4. Phylogénie

4.1. Comparaison de séquences

La première question que se pose le biologiste lorsqu'il a obtenu une séquence est : « Y a-t-il dans la banque de données une ou plusieurs séquences qui ressemblent à la mienne? ». La réponse à cette question nécessite de définir la ressemblance entre séquences. L'alignement de deux séquences est la base de cette comparaison.

La comparaison de séquences est la tâche informatique la plus utilisée par les biologistes. Il s'agit dans quelle mesure deux séquences, génomiques, se ressemblent.

Ainsi, si deux séquences sont très similaires et si l'une est connue pour être codante, l'hypothèse que la seconde le soit aussi peut être avancée. Un biologiste qui détient une nouvelle séquence s'intéresse en premier temps à parcourir ces bases de données, afin d'y trouver les séquences similaires et de faire hériter à la nouvelle séquence les connaissances qui leur sont associées. C'est également en comparant des séquences de génomes d'espèces actuelles qu'il est possible de reconstruire des arbres phylogénétiques qui rendent compte de l'histoire évolutive.

Confus par la variété de la vie, parmi les premières activités biologiques de l'homme était la classification. Les biologistes étaient impliqués dans la question d'obtenir une classification hiérarchique de toutes les espèces en cohérence avec leur relation évolutionnaire, aussi connue sous le non de l'arbre de la vie. Ce qui a fait de la construction d'arbres une activité centrale des biologistes, mais aussi pour comprendre les,similarités fonctionnelles des organismes. L'évolution requière trois ingrédients basiques:,reproduction, avec variation et sélection.

La Figure 5 définie l'évolution par la variabilité génétique, son mécanisme par la mutation, son explication par la variabilité écologique et son objectif pour l'adaptation, avec quelques exemples.

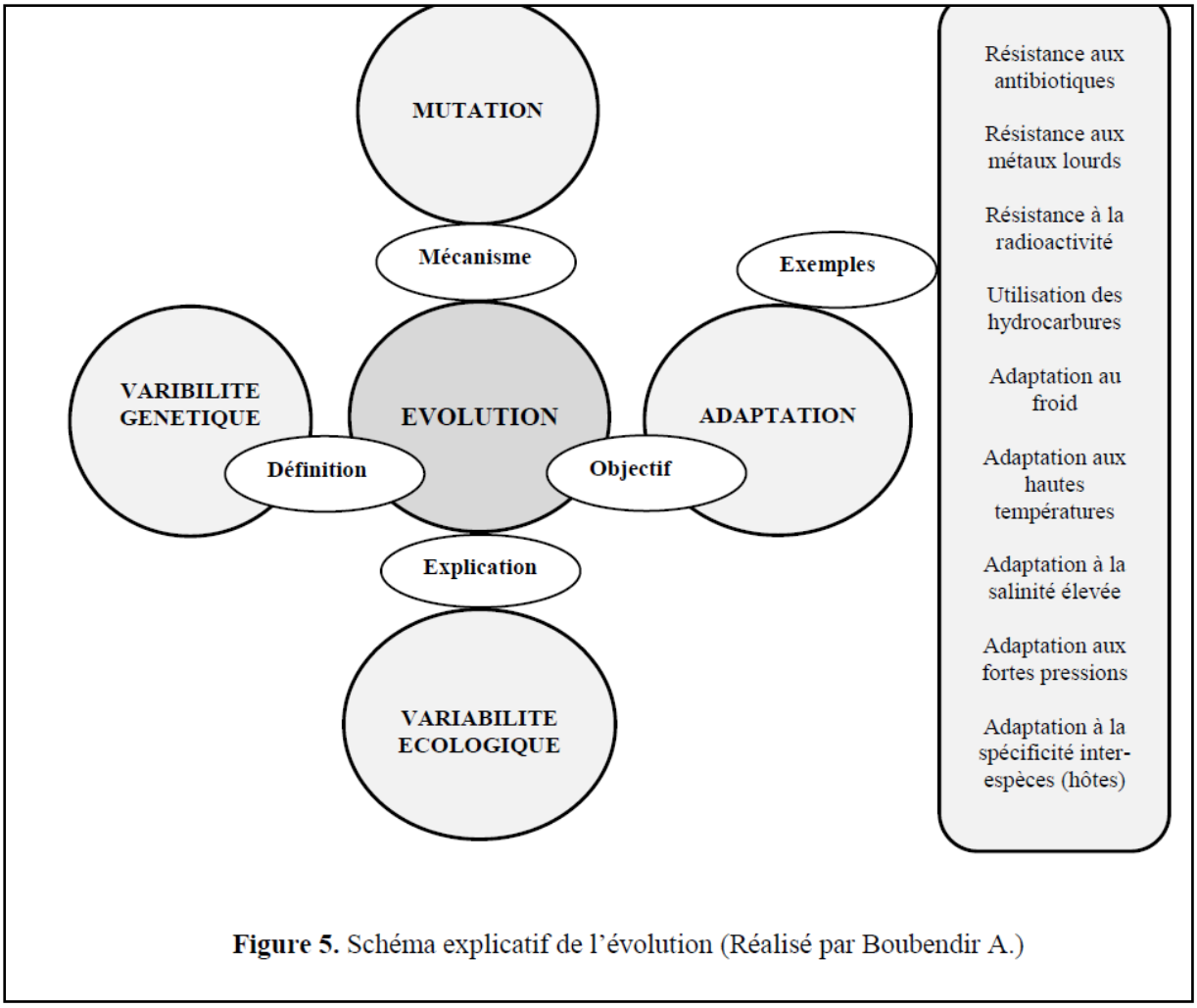


Figure 5. Schéma explicatif de l'évolution (Réalisé par Boubendir A.)

Figure 5. Schéma explicatif de l'évolution (Réalisé par Boubendir A.)

4.2. Les données de la phylogénie

La phylogénie a été dénommée par Ernst Haeckel, c'est un mot latin composé de « fulon » (tribu, race) et « genus » (naissance, origine), donc la phylogénie signifie à la base l'ancêtre (origine) commun d'un groupe de gènes ou autres séquences.

La phylogénie se base sur le principe de la comparaison de caractères spécifiques pour un ensemble d'individus. Ces caractères sont en général homologues et appartiennent à des organismes contemporains.

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.

4.2.1. Les données phénotypiques

Comprennent les caractères observables (aux différents états: morphologiques,

biochimiques et physiologiques) et les patterns binaires (de type présence d'un caractère donné / absence de ce même caractère). Dans le cas des bactéries, par exemple, les caractères peuvent être :

- Biochimiques et enzymatiques,
- Antigéniques
- Sensibilité vis-à-vis des antibiotiques
- Sensibilités aux phages,
- Profils électophorétiques de systèmes enzymatiques, etc.

4.2.2. Les données moléculaires

Dans ce cas, ce sont des séquences biologiques de type acides nucléiques telles que les séquences de gènes particuliers, d'ARNm, RFLPs, Microsatellites, SNPs, IGS (ARNr et mitochondries), ITS (ARNr et mitochondries), séquences des cytochromes C, séquences des facteurs d'élongation alpha, ou encore des séquences de protéines enzymatiques ou de structure .

Les données les plus employées pour les constructions phylogénétiques sont les marqueurs suivants :

- ADNr 16S : Bactéries
- ADNr 18S, actine, EF1, RPB1 : Eucaryotes
- ADNr 18S, RBCL : Végétaux

Traditionnellement, les arbres phylogénétiques sont construits par comparaison des caractères phénotypiques, on parle alors de *phénogramme*, et sa continue un jouer un rôle dominant dans l'analyse des données telles que les fossiles.

Cependant, les arbres phylogénétiques sont basés actuellement sur l'alignement multiple de séquences nucléotidiques ou d'acides aminés, on parle alors de *phylogramme*, et on appel sa la phylogénie moléculaire.

4.3. La construction d'un arbre phylogénétique

4.3.1. La matrice de distances

La distance évolutionnaire est définit étant le pourcentage de substitution de nucléotides ou d'acide aminés, elle est estimée par plusieurs modèles à savoir modèle le p-distance, Poisson, Dayhoff, Jones-Taylor-Thomson (JTT), etc. La distance est calculée entre les séquences deux à deux pour donner enfin la matrice de distance (Tableau 2).

	1	2	3	4	5	6	7	8	9	10
1. Synechocys										
2. Odontella	0.387									
3. Porphyra	0.305	0.326								
4. Cyanophora	0.304	0.366	0.291							
5. Euglena	0.496	0.493	0.469	0.474						
6. Marchantia	0.402	0.421	0.371	0.366	0.457					
7. Pinus	0.432	0.459	0.414	0.407	0.486	0.193				
8. Nicotiana	0.435	0.462	0.409	0.412	0.491	0.204	0.187			
9. Zea	0.455	0.478	0.429	0.432	0.500	0.241	0.224	0.123		
10. Oryza	0.454	0.478	0.430	0.432	0.500	0.241	0.223	0.122	0.025	

Tableau 2. Estimation de la divergence évolutionnaire entre les séquences des protéines de chloroplaste de 10 espèces végétales

4.3.2. La topologie de l'arbre phylogénétique

Les différentes méthodes de constructions d'arbres phylogénétiques diffèrent à la fois par les hypothèses évolutives qu'elles impliquent et par les algorithmes qu'elles utilisent. Elles peuvent être regroupées en deux catégories :

 Les méthodes de distances : Les distances génétiques (% de substitutions des nucléotides ou des acides aminés par exemple) sont mesurées entre toutes les séquences prises deux à deux. Ces méthodes sont rapides et donnent de bons résultats.

 Les méthodes basées sur les caractères : S'intéressent aux caractères phénotypiques qui présentent des états supérieurs à deux. Elles regroupent les méthodes de "parcimonie" et les méthodes de "Maximum de vraissemblance".

Pour les méthodes de distances (qui intéresseront notre cours), il s'agit tout d'abord de choisir le critère de distance entre les futures feuilles de l'arbre (individus ou OTUs). Par exemple, si ces individus sont des séquences d'ADN, on peut choisir comme distance entre deux d'entre elles le nombre de nucléotides qui diffèrent. Pour déterminer cette valeur, on est amené à en effectuer un alignement multiple. Puis on peut utiliser la méthode **UPGMA** (unweighted pair group method with arithmetic mean) ou celle de **NJ** (Neighbor-Joining) pour en déduire la topologie de l'arbre. Par contre, si ces individus ont été étudiés sur les plans morpho-physico-biochimiques, alors les distances découleront des coefficients de similarité. Les méthodes de distances utilisent deux algorithmes distincts pour construire des dendrogrammes :

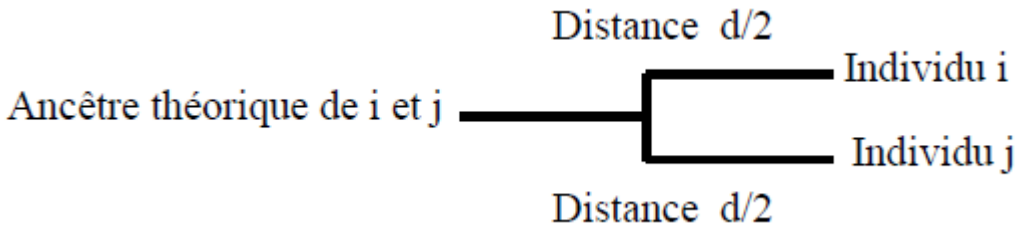
4.3.2.1. La méthode UPGMA

UPGMA utilise un algorithme de clustérisation 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 individus (OTUs) les plus proches et ce groupe est ensuite traité comme un seul individu, puis on recherche l'individu le plus proche et ainsi de suite jusqu'à ce qu'il n'y ait plus que deux groupes. Cet algorithme permet de calculer un *arbre ultra métrique*.

La méthode UPGMA s'effectue selon les étapes suivantes :

-**Etape 1:** Dans la matrice des distances (symbolisées par dij), trouver les taxons i et j pour lesquels la distance dij est la plus petite. On clustérise tout d'abord les deux OTUs avec la distance la plus petite.

-**Etape 2 :** Mettre la racine (ancêtre théorique des deux OTUs choisis) à égale distance des deux OTU i et j c'est-à-dire à d = dij/ 2. Cette distance sera égale à la longueur de la branche du clade qui regroupe les individus i et j :



-**Etape 3 :** Créer un nouvel ensemble incluant i et j.

-**Etape 4 :** Calculer la distance entre le nouveau groupe (ij) et chaque autre taxon (k), en appliquant la formule suivante : (dki + dkj) / 2

-**Etape 5 :** A partir de cette nouvelle matrice, répéter l'opération depuis l'étape 1.

4.3.2.2. La méthode NJ

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 (*arbre non ultra métrique*). La matrice de distances permet de prendre en compte la divergence moyenne de chacun des individus avec les autres taxons. L'arbre est alors construit en reliant les individus les plus proches dans cette nouvelle matrice.

La méthode NJ s'effectue selon les étapes suivantes :

-**Etape 1 :** Calcul de la divergence nette r(i) de chacun des N OTU par apport aux autres

-**Etape 2 :** calcul de la nouvelle matrice des distances en utilisant la formule suivante :

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

Etape 3 : choix des plus proches voisins, c'est-à-dire des deux OTUs ayant le M (i,j) le plus petit. Les deux premiers OTUs forment un nouveau noeud u.

-Etape 4 : calcul de la distance de chacun des deux OTUs par rapport au noeud u.

$S (i,u) = d(i,j)/2 + [r(i) - r(j)]/2(N-2)$

d'où $S (j,u) = d(i,j) - S(i,u)$

-Etape 5 : Calcul des distances entre u et toutes les OTUs.

-Etape 6 : Créer une nouvelle matrice et répéter l'opération depuis l'étape 1.

4.4. Evaluation d'un arbre phylogénétique

Après la construction avec succès de l'arbre phylogénétique, l'étape suivante requière l'évaluation de la topologie de l'arbre. Ce processus peut être performé par l'usage de deux méthodes d'évaluation, nommées la méthode bootstrap et le test des branches internes.

4.4.1. La méthode bootstrap

Le concept de base de la méthode bootstrap est l'évaluation de la topologie de l'arbre par la construction d'arbres phylogénétiques égale au nombre de pseudo-données répétées. Les noeuds de l'arbre montrant des valeurs >70% de bootsrap sont généralement considérés comme consistants.

4.4.2. Le test des branches internes

Ce test est calculé en utilisant la procédure bootstrap, sa construction est basée sur la longueur des branches internes, il est valable seulement dans les arbres NJ. Dans ce test la confidence de la longueur des branches internes est non-zéro.

4.5. Exemples d'arbres phylogénétiques

Les valeurs des distances évolutionnaires obtenues précédemment dans la matrice de distance (les séquences des protéines de chloroplaste de 10 espèces végétales, Tableau 3), sont projetées dans l'espace et permettent de construire l'arbre phylogénétique avec :

- La méthode NJ et test boostrap (Figure 6),
- La méthode NJ et test des branches internes (Figure 7).
- La méthode UPGMA et test boostrap (Figure 8),

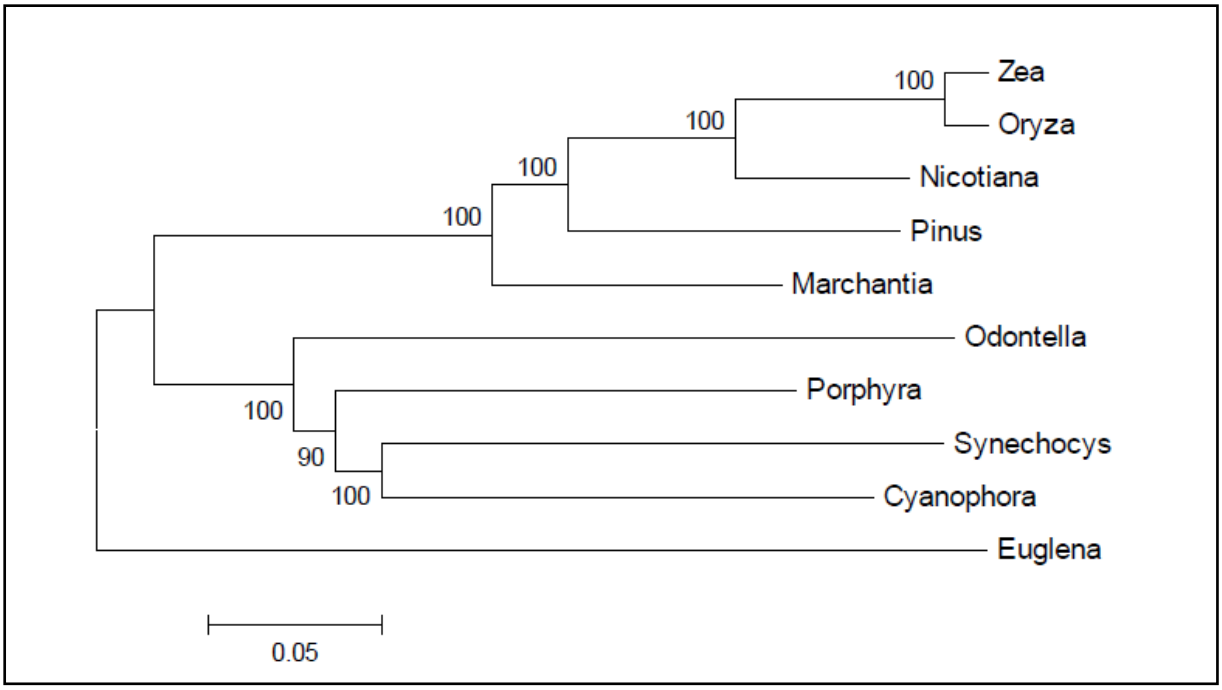


Figure 6. Arbre phylogénétique des séquences des protéines de chloroplaste de 10 espèces végétales par la méthode **NJ** avec test **bootstrap**, réalisée par le logiciel MEGA6.

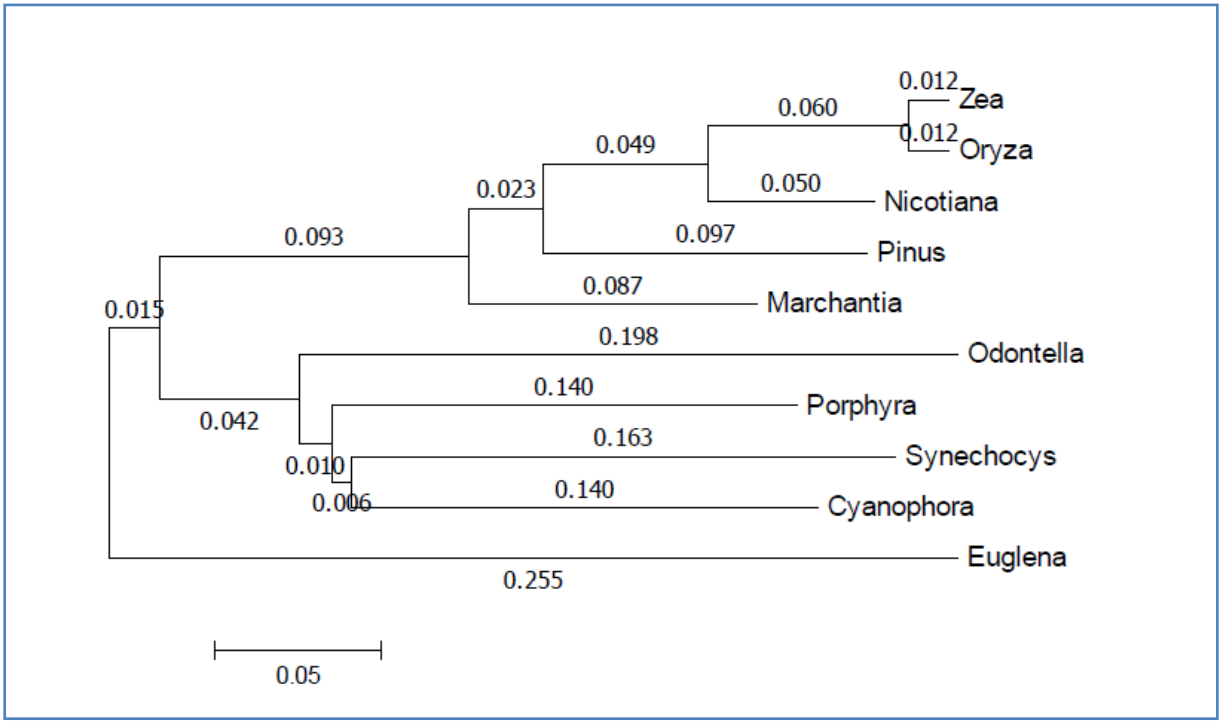


Figure 7. Arbre phylogénétique des séquences des protéines de chloroplaste de 10 espèces végétales par la méthode **NJ** avec test des **branches internes**, réalisée par le logiciel MEGA6.

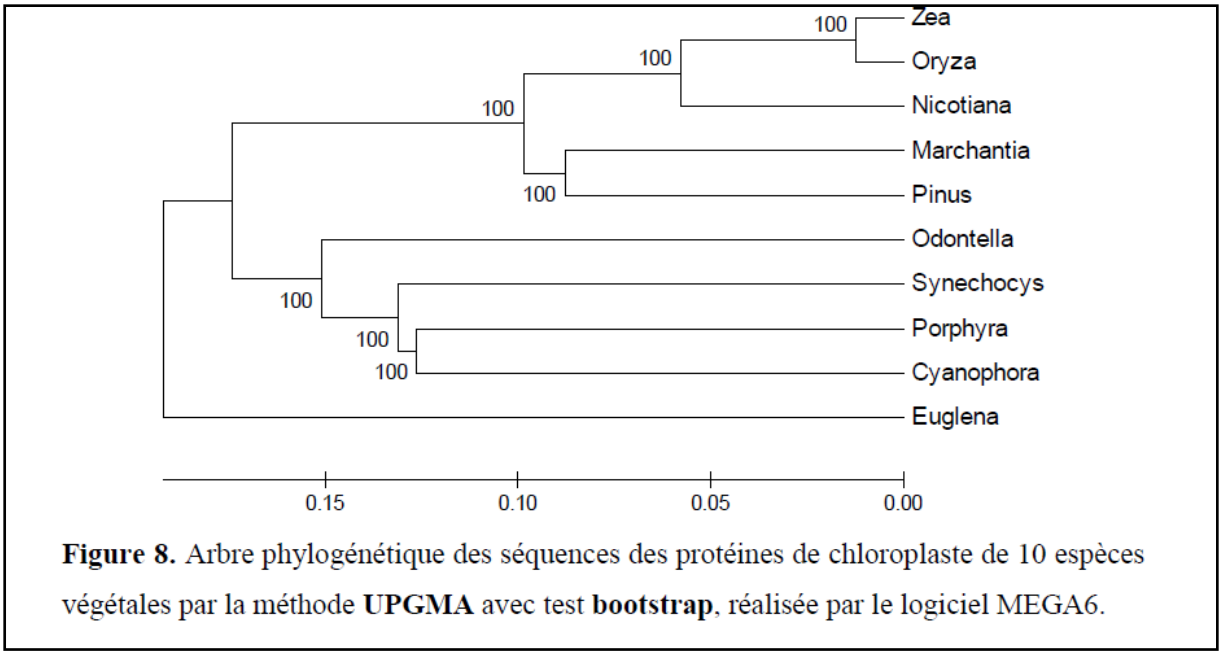


Figure 8. Arbre phylogénétique des séquences des protéines de chloroplaste de 10 espèces végétales par la méthode **UPGMA** avec test **bootstrap**, réalisée par le logiciel MEGA6.

Figure 8. Arbre phylogénétique des séquences des protéines de chloroplaste de 10 espèces végétales par la méthode **UPGMA** avec test **bootstrap**, réalisée par le logiciel MEGA6