I. INTRODUCTION

Depuis Darwin, il est communément admis que les êtres vivants descendent tous les uns des autres. Jusqu'aux années 1960, les comparaisons entre des morphologies, des comportements et des répartions géographiques des espèces étaient les seuls moyens disponibles pour contruire 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 : la phylogénie.

1. L'évolution selon LAMARK (1744-1829)

Pour Lamark, l'évolution était due à une adaptation continue au milieu ambiant : un environnement changeant altère les besoins de l'organisme vivant qui s'adapte en modifiant son comportement et en utilisant certains organes plus que d'autres.

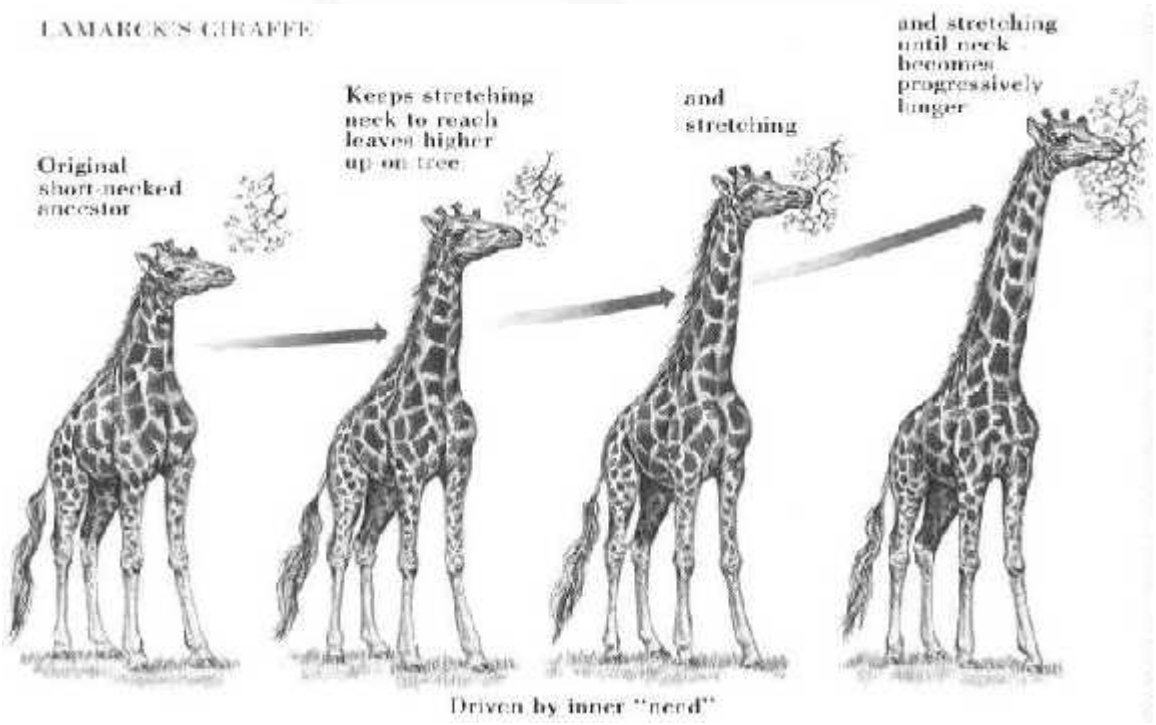


Figure 1 : l'évolution selon Lamark

2. L'évolution selon Darwin (1809-1882)

La théorie défendue par Darwin est l'évolution par sélection naturelle. Au sein d'une même lignée, tous les individus sont différents et la nature favorise la multiplication de ceux qui jouissent d'un quelconque avantage.

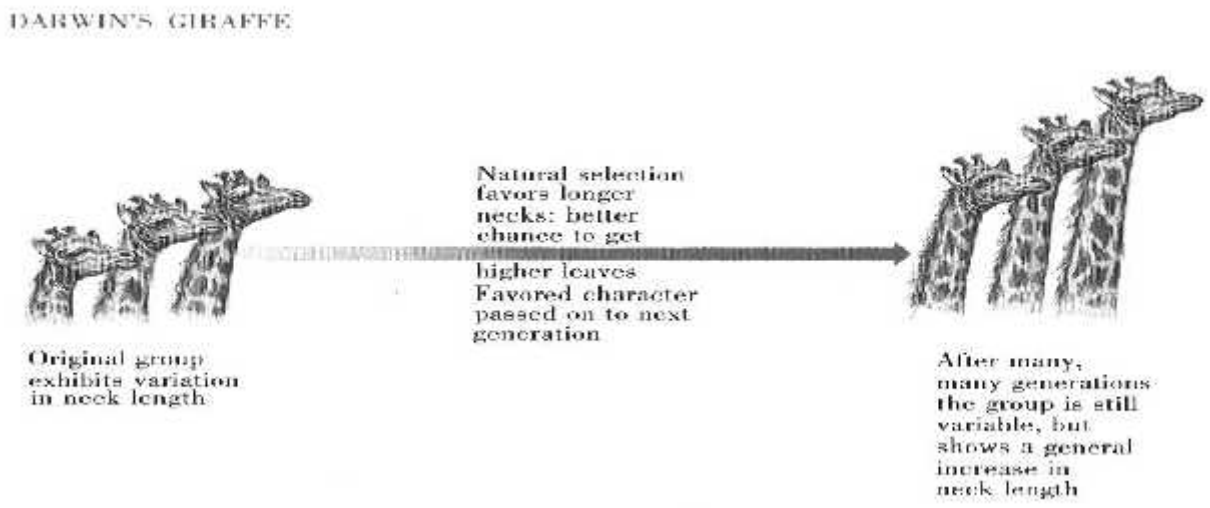


Figure 2 : l'évolution selon Darwin

3. Les concepts modernes : la théorie neutraliste vs la théorie sélectionniste

Quand les lois de la génétique ont été connues, il est né un paradoxe entre la sélection (disparition de certains caratères) et le polymorphisme génétique (variabilité). Il faut cependant rappeler que ce sont les phenotypes qui se heurtent à la pression de sélection et non les génotypes (avec le jeu des dominances et récessivités, à un phenotype correspond plusieurs génotypes). Pour l'expliquer, 2 théories s'opposent : la théorie neutraliste et la théorie sélectionniste.

La théorie neutraliste (Kimura)

La plupart des mutations restent neutres, se fixent au hasard (seules les mutations très défavorisantes ou létales pour l'individu sont éliminées) et le milieu n'a pas de rôle sélectif.

La théorie sélectionniste

la plupart des nouveaux allèles apparus par mutations se fixent dans les populations parce qu'ils sont avantageux pour les porteurs dans le milieu où ils vivent (sélection darwinienne).

4. Evolution convergente et évolution divergente

-L'évolution convergente correspond à des solutions trouvées de manière indépendante chez des organismes différents pour résoudre le même problème.

- L'évolution divergente correspond au contraire à des protèines ayant le même ancêtre commun mais qui se sont spécialisées dans des fonctions différentes.

Gènes orthologues

Paire de gènes nés de la divergence de leur ancêtre commun (spéciation)

Gènes paralogues

Paire de gènes nés de la duplication de leur ancêtre commun

La taxinomie est la science qui a pour objet de décrire les organismes vivants et de les regrouper en entités appelées taxons afin de pouvoir les identifier puis les nommer, et enfin les classer.

Taxon: entité conceptuelle qui est censée regrouper tous les organismes vivants possédant en commun certains caractères bien définis. Les taxons sont organisés en clades qui s'emboîtent les uns à l'intérieur des autres.

Clade: est une partie d'un cladogramme, une branche contenant deux éléments plus proches entre eux qu'avec n'importe quel autre élément. L'espèce constitue le taxon de base de la classification. Plus le rang du taxon est élevé et plus le degré de ressemblance (le nombre de caractères qu'ils ont en commun) entre les individus concernés (plantes, animaux, champignons, bactéries) diminue, et inversement.

Phylogénèse: (phûlon = tribu et genesis = origine)

Histoire évolutive des espèces, suite des évènements évolutifs qui ont mené à la diversification des êtres vivants

Phylogénie: (néologisme d'origine anglaise, phylogeny)

« a diagram representing a phylogenesis » (Oxford Dictionary)

II. LA PHYLOGENIE MOLECULAIRE

Cela correspond à de la phylogénie par comparaison de gènes [les gènes utilisés doivent être choisis avec soin : il faut que cela soit des gènes subissant de fortes contraintes fonctionnelles donc ayant un taux de mutation faible. Un bon exemple est le cytochrome B intervenant dans les chaines d'oxydation cellulaire de tous les êtres vivants (les êtres vivants actuels l'ont sans doute hérité d'un ancêtre commun il y a trois milliards d'années)].
Il y a cependant une accumulation des mutations au cours du temps et pour rendre compte de ce phénomène, Zuckerland et Pauling (1962) ont développé la théorie de l'horloge moléculaire.

1. L'horloge moléculaire

1.1 <u>Définition</u>

En résumé, on constate que le taux d'accumulation des mutations dans le génome d'organismes différents est du même ordre de grandeur dans des régions homologues (régions soumises à la même pression de sélection).
L'accumulation sera maximale pour des régions qui ne sont pas soumises à la pression de sélection naturelle (ne codant pas pour des gènes) et minimale dans les parties du génome soumises à une forte pression (c'est à dire les régions codant pour des fonctions essentielles à la survie de l'organisme).
Chaque séquence accumule les mutations à un rythme qui lui est propre et qui est dicté par l'intensité de la pression de sélection à laquelle elle est soumise. Pour reconstituer des phylogénies (dater la divergence entre deux espèces), on peut utiliser différentes molécules comme on utilise les aiguilles d'une montre pour calibrer l'horloge :
- la trotteuse des secondes (taux de mutation important, par exemple un pseudogène) pour des évènements récents (études des sous populations au sein d'une espèce).
- l'aiguille des minutes (taux de mutation moyen, par exemple le cytochrome C) pour l'analyse d'un passé proche.
- l'aiguille des heures (taux de mutations faible : les histones) pour l'étude d'un passé lointain.
La vitesse d'évolution de la séquence est du même ordre de grandeur au sein d'une même classe fonctionnelle de protéines et elle est différente pour des protéines qui ont des fonctions

différentes : la vitesse d'évolution de la sérum albumine est toujours plus importante que celle du cytochrome C. Ces différences de vitesse dépendent à la fois de la probabilité qu'une substitution apparaisse et de sa compatibilité avec la survie de l'organisme.
Si l'on admet cette théorie, et que l'on connaît le taux d'accumulation des mutations, il est possible d'estimer le temps de divergences d'espèces en comparant leur diversité moléculaire.

1.2 Arguments contre l'horloge moléculaire

La théorie de l'horloge moléculaire est remise en cause et plusieurs arguments ont été développés :

- L'horloge moléculaire ne serait pas constante (Goodman): les mutations avantageuses se fixeraient plus rapidement lors de la formation de nouvelles espèces.
- L'horloge moléculaire serait épisodique (Gillepsie) et les mutations ne se produiraient pas de façon indépendante au cours de l'évolution: il y aurait des épisodes d'accumulation suivis d'arrêts évolutifs.

1.3 Conclusion

Bien que le débat persiste, il semble que l'horloge moléculaire fonctionne assez bien sur de longues périodes évolutives, pour des gènes ayant un taux de mutation relativement faible où même si l'horloge ne bat pas très régulièrement, les ralentissements et les accélérations se compensent.
Il faut également se méfier des estimations de temps de divergence basées sur un petit nombre de gènes.

Etapes d’une phylogénèse moléculaire

1) Choix de la séquence cible
2) Collecter les séquences
3) Aligner correctement les séquences
4) Choisir un algorithme de construction d’un
arbre phylogénétique

5) Évaluer statistiquement la robustesse de l'arbre (bootstrap)

2.Méthodes de reconstructions

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.

2.1. 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.
Programmes DNADIST et PROTDIST de Phylip

- **UPGMA (Unweight Pair Group Method with Arithmetic mean)**

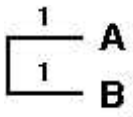
Cette méthode est utilisée pour reconstruire des arbres phylogénétiques si les séquences ne sont pas trop divergentes.
UPGMA utilise un algorithme de clusterisation séquentiel dans lequel les relations sont identifiées dans l'ordre de leur similarité et la reconstruction de l'arbre se fait pas à pas 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

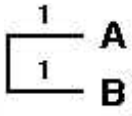
dist(A,B),E = (distAE + distBE) / 2 = 6

dist(A,B),F = (distAF + distBF) / 2 = 8

MATRICE **ARBRE**

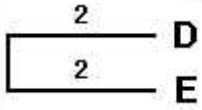
Cycle 1

	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



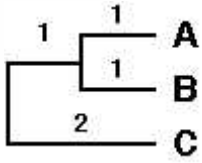
Cycle 2

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



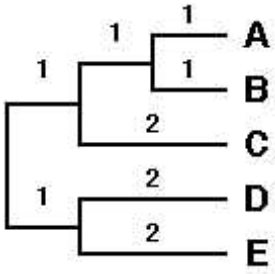
Cycle 3

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



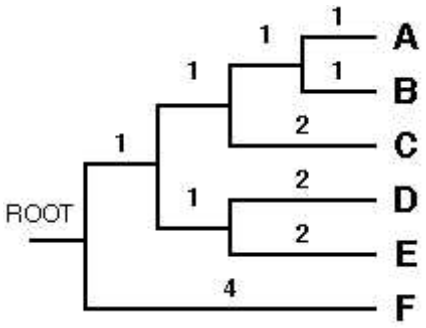
Cycle 4

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



Cycle 5 **ABC,DE**

F 8



NJ(Neighbor-Joining)

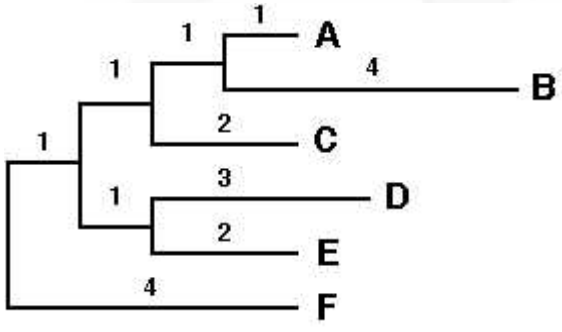
Cette méthode développée par Saitou et Nei (1987) tente de corriger la méthode UPGMA afin d'autoriser un taux de mutation différent sur les branches.

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.

Programme NEIGHBOR de Phylip

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

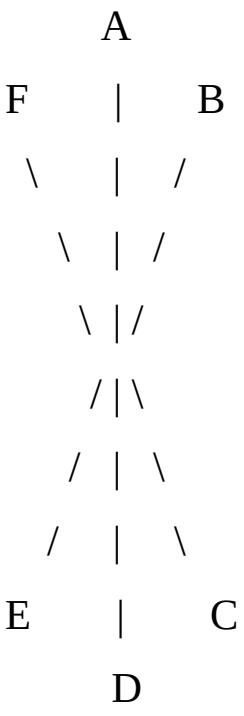
	A	B	C	D	E
B	-13				
C	-11.5	-11.5			
D	-10	-10	-10.5		
E	-10	-10	-10.5	-13	
F	-10.5	-10.5	-11	-11.5	-11.5

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

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

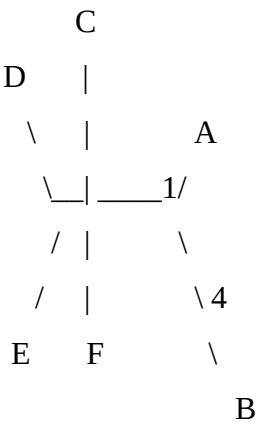
d (EU) = d(AE) + d (BE) - d(AB) / 2 = 5

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.

2.2. Méthodes fondées sur les caratères

Ces méthodes sont très lentes mais elles sont précises.

Parcimonie

La parcimonie consiste a 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.

Cette méthode, quand elle est appliquée à des séquences protéiques, utilise le code génétique pour comptabiliser le nombre de substitutions nécessaires (changements de bases) pour passer d'un site à l'autre d'une séquence à l'autre.

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 :

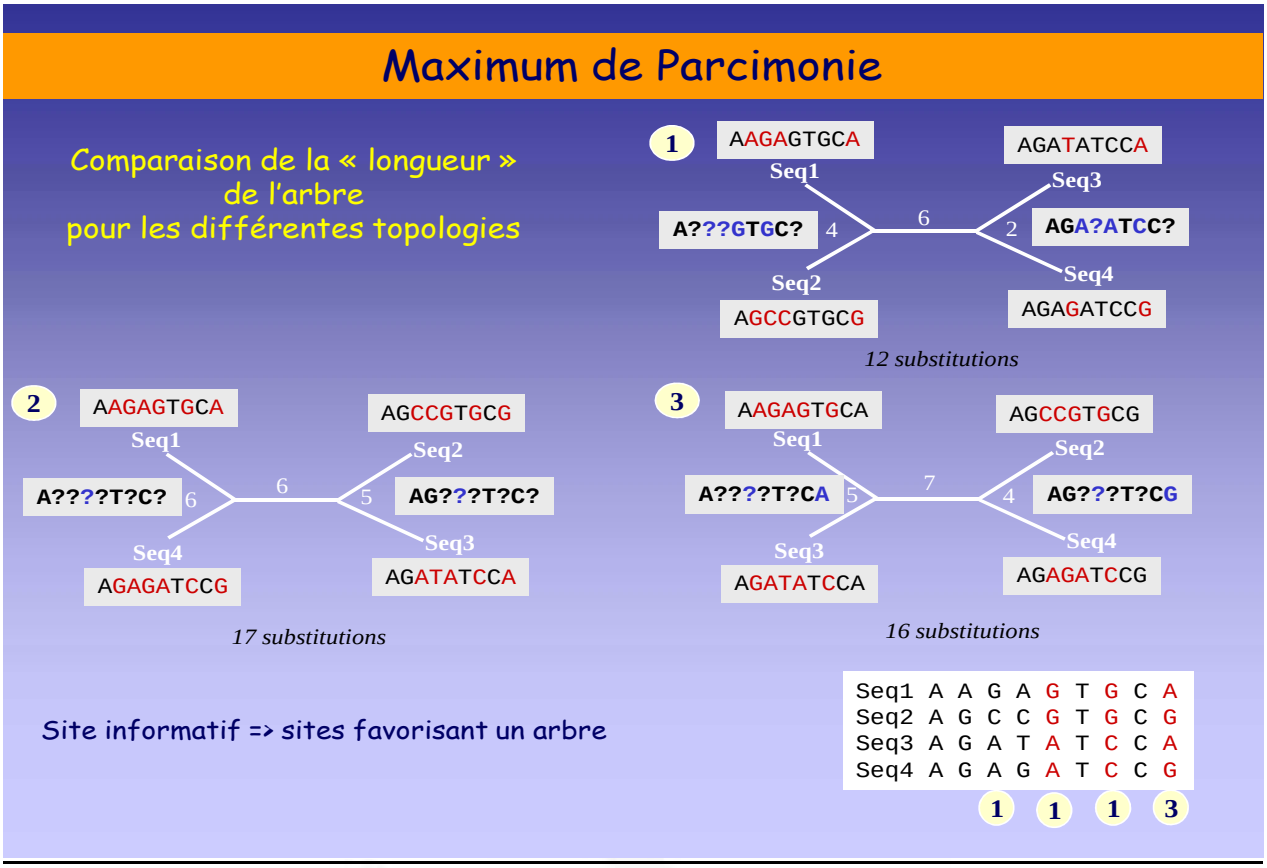
le nombre d'arbres enracinés possibles pour n OTUs : Nr = (2n - 3)! / (2exp(n-2))(n-2)!

le nombre d'arbres non enracinés possibles pour n OTUs : Nu = (2n -5)! / (2exp(n-3))(n-3)!

Programme DNAPARS et PROTPARS de Phylip

Nombre d'OTUs	**Nb d'arbres non enracinés**	**Nb d'arbres enracinés possibles**
2	1	1
3	1	3
4	3	15
5	15	105
6	105	945
7	954	10 395
8	10 395	135 135

METHODE 1



METHODE 2

Sequence 1 2 3 4 5 6 7 8 9

1 A A G A G T G C A

2 A G C C G T G C G

3 A G A T A T C C A

4 A G A G A T C C G

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)

(1) AAGAGTGCA AGATATCCA (3)
\ 4 / 2
\ 4 /
AGCCGTGCG --- AGAGATCCG Nombre de mutations : 10
/ \
/ 0 \ 0
(2) AGCCGTGCG AGAGATCCG (4)

(1) AAGAGTGCA AGCCGTGCG (2)
\ 1 / 3
\ 5 /
AGGAGTGCA --- AGAGGTCCG Nombre de mutations : 14
/ \
/ 4 \ 1
(3) AGATATCCA AGAGATCCG (4)

(1) AAGAGTGCA AGCCGTGCG (2)
\ 1 / 3
\ 5 /
AGGAGTGCA --- AGATGTCCG Nombre de mutations : 16
/ \
/ 5 \ 2
(4) AGAGATCCG AGATATCCA (3)

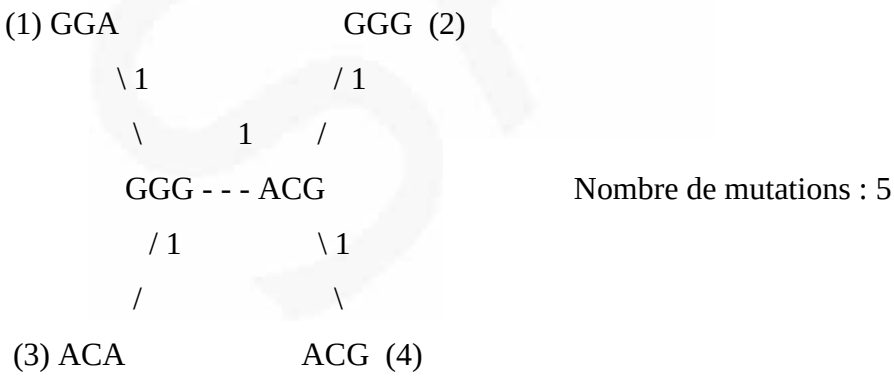
L'arbre I est celui nécessitant le moins de mutations, c'est donc le plus parcimonieux.
Cette analyse prend en compte tous les sites des séquences mais l'analyse peut également se faire uniquement sur les sites informatifs, c'est à dire quand à cette position il y a au moins 2 nucléotides différents,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

Séquence	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
1	A	A	G	A	G	T	G	C	A
2	A	G	C	C	G	T	G	C	G
3	A	G	A	T	A	T	C	C	A
4	A	G	A	G	A	T	C	C	G
					*		*		*

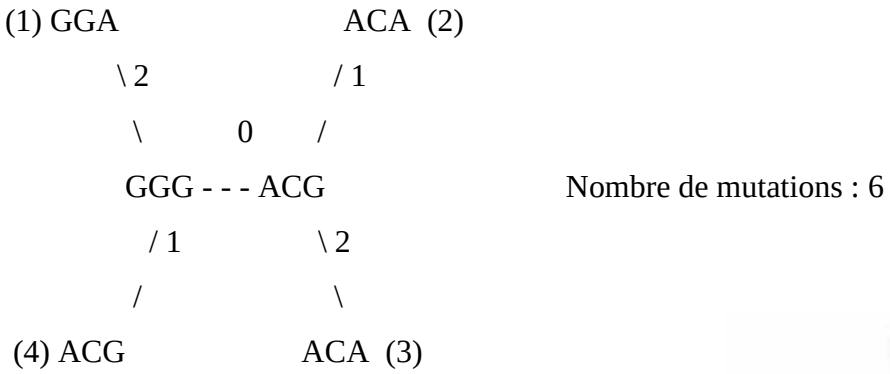
On peut donc "réduire" les séquences aux seuls sites informatifs :

1	G	G	A
2	G	G	G
3	A	C	A
4	A	C	G
	*****	*****	*****

Nombre de mutations : 4



Nombre de mutations : 5

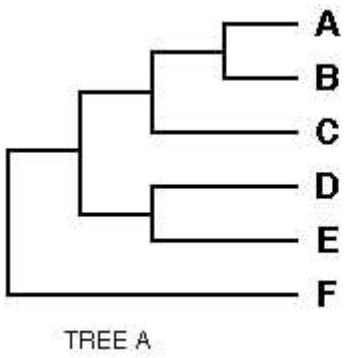


Dans le cas de 4 séquences, un site informatif favorise seulement un arbre : le site 5 favorise l'arbre I plus que les arbres II et III (il supporte l'arbre I). L'arbre le plus parcimonieux est celui qui est supporté par le plus grand nombre de sites informatifs.
Le maximum de parcimonie recherche l'arbre optimal et dans ce processus, il est possible de trouver plusieurs arbres optimaux (= arbres ex-aequo = configuration comptabilisant le même nombre minimal de substitutions nécessaires pour passeer d'une séquence à l'autre dans l'ensemble de l'arbre).
Afin de garantir de trouver l'arbre le meilleur possible, il faut faire une évaluation de toutes les topologies possibles mais cela devient impossible lorsque l'on a plus de 12 séquences.

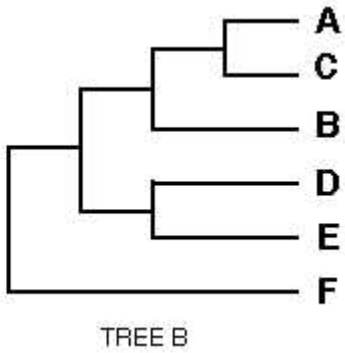
Branch and Bound : cette méthode est dérivée du maximum de parcimonie, elle garantit de trouver le meilleur arbre mais sans évaluer tous les arbres possibles. Elle permet de traiter un plus grand nombre de séquences mais reste limitée.

Recherche heuristique : il y a un réarrangement des branches à chaque étape, cette méthode ne garantit pas de trouver l'arbre optimal.

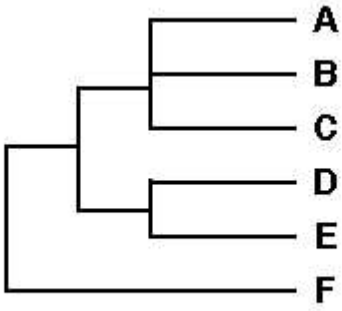
Arbre consensus : comme la méthode du maximum de parcimonie peut conduire à trouver plusieurs arbres équivalents, on peut créer un arbre consensus (avec utilisation du bootstraping). Cet arbre consensus est construit à partir des noeuds les plus fréquemment rencontrés sur l'ensemble des arbres possibles.



TREE A



TREE B



CONSENSUS TREE

<u>Avantages et inconvénients de la parcimonie</u>

Avantages :

- Méthode basée sur les caractères : méthode cladistique plutôt que phénétique.
- Méthode ne réduisant pas la séquence à un simple nombre.
- Méthode essayant de donner une information sur les séquences ancestrales.
- Méthode évaluant différents arbres.

Inconvénients :

- Méthode très lente par rapport aux méthodes basées sur les distances.
- Méthode n'utilisant pas toute l'information disponible (seuls les sites informatifs sont pris en compte)
- Méthode ne faisant pas de corrections pour les substitutions multiples
- Méthode ne donnant aucune information sur la longueur des branches
- Méthode connue pour être très sensible au biais des codons

III. CONCLUSION

L'étude de la phylogénie est un vaste domaine et quelque soit la méthode utilisée, des hypothèses très simplificatrices sont faites sur l'évolution biologique des séquences. Actuellement, pour reconstruire une bonne phylogènie, la qualité et le nombre des données provoquent plus de variations au sein d'un arbre qu'un changement de méthode.

Pour construire de bons arbres, il faut :

- Avoir le plus grand nombre de gènes homologues possibles
- Aligner les séquences très soigneusement
- Eliminer les régions ambigues, les régions hypervariables, les gaps des alignements
- Utiliser si possibles plusieurs méthodes de reconstruction, prendre NJ plutôt que UPGMA (le neighbor-joining autorise des taux de mutations différents sur les branches) et incorporer des biais dans les taux de mutations / substitutions.
- Evaluer l'arbre statistiquement : bootstrapping.

Souvent les arbres obtenus sont différents selon le gène considéré. Cela est du à plusieurs causes :

- Tous les gènes n'ont pas la même vitesse d'évolution
- L'évolution convergente
- Les phénomènes de recombinaison
- Les transferts de gènes
- La confusion gènes paralogues (duplication au sein d'une espèce) / gènes orthologues (même gène dans des espèces différentes):