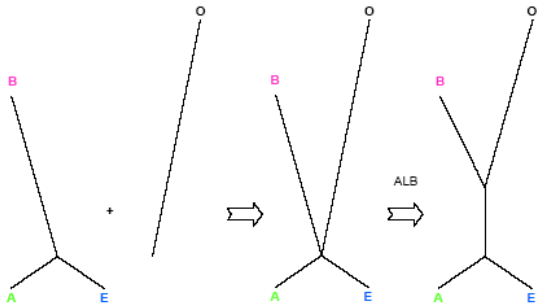
Zone de Felsenstein

- Si la condition de consistance est vérifiée
  - ⇒ P(xxyy) -> 1 (quand L -> +∞)
- Si la condition de consistance n'est pas vérifiée (p >> q)
  - ⇒ P(xxyy) < P(xyxy)
  - ⇒ Inconsistance
  - ⇒ la topologie inférée sera fausse.
- La zone d'inconsistance appelée zone de Felsenstein est d'autant plus grande que les séquences sont courtes.



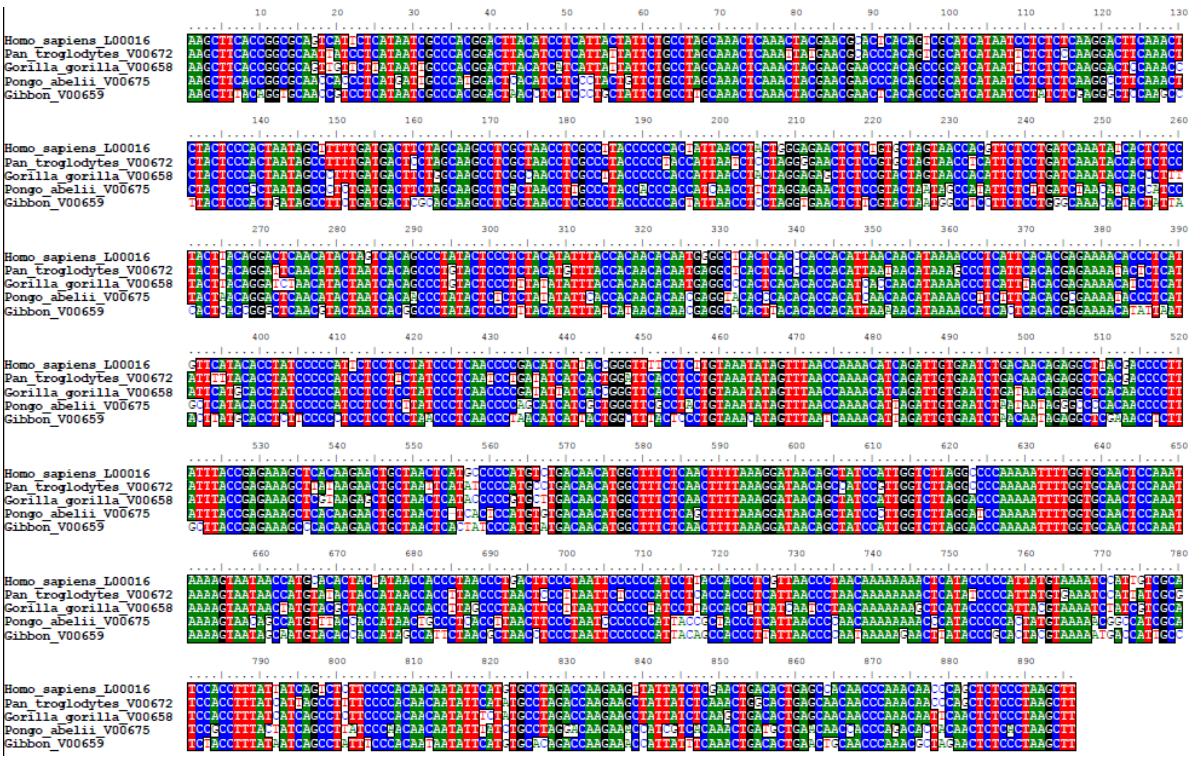
Attraction des longues branches





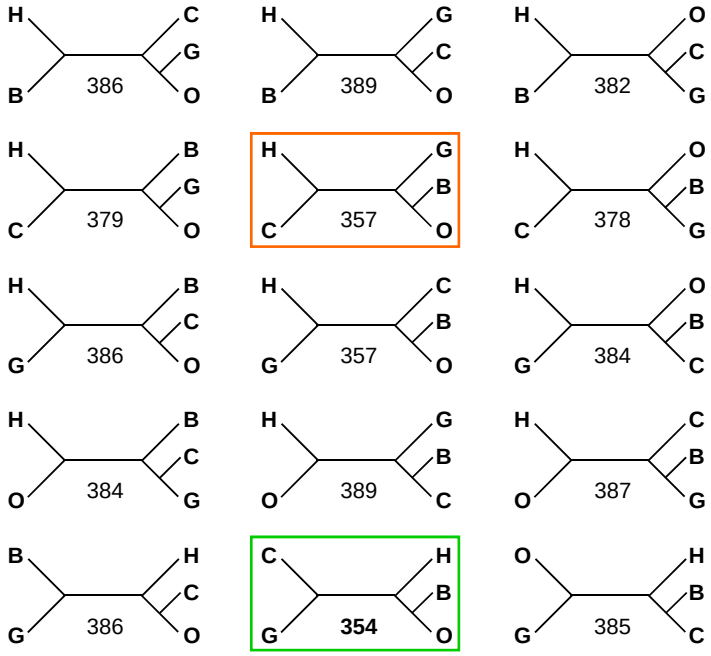
Analyse de séquences et phylogénie moléculaire (Céline Brochier-Armanet 2015-2016)

Application à la phylogénie des Hominoïdes



(Brown et al (1982) J Mol Evol) (Données disponibles sur http://www.frangun.org/sequences_A.fasta)

Application à la phylogénie des Hominoïdes



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

(Perrière & Brochier-Armanet, (2010) Concepts et méthodes en phylogénie moléculaire, Springer)

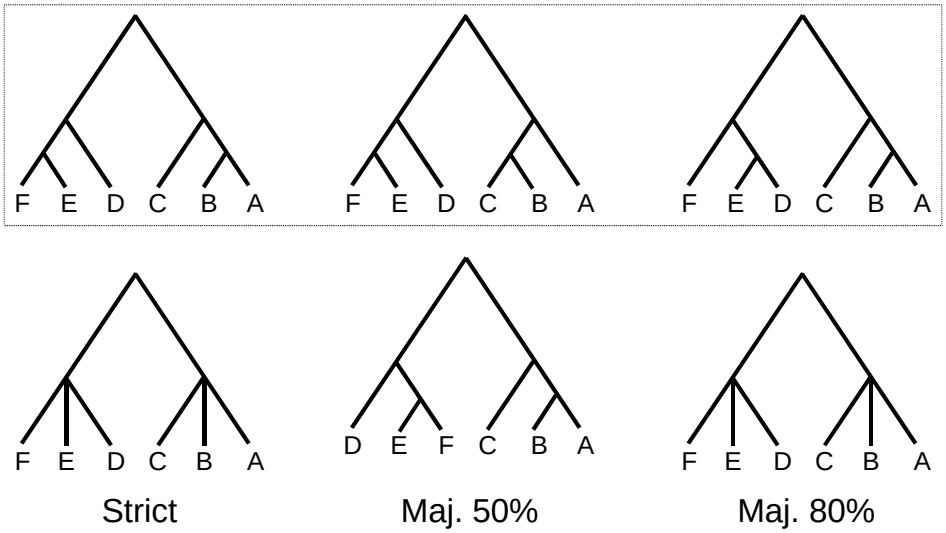
Analyse de séquences et phylogénie moléculaire (Céline Brochier-Armanet 2015-2016)

Parcimonie: Récapitulatif & propriétés

- Produit des arbres non racinés
- Le positionnement des changements dans un arbre n'est pas unique
  - ne permet pas d'inférer des longueurs de branches de manière unique
- Plusieurs arbres équiparcimonieux peuvent être trouvés
  - Inférence de consensus
- Le nombre d'arbre croissant de manière rapide avec le nombre de séquences, seule un sous-ensemble des topologies est testé pour identifier l'arbre le plus parcimonieux
  - Utilisation d'heuristiques pour explorer l'espace des arbres de manière rationnelle
  - Aucune certitude d'identifier l'arbre le plus parcimonieux à la fin de l'analyse
- Absence de critère pour discriminer le(les) arbre(s) le(s) plus parcimonieux des arbres légèrement moins parcimonieux
  - ex. est-ce qu'un arbre comptant 2504 pas est significativement meilleurs que les 20 arbres comptant 2502 pas ?
- La parcimonie classique (algorithme de Fitch) considère toutes les substitutions comme équivalentes
  - Parcimonie pondérée (algorithme de Sankoff) permet de pondérer les types de changements

Pour approfondir ces notions: Concepts et méthodes en phylogénie moléculaire (2010) Perrière & Brochier-Armanet (Springer)

Consensus d'arbres



(Perrière & Brochier-Armanet, (2010) Concepts et méthodes en phylogénie moléculaire, Springer)

Explorer l'espace des topologies

- n < 12: Exploration exhaustive
- n < 20: branch-and-bound
- n > 20: heuristiques

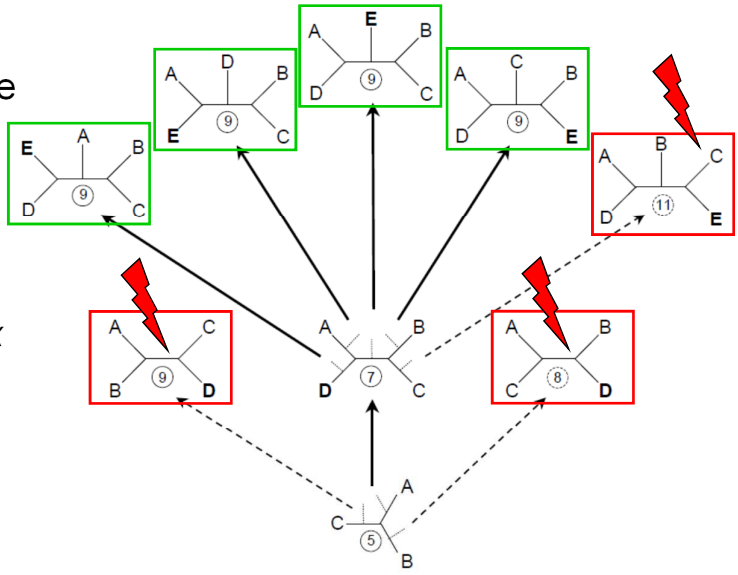
- Utilisé pour la parcimonie, mais aussi les moindres carrés, le maximum de vraisemblance, etc.

- Topologie de départ?
  - Topologie aléatoire
  - Meilleure topologie issue d'une recherche séquentielle

Recherche séquentielle

- Arbre à 3 feuilles
- Choix du 4ième taxon à ajoute
  - ordre des taxa dans l'alignement
  - aléatoirement
  - maximum du minimum (taxon qui induit un *Lmax* minimal)

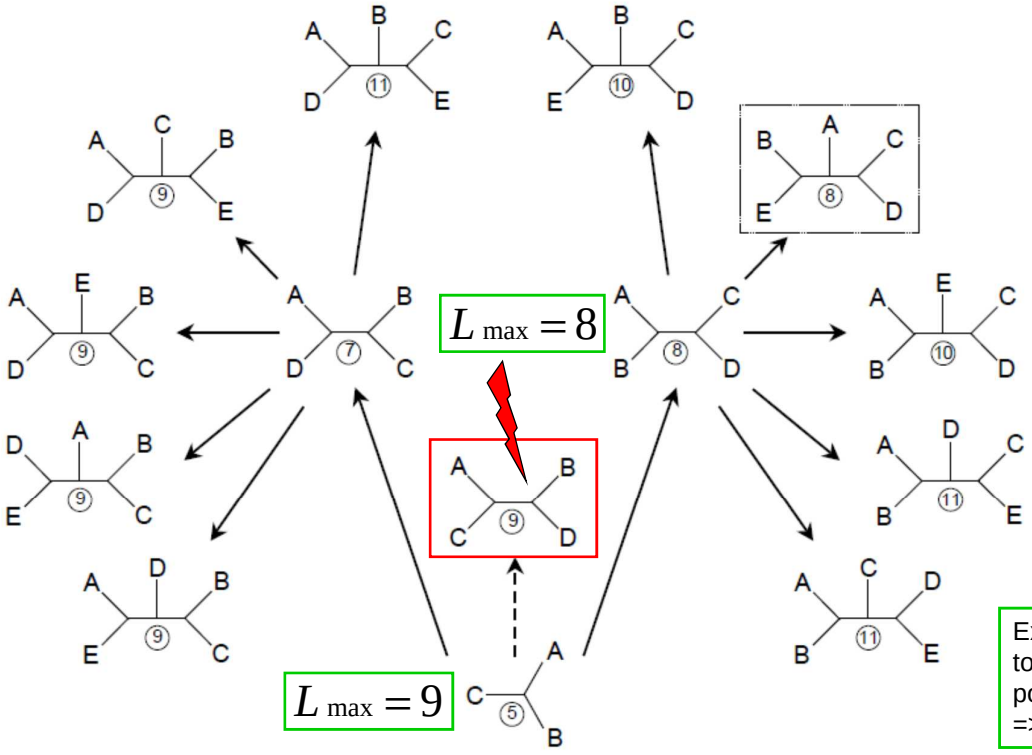
UTO	1	2	3	4	5	6
A	A	T	T	A	A	T
B	T	T	A	T	T	T
C	A	A	T	T	T	T
D	A	A	T	A	A	A
E	T	T	A	A	A	T



$L_{max} = 9$

(Perrière & Brochier-Armanet, (2010) Concepts et méthodes en phylogénie moléculaire, Springer)

Branch-and-bound



$L_{max} = 8$

$L_{max} = 9$

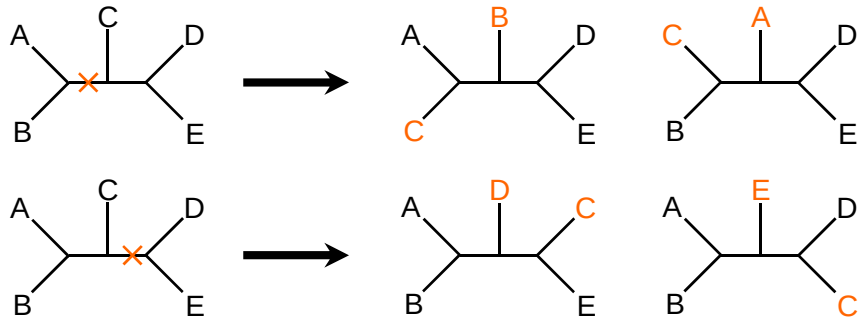
Examen de 10/15 topologies possibles
=> Gain de 1/3

(Perrière & Brochier-Armanet, (2010) Concepts et méthodes en phylogénie moléculaire, Springer)

Nearest Neighbor Interchange (NNI)

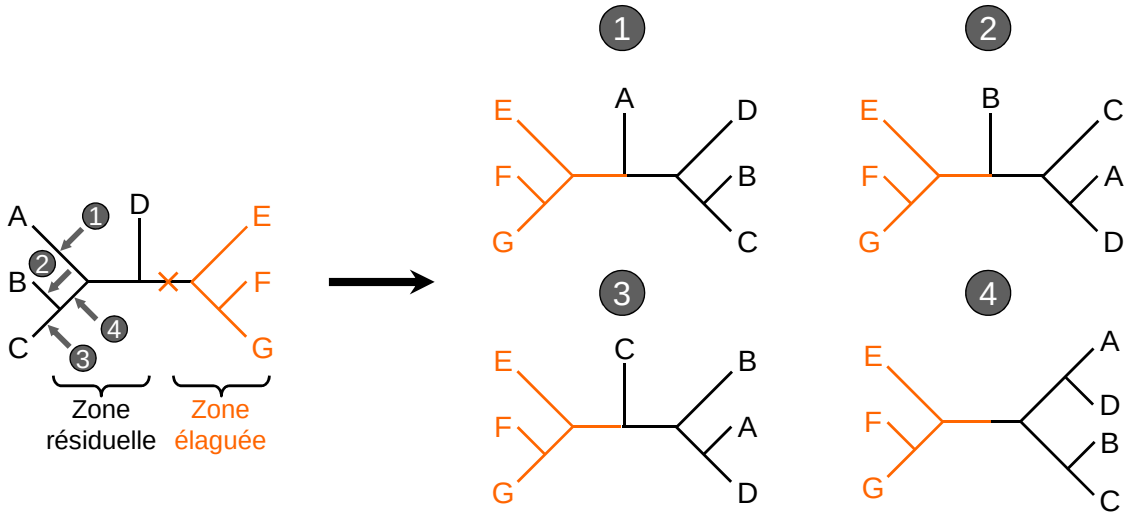
- Examen des topolgies se situant à une distance topologique $d_T = 2$ de l'arbre de départ
- $2(n-3)$ arbres situés à une distance topologie $d_T = 2$

Complexité en *O(n)*



(Perrière & Brochier-Armanet, (2010) Concepts et méthodes en phylogénie moléculaire, Springer)

Subtree pruning and regrafting (SPR)



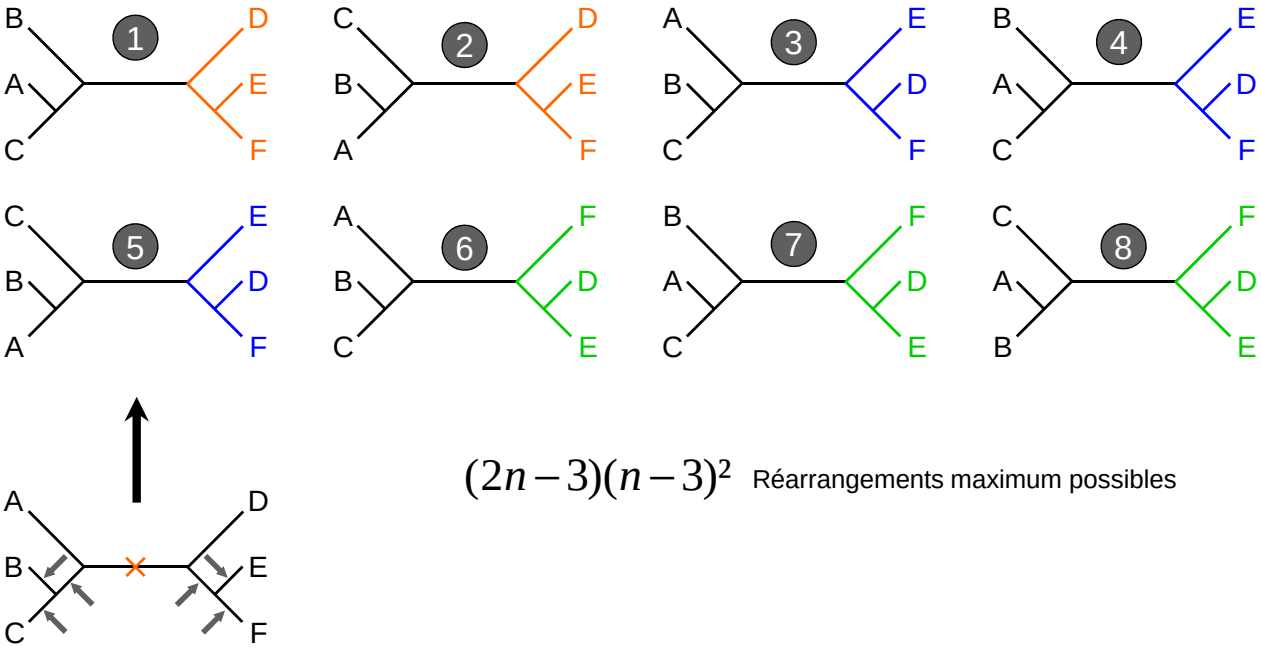
Si coupure au niveau d'une branche interne: (2n - 8) arbres voisins
Si coupure au niveau d'une branche externe: (2n - 6) arbres voisins
Un arbre non raciné compte: (n – 3) branches internes et n branches externes

⇒ Nombre de voisins explorables: $nx(2n-6)+(n-3)(2n-8)$
$= 4(n-3)(n-2)$

Complexité en $O(n^2)$

(Perrière & Brochier-Armanet, (2010) Concepts et méthodes en phylogénie moléculaire, Springer)

Tree Bisection and Reconnection (TBR)



$(2n-3)(n-3)^2$ Réarrangements maximum possibles

Complexité en $O(n^3)$

(Perrière & Brochier-Armanet, (2010) Concepts et méthodes en phylogénie moléculaire, Springer)

Analyse de séquences et phylogénie moléculaire (Céline Brochier-Armanet 2015-2016)

Lectures conseillées

- « Biologie évolutive » Thomas, Lefèvre, Raymond (2010) de boeck
- « Evolution » Barton, Briggs, Eisen, Goldstein, Patel (2007) Cold Spring Harbor Laboratory Press
- « Concepts et méthodes en phylogénie moléculaire » Perrière & Brochier-Armanet (2010) Springer collection IRIS