Phylogénie

Anne Lopes

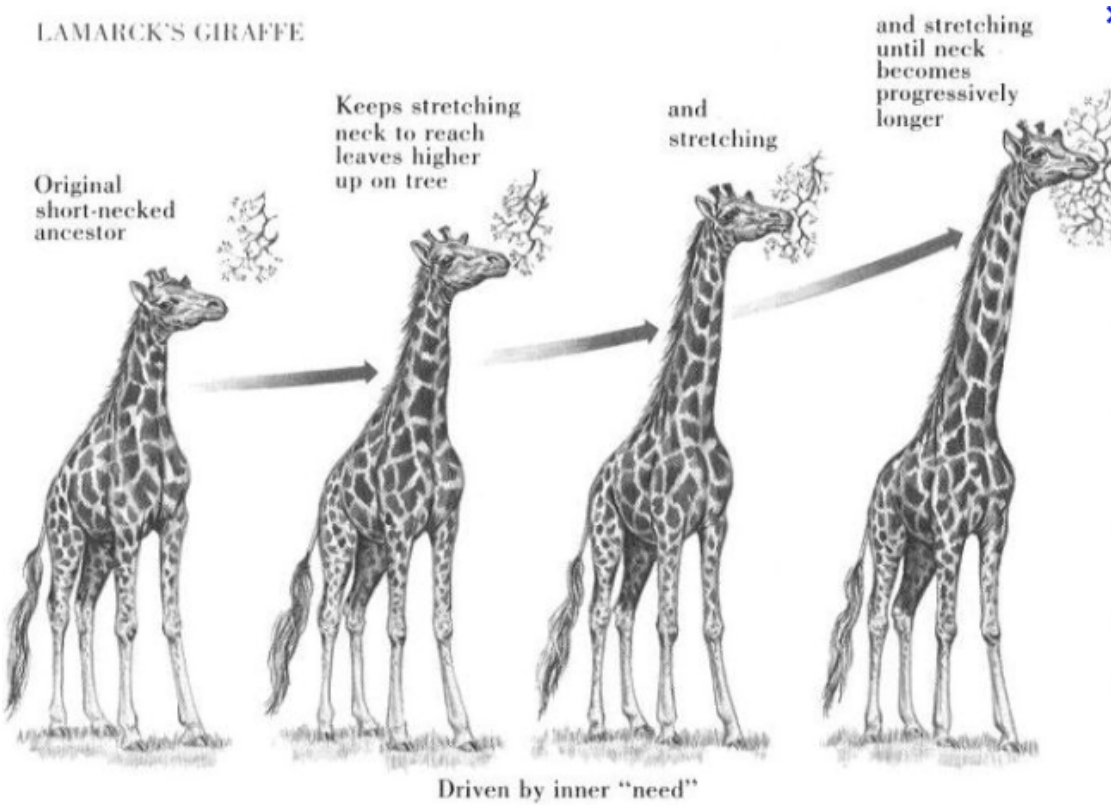
Analyse de Séquences Génomiques

Mars 2019

introduction

- étude de la diversité du vivant depuis l'antiquité : Lucrèce, Démocrite…

- Lamarck (1744-1829) : adaptation continue au milieu ambiant

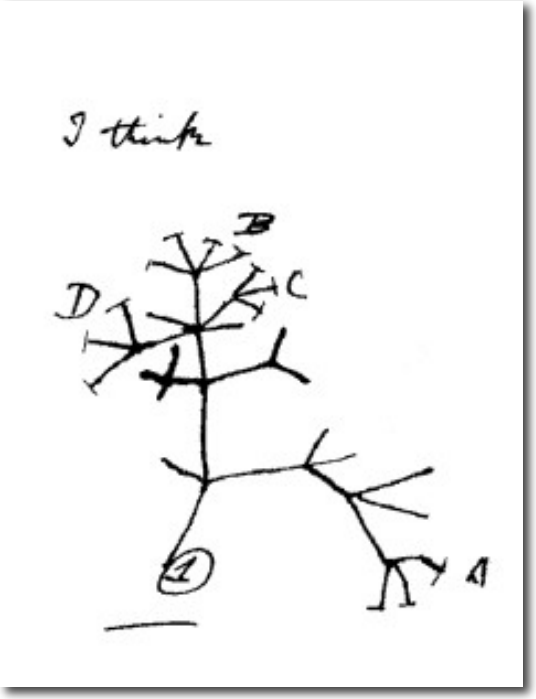


introduction

- étude de la diversité du vivant depuis l'antiquité : Lucrèce, Démocrite…

- Lamarck (1744-1829) : adaptation continue au milieu ambiant

- Darwin (1809-1882) : sélection naturelle



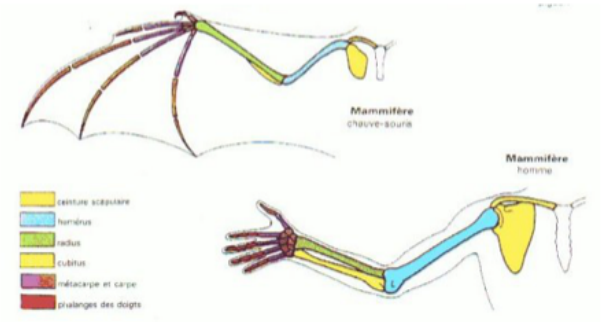
Biston betularia

Etant donné un environnement, seuls les individus capables de se reproduire vont transmettre leur matériel génétique. Les individus les plus adaptés seront sélectionnés par l'environnement.

introduction

• jusqu'aux années 60 :

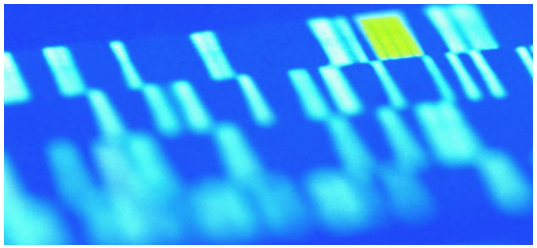
. comparaison de morphologies, comportements, répartition géographiques



• depuis les années 60 :

. 1962 (E. Zuckerkandl & L. Pauling) : les mutations s'accumulent dans un génome à vitesse proportionnelle au temps géologique

. utilisation de l'information contenue dans les séquences protéiques et nucléotidiques



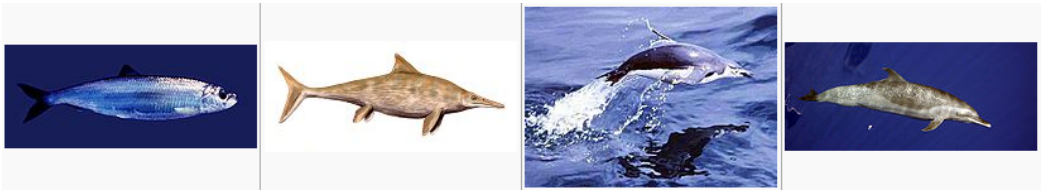
introduction

• jusqu'aux années 60 : comparaison de morphologies, comportements, répartition géographiques

• depuis les années 60 : utilisation de l'information contenue dans les séquences protéiques et nucléotidiques → naissance de la phylogénie moléculaire

• évolution convergente VS divergente :

. *convergence* : solutions trouvées de façon indépendante chez différents organismes pour résoudre le même problème



. *divergence* : protéines ayant le même ancêtre commun qui ont divergé

horloge moléculaire

- 1962 (E. Zuckerkandl & L. Pauling) :
  . les mutations s'accumulent dans un génome à vitesse proportionnelle au temps géologique

- amélioration du modèle :
  . taux d'accumulation de mutation entre organismes différents est du même ordre dans les régions homologues & toutes les régions ne sont pas soumises à la même horloge

trotteuse : pseudogènes, régions intergéniques, évènements récents (étude de sous-populations)

aiguille des minutes : taux de mutation moyen, ex: cytochrome C, passé proche

aiguille des heures : taux de mutation faible, ex: histones, polymérases, passé lointain

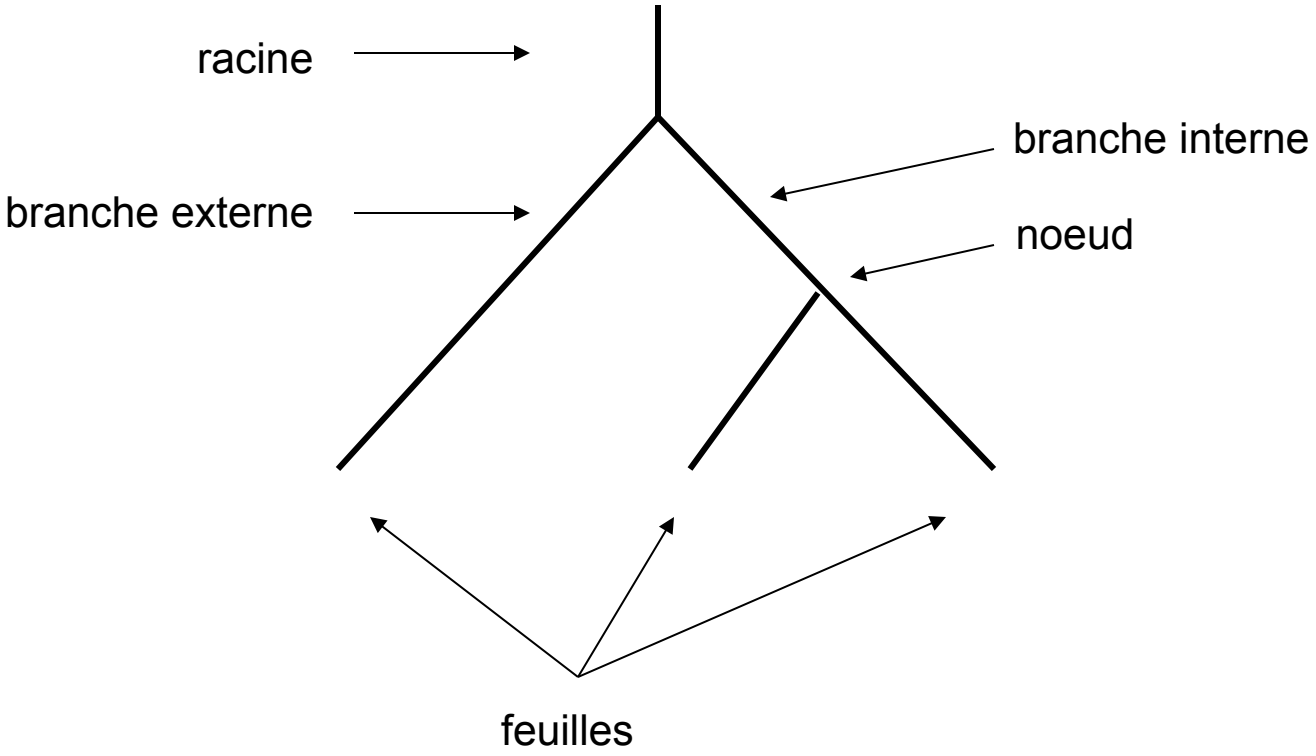
horloge moléculaire

- 1962 (E. Zuckerkandl & L. Pauling) :
  . les mutations s'accumulent dans un génome à vitesse proportionnelle au temps géologique

- amélioration du modèle :
  . taux d'accumulation de mutation entre organismes différents est du même ordre dans les régions homologues & toutes les régions ne sont pas soumises à la même horloge

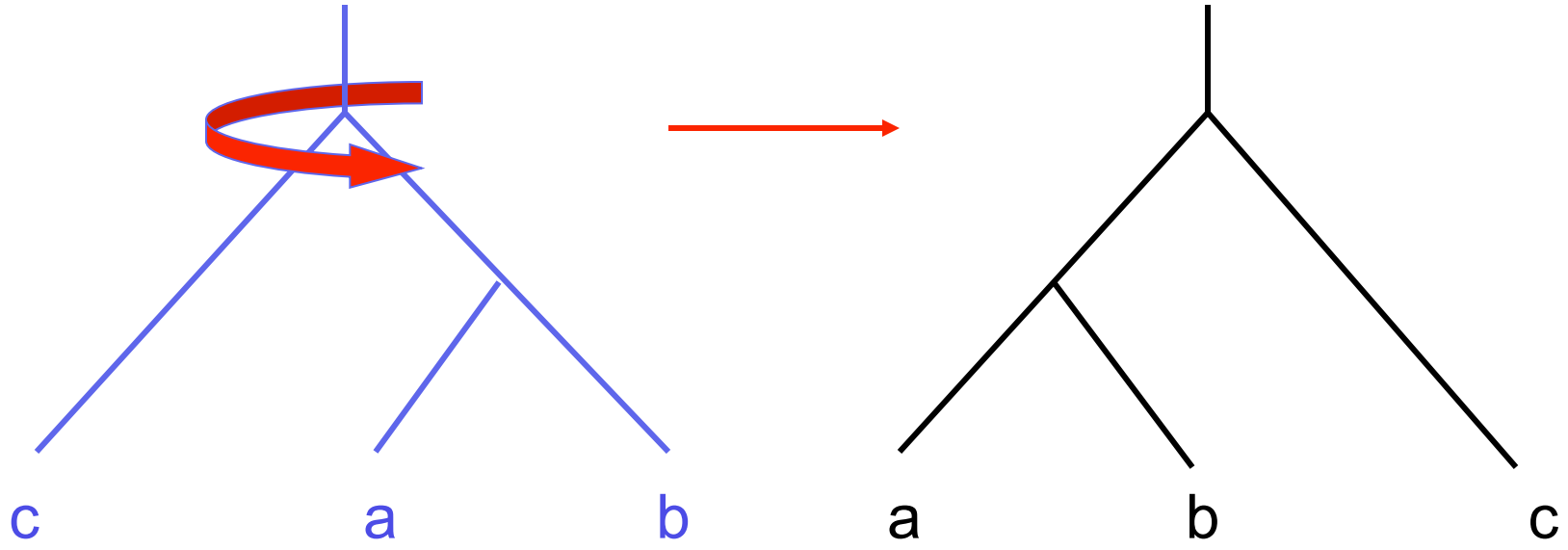
- limites de l'horloge moléculaire :
  . horloge moléculaire ne serait pas constante : les mutations avantageuses se fixeraient plus vite que les autres
  . horloge moléculaire épisodique : épisodes d'accumulation suivis d'arrêts évolutifs

vocabulaire



1 arbre = 1 topologie + longueurs de branches associées

propriétés



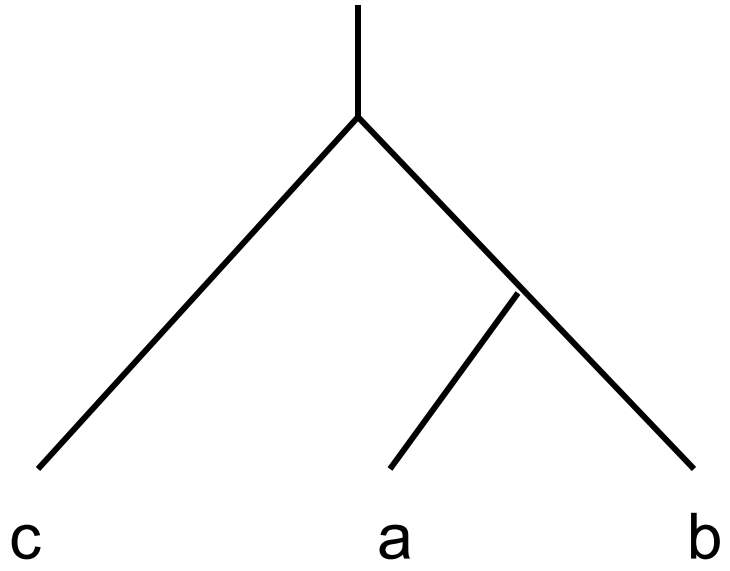
arbre 1

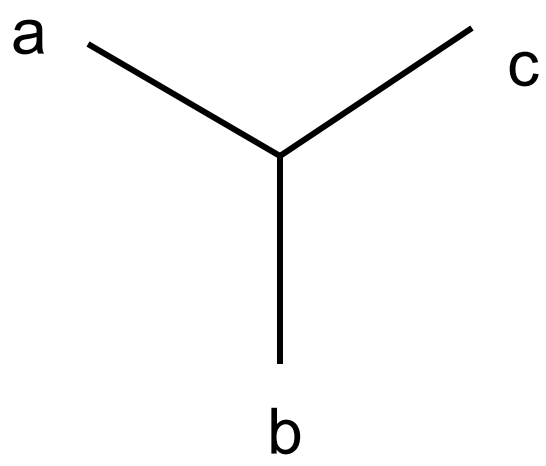
arbre 2

arbre 1 = arbre 2

arbres racinés/non racinés

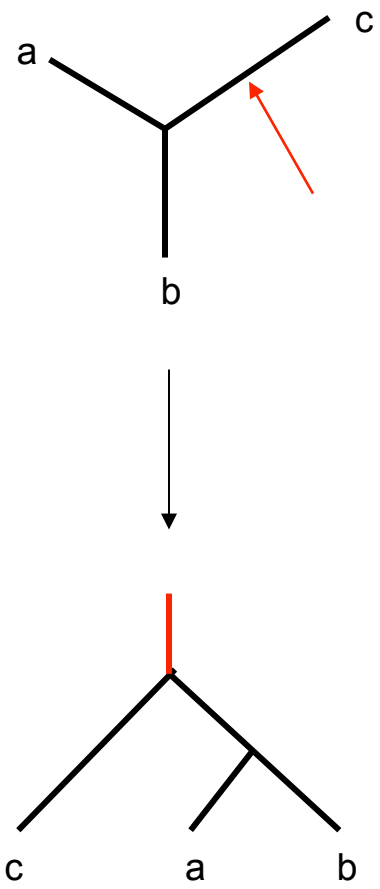
arbre enraciné
(graphe orienté dans le temps)





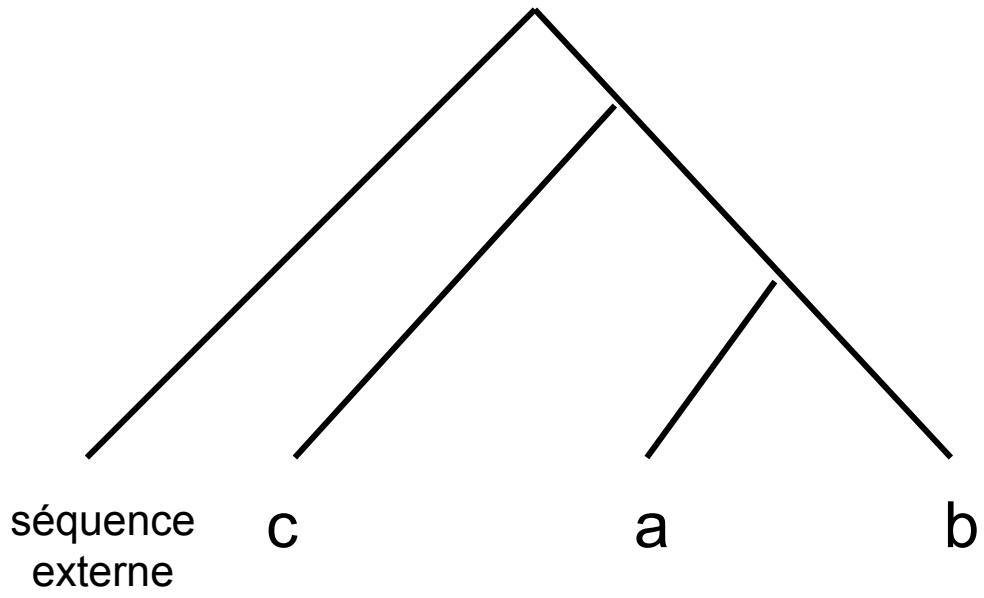
- 1 arbre = 1 topologie + longueurs de branches associées
- longueur des branches proportionnelle à la quantité d'évolution entre les séquences (feuilles) et leur ancêtre (nœud)

comment enraciner un arbre ?



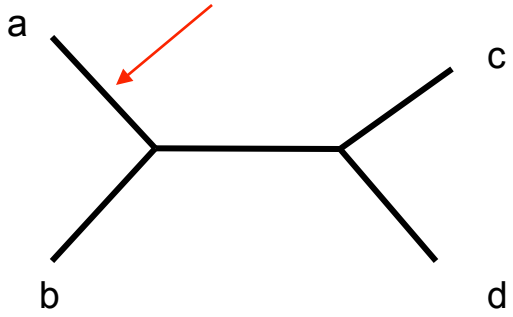
- **hypothèse de l'horloge moléculaire :**
  toutes les séquences ont évolué à la même vitesse depuis leur divergence.
  La racine est le barycentre des feuilles.

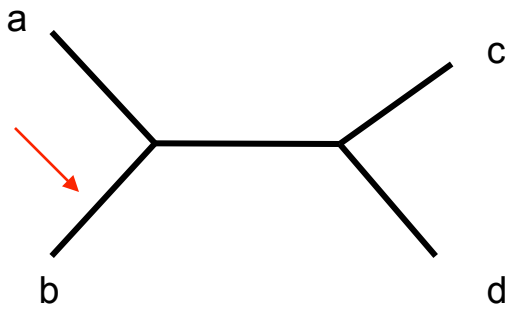
comment enraciner un arbre ?

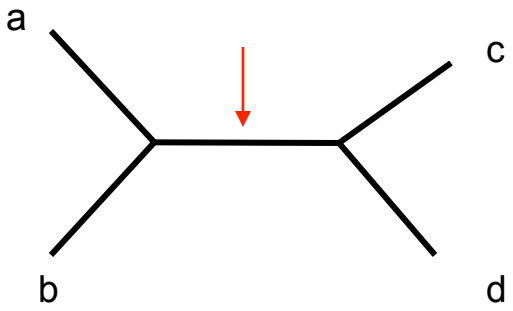


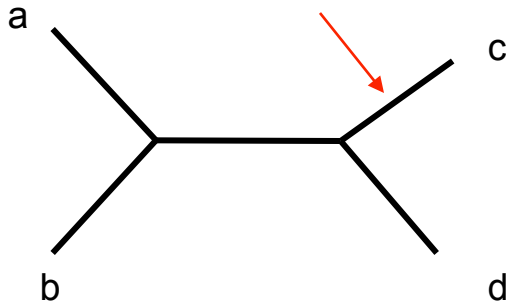
• **méthode du groupe externe :** inclure un groupe de séquences *a priori* externe au groupe d'intérêt. La racine devient la branche reliant le groupe externe aux autres séquences.

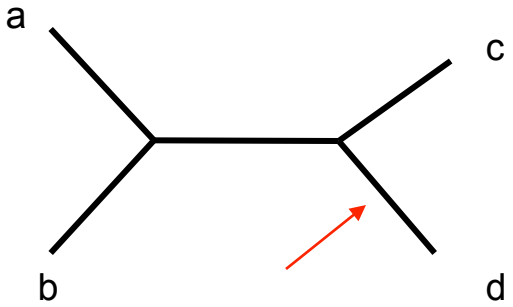
comment enraciner un arbre ?







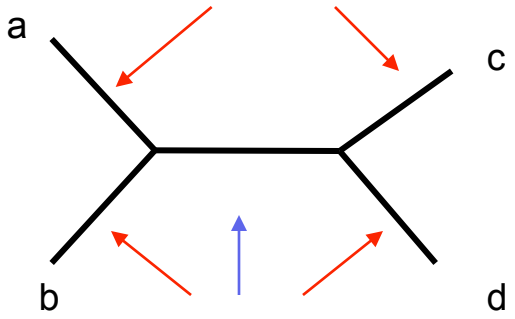




5 solutions différentes !

Il existe autant de racines possibles que de branches dans un arbre non raciné !

nombre de topologies



- un arbre avec (n-1) taxons, possède :
  - . (n-1)-3 branches internes
  - . (n-1) branches externes

- on peut lui ajouter un $n^{ième}$ taxon, sur les branches internes ou externes :

(n-1)-3 + (n-1) = 2n-5 possibilités

- le nombre total d'arbres avec n taxons, T_n est donc défini par :

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

par récurrence on a : $T_n = \prod_{k=3}^{n} (2k-5)$

nombre de topologies

Combien d'arbres non racinés ?

Taxons	Nombre d'arbres
3	1
4	3
5	15
6	105
7	945
8	10395
9	135135
10	$2{,}03.10^{6}$
20	$2{,}22.10^{20}$
30	$8{,}69.10^{36}$
40	$1{,}31.10^{55}$
50	$2{,}84.10^{74}$
100	$1{,}7.10^{182}$

Combien d'arbres racinés ?

Taxons	Nombre d'arbres	
3	1	* nb de branches = 3
4	3	* nb de branches = 15
5	15	* nb de branches = 105
6	105	
7	945	
8	10395	
9	135135	
10	$2{,}03.10^{6}$	
20	$2{,}22.10^{20}$	
30	$8{,}69.10^{36}$	
40	$1{,}31.10^{55}$	
50	$2{,}84.10^{74}$	
100	$1{,}7.10^{182}$	

Donc nombre d'arbres racinés pour n taxons = nb d'arbres non racinés pour n+1 taxons

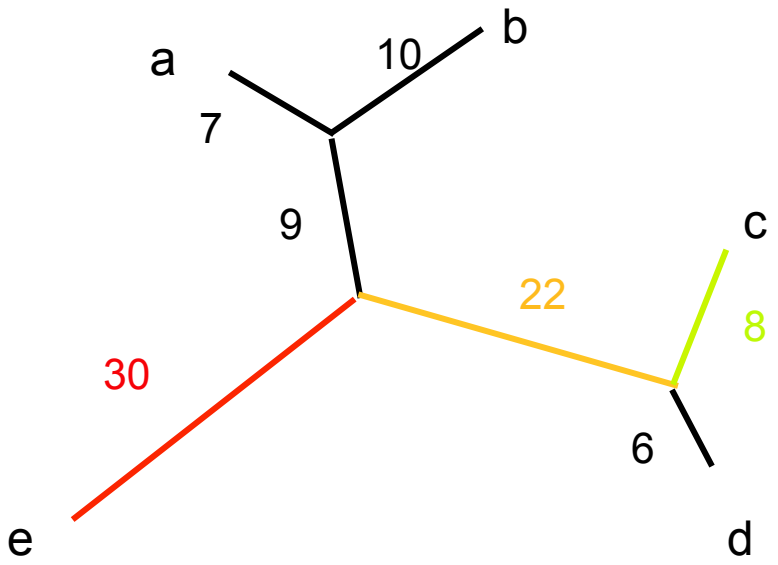
longueur de branches

- les distances entre nœuds/feuilles peuvent se calculer de différentes façons :

  . nombre de mutation par site pour aller d'un nœud à l'autre

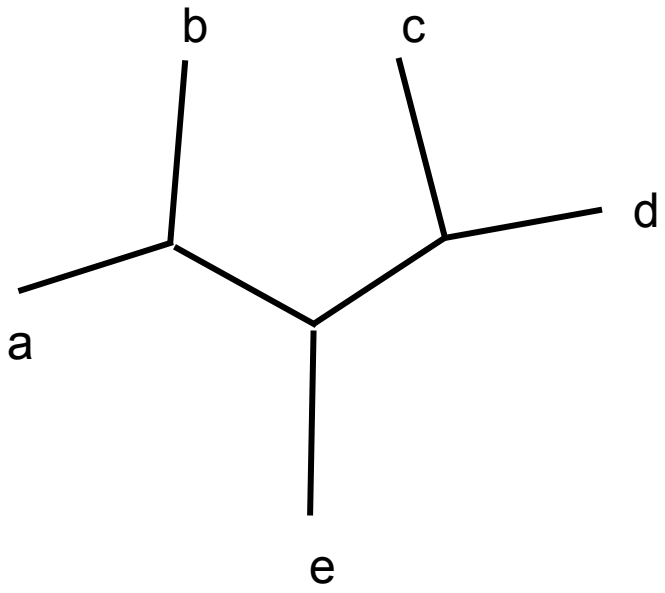
  . estimation du temps d'évolution pour aller d'un nœud à l'autre (modèle évolutif)

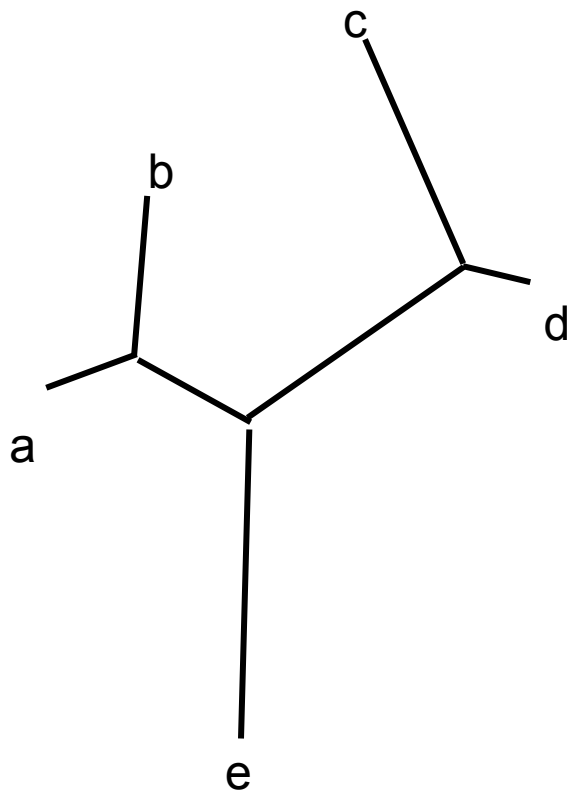
  . $d_{ij}(T)$ correspond à la longueur du chemin pour aller entre i et j dans l'arbre T



$d_{ec} = 30 + 22 + 8 = 60$

longueur de branches





- cladogrammes : longueur de branches arbitraire, ne reflète pas la distance évolutive

- phylogrammes : longueur des branches proportionnelle à la distance évolutive entre séquences (nb mut/site)
- phénogrammes : longueur des branches proportionnelle à un % de divergence
- chronogrammes : longueur des branches proportionnelle au temps écoulé

reconstruction d'un arbre

- la succession d'évènements de spéciation, de duplications et de transferts horizontaux est unique !

  . un seul arbre correspond à CETTE histoire évolutive : l'arbre VRAI !

- le (ou les) arbres obtenu(s) à partir d'une méthode de reconstruction et un jeu de données particulier est un (des) arbre(s) INFERE(S)

  . il correspond à un modèle évolutif et un jeu de données

la plupart du temps : arbre VRAI ≠ arbre INFERE !

reconstruction d'un arbre

- deux grands types de données utilisées pour reconstruire les arbres

  . caractères (morphologiques, acides aminés…)

  . distances

reconstruction d'un arbre

trois grands types de méthodes utilisées pour reconstruire les arbres

- méthodes de distances ou phénétiques (distances)

  . recherche de l'arbre qui représente au mieux les distances évolutives entre séquences
  . la séquence est réduite à un nombre (la distance). Cette distance est calculée à partir d'un modèle évolutif
  . exemples : UPGMA, Neighbor Joining…

- méthodes cladistiques (caractères)

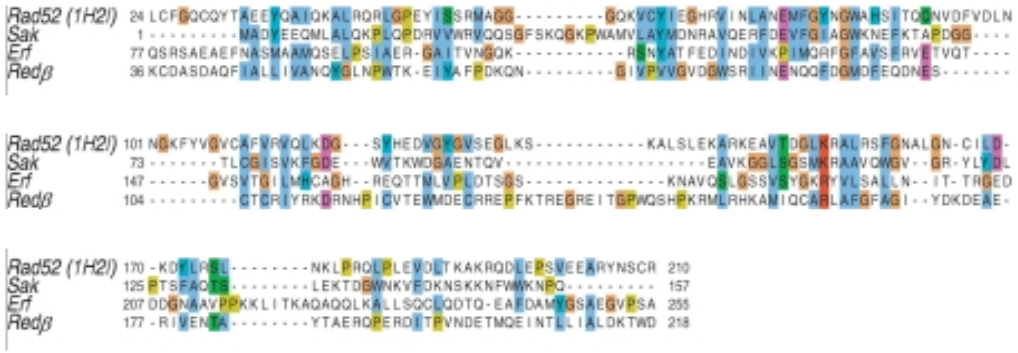
  . recherche de l'arbre impliquant le moins de changements évolutifs cohérent avec les données
  . considère les sites individuellement
  . exemple : maximum de parcimonie

- méthodes probabilistes (caractères)

  . recherche de l'arbre le plus vraisemblable étant donné un modèle évolutif et des données
  . considère les sites individuellement
  . exemples : Maximum de Vraisemblance

reconstruction d'un arbre

En phylogénie moléculaire, toutes ces méthodes partent du même point de départ !

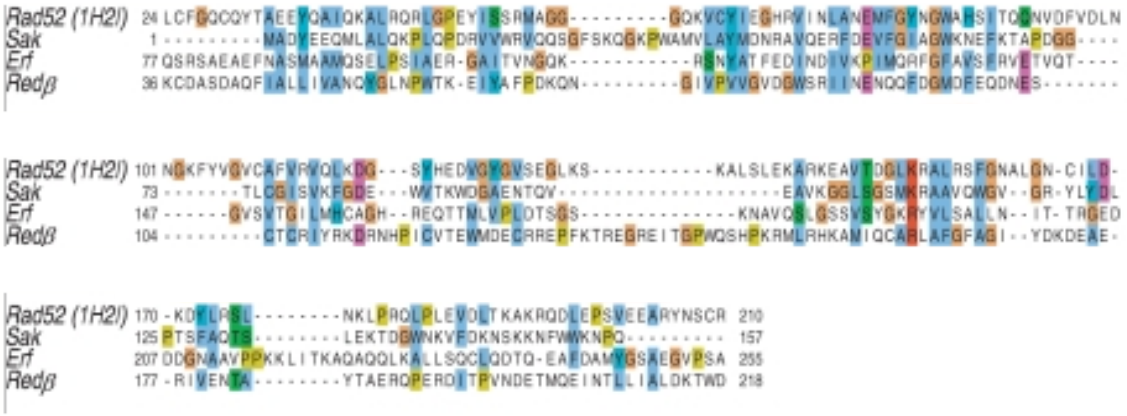


- point de départ :
  un alignement de séquences homologues

- arrivée :
  arbre(s) décrivant les liens évolutifs entre les séquences de l'alignement

reconstruction d'un arbre

En phylogénie moléculaire, toutes ces méthodes partent du même point de départ !



- chaque colonne de l'alignement correspond à une position de l'alignement ou site
- la qualité de l'alignement est déterminante : les régions ambiguës doivent être retirées avant l'analyse
  . nettoyage manuel ou automatique (Gblocks)
- la plupart des méthodes de reconstruction ne prend pas en compte les évènements d'insertion/délétion

méthodes de distances

- idée de départ : plus la ressemblance globale entre deux séquences est importante, plus il y a de chances que ces séquences soient proches

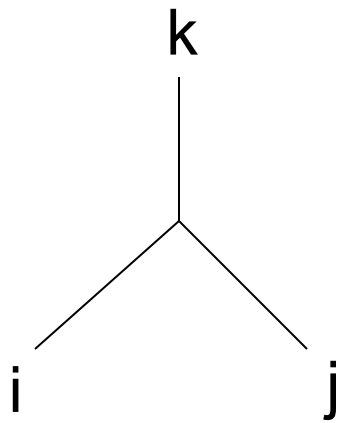
- propriétés des métriques utilisées :

  . $d_{ij} > 0$ si i différent de j

  . $d_{ij} = 0$ si i = j

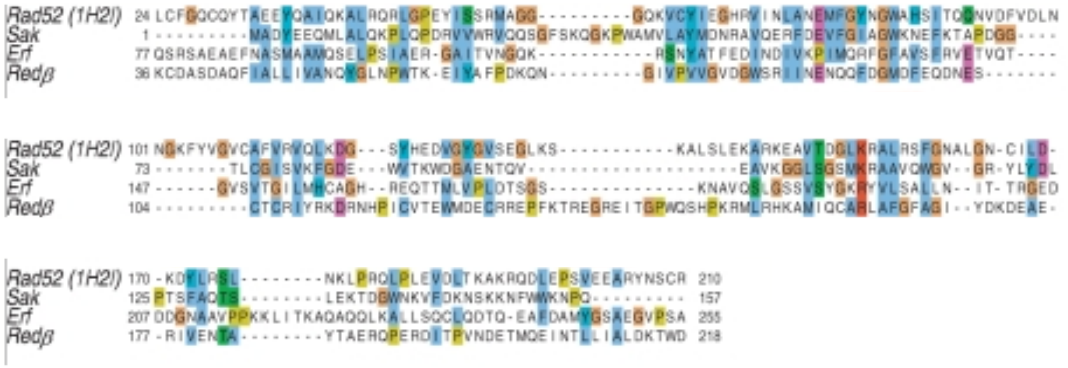
  . $d_{ij} = d_{ji}$

  . $d_{ik} + d_{kj} >= d_{ij}$



méthodes de distances

- principe général :
  - . Construction de la matrice de distances
  - . Reconstruction de l'arbre dont les distances sont les plus proches de celles calculées dans la matrice

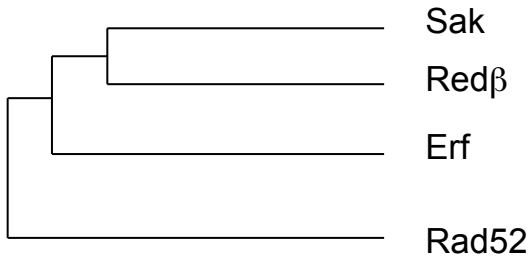


Calcul de distance à l'aide d'un modèle évolutif

	Rad52	Sak	Erf	Redβ
Rad52				
Sak	d12			
Erf	d13	d32		
Redβ	d14	d24	d34	

matrice de distances

Méthode de clustering



méthodes de distances : UPGMA

principe

Algorithme itératif de clustering : création à chaque étape d'un nouveau cluster regroupant deux clusters plus proches

L'arbre est construit de bas en haut (feuilles vers racine). A chaque étape on rajoute un nœud au dessus des précédents

méthodes de distances : UPGMA

Calcul de la matrice de distance à partir de l'alignement de séquence

Choix de la paire de séquence la plus proche (A,B)

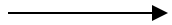
Calcul de la distance entre (A,B) et le reste

Définition d'une nouvelle matrice de distance

```
Sequence A: AAAAA
Sequence B: AAAAG
Sequence C: AACGC
Sequence D: ATGGT
```

Calcul de la matrice de distance

	A	B	C	D
A	/	/	/	/
B	?	/	/	/
C	?	?	/	/
D	?	?	?	/



```
Sequence A: AAAAA
Sequence B: AAAAG
Sequence C: AACGC
Sequence D: ATGGT
```

Calcul de la matrice de distance

	A	B	C	D
A	/	/	/	/
B	0.2	/	/	/
C	0.6	0.6	/	/
D	0.8	0.8	0.4	/

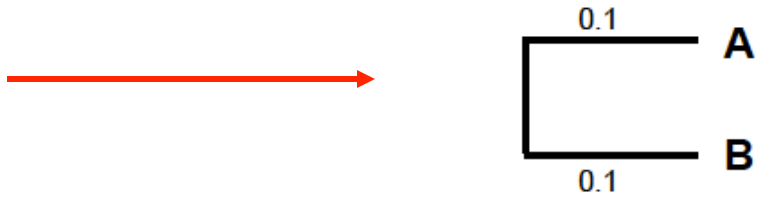
méthodes de distances : UPGMA

Calcul de la matrice de distance à partir de l'alignement de séquence
Choix de la paire de séquence la plus proche (A,B)
Calcul de la distance entre (A,B) et le reste
Définition d'une nouvelle matrice de distance

```
Sequence A: AAAAA
Sequence B: AAAAG
Sequence C: AACGC
Sequence D: ATGGT
```

Calcul de la matrice de distance

	A	B	C	D
A	/	/	/	/
B	0.2	/	/	/
C	0.6	0.6	/	/
D	0.8	0.8	0.4	/

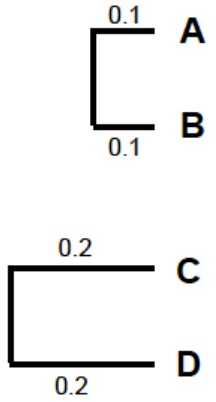


méthodes de distances : UPGMA

Calcul de la matrice de distance à partir de l'alignement de séquence
Choix de la paire de séquence la plus proche (A,B)
Calcul de la distance entre (A,B) et le reste
Définition d'une nouvelle matrice de distance

dist(AB),C = (distAC + distBC) / 2
dist(AB),D = (distAD + distBD) / 2

	AB	C	D
AB	/	/	/
C	0.6	/	/
D	0.8	0.4	/

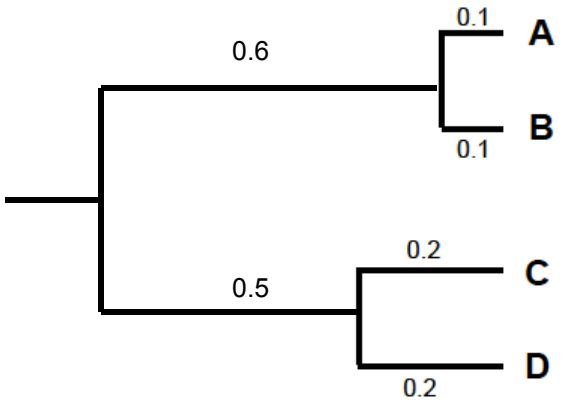


méthodes de distances : UPGMA

Calcul de la matrice de distance à partir de l'alignement de séquence
Choix de la paire de séquence la plus proche (A,B)
Calcul de la distance entre (A,B) et le reste
Définition d'une nouvelle matrice de distance

dist(CD),(AB) = (distC(AB) + distD(AB)) / 2

	AB	CD
AB	/	/
CD	0.7	/



méthodes de distances : UPGMA

- conditions d'application :

  . hypothèse d'horloge moléculaire : constance des taux d'évolution le long des lignées

- caractéristiques de l'arbre obtenu :

  . longueurs des branches entre la racine et chacune des feuilles identiques

- avantages :

  . rapidité et simplicité

- limites :

  . hypothèse d'horloge moléculaire pourtant contreversée

  . résultats faux si les distances de la matrices ne correspondent pas au critère d'horloge moléculaire

méthodes de distances : Neighbor Joining (Saitou & Nei, Mol. Biol. Evol., 1987)

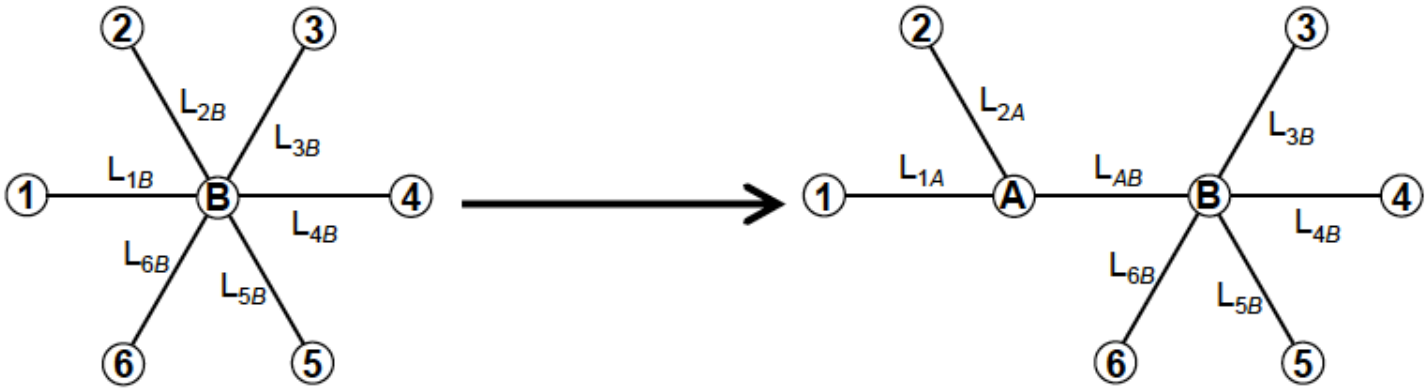
- amélioration d'UPGMA :

. Autorisation de taux de mutations différents sur les branches, donc n'implique l'hypothèse d'horloge moléculaire

. Au début on part d'un arbre en étoile. Cet arbre sera corrigé itérativement pour tenir compte de la divergence moyenne de chaque séquence avec les autres

. L'arbre est reconstruit en reliant les séquences les plus proches. Lorsque deux séquences sont reliées, le nœud représentant leur ancêtre commun est ajouté à l'arbre et remplace les deux séquences venant d'être fusionnées.

. L'algorithme continue jusqu'à ce que toutes les séquences aient été regroupées.



méthodes de distances : Neighbor Joining (Saitou & Nei, Mol. Biol. Evol., 1987)

- conditions d'application :

  . hypothèse d'horloge moléculaire non nécessaire

- avantages :

  . rapidité (permet de travailler avec un grand nombre de séquences) et simplicité

- limites :

  . réduit la séquence à un nombre. Perte d'information

  . ne permet pas de combiner des caractères différents

méthodes cladistiques : méthode de parcimonie

- principe : méthode basée sur les caractères

  . Identifier la topologie impliquant le nombre minimal et suffisant de changements évolutifs afin de rendre compte des données (différences entres séquences)

  . L'arbre le plus parcimonieux correspond au chemin le plus court conduisant aux caractères observés

- condition :

  . Un site (colonne) est informatif ssi il y a au moins deux types de nucléotides dans le site et si chacun d'entre eux est représenté au moins deux fois

méthodes cladistiques : méthode de parcimonie

- étapes principales :

  . Identifier toutes les topologies possibles

  . Identifier tous les sites informatifs

  . Identifier le nombre de changements évolutifs nécessaires pour expliquer les caractères observés pour un site étant donnée une topologie

  . Répeter l'étape précedente pour tous les sites informatifs

  . Répeter les deux étapes précédentes pour toutes les topologies étudiées

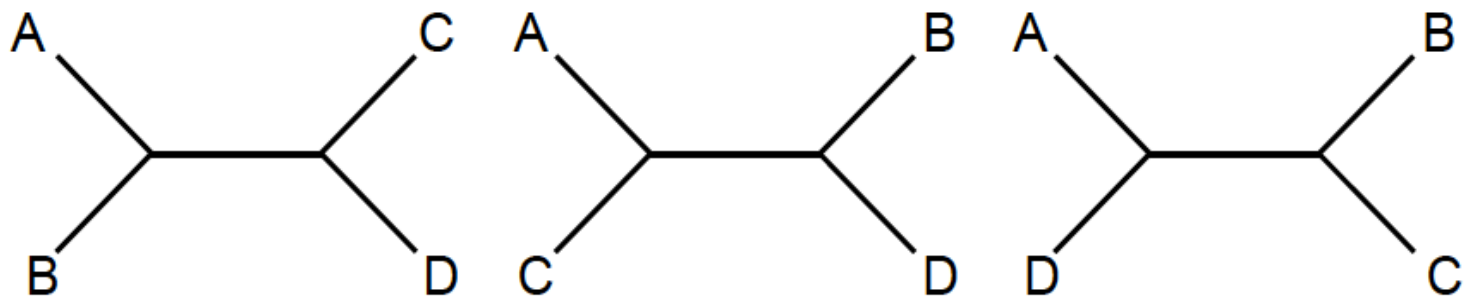
  . Etape finale : choisir l'arbre le plus parcimonieux càd : identifier parmi toutes les topologies testées, celle minimisant le nombre de changements évolutifs

méthodes cladistiques : méthode de parcimonie

- Identifier toutes les topologies possibles

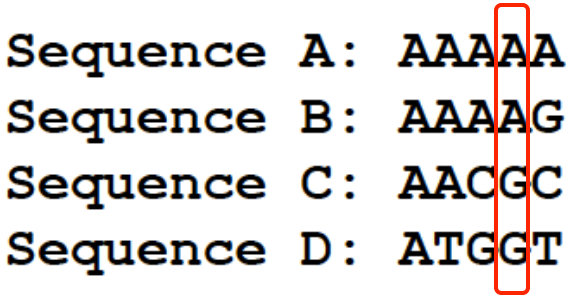
```
Sequence A: AAAAA
Sequence B: AAAAG
Sequence C: AACGC
Sequence D: ATGGT
```

. 4 séquences : 3 topologies possibles !



méthodes cladistiques : méthode de parcimonie

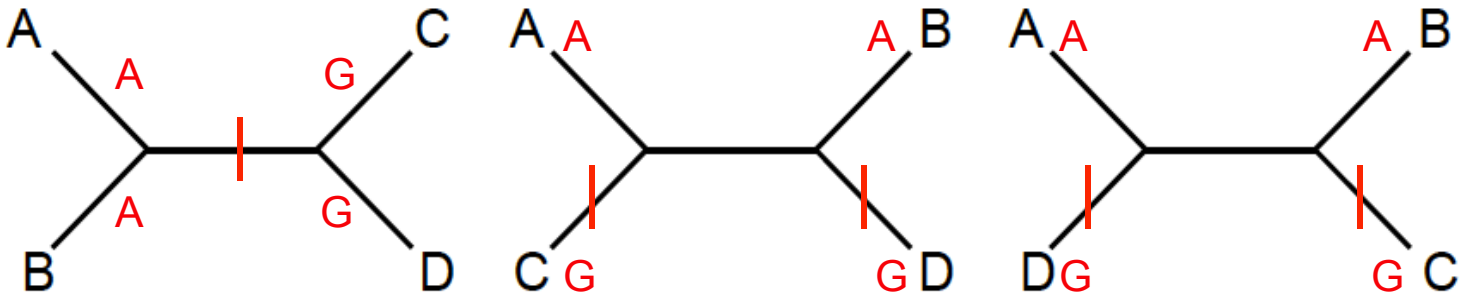
- Identifier tous les sites informatifs



méthodes cladistiques : méthode de parcimonie

- Identifier le nombre de changements évolutifs nécessaires pour expliquer les caractères observés pour un site étant donnée une topologie

```
Sequence A: AAAAA
Sequence B: AAAAG
Sequence C: AACGC
Sequence D: ATGGT
```



Nb évènements : 1

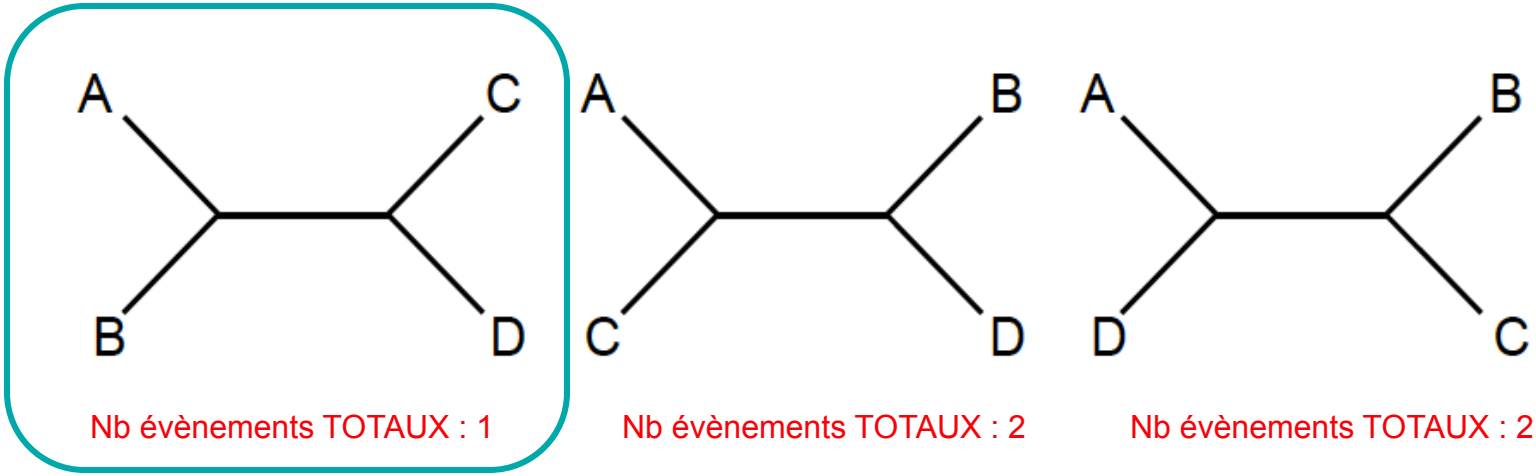
Nb évènements : 2

Nb évènements : 2

méthodes cladistiques : méthode de parcimonie

- Répeter l'étape précedente pour tous les sites informatifs : ✔
- Répeter les deux étapes précédentes pour toutes les topologies étudiées : ✔
- Sommer sur tous les sites et toutes les topologies et choisir celle minimisant le nombre de changements évolutifs

arbre retenu



méthodes cladistiques : méthode de parcimonie

- Exploration des topologies

Taxons	Nombre d'arbres
3	1
4	3
5	15
6	105
7	945
8	10395
9	135135
10	$2{,}03.10^{6}$
20	$2{,}22.10^{20}$
30	$8{,}69.10^{36}$
40	$1{,}31.10^{55}$
50	$2{,}84.10^{74}$
100	$1{,}7.10^{182}$

- Le nombre de topologies augmente très vite avec le nombre de taxons

- pour n < 12 : exploration exhaustive

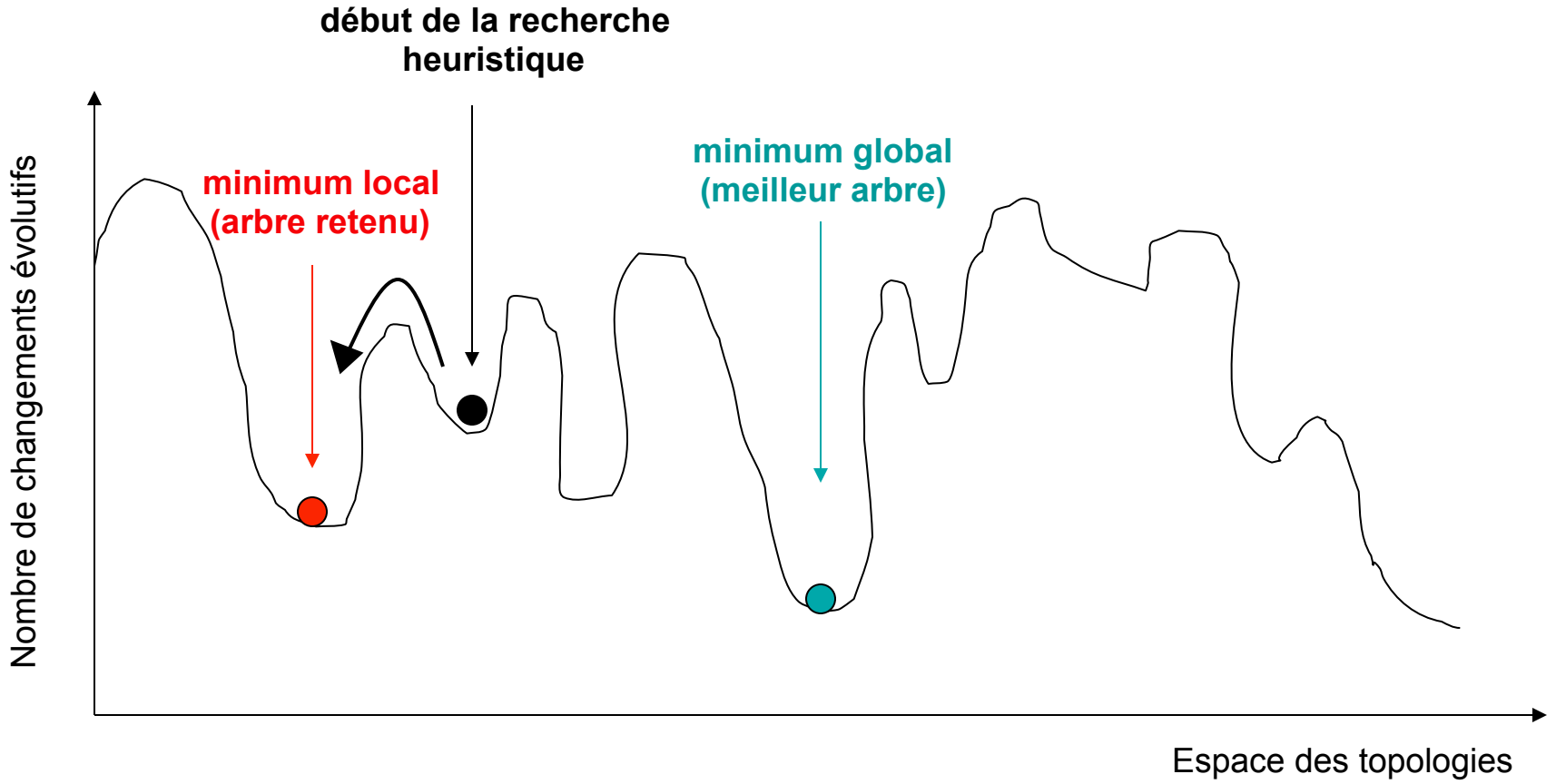
- pour n > 20 : heuristiques

- topologie de départ :
  - . aléatoire
  - . meilleure topologie issue d'une autre approche

exploration des topologies

- approches heuristiques

  . Exloration d'une portion de l'espace des solutions
  . Ne garantit pas de trouver l'arbre le plus parcimonieux



méthodes cladistiques : méthode de parcimonie

- conditions d'application :

  . hypothèse d'horloge moléculaire non nécessaire

- caractéristiques de l'arbre obtenu :

  . plusieurs arbres possibles

- avantages :

  . la séquence n'est pas réduit à un nombre
  . méthode donnant une information sur les intermédiaires évolutifs

- limites :

  . très lente
  . seuls les sites informatifs sont pris en compte (perte d'information)
  . pas de correction pour les substitutions multiples
  . ne donne pas de bons résultats si les séquences sont trop divergentes

méthodes probabilistes : Maximum de Vraisemblance

- principe :

  . Le but est de maximiser la vraisemblance des données observées étant donnée une hypothèse évolutive

  . Ici les données (D) sont l'alignement multiple et l'hypothèse évolutive (H) correspond à l'arbre proposé (topologie + longueur de branches) et le modèle évolutif utilisé

vraisemblance de D $L_D = P(D|H)$

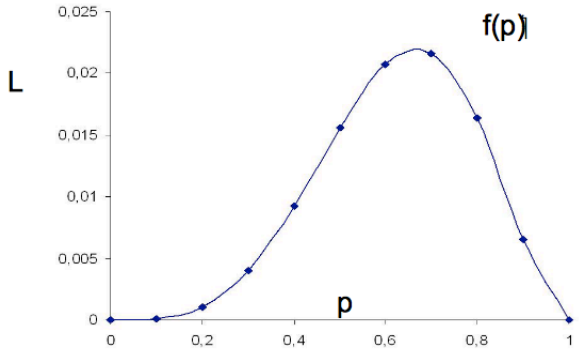
L'hypothèse pour laquelle cette probabilité est maximale est celle qui explique le mieux les données

- exemple :

  . Soit p la probabilité d'obtenir pile, quelle est la vraisemblance d'observer le résultat suivant si on réalise 6 lancés ?

D : Pile Face Face Pile Pile Pile

$L = P(D|H) = P(D|p) = p\,(1-p)\,(1-p)\,p\,p\,p = p^4(1-p)^2$



méthodes probabilistes : Maximum de Vraisemblance

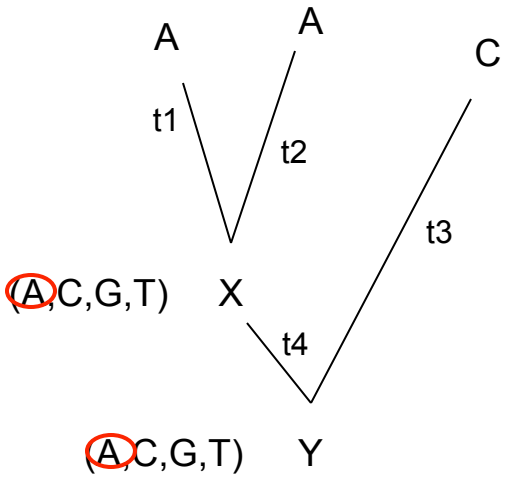
- données = alignement & hypothèse = arbre + modèle évolutif

Seq 1 AAA
Seq 2 ACT
Seq 3 CAT

site i

La vraisemblance L(T|i) d'une topologie T pour le site i est la probabilité P(i|T)
Que le profil {AAC} ait été généré chez les séquences (1,2,3) par la topologie T

- calcul de la vraisemblance pour le site i



P(A,A,C,Ay,Ax|T) =
P(Ay)*P(Ax|Ay, t4)*P(A|Ax,t1)*P(A|Ax, t2)*P(C|Ay,t3)

méthodes probabilistes : Maximum de Vraisemblance

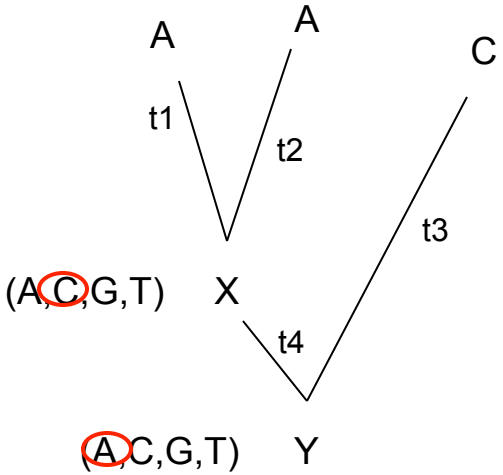
- données = alignement & hypothèse = arbre + modèle évolutif

Seq 1 AAA
Seq 2 ACT
Seq 3 CAT

site i

La vraisemblance L(T|i) d'une topologie T pour le site i est la probabilité P(i|T)
Que le profil {AAC} ait été généré chez les séquences (1,2,3) par la topologie T

- calcul de la vraisemblance pour le site i



P(A,A,C,Ay,Cx|T) =
P(Ay)*P(Cx|Ay, t4)*P(A|Cx,t1)*P(A|Cx, t2)*P(C|Ay,t3)

etc sur toutes les possibilités

méthodes probabilistes : Maximum de Vraisemblance

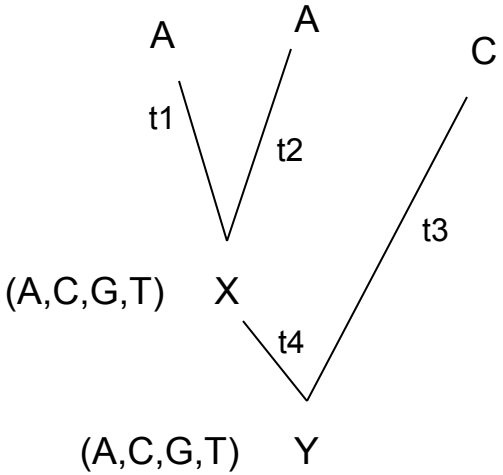
- données = alignement & hypothèse = arbre + modèle évolutif

Seq 1 AAA
Seq 2 ACT
Seq 3 CAT

site i

La vraisemblance L(T|i) d'une topologie T pour le site i est la probabilité P(i|T)
Que le profil {AAC} ait été généré chez les séquences (1,2,3) par la topologie T

- calcul de la vraisemblance pour le site i



Parmi toutes les combinaisons possibles, laquelle correspond à la plus grande probabilité ?

P(A,A,C,Ay,Ax|T) > P(A,A,C,Cy,Ax|T) > …

Cette probabilité (max) correspond à la vraisemblance de cette topologie pour le site étudié

méthodes probabilistes : Maximum de Vraisemblance

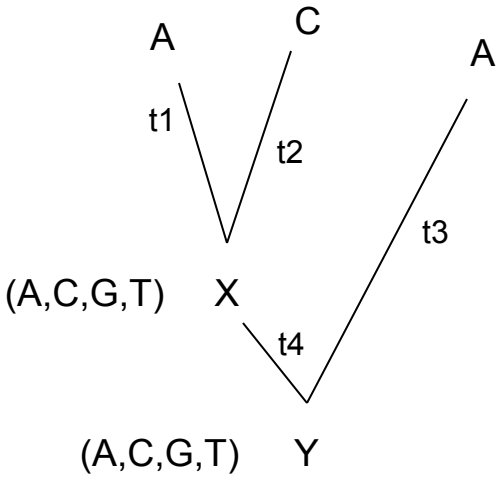
- données = alignement & hypothèse = arbre + modèle évolutif

Seq 1 AAA
Seq 2 ACT
Seq 3 CAT

site i

La vraisemblance L(T|i) d'une topologie T pour le site i est la probabilité P(i|T)
Que le profil {AAC} ait été généré chez les séquences (1,2,3) par la topologie T

- calcul de la vraisemblance pour le site i + 1



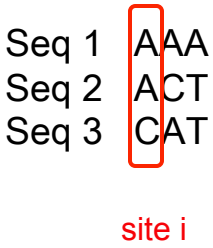
Cette même topologie est évaluée pour le site i + 1
…

La vraisemblance de cette topologie pour tout l'alignement est le produit des vraisemblances de tous les sites

$L(T) = \pi L(T|i)$ avec i allant de 1 à n (taille de l'ali.)
$\ln L(T) = \sum \ln L(T|i)$

méthodes probabilistes : Maximum de Vraisemblance

- données = alignement & hypothèse = arbre + modèle évolutif



La vraisemblance L(T|i) d'une topologie T pour le site i est la probabilité P(i|T)
Que le profil {AAC} ait été généré chez les séquences (1,2,3) par la topologie T

- calcul de la vraisemblance pour tous les sites étant donnée une topologie
  . ajustement des longueurs de branches (modèle évolutif, ajustement des probas)
- répéter le calcul pour toutes les topologies
- choisir la topologie maximisant la vraisemblance càd possédant la plus forte probabilité d'expliquer les données étant donné le modèle évolutif
- exploration des topologies : exhaustive (<10 seq), branch & bound (<20 seq) ou heuristique
- l'arbre de départ peut être pris aléatoirement ou à partir d'une autre approche

méthodes probabilistes : Maximum de Vraisemblance

- avantages :

  . une des méthodes les plus justifiées d'un point de vue théorique

  . tient compte de l'information contenue dans tous les sites

  . les simulations montrent que c'est la meilleure

  . moins sensible aux artefacts d'attraction des longues branches

- limites :

  . très lente surtout avec l'augmentation du nombre de sites

- progrès :

  . PhyML très rapide aujourd'hui. Applicable à des jeux de séquences relativement important (500 et 1000 séquences pour des séquences de 100 à 400 résidus)

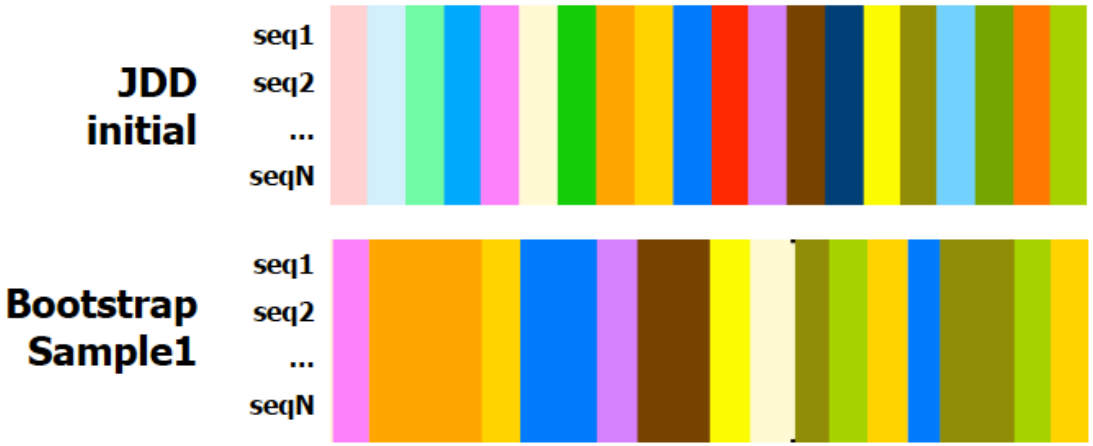
- précautions :

  . L'arbre le plus vraisemblable ne correspond pas forcément au vrai arbre !

Bootstrap

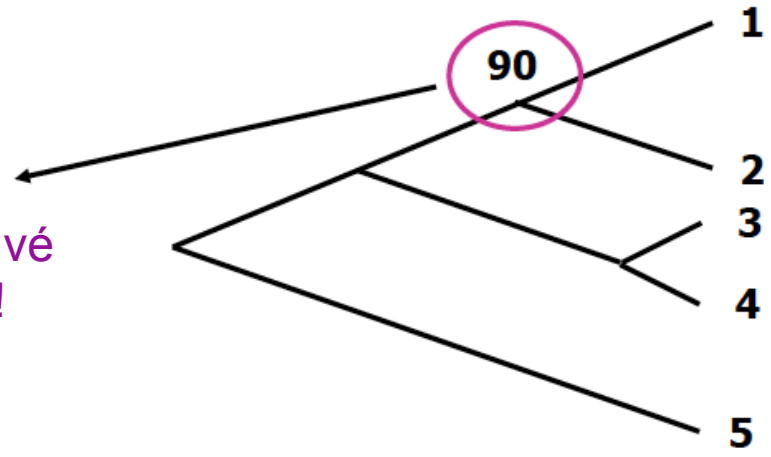
On estime les phylogénies obtenues à partir d'un certain nombre de ré-échantillonnages de même taille que notre jeu de données initial

On réalise X tirages avec remise de n sites parmi les n sites au sein du JDD initial



robustesse des arbres

Comparaison des arbres T et T' : pour chaque sous-arbre de T, on regarde s'il est présent dans T'.
On compte ensuite pour chaque sous-arbre le nombre de fois où il est présent dans les T'. Cette fréquence avec laquelle on retrouve un sous-arbre est la valeur de bootstrap (plus elle est élevée plus la fiabilité de la branche est importante).



Cet arbre est retrouvé dans 90% des cas !