Phylogénie

Informatique Génomique - Master 1

Guillaume Blin

IGM-LabInfo UMR 8049,
Bureau 4B066
Université de Marne La Vallée
gblin@univ-mlv.fr
http://igm.univ-mlv.fr/∼gblin

2007-08

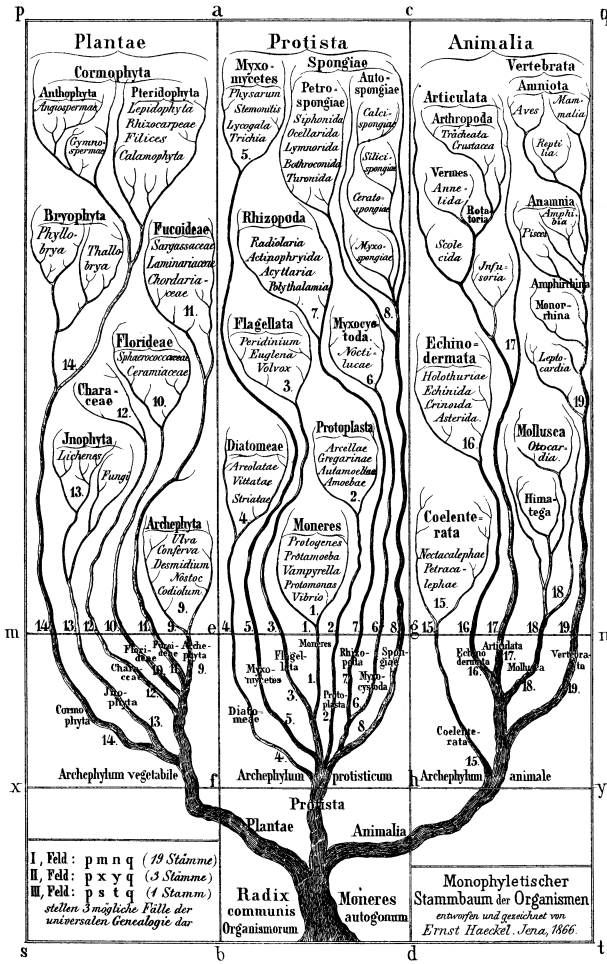
Plan

Plan

Phylogénie

Définition

- Le terme *phylogénie* fut inventé par Ernst Haeckel en 1866 pour définir l'enchaînement des espèces animales et végétales au cours du temps
- Jusqu'alors le concept était exprimé par le terme *généalogie*
- Dans l'Origine des espèces (1872), Charles Darwin introduisit le mot phylogeny avec la définition suivante : *les lignes généalogiques de tous les êtres organisés*
- Nous définirons la phylogénie comme *l'histoire des espèces à partir d'évolutions observées*



Phylogénie

L'objectif de la reconstruction phylogénétique

- Comprendre l'origine de la vie
- Étudier la biodiversité
- Déterminer l'origine géographique des espèces
- Comprendre les mécanismes moléculaires
- ...

Théorie de l'évolution

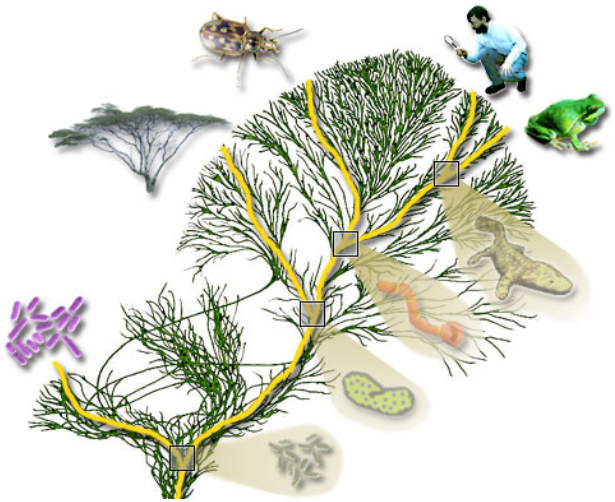
La vie est monophylétique

- *Groupe qui comprend une espèce ancestrale et tous ses descendants*
- Tous les organismes sur terre ont un ancêtre commun
- Tout couple d'organismes partage un ancêtre commun

Théorie de l'évolution

La vie est monophylétique

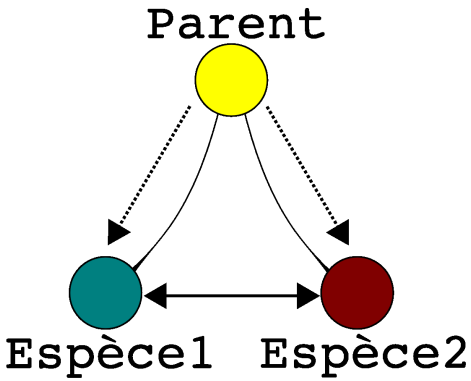
- Des évènements dits de *spéciation* conduisent à la création de deux espèces distinctes
- Les spéciations sont causées par l'éloignement physique des espèces en groupes où différentes variations génétiques deviennent dominantes



Théorie de l'évolution

Distance évolutive

- La mesure de l'évolution entre les espèces se caractérise
  - soit en se basant sur leurs *phénotypes*, i.e. l'ensemble des caractèristiques qu'elles expriment
  - soit en se basant sur leurs *génotypes*, i.e. l'ensemble des caractéristiques comprisent dans leurs génomes et qu'elles peuvent éventuellement exprimer

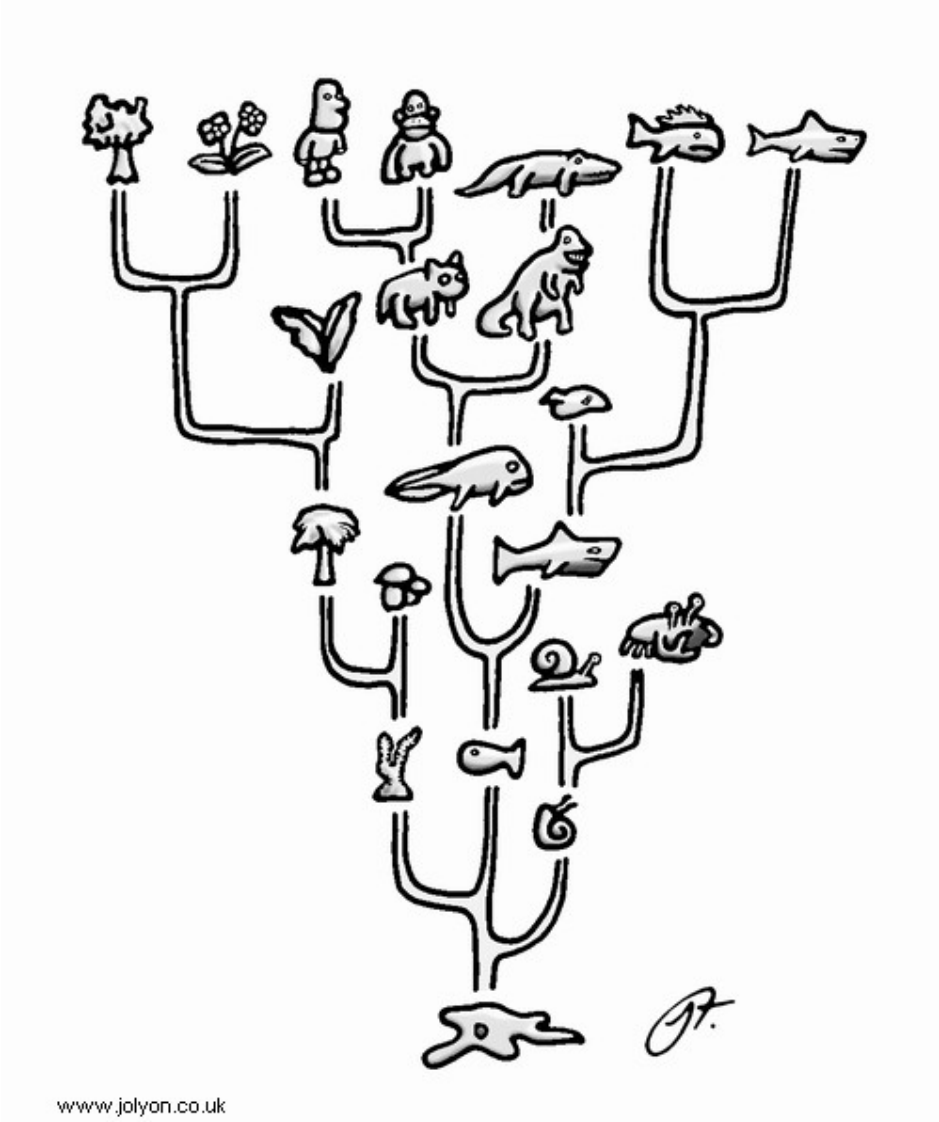


- L'approche morphologique est utilisée lorsqu'on ne dispose d'aucune information sur les génomes des espèces.
- Depuis le séquençage des génomes on a de plus en plus tendance à utilisé le second type d'approche

Théorie de l'évolution

Arbre phylogénétique

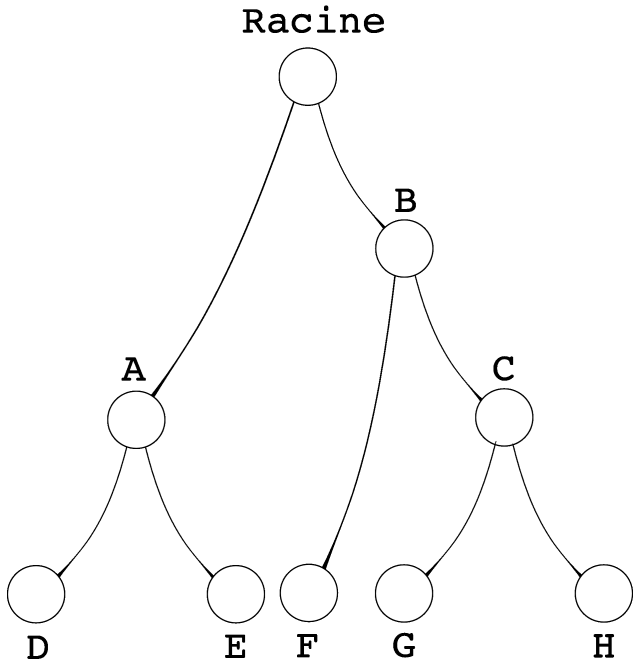
- L'évolution entre les espèces est représentée sous forme d'un arbre phylogénétique dont les branches indiquent le degré de proximité entre les espèces



Arbre phylogénétique

Terminologie

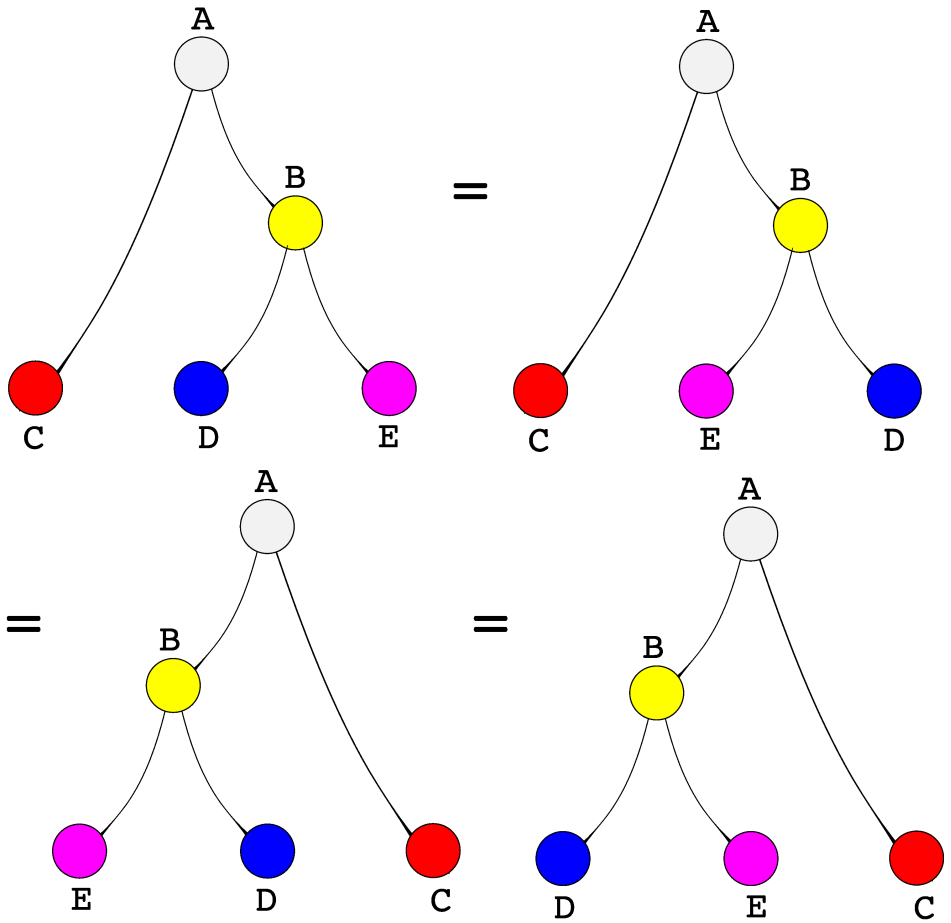
- Graphe composé de noeuds et de branches
- Chaque branche connecte deux noeuds adjacents
- Les noeuds ≡ unités de taxomomie
- Unité de taxomonie ≡ espèce/gène /individu
- Les branches ≡ relations d'héritage



- La longueur des branches ≡ mesure de l'évolution
- Noeud interne ≡ le plus petit ancêtre commun hypothétique
- Feuille ≡ une espèce actuelle ou un gène (taxa)

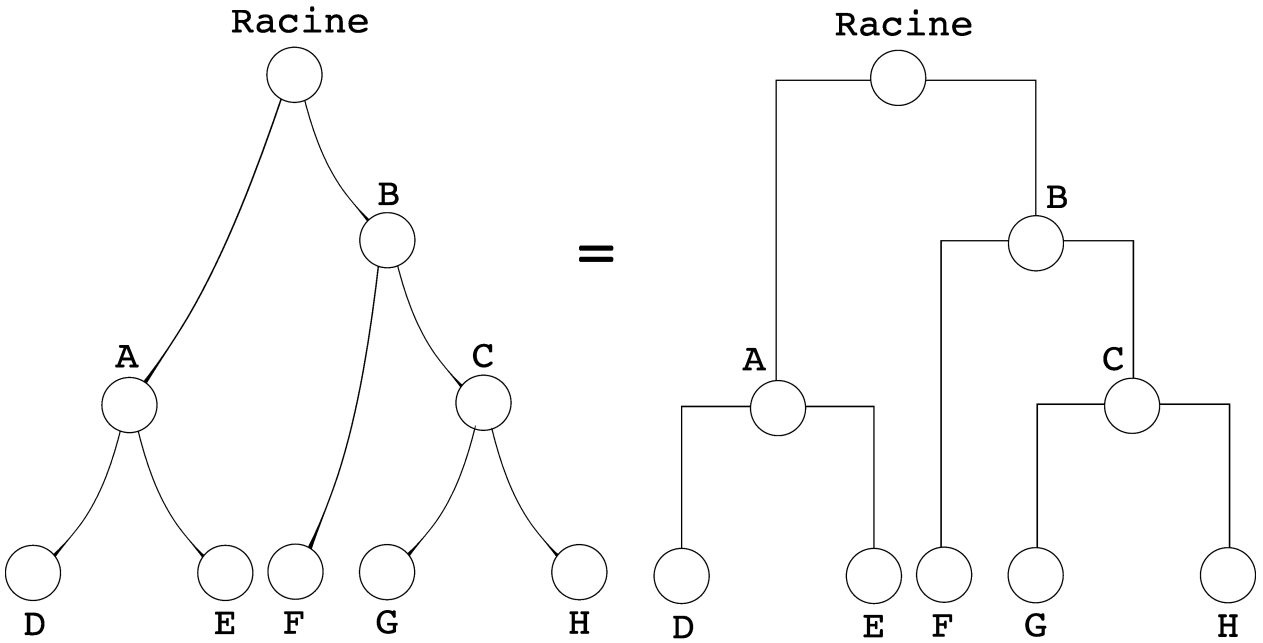
Arbre phylogénétique

Equivalences des arbres



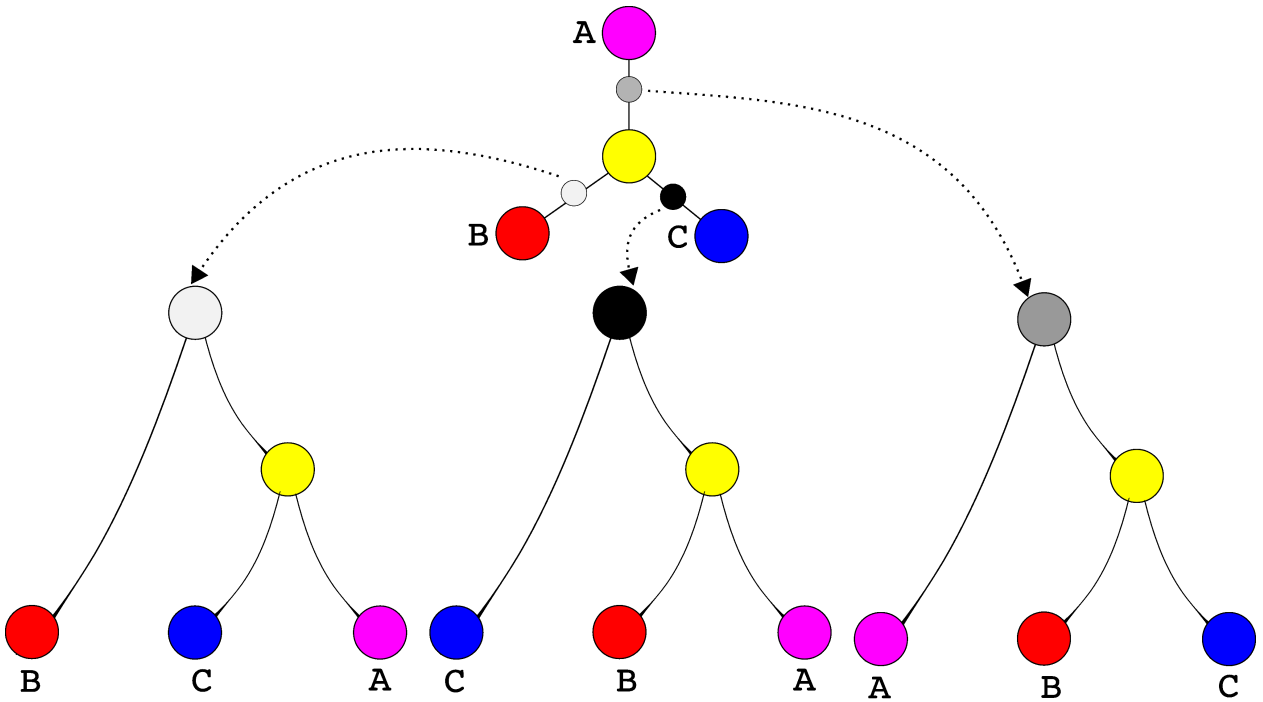
Arbre phylogénétique

Dendogramme



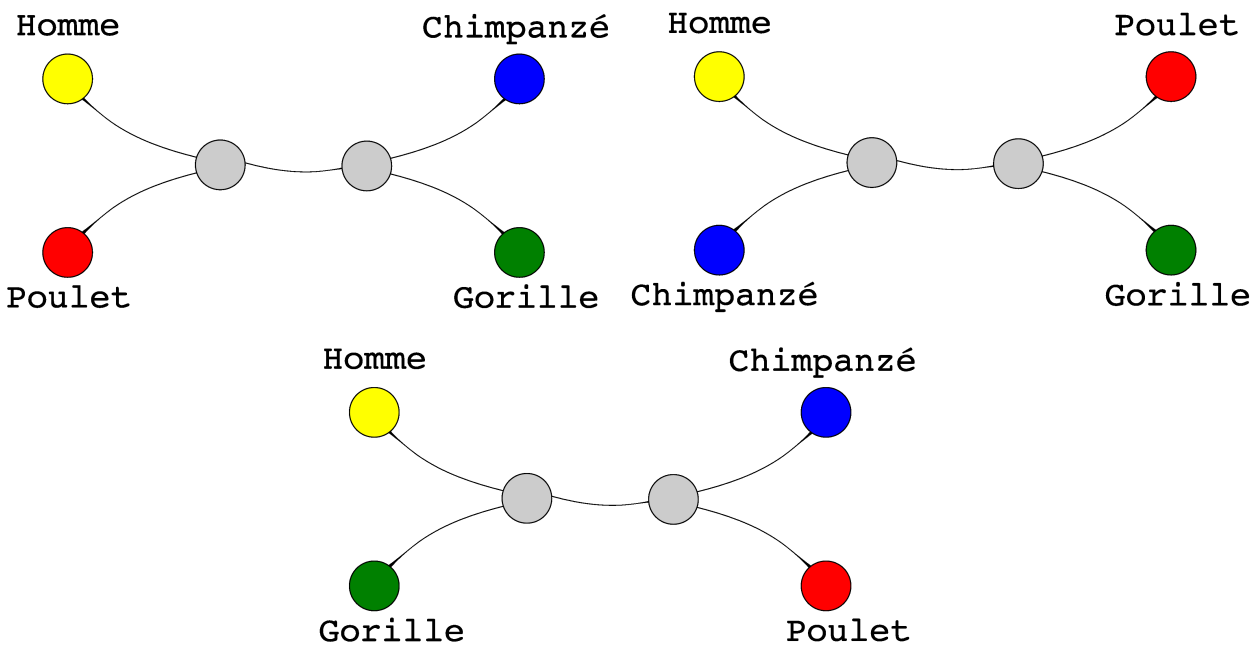
Arbre phylogénétique

Arbres enraciné vs non-enraciné



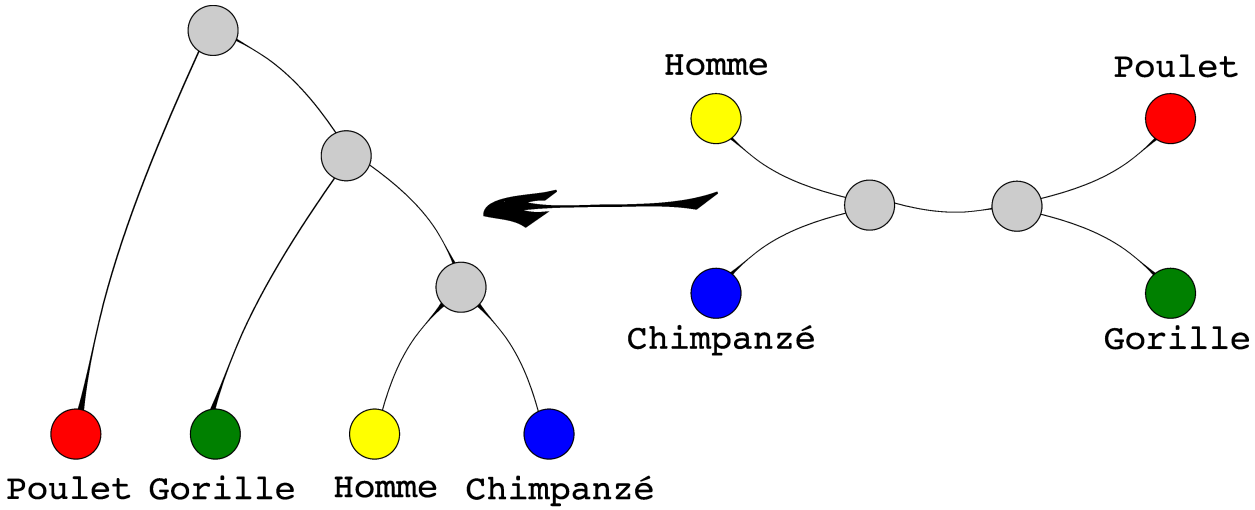
Arbre phylogénétique

Arbres non-enracinés possibles pour 4 taxons



Arbre phylogénétique

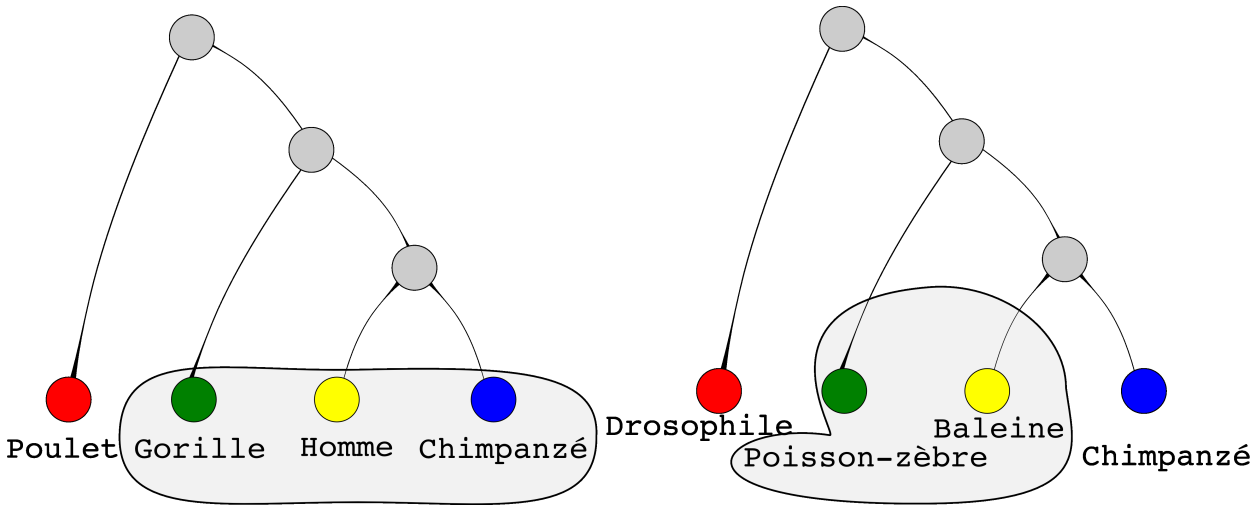
Enracinement basé sur connaissances a priori



Arbre phylogénétique

Groupes monophylétiques : les clades

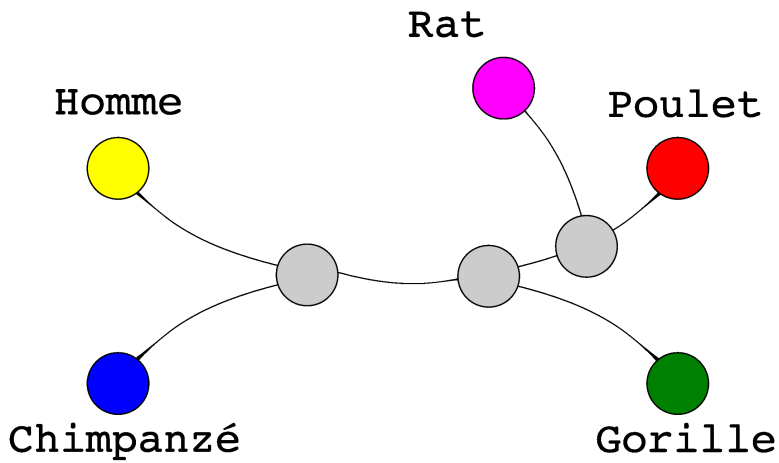
- Un groupe est dit monophylétique (nommé clade) si
  - il existe un ancêtre commun
  - tous les descendants de cet ancêtre appartiennent au groupe



- L'adaptation à l'eau est apparue plus d'une fois durant l'évolution, et indépendamment (ou perdue dans le cas du chimpanzé)

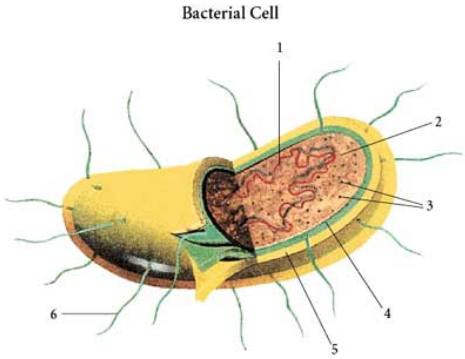
Arbre phylogénétique

Groupes monophylétiques



- Étant donné un arbre non enraciné, on ne peut pas determiner quels sont les groupes monophylétiques
- On peut uniquement désigner ceux qui n'en sont pas
- {Poulet,Rat} peut être monophylétique si la racine est placée entre leur ancêtre commun et le reste
- En réalité, la racine se trouve entre le Poulet et le reste ...
- {Homme,Gorille} n'est pas monophylétique quelque soit la racine choisie

Arbre phylogénétique



Quelles données utiliser ?

- Données moléculaires (ADN, ARN, protéine)
- Données morphologiques
- Données moléculaires sont plus adaptées car :
  - elles sont héritables
  - la description de caractères est ambigüe
  - les données moléculaires peuvent être utilisées dans un traitement quantitatif
  - elles permettent de déduire des relations évolutionnaires entre des organismes distants (ARN ribosomal)
  - elles sont en plus grand nombre (e.g. bactérie)

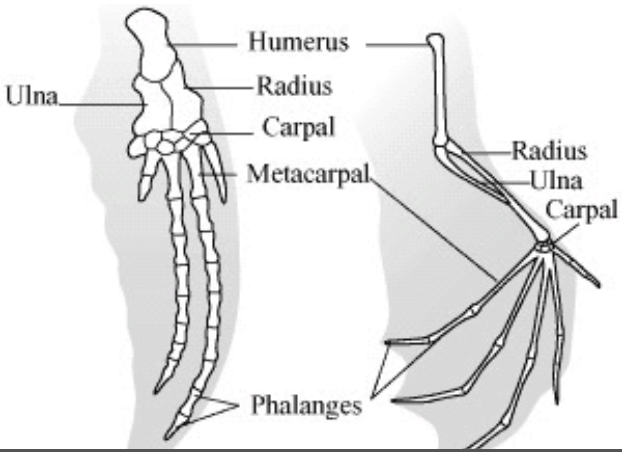
Homologie

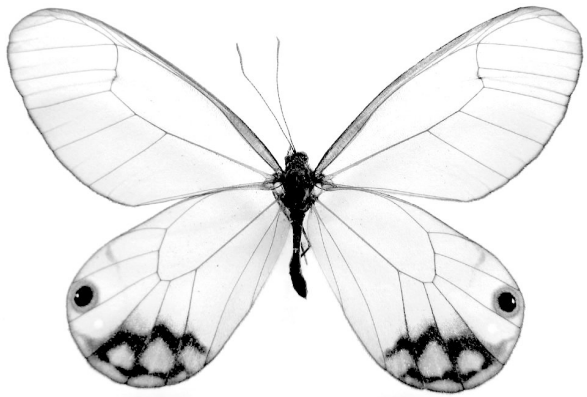
Définition

- Avant Darwin, l'homologie était définie morphologiquement

Exemple - La similarité entre des propriétés de divers espèces

- La chauve souris et le papillon volent ; la baleine nage
- Les ailes de chauves souris et de papillons ne sont pas homologues
- Les ailes de chauves souris et la nageoire de la baleine le sont

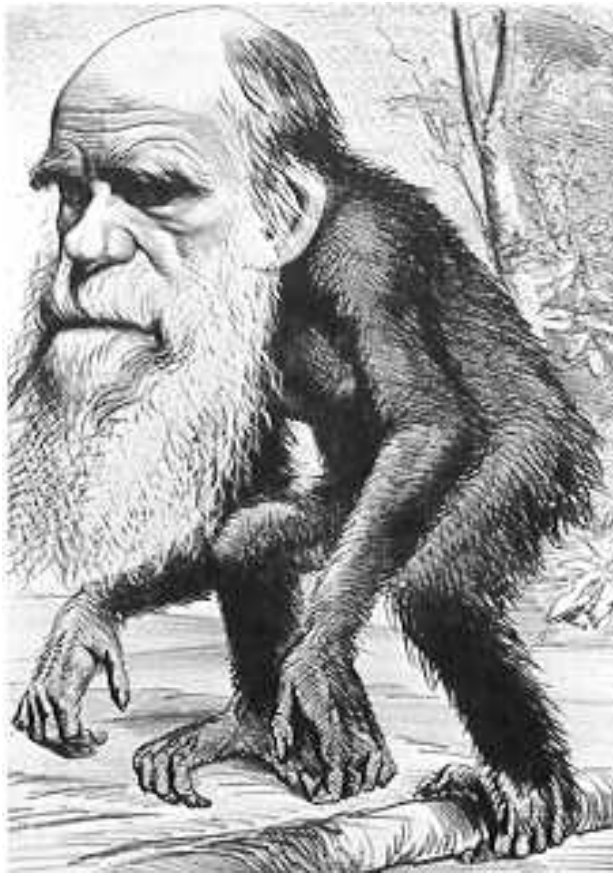




Homologie

Interprétation : de Darwin à nos jours

- Darwin (1859) : L'homologie est le résultat de la descendance avec des modifications d'un ancêtre commun
- La génétique moderne : L'homologie est déterminée par les gènes
- Deux séquences sont homologues si elles sont similaires et partagent un ancêtre commun (la similarité seule n'est pas suffisante)
- De grandes similarités impliquent généralement homologie



Homologue/Orthologues/Paralogues

Orthologie/Paralogie

- Une orthologie désigne un degré de similarité entre deux gènes existants chez deux espèces différentes
- Deux gènes sont considérés comme orthologues, s'ils sont homologues (i.e. descendant d'un même gène ancestral) et issus d'un évènement de spéciation à partir d'un ancêtre commun
- Une paralogie désigne un degré de similarité entre deux gènes existants chez une même espèce
- Deux gènes sont considérés comme paralogues, s'ils sont homologues et issus d'un évènement de duplication à partir d'un ancêtre commun

Plan

Reconstruction phylogénétique

Méthodes

- Il existe deux grands types de méthodes permettant la reconstruction d'arbres phylogénétiques
  - les méthodes basées sur les mesures de distances entre séquences prises deux à deux, c'est à dire le nombre de substitutions de nucléotides ou d'acides aminés entre ces deux séquences.
  - les méthodes basées sur les caractères qui s'intéressent au nombre de mutations (substitutions / insertions /délétions) qui affectent chacun des sites (positions) de la séquence.

Méthodes fondées sur les distances

Recherche d'OTU

- Méthodes de reconstruction d'arbre phylogénétique sans racine basée sur la recherche d'OTU (operationnal taxonomic units, e.g. une séquence) les plus proches et ceci à chaque étape de regroupement.
- Méthodes rapides et donnant de bons résultats pour des séquences ayant une forte similarité.
- Programmes : DNADIST et PROTDIST de Phylip

Méthodes fondées sur les distances

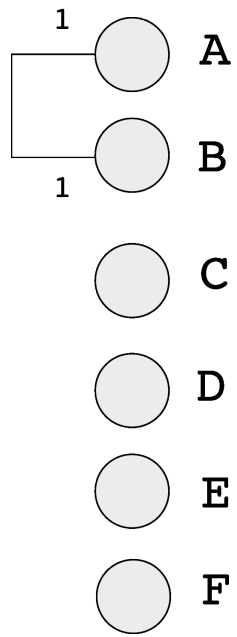
UPGMA (Unweight Pair Group Method with Arithmetic mean)

- Utilisée pour reconstruire des arbres phylogénétiques si les séquences ne sont pas trop divergentes.
- UPGMA utilise un algorithme de clusterisation séquentiel dans lequel les relations sont identifiées dans l'ordre de leur similarité et la reconstruction de l'arbre se fait pas à pas grâce à cet ordre.
- Il y a d'abord identification des deux séquences les plus proches et ce groupe est ensuite traité comme un tout, puis on recherche la séquence la plus proche et ainsi de suite jusqu'à ce qu'il n'y ait plus que deux groupes.

Méthodes fondées sur les distances

UPGMA (Unweight Pair Group Method with Arithmetic mean)

- Exemple : On considère la matrice de distances associée à un groupe de 6 OTUs



	A	B	C	D	E
B	2				
C	4	4			
D	6	6	6		
E	6	6	6	4	
F	8	8	8	8	8

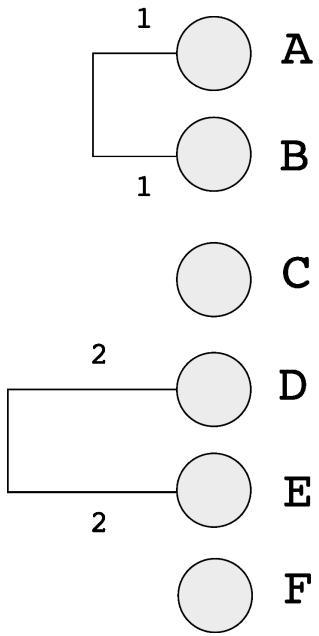
- On clusterise tout d'abord les deux OTUs avec la distance la plus faible (A et B). Le point de branchement est positionné à la distance $\frac{2}{2} = 1$.

Méthodes fondées sur les distances

UPGMA (Unweight Pair Group Method with Arithmetic mean)

- Exemple : On considère la matrice de distances associée à un groupe de 6 OTUs

	A,B	C	D	E
C	4			
D	6	6		
E	6	6	4	
F	8	8	8	8

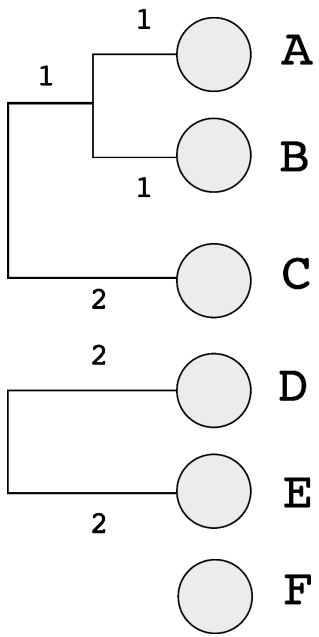


Méthodes fondées sur les distances

UPGMA (Unweight Pair Group Method with Arithmetic mean)

- Exemple : On considère la matrice de distances associée à un groupe de 6 OTUs

	A,B	C	D,E
C	4		
D,E	6	6	
F	8	8	8

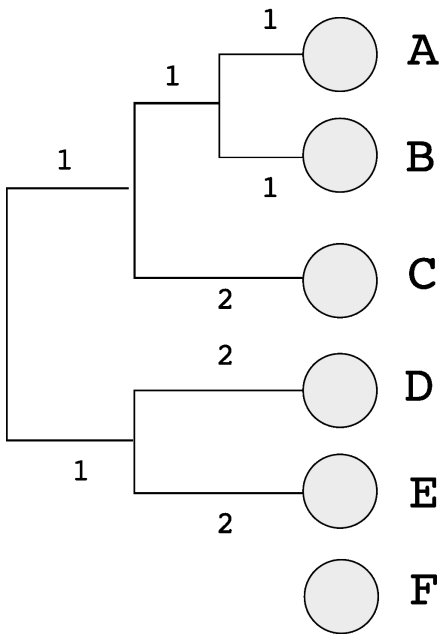


Méthodes fondées sur les distances

UPGMA (Unweight Pair Group Method with Arithmetic mean)

- Exemple : On considère la matrice de distances associée à un groupe de 6 OTUs

	A,B,C	D,E
D,E	6	
F	8	8

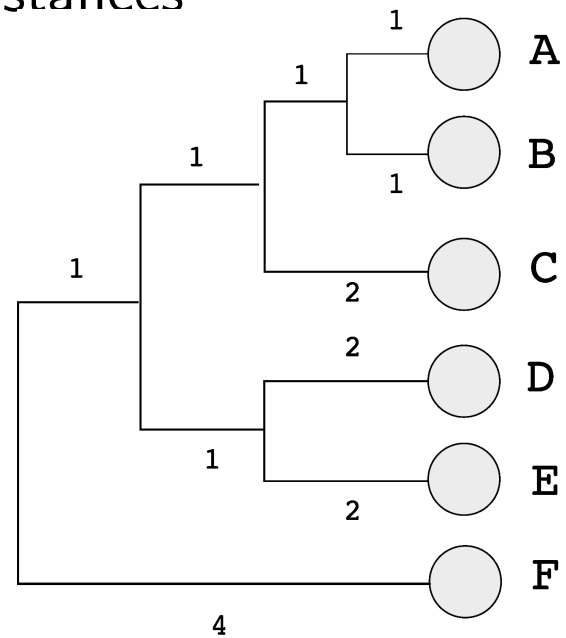


Méthodes fondées sur les distances

UPGMA (Unweight Pair Group Method with Arithmetic mean)

- Exemple : On considère la matrice de distances associée à un groupe de 6 OTUs

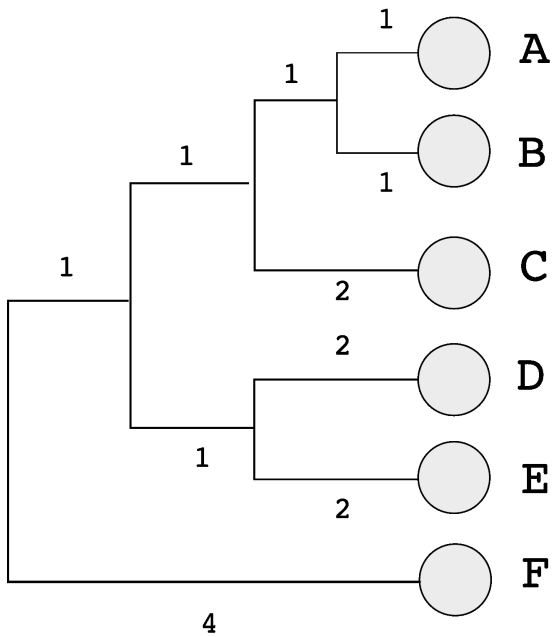
	A,B,C,D,E
F	8



Méthodes fondées sur les distances

UPGMA (Unweight Pair Group Method with Arithmetic mean)

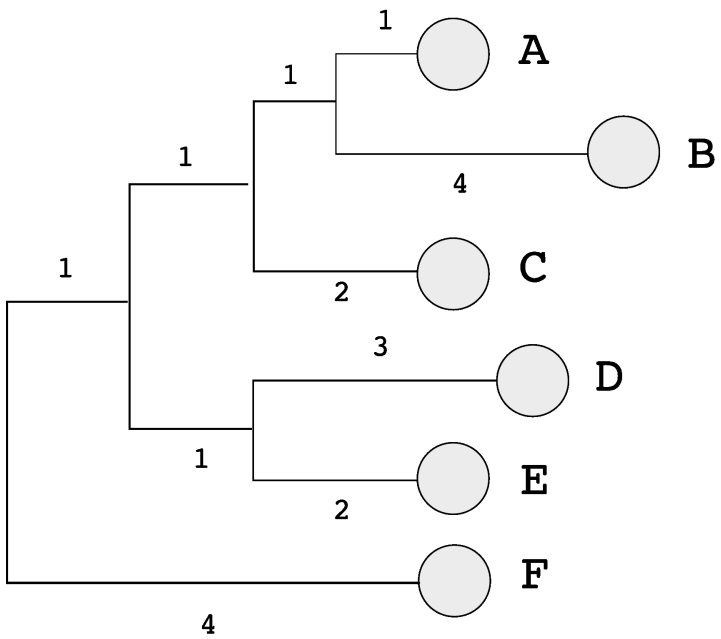
- Cette méthode conduit essentiellement à un arbre non enraciné. Si on veut enraciner l'arbre, on peut appliquer la méthode du "mid-point rooting" : la racine de l'arbre est à équidistance de tous les OTUs soit (ABCDE),F / 2 = 4



Méthodes fondées sur les distances

Les inconvénients de la méthode UPGMA

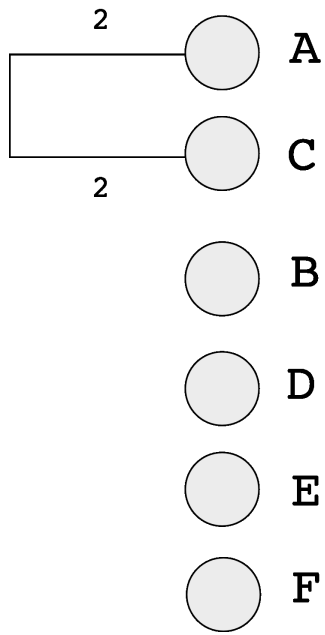
- L'inconvénient majeur est la sensibilité de la méthode à des taux de mutations différents sur les différentes branches
- Supposons que l'on veuille reconstruire l'arbre suivant à partir de la matrice de distances associée aux séquences



Méthodes fondées sur les distances

Les inconvénients de la méthode UPGMA

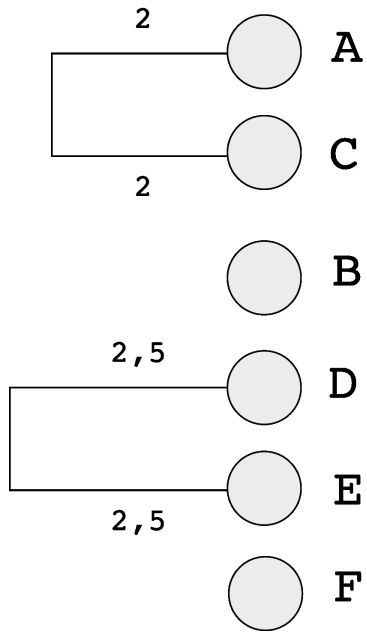
	A	B	C	D	E
B	5				
C	4	7			
D	7	10	7		
E	6	9	6	5	
F	8	11	8	9	8



Méthodes fondées sur les distances

Les inconvénients de la méthode UPGMA

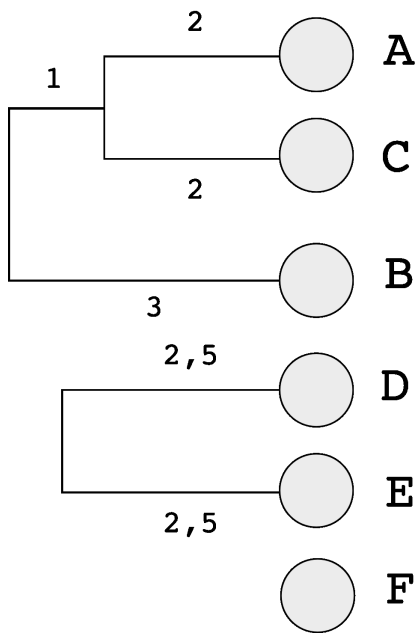
	A,C	B	D	E
B	4			
D	7	10		
E	6	9	5	
F	8	11	8	9



Méthodes fondées sur les distances

Les inconvénients de la méthode UPGMA

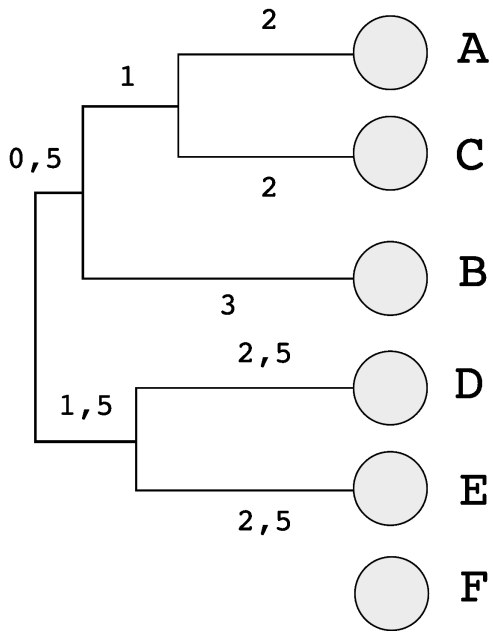
	A,C	B	D,E
B	6		
D,E	6,5	9,5	
F	8	11	8,5



Méthodes fondées sur les distances

Les inconvénients de la méthode UPGMA

	A,C,B	D,E
D,E	8	
F	9,5	9,5

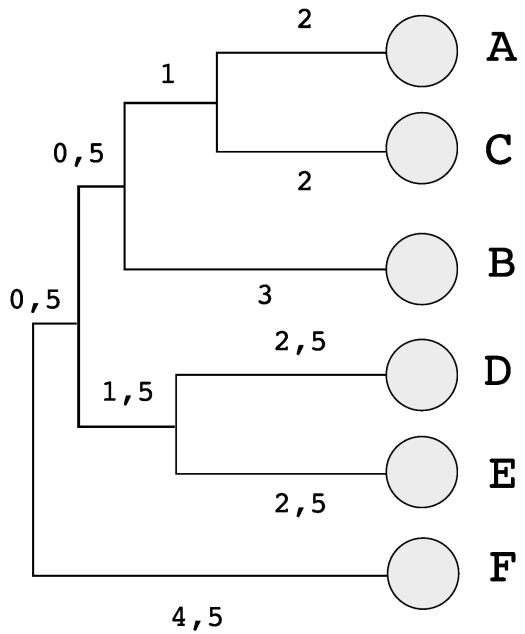


Méthodes fondées sur les distances

Les inconvénients de la méthode UPGMA

	A,C,B,D,E
F	9

- Topologie fausse ! !



Méthodes fondées sur les distances

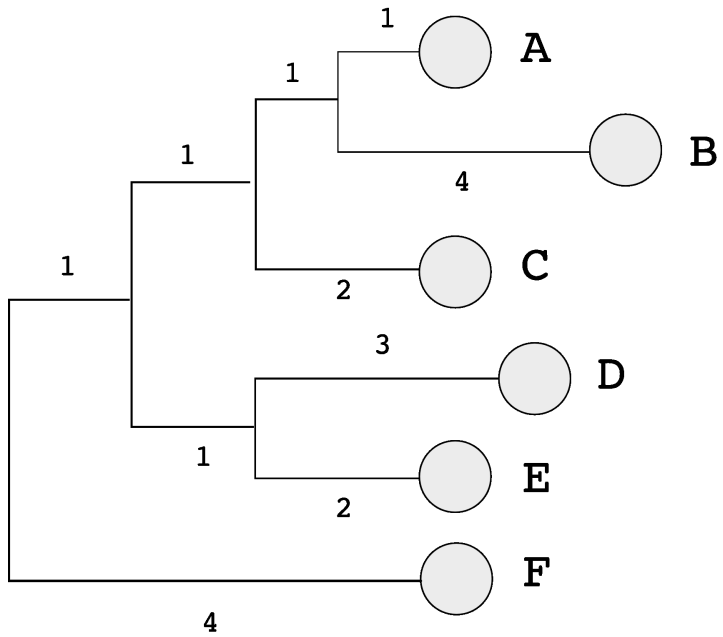
NJ(Neighbor-Joining)

- Méthode – Saitou et Nei (1987) – tentant de corriger UPGMA afin d'autoriser un taux de mutation différent sur les branches.
- Les données initiales permettent de construire une matrice qui donne un arbre en étoile.
- Cette matrice de distances est ensuite corrigée afin de prendre en compte la divergence moyenne de chacunes des séquences avec les autres. L'arbre est alors reconstruit en reliant les séquences les plus proches dans cette nouvelle matrice.
- Lorsque deux séquences sont liées, le noeud représentant leur ancêtre commun est ajouté à l'arbre tandis que les deux feuilles sont enlevées. Ce processus convertit l'ancêtre commun en un noeud terminal dans un arbre de taille réduite.
- Programme NEIGHBOR de Phylip

Méthodes fondées sur les distances

NJ(Neighbor-Joining)

	A	B	C	D	E
B	5				
C	4	7			
D	7	10	7		
E	6	9	6	5	
F	8	11	8	9	8



Méthodes fondées sur les distances

NJ(Neighbor-Joining)

	A	B	C	D	E
B	5				
C	4	7			
D	7	10	7		
E	6	9	6	5	
F	8	11	8	9	8

- Etape 1 : Calcul de la divergence de chacun des N OTUs par rapport aux autres (N= 6)
- $r(A) = 5 + 4 + 7 + 6 + 8 = 30$
- $r(B) = 42$, $r(C) = 32$, $r(D) = 38$, $r(E) = 34$, $r(F) = 44$

Méthodes fondées sur les distances

NJ(Neighbor-Joining)

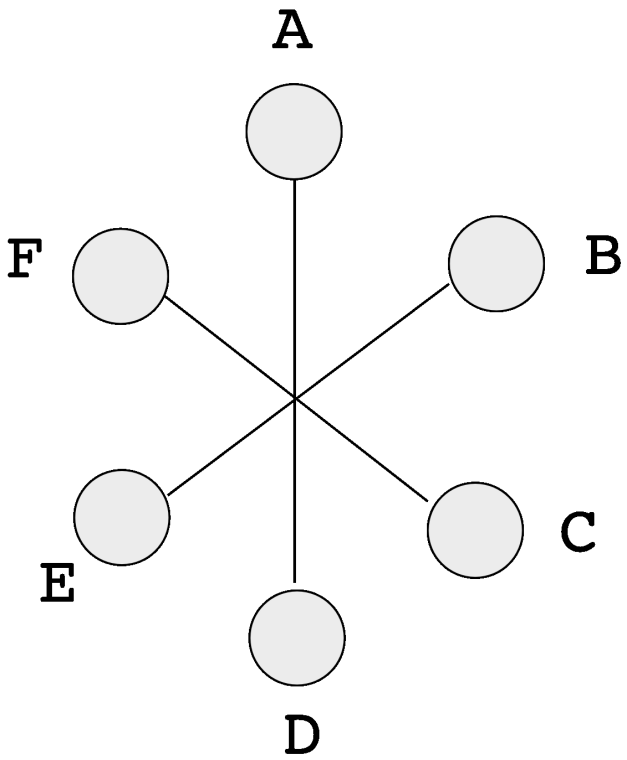
	A	B	C	D	E
B	-13				
C	-11,5	-11,5			
D	-10	-10	-10,5		
E	-10	-10	-10,5	-13	
F	-10,5	-10,5	-11	-11,5	-11,5

- Etape 2 : cacul de la nouvelle matrice en utilisant la formule
- $M(i,j) = d(ij) - [r(i) + r(j)]/(N-2)$
- Ce qui donne pour la paire
  AB : $M(AB) = 5 - [30 + 42]/4 = -13$

Méthodes fondées sur les distances

NJ(Neighbor-Joining)

- Ceci permet de construire l'arbre en étoile suivant



Méthodes fondées sur les distances

NJ(Neighbor-Joining)

	A	B	C	D	E
B	-13				
C	-11,5	-11,5			
D	-10	-10	-10,5		
E	-10	-10	-10,5	-13	
F	-10,5	-10,5	-11	-11,5	-11,5

- Etape 3 : choix des plus proches voisins, c'est à dire des deux OTUs ayant le M(i,j) le plus petit, donc soit A et B ou D et E.
- On prend A et B et on forme un nouveau noeud U et on calcule la longueur de la branche entre U et A ainsi qu'U et B :
  - $S(AU) = d(AB)/2 + [r(A) - r(B)]/2(N-2) = 1$
  - $S(BU) = d(AB) - S(AU) = 5 - 1 = 4$

Méthodes fondées sur les distances

NJ(Neighbor-Joining)

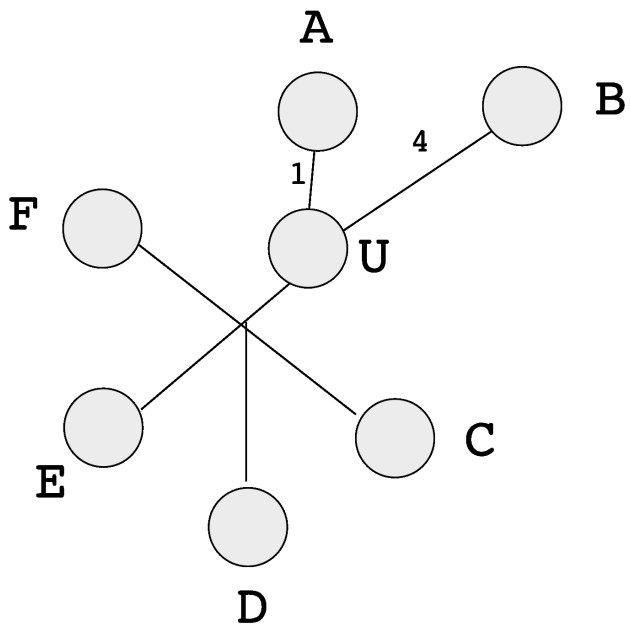
	U	C	D	E
C	3			
D	6	7		
E	5	6	5	
F	7	8	9	8

- Etape 4 : on définit les nouvelles distances entre U et les autres OTUs
  - $d(CU) = d(AC) + d(BC) - d(AB)/2 = 3$
  - $d(DU) = d(AD) + d(BD) - d(AB)/2 = 6$
  - $d(EU) = d(AE) + d(BE) - d(AB)/2 = 5$
  - $d(DU) = d(AF) + d(BF) - d(AB)/2 = 7$

Méthodes fondées sur les distances

NJ(Neighbor-Joining)

- La procédure complète repart de l'étape 1 avec $N = N - 1 = 5$.



Méthodes fondées sur les caratères

Parcimonie

- Méthode très lente mais précise
- La parcimonie consiste à minimiser le nombre de "pas" (mutations / substitutions) nécessaires pour passer d'une séquence à une autre dans une topologie de l'arbre.
- Pour cela, cette méthode s'appuie sur les hypothèses suivantes :
  - les sites évoluent indépendemment les uns des autres (la séquence peut être considérée comme une suite de caractères non ordonnés)
  - la vitesse d'évolution est lente et constante au cours du temps.

Méthodes fondées sur les caratères

Parcimonie

- La méthode de maximum de parcimonie recherche toutes les topologies possibles afin de trouver l'arbre optimal (mimimum) et le temps nécessaire pour cette exploration croit rapidement avec le nombre de séquences :
  - le nombre d'arbres enracinés possibles pour n OTUs : $Nr = (2n-3)!/(2exp(n-2))(n-2)!$
  - le nombre d'arbres non enracinés possibles pour n OTUs : $Nu = (2n-5)!/(2exp(n-3))(n-3)!$
- e.g. pour $n = 9$, 135135 arbres non enracinés et 34459425 arbres enracinés possibles
- Programme : DNAPARS et PROTPARS de Phylip

Méthodes fondées sur les caratères

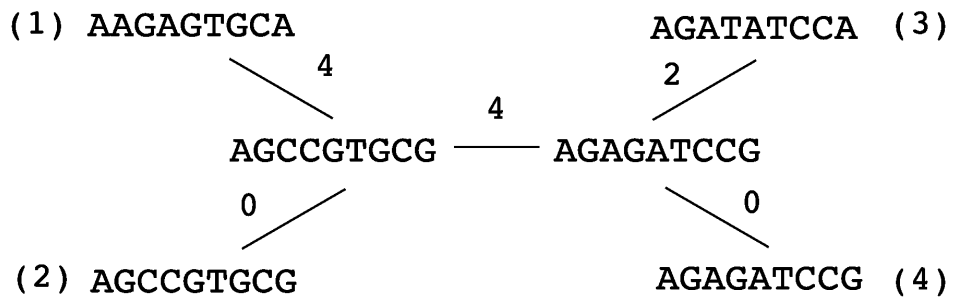
Exemple

- 1 – AAGAGTGCA
- 2 – AGCCGTGCG
- 3 – AGATATCCA
- 4 – AGAGATCCG
- Pour 4 séquences, il y a 3 arbres non enracinés possibles. Ces trois arbres sont analysés (recherche de la séquence ancestrale et comptage du nombre de mutations)

Méthodes fondées sur les caratères

Exemple

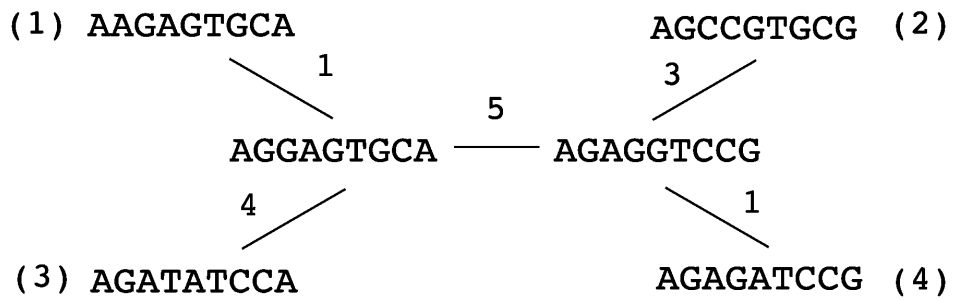
- Nombre de mutations : 10



Méthodes fondées sur les caratères

Exemple

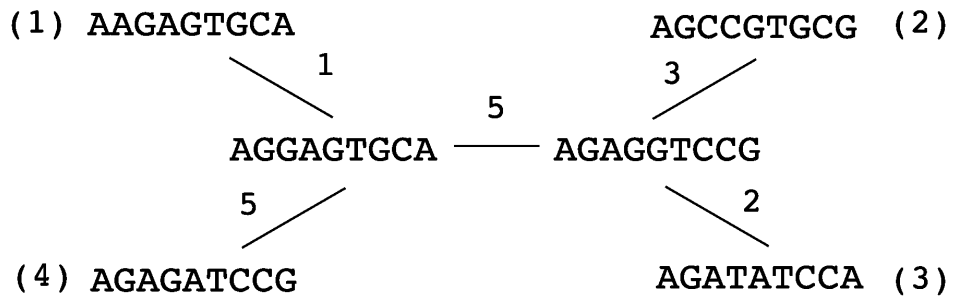
- Nombre de mutations : 14



Méthodes fondées sur les caratères

Exemple

- Nombre de mutations : 16



- Le premier arbre est celui nécessitant le moins de mutations, c'est donc le plus parcimonieux.
- Cette analyse prend en compte tous les sites des séquences mais l'analyse peut également se faire uniquement sur les sites informatifs, c'est à dire quand à cette position il y a au moins 2 nucléotides différents, représentés chacun dans au moins deux séquences.

Méthodes fondées sur les caratères

Maximum de parcimonie

- Le maximum de parcimonie recherche l'arbre optimal et dans ce processus, il est possible d'en trouver plusieurs
- Une évaluation de toutes les topologies possibles est néccessaire mais impossible dès 12 séquences.
- Branch and Bound : cette méthode est dérivée du maximum de parcimonie, elle garantit de trouver le meilleur arbre mais sans évaluer tous les arbres possibles. Elle permet de traiter un plus grand nombre de séquences mais reste limitée
- Recherche heuristique : il y a un réarrangement des branches à chaque étape (rien ne garantit de trouver l'arbre optimal)
- Arbre consensus : l'arbre consensus est construit à partir des noeuds les plus fréquemment rencontrés sur l'ensemble des arbres possibles.

Méthodes fondées sur les caratères

Avantages et inconvénients de la parcimonie

- Avantages :
  - Basée sur les caractères : méthode cladistique plutôt que phénétique.
  - Ne réduisant pas la séquence à un simple nombre.
  - Essayant de donner une information sur les séquences ancestrales.
  - Évaluant différents arbres.
- Inconvénients :
  - Très lente par rapport aux méthodes basées sur les distances.
  - N'utilisant pas toute l'information disponible (seuls les sites informatifs sont pris en compte)
  - Ne faisant pas de corrections pour les substitutions multiples
  - Ne donnant aucune information sur la longueur des branches

Méthodes fondées sur les caratères

Enraciner un arbre

- Le plus souvent, les méthodes de reconstruction phylogénétiques aboutissent à des arbres non enracinés. Pour enraciner un arbre, on peut ajouter une séquence dont on sait qu'elle est beaucoup plus ancienne que toutes les autres
- Cependant, il ne faut pas que la séquence choisie soit
  - trop éloignée des autres données. En effet, cela peut conduire à des erreurs dans la topoolgie de l'arbre.
  - soit trop proche des séquences car dans ce cas, cela n'est peut-être pas un vrai "outgroup".
- L'utilisation de plus d'un "outgroup" améliore en général l'évaluation de l'arbre.
- Enfin, en l'absence d'un bon "outgroup", la racine peut être positionnée approximativement à égale distance de toutes les séquences : on parle alors de mid-point rooting.

Récapitulatif

Distances

- SEQUENCES : Très proches
- AVANTAGES : Rapides et faciles à mettre en oeuvre
- INCONVENIENTS : Tous les sites sont traités de manière équivalente d'où un perte d'informations ; non applicables à des séquences éloignées
- PROGRAMMES : DNAdist, Protdist, FITCH, KITSCH
- REMARQUES : Il vaut mieux utiliser le Neigbor-joining plutôt qu'UPGMA car Nj autorise des taux de mutations différents le long des branches

Récapitulatif

Parcimonie

- SEQUENCES : Relativement éloignées
- AVANTAGES : Evaluation de différents arbres – Essaie de donner des informations sur les séquences ancestrales
- INCONVENIENTS : Lente – Inutilisable lorsque l'on a un grand nombre de séquences
- PROGRAMMES : DNApars, PROTpars
- REMARQUES : On peut obtenir plusieurs arbres équivalents et dans ce cas le choix de l'un par rapport aux autres peut être difficile à justifier

Plan

Bibliographie

Références

1. Touzet, H and Varré, S.. *Cours*, `http ://www2.lifl.fr/SEQUOIA/members.php`.
2. Tompa, M.. *Computational Biology*, Course in english, `http ://www.cs.washington.edu/education/courses/527/00w`
3. J. Setubal and J. Meidanis. *Introduction to Computational Molecular Biology*. PWS Publishing Co, 1997.