Université des frères Mentouri Constantine 1

Faculté : SNV

Département : Biologie Appliquée

Licence : Biotechnologie microbienne

Matière : Bioinformatique

Enseignante : Dr BENHAMDI A.

1. Théorie de l'évolution et phylogénie

- Tous les organismes vivants dérivent d'un ancêtre commun.
- La diversité est due à la séparation d'une espèce en deux espèces différentes.
- Idée de base : les caractères sont transmis d'une génération à l'autre, au cours de l'évolution, et ces caractères subissent une série de mutations

La phylogénie est l'étude des relations de parenté entre différents êtres vivants en vue de comprendre l'évolution des organismes vivants. D'après Darwin (1859) les êtres vivants descendent tous les uns des autres et la phylogénie désigne les lignes généalogiques de tous les êtres organisés. Le terme de phylogénie en lui-même date de la fin du 19^{e} siècle (inventé par Haeckel, 1866) et avait le sens de « l'enchaînement des espèces animales ou végétales au cours du temps ».

Le but de la reconstruction phylogénétique est de :

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

etc.

En phylogénie, on représente couramment les liens de parenté par un arbre phylogénétique. On peut diviser les données qui vont nous servir pour la construction d'arbres phylogénétiques en deux groupes distincts :

- Les données liées aux caractères phénotypiques.
- Les données moléculaires telles que les séquences d'ADN ou de protéines.

En fait ces données concernent les caractères morphologiques, physiologiques, génétiques et génomiques.

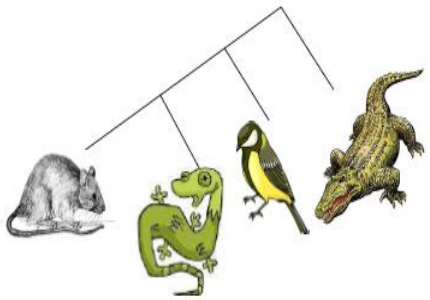
1.1. **Les données phénotypiques** : elles comprennent les caractères observables (liés aux différents états : morphologiques, biochimiques et physiologiques) et les patterns binaires (présence d'un caractère donné / absence de ce même caractère).

- **Exemple**

	écailles	Ovipares	membrane nictitante (paupière de l'œil)	Gésier
rat	0	0	0	0
oiseau	0	1	1	1
Lézard	1	1	0	0

crocodile	1	1	1	1

	rat	oiseau	lézard	croco
écailles	0	0	1	1
ovipare	0	1	1	1
oeil	0	1	0	1
gésier	0	1	0	1

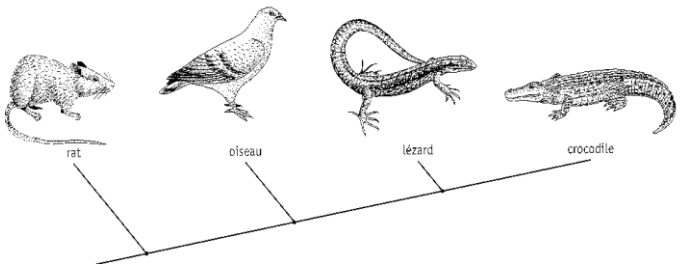


On voit que l'arbre est construit à partir de 4 caractères pour 4 espèces qui ont toutes une origine évolutive commune : ce sont des vertébrés. Dans cet exemple, et en se basant sur la théorie de l'évolution, le rat est considéré comme l'espèce la plus ancienne de ces vertébrés mais au cours du temps, il y a eu une séparation des lignées par l'apparition d'autres caractères : le lézard a l'oviparité de plus, c'est donc l'espèce qui vient après le rat puis l'oiseau et enfin le crocodile.

D'après les données précédentes, les vertébrés les plus proches du lézard sont le rat et l'oiseau.

Limites des données phénotypiques

	rat	oiseau	lézard	croco
écailles	0	0	1	1
ovipare	0	1	1	1



En ne prenant en considération que les deux caractères (présence d'écailles et oviparité) on constate que les vertébrés les plus proches du lézard sont l'oiseau et le crocodile. Ainsi selon cette démarche, la position des espèces dans cet arbre et le degré de parenté existant entre les différents individus peuvent être changés avec la suppression ou l'addition d'un seul caractère.

Un arbre est de ce fait réfutable, c'est-à-dire qu'il est considéré comme juste jusqu'au moment où de nouveaux arguments le complètent voire le modifient, actuellement les progrès de la biologie moléculaire, la découverte de nouvelles espèces, l'importance que l'on donne aux caractères…. permettent d'apporter de nouvelles informations sur les liens de parenté entre espèces.

Exercice d'application :

Taxon / Caractère	Lamproie	Bacille	Myxine	Tortue	Kangourou	Méduse	Chauve-souris
Vertèbres	+	-	-	+	+	-	+
Crâne	+	-	+	+	+	-	+
Mâchoires	-	-	-	+	+	-	+
Poils	-	-	-	-	+	-	+
Cellule eucaryote	+	-	+	+	+	+	+

+ : présence de l'état dérivé du caractère

- : absence de l'état dérivé du caractère

- A partir du tableau ci-dessus, tracer l'arbre phylogénétique correspondant.
- Quelles sont les caractéristiques de l'ancêtre commun le plus ancien.
- Quelles sont les espèces les plus proches à la tortue.
- En ne prenant en considération que les deux caractères : vertèbres et crâne établir le nouvel arbre.
- Qu'est ce que vous remarquez ?

1.2. Les données moléculaires

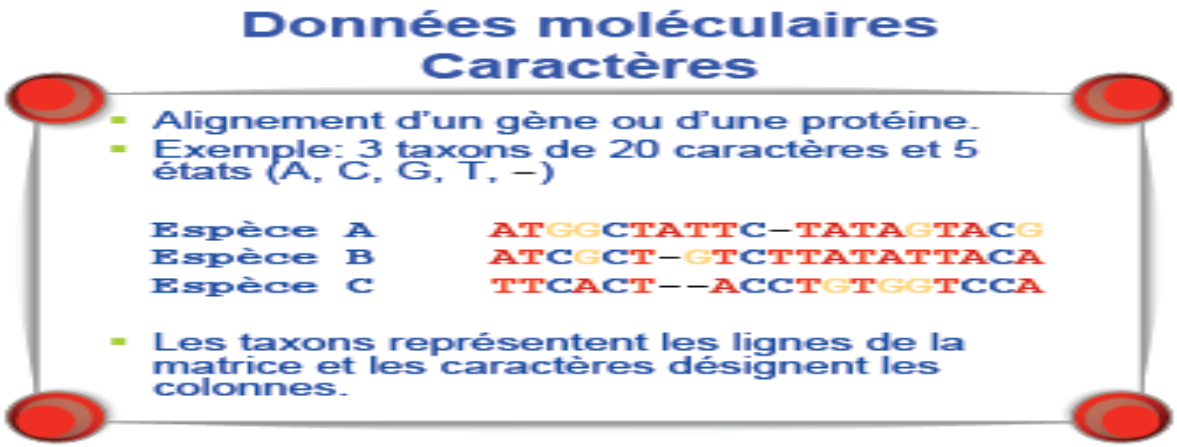
Jusqu'aux années 1960, les comparaisons entre les morphologies, les comportements et les répartitions géographiques des espèces étaient les seuls moyens disponibles pour construire des classifications d'espèces. La découverte que des protéines homologues (ou acides nucléiques) avaient des séquences en acides aminés (ou en bases) qui variaient d'une espèce à l'autre a fourni un nouveau moyen d'étude. L'évolution moléculaire est une extension de la théorie darwinienne qui a donné naissance à la phylogénie moléculaire. E. Zuckerkandl et L. Pauling en 1965, puis W. Fitch et E. Margoliash en 1967 apparaissent comme les pionniers des phylogénies moléculaires. La progression des séquences et des programmes informatiques disponibles a fait paraître, à partir de la fin des années 1980, de nombreuses études phylogénétiques moléculaires portant sur des données se rapportant aux : séquences moléculaires (phylogénie moléculaire), caractères discrets, fréquences des gènes, sites de restriction et microsatellites.

1.1.1. **Définition :** La phylogénie moléculaire est l'utilisation de séquences de gènes et de protéines pour obtenir des informations sur l'histoire évolutive des êtres vivants et notamment sur leurs liens de parenté (leur phylogénie). Elle renseigne sur les changements survenus au niveau des espèces au cours de l'évolution et tente d'établir des filiations et des liens de parenté entre espèces différentes à partir de l'analyse de séquences.

Elle permet notamment de distinguer les espèces orthologues et paralogues :

- Orthologue : séquences homologues de même origine et provenant d'espèces différentes.
- Paralogues séquences homologues d'un même organisme ayant divergé par duplication d'un même gène ancestral.

1.1.2. Exemple :



Définition:
• La phylogénie moléculaire est la discipline ayant pour objectif la reconstruction de l'histoire évolutive des espèces par comparaison des séquences de leurs gènes ou de leurs protéines.

Données:
• Un ensemble d'organismes (taxa) et pour chacun un ensemble de données moléculaires (séquences par exemple)

Lexique phylogéniques :

- **Un ancêtre commun :** à deux ou plusieurs espèces correspond à l'organisme parent hypothétique le plus proche dans le temps desdites espèces.

- **Caractère :** tout attribut (propriété) observable d'un organisme (écologique, morphologique, comportemental, moléculaire, physiologique…).

- **État de caractère** : forme particulière d'un caractère

Ex : Caractère : couleur de l'oeil : état de caractère : bleu, marron, vert, absence d'œil.

- **Etat ancestral ou primitif** état le plus ancien d'un caractère

- **Taxon** groupe au sens large : espèce, famille, classe…

- **Matrice Taxons caractères** tableau à double entrée présentant les états des caractères pour chacun des taxons

- **Matrice (de caractères) n.f.** Tableau à double entrée comportant en général verticalement une série d'espèces ou de taxons et horizontalement une série de caractères

- **Matrice de distances (génétiques)** tableau à double entrée Il comporte verticalement une série d'espèces ou de taxon et horizontalement cette même série dans chacune indiquant le nombre (ou le %) de différences existant dans la séquence de molécules homologues de différents taxons comparés 2 par 2.

- **Alignement de séquences** (ou alignement séquentiel) est une manière de représenter deux ou plusieurs séquences de macromolécules biologiques (ADN, ARN ou protéines) les unes sous les autres, de manière à en faire ressortir les régions homologues ou similaires.

- **Identité :** égalité parfaite entre deux séquences.

- **Similitude** : ressemblance entre 2 ou plusieurs séquences, on peut mesurer un taux de ressemblance : pourcentage de similarité = pourcentage d'identité + pourcentage de substitutions conservatives par exemple acide aspartique -> acide glutamique

- **Homologie** : Deux séquences sont dites homologues si elles possèdent un ancêtre commun.

- **Arbre phylogénétique** schéma illustrant les liens de parenté entre des taxons.

- **Clade (ou groupe monophylétique)** ensemble constitué par un ancêtre et tous ses descendants.

- **Mutation n.f.** Désigne tout type d'altération du génome* qui n'est pas réparée. C'est en général une erreur de réplication dans une séquence d'ADN*

Rappel sur les méthodes de comparaison de séquences et d'alignements

La comparaison de deux séquences protéique ou nucléotidiques a pour but de repérer les régions identiques ou très proches de ces deux séquences, elle se fait soit par la :

1 : recherche de segments identiques : Exemple : méthode de Dumas et Ninio (1982) codification numérique des séquences.

2 : recherche de segments similaires : Exemple la méthode de Dotplot : matrices de points.

3: recherche d'alignements optimaux entre 2 séquences :

- Alignement global : alignement de 2 séquences sur la totalité de leur longueur en tenant compte de tous les résidus. Il permet de mesurer le degré de similarité entre deux séquences ex : Needleman-Wunsch (1970)
- Alignement local : il s'intéresse à une partie de la séquence à comparer au lieu de sa totalité. Il permet de mettre en évidence des zones fortement similaires voire identiques ex : l'Algorithme de Smith & Waterman (1981).

Illustration du codage numérique

La comparaison matricielle des deux séquences sous forme de chaîne d'entiers permet de localiser ensuite sur les séquences tous les endroits possédant des segments communs de longueur rédéfinie par le codage. Pour cela il suffit de repérer les positions des séquences où les codes sont identiques.

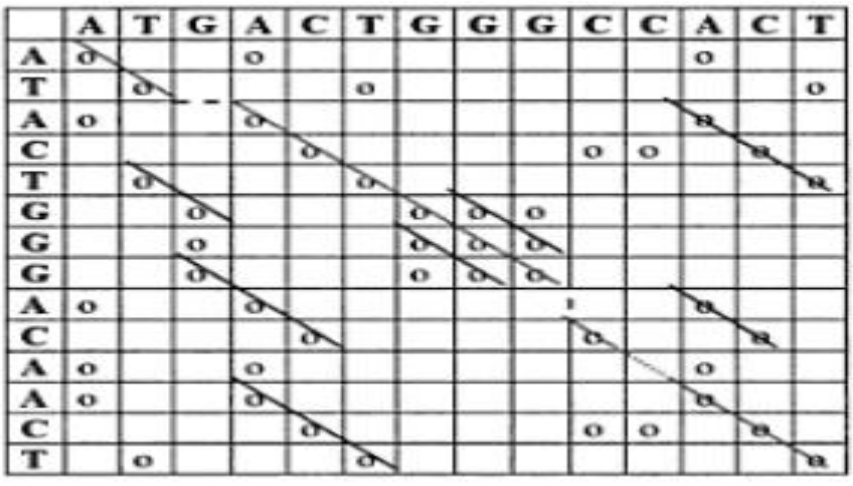
Mots de 4 caractères

```
SEQ TEST      A  C  G  T  C  G  T  T  C  G  A  T  T  A  (N=14)
.  .  .  .
1 ACGT       ----------
2 CGTC         ----------
3 GTCG           ----------
4 TCGT             ----------
5 CGTT               ----------
6 GTTC                 ----------
7 TTCG                   ----------
8 TCGA                     ---------
9 CGAT                       ---------
10 GATT                        ---------
```

11 ATTA ----------

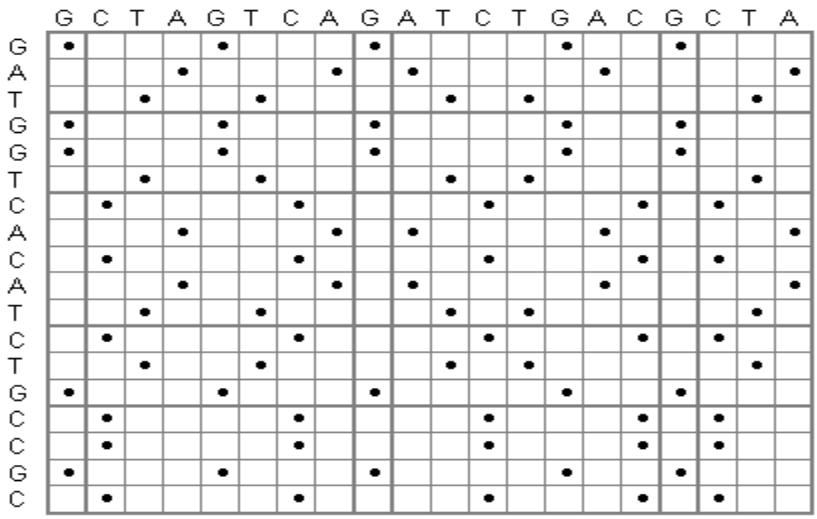
SEQ BANQUE T C G A C G C G G A T (M=11)
Le mot TCGA est commun aux deux séquences
Mots de 5 caractères
SEQ TEST A C G T C G T T C G A T T A (N=14)
. . . .
1 ACGTC ----------
2 CGTCG ----------
3 GTCGT ----------
4 TCGTT ----------
5 CGTTC ----------
6 GTTCG ----------
7 TTCGA ----------
8 TCGAT ----------
9 CGATT ----------
10 GATTA ----------
SEQ BANQUE T C G A C G C G G A T (M=11)
On ne retrouve plus le motif commun entre les 2 séquences ce qui montre bien que plus la taille du mot est importante, plus la probabilité de trouver un motif commun entre les deux séquences est faible.

La méthode de Dot plot : Elle permet de mettre en évidence toutes les portions identiques entre deux séquences comparées. Dans le cas d'une identité parfaite des 2 séquences, le résultat de DotPlot sera une diagonale



Exercice d'application

Déterminer les régions similaires entre les deux séquences suivantes :

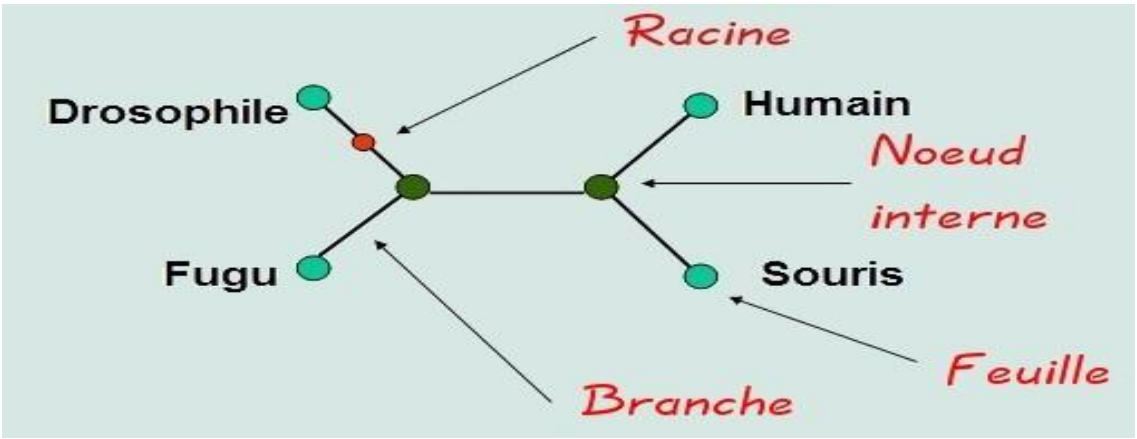


Les arbres phylogénétiques :

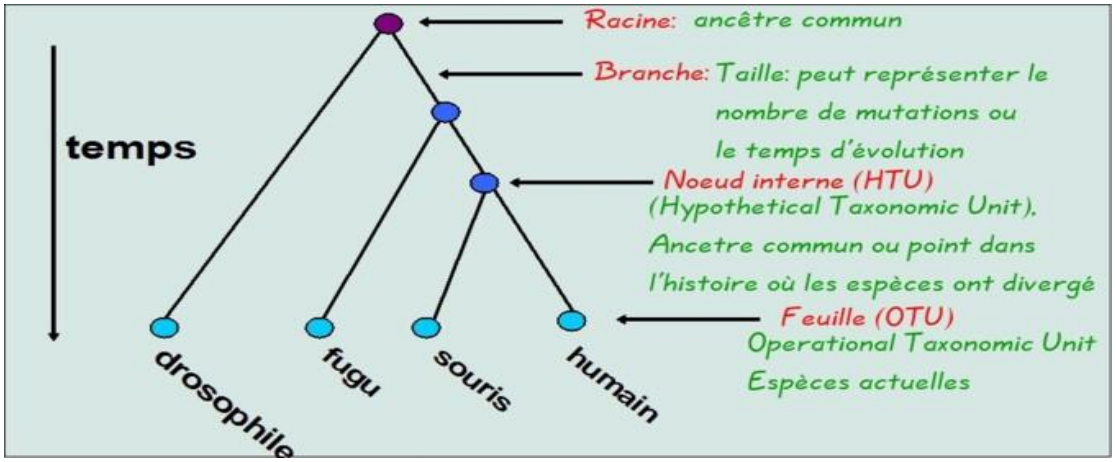
Arbres phylogénétique

Arbre : = Représentation graphique de la phylogenèse d'un groupe de taxons (séquences).

Arbre non enraciné : graphe connexe non cyclique. Il n'y a qu'un seul et unique chemin pour passer d'un sommet à l'autre,



Arbre enraciné : possède une contrainte supplémentaire par rapport au précédent. Présence de liens orientés depuis une origine ou un ancêtre,



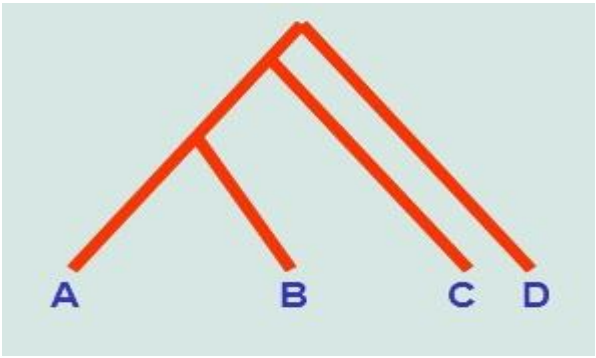
Branche : relation entre deux taxons (UEH et/ou UE),

Les variétés d'arbres

Dendrogramme : arbre exprimant les liens entre taxons sous la forme d'une succession de branchements. Il existe plusieurs types de dendrogrammes selon les méthodes avec lesquelles ils ont été construits.

Cladogramme : dendrogramme exprimant les relations phylogénétiques (de parenté) entre taxons et construit à partir d'une analyse cladistique. Chacun des points de branchements ou noeuds, est défini par une ou plusieurs synapomorphies.

Cladogramme

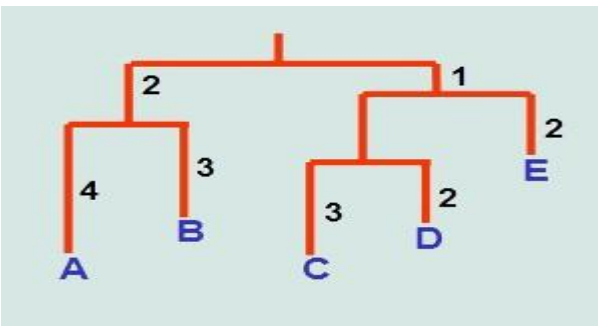


Indique simplement les relations d'ancêtre entre les espèces.

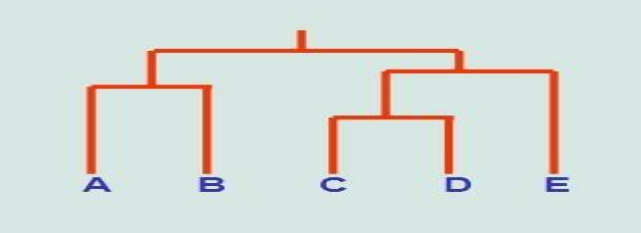
Ex: Les espèces A et B ont un ancêtre commun plus récent que les espèces A et C

Phénogramme : un dendrogramme obtenu par méthodes de distance où les relations entre taxa expriment des degrés de similitude globale;

Phylogramme : C'est un cladogramme dont la longueur des branches est proportionnelle au nombre de changements évolutifs.



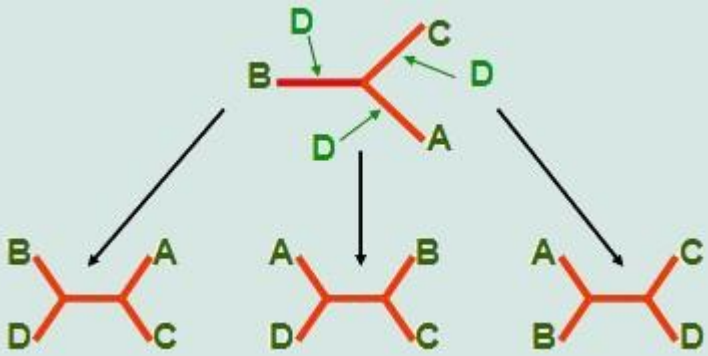
Arbres additifs: Longueur des branches proportionnelle au nombre de changements évolutifs



Arbres additifs où les feuilles sont équidistantes de la racine.

Détermination du nombre d'arbres

Énumération de tous les arbres possibles



3 arbres non enracinés pour 4 espèces:

Enracinement des arbres phylogénétiques

Pour chacun de ces arbres, on a 5 arbres avec racines donc: 3 x 5 = 15 arbres racinés pour n= 4 espèces

Il y a autant de racines possibles que de branches dans un arbre non raciné

Nombre d'arbres possibles		
Nombre de taxons	**Nombre d' arbres non enracinés**	**Nombre Arbres enracinés**
2	1	1
3	1	3
4	3	15
5	15	105
6	105	945
7	945	10 395
8	10 395	135 135
9	135 135	2 027 025

Nombre d'arbres : Le nombre théorique d'arbres phylogénétiques dépend du nombre d'OTUs qui entrent dans la construction phylogénétique

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)!

Autre méthode :

$$N_{\text{arbres non-enracinés}} = \prod_{i=3}^{s} (2i - 5) \qquad N_{\text{arbres enracinés}} = \prod_{i=2}^{s} (2i - 3)$$

S : étant le nombre d'individus ou de taxons

Exemple pour S = 4 taxons

- N non raciné = (2x3 – 5) x (2x4 – 5) = (6-5) x (8-5) = 1x3 = 3 arbres possibles
- N raciné = (2x2 -3) x (2x3 – 3) x (2x4 – 3) = 1 x 3 x 5 = 15 arbres possibles

Exemple pour S = 5 taxons

- N non raciné = (2x3 – 5) x (2x4 – 5) x (2x5 – 5) = 1x3x5 = 15 arbres possibles
- N raciné = (2x2 -3) x (2x3 – 3) x (2x4 – 3) x (2x5 – 3) = 1 x 3 x 5 x 7 = 105 arbres possibles

Nombre d'OTUs	Nombre d'arbres non racinés	Nombre d'arbres racinés
3	1	3
4	3	15

5	15	105
6	105	945
7	945	10 395
8	10 395	135 135
9	135 135	2 027 025
10	2 027 025	34 459 425

Méthodes de reconstruction des arbres

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, ex : UPGMAet NJ
- 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, ex : la parcimonie

Méthodes fondées sur les distances

Ce sont des méthodes de reconstruction d'arbre phylogénétique sans racine basée sur la recherche d'OTU (operationnal taxonomic units, le plus souvent équivalent à une séquence) les plus proches et ceci à chaque étape de regroupement.
Ces méthodes sont rapides et donnent de bons résultats pour des séquences ayant une forte similarité.

UPGMA (Unweight Pair Group Method with Arithmetic mean)

Cette méthode est 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.

Exemple

On considère la matrice de distances associé 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 2/2=1.
On peut alors construire le sous arbre suivant :
Dans la suite, le cluster (A,B) est considéré comme un tout et on peut calculer une nouvelle matrice de distance :
dist(A,B),C = (distAC + distBC) / 2 = 4
dist(A,B),D = (distAD + distBD) / 2 = 6
dist(A,B),E = (distAE + distBE) / 2 = 6
dist(A,B),F = (distAF + distBF) / 2 = 8

MATRICE **ARBRE**

Cycle 1

	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

Cycle 2

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

Cycle 3

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

Cycle 4

	AB,C	**D,E**
D,E	6	
F	8	8

Cycle 5

	ABC,DE
F	8

Cette méthode conduit essentiellement à unn 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

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.

NJ(Neighbor-Joining)

Cette méthode développée par Saitou et Nei (1987) tente de corriger la méthode 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.

Exemple

La matrice de distance associée à cet arbre est la suivante :

	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

Etape 2 : cacul de la nouvelle matrice en utilisant la formule

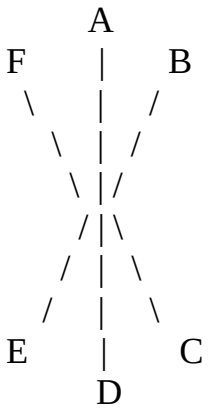
M(i,j) = d(ij) -[r(i) + r(j)] / (N-2)

ce qui donne pour la paire AB : M(AB) = 5 - [30 + 42] / 4 = -13

	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

Ceci permet de construire l'arbre en étoile suivant :



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 soit 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'entre U et B :

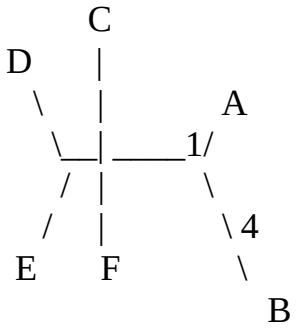
S (AU) = d (AB) / 2 + [r(A) - r(B)] / 2 (N-2) = 5/2 + [30-42] /2(6-4) = 1
S(BU) = d (AB) - S(AU) = 5 - 1 = 4

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

création d'une nouvelle matrice :

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

Et d'un arbre en étoile :



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

La parcimonie