Méthodes phénétiques

C'est en 1963 que Edwards et Cavalli-Sforza ont explicitement invoqué le principe de parcimonie à propos de génétique des populations : l'estimation la plus plausible d'un arbre évolutif est celle qui fait appel à la quantité minimale d'évolution.

Exemple :si tel est le cas, les arbres les plus courts pour un jeu de données devraient l'être aussi pour d'autres jeux de données sur les mêmes taxons.
La séquence de la β-globine pour 11 taxons est utilisée pour construire les arbres les plus parcimonieux. Des arbres de 124 à 133 changements ont été obtenus . On compare ensuite la longueur de ces mêmes arbres en prenant comme protéines soit le cytochrome C soit les fibrinopeptides A et B et l'α-globine. On montre qu'en moyenne les arbres qui nécessitent le moins de pas avec la β-globine en nécessitent également moins avec les autres jeux de données. .(Testing the theory of descent, Penny, Hendy and Steel (1991) in Phylogenetic analysis of DNA sequences ed Miyamoto and Cracraft).

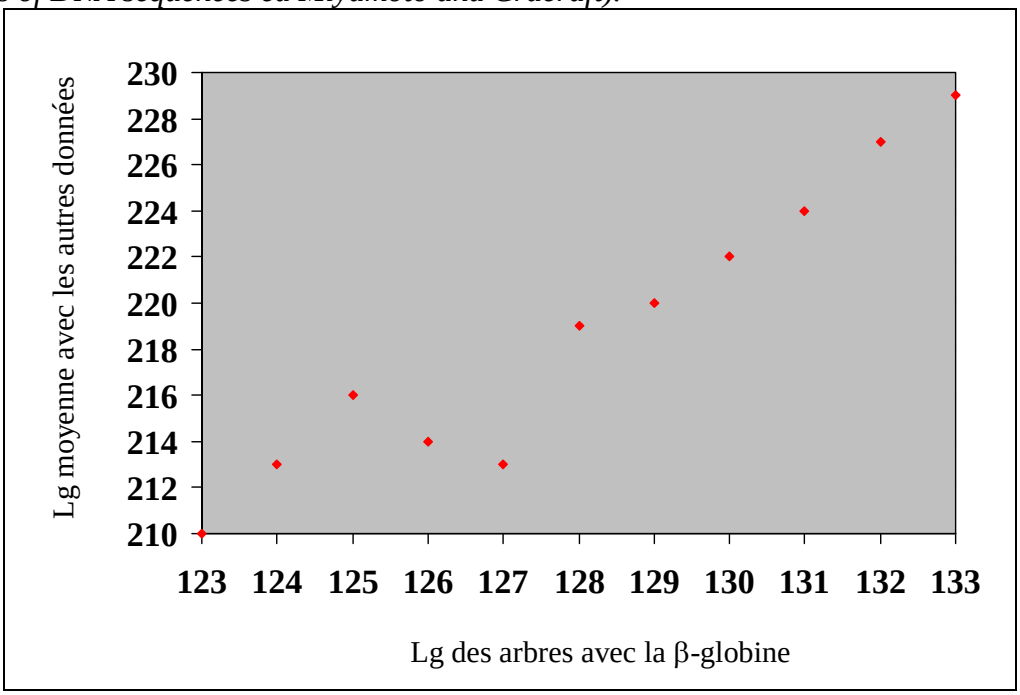


Figure II- 1. Les mêmes arbres sont les plus parcimonieux avec deux jeux de données

Dans les années suivantes, ce principe fut repris tant pour des analyses de distance (Cavalli-Sforza et Edwards 1967 ; Fitch et Margoliash 1967) que pour des analyses cladistiques (Camin et Sokal 1965 ;Kluge et Farris 1969).
Popper qui n'était cependant pas anti évolutionniste dit en 1976 : "Le Darwinisme n'est pas une hypothèse testable, mais un programme de recherche métaphysique – un cadre possible pour une théorie de l'évolution testable".
Si l'on applique un programme de recherche d'arbres les plus parcimonieux sur 5 protéines différentes pour 11 taxa différents, les arbres obtenus doivent être similaires s'ils reflètent les relations de parenté (ou évolutives) entre ces 11 taxa.

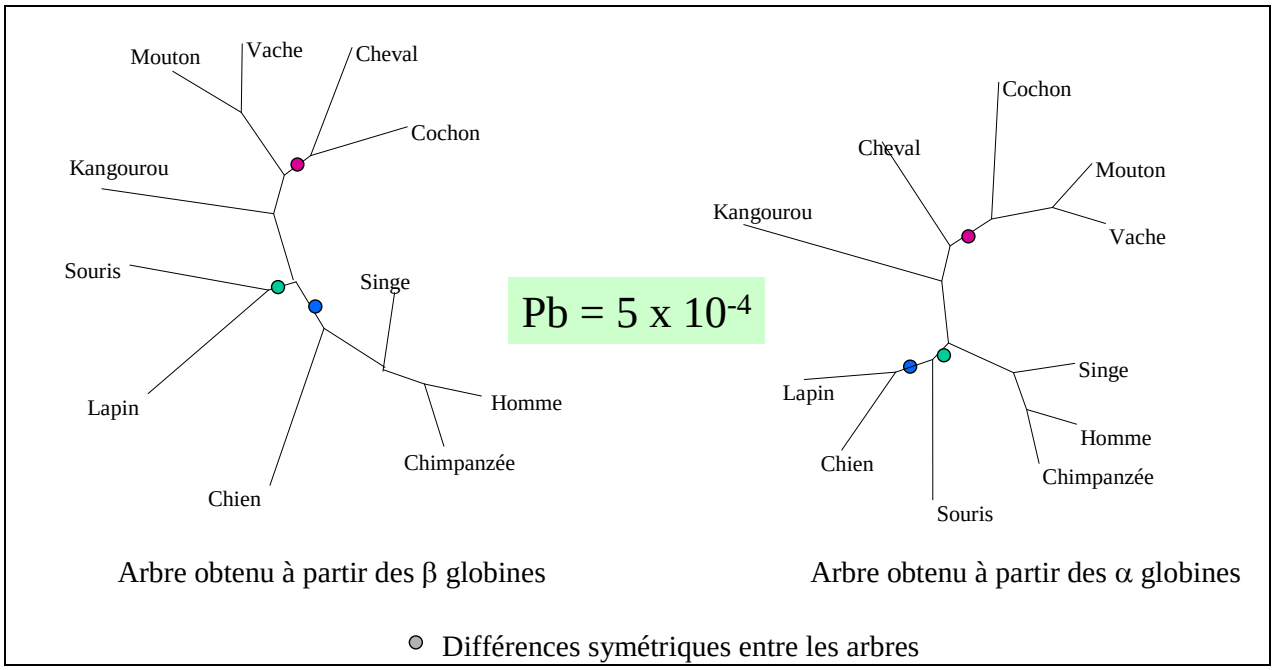


Figure II- 2. La comparaison des arbres construits avec les a et b globines montre que la probabilité de retrouver des arbres avec si peu de différences est de $5*10^{-4}$

Les méthodes phénétiques (ou de distance) ont leur origine dans les méthodes de la taxinomie numérique conçues en 1957 par Michener et Sokal. Elles s'opposèrent aux pratiques des systématiciens évolutionnistes comme Mayr car se voulaient libres de toute spéculation phylogénétique. Les techniques employées sont celles de la classification d'organismes sur la base de similitude globale. Les conséquences phylogénétiques que l'on peut en tirer ne sont qu'accessoires.

Concepts de base

- Les relations entre taxons sont des relations phénétiques et non des relations phylogénétiques
- Plus on a d'informations, plus de caractères décrits plus on peut être prédictif
- Chaque caractère a le même poids (pas d'a priori phylogénétique). Cependant une pondération peut se faire dans certains cas sur la base de critères opérationnels
- La ressemblance est calculée entre chaque paire d'unités taxinomiques et s'exprime par des coefficients de similitude qui forment les éléments d'une matrice de similitude. Des taxons différents ont des associations différentes de caractères donc on mesure une similitude globale.
- La représentation des relations taxinomiques (restituées au moyen de techniques numériques variées) se fait au moyen de schémas, les phénogrammes indiquant les relations phénétiques
- Les mesures de similitude phénétique entre les organismes appartenant à différentes époques géologiques fournissent une information objective sur la vitesse et la direction de l'évolution.
- Les inférences phylogénétiques s'effectuent en dernier en intégrant des hypothèses sur l'histoire et les mécanismes de l'évolution.

La taxinomie numérique est une science empirique qui base la classification sur la similitude globale, rappelant en cela la méthode d'Adanson (1726-1806) qui créa les principales familles d'angiospermes et ne suivi jamais le Sexual System de Linnée.

Les caractères utilisés ne sont que des caractères homologues qu'il est préférable de rendre binaires. Ce qui pose un problème de codage des caractères (exemple de couleurs de la fleur).

Il ressort de tout cela que les méthodes phénétiques sont dénuées de contenu évolutif. Cependant on rencontre souvent dans la littérature des phénogrammes interprétés comme des arbres phylogénétiques. Des arbres phénétiques peuvent être assimilés à des arbres phylogénétiques si des hypothèses concernant les phénomènes évolutifs sont posées. Enfin certaines sources d'information ne peuvent être interprétées qu'au moyen de méthodes phénétiques (données immunologiques, hybridation d'ADN).

Distance

Similitude et distance

La ressemblance s'établit à partir d'informations biologiques variées qui ont des formulations variées. Il peut s'agir d'un caractère présentant deux états possibles et mutuellement exclusifs (caractère morphologique ou une base ou un AA en une position donnée).

$$S_{ij} = \frac{n_{aa} + n_{bb}}{K}$$

Indice de concordance simple de Sokal et Michener (1958)

$$S_{ij} = \frac{n_{bb}}{K - n_{aa}}$$

Indice de similitude de Jaccard (1908);cas des RFLP ou des caractères manquants desfossiles.

Avec l'indice de similitude de Jaccard les caractères qui présentent tous deux un certain état sont considérés comme non informatifs (exemple des caractères manquants avec des fossiles ou des bandes absentes dans des analyses RFLP)

Ce peut être la présence ou l'absence d'un caractère (que l'on peut ramener au cas précédent) ou encore des valeurs continues telles que fréquences géniques, mesures morphométriques etc. On compare toutes ces variables en notant la fraction qui est identique d'une UE à l'autre (similitude). Dans d'autres cas (hybridation d'ADN)la comparaison s'exprime par une seule valeur : le % d'hybridation croisée.

Plus la similitude entre deux taxons est grande moins la distance qui les sépare est grande.

$$d_{ij} = 1 - S_{ij}$$

Propriétés des distances

Les distances sont toujours positives, commutatives, la distance d'une UE à elle même est nulle. Elles peuvent être métriques (il est plus court d'aller directement d'une UE à une autre que de passer par un ancêtre), si de plus les deux plus grandes distances sont égales elles sont ultra métriques. En revanche si la distance ij est égale à ik+jk les distances sont additives

$\delta_{ij} > 0$ si i<>j (positivité)

$\delta_{ij} = 0$ si i=j (la distance de l'UE à elle même est nulle)

$\delta_{ij} = \delta_{ji}$ (commutativité)

Distances métriques 1(Propriété de l'inégalité triangulaire)

$\delta_{ij} \leq \delta_{ik} + \delta_{jk}$

Distances ultramétriques (donc les 2 plus grandes distances sont égales 2)

$\delta_{ij} \leq \max(\delta_{ik}, \delta_{jk})$ avec $\delta_{jk} = \delta_{ik}$

Distances additives

$\delta_{ij} = \delta_{ik} + \delta_{jk}$

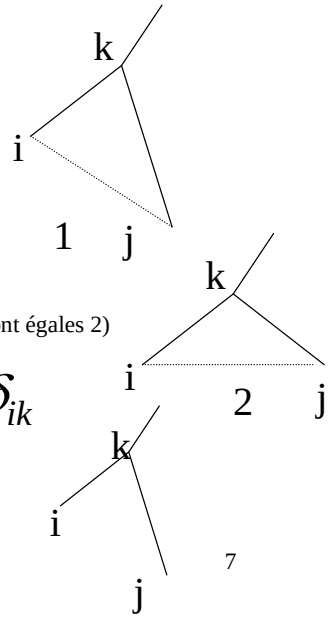


Figure II- 3. Propriétés des distances.

Distances observées et évaluées

Afin de simplifier l'analyse nous considérerons un caractère qui peut être sous deux états a et b. Si l'on observe le même état pour ce caractère dans deux taxons, cela peut résulter de différents événements.

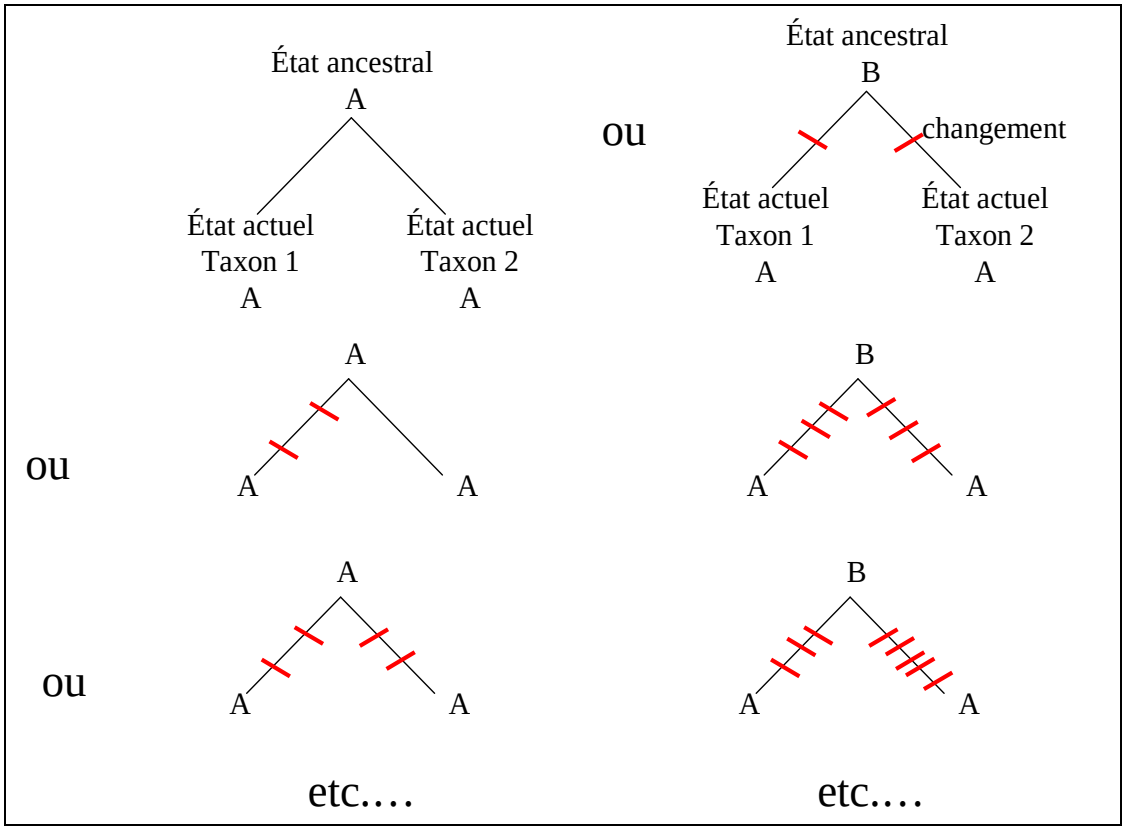


Figure II- 4. Différents événements qui rendent compte du même état de caractère pour les taxons 1 et 2.

L'ndice de concordance simple de Sokal et Michener donne comme distance observée sur K sites au total

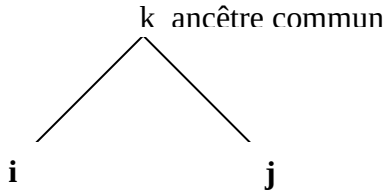
$$s_{ij} = \frac{n_{aa} + n_{bb}}{K}$$

Suivant le scénario évolutif choisi, la distance n'est pas la même car certains événements sont cachés. Est-il possible de corriger les distances observées afin de tenir compte des événements cachés (événements multiples avec les caractères moléculaires) ?

Principe du calcul

Sij = $\frac{naa + nbb}{K}$

dij = 1 – Sij



$$D_{obs} = \frac{n_{ab} + n_{ba}}{K}$$

Soit π la probabilité d'une différence entre une UE (i ou j) et son ancêtre (k) sur un caractère
Soit f la probabilité que le caractère K soit dans l'état a et (1-f) qu'il soit dans l'état b.

Probabilité d'avoir a à la fois chez i et j
Si k est sous la forme a Pb=f

aa — Pas de changement entre k et i Pb 1-π
Pas de changement entre k et j Pb 1-π
$f*(1-\pi)^2$

bb — ou un changement entre k et i Pb π
et un changement entre k et j Pb π
$f*\pi^2$

ab — ou Pas de changement entre k et i Pb 1-π
et un changement entre k et j Pb π
$f*(1-\pi)*\pi$

États de i et j	État de l'ancêtre k : A, Pb f	État de l'ancêtre k : B, Pb 1-f	Probabilité de l'observation
A et A	$f*(1-\pi)^2$	$(1-f)*\pi^2$	$f*(1-\pi)^2+(1-f)*\pi^2$
B et B	$(f)*\pi^2$	$(1-f)*(1-\pi)^2$	$(1-f)*(1-\pi)^2+(f)*\pi^2$
A et B	$f*(1-\pi)*\pi$	$(1-f)*(1-\pi)*\pi$	$(1-\pi)*\pi$
B et A	$(1-f)*(1-\pi)*\pi$	$f*(1-\pi)*\pi$	$(1-\pi)*\pi$

$$D_{obs} = 2\pi(1-\pi)$$ pour chaque position

Si la fréquence de changements de a en b reste constante par unité de temps (hypothèse de l'horloge moléculaire) et puisque les changements sont un événement rare, cette probabilité peut s'écrire sous forme d'une loi de Poisson :

$$P_r = \frac{n^r e^{-n}}{r!}$$

avec r nombre d'événements et n moyenne de ces changements
On peut remplacer n par mt avec m = nombre de changements par unité de temps.

$$P_r = \frac{mt^r e^{-mt}}{r!}$$

Pour 1 changement $a \rightarrow b$
Pour 2 changements $a \rightarrow a$
Pour 2n+1 changements $a \rightarrow b$
Pour 2n changements $a \rightarrow a$

La probabilité π que i et k soient sous 2 états différents est égale à la somme des probabilités de changements impairs P(1) +P(3) +P(5)+……

$$\pi = \frac{mt^1 e^{-mt}}{1!} + \frac{mt^3 e^{-mt}}{31!} + \frac{mt^{51} e^{-mt}}{51!} + \cdots$$

$$\pi = e^{-mt}\left(\frac{mt^1}{1!} + \frac{mt3}{3!} + \frac{mt5}{5!} + \cdots\right)$$

or $e^x = 1 + \frac{x^1}{1!} + \frac{x^2}{2!} + \frac{x^3}{3!} + ...$

Et on sait que $\frac{x^1}{1!} + \frac{x^3}{3!} + \frac{x^5}{5!} + \cdots = \frac{e^x + e^{-x}}{2}$

(ainsi que $\frac{x^2}{2!} + \frac{x^4}{4!} + \frac{x^6}{6!} + \cdots = \frac{e^x - e^{-x}}{2}$ qui est inutile ici)

En remplaçant x par mt

$$\pi = \frac{e^{-mt}}{2}\left(e^{mt} + e^{-mt}\right) = \frac{1}{2}\left(1 + e^{-2mt}\right)$$

$$D_{obs(ij)} = \frac{n_{ab} + n_{ba}}{K} = P_{ab} + P_{ba} = 2\left(\pi\left(1 - \pi\right)\right)$$

$$D_{obs(ij)} = \left(\frac{1 + e^{-2mt}}{2}\right)\left(1 - \left(\frac{1 + e^{-2mt}}{2}\right)\right)$$

$$D_{obs(ij)} = \left(\frac{1 + e^{-2mt}}{2}\right)\left(\frac{1 - e^{-2mt}}{2}\right) * 2$$

$D_{obs(ij)} = 1/2\ (1 - e^{-4mt})$ pour la distance observée au temps t
Or on veut estimer la distance réelle
$D_{est\ (ij)} = 2mt$ par la distance observée $D_{obs(\ ij)}$
$1 - 2\ D_{obs(\ ij)} = e^{-4mt}$
$Log\ (1 - 2\ D_{obs(\ ij)} = -4mt$
$D_{est\ (ij)} = 2mt = -1/2\ Log\ (1 - 2\ D_{obs(\ ij)})$

Corrections utilisées

Les différentes corrections possibles sont calculées d'une façon analogue mais avec 4 états au lieu de 2 (AGCT).
Ces corrections supposent donc l'existence d'une horloge moléculaire.
Une autre hypothèse implicite est que tous les changements de caractères sont indépendants les uns des autres (produit de probabilités). Si ces points ne sont pas respectés par les données la correction de la distance n'est pas justifiée.

Dans ces différentes méthodes de correction, on distingue souvent les transitions (remplacement d'une base par une autre de même type, une purine par une autre purine, une pyrimidine par une autre pyrimidine) des transversions (une base d'un type est remplacée par une base de l'autre type. Ces deux types de substitution ont des effets différents sur une séquence codant une protéine. Du fait de la dégénérescence du code génétique, les transitions donneront plus souvent des mutations silencieuses alors que les transversions entraîneront plus souvent des substitutions d'acides aminés dans la séquence protéique. Ces deux types de substitution ne sont pas soumises aux mêmes pressions sélectives.

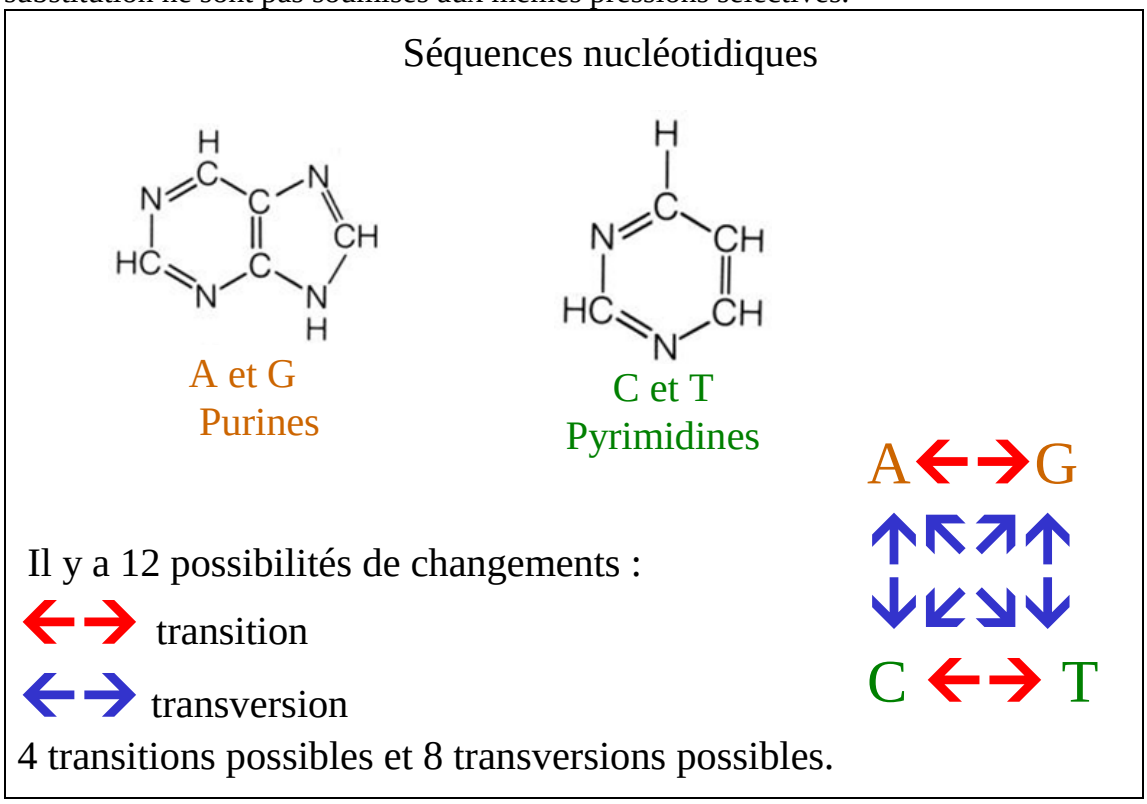


Figure II- 5. Transitions et transversions.

- Jukes et Cantor: un caractère peut se présenter sous 4 états différents avec des probabilités de changement toutes égales entre elles, les distances observées sont corrigées suivant la formule suivante :

$$D_{est(ij)} = -\frac{3}{4}\ln\left(1-\frac{4}{3}\left(1-s_{ij}\right)\right)$$

- Indice de Kimura (1980) où P et Q sont respectivement les fréquences de transitions et transversions (les quatre types de transition sont équiprobables, il en est de même des huit types de transversion):

$$D_{est(ij)} = -\frac{1}{2} Ln\left[(1-2P-Q)\left(\sqrt{(1-2Q)}\right)\right]$$

- Quelques autres modèles

Modèle	A/G/C/T	Pb SI	Pb Ve
Jukes et Cantor	A=G=C=T=25%	λ	λ
Kimura 2p	A=G=C=T=25%	α	β
Tamura 3p	A+T=1-θ, G+C=θ	α	β
Tajima et Nei 1p	A≠G≠C≠T	λ	λ
Hasegawa HKY85 2p	A≠G≠C≠T	α	β
Tamura et Nei 3p	A≠G≠C≠T	α1(Pyr) et α2(Pur)	β
Modèle à 8 p	A≠G≠C≠T	α1, α2, α3 et α4	β1, β2, β3 et β4

Cas du modèle à 8 paramètres

Mutant / Normal	A	T	C	G4
A	-	β2	β3	α4
T	β1	-	α3	β4
C	β1	α2	-	β4
G	α1	β2	β3	-

Parmi ces modèles, dans les deux premiers la fréquence à l'équilibre des 4 nucléotides est 25% ; la fréquence initiale en est quelconque, ces modèles sont dits non stationnaires. Par contre les estimateurs de distance dans tous les autres modèles nécessitent que les fréquences des 4 nucléotides restent les mêmes tout au long du processus évolutif : modèles stationnaires.

Des tests statistiques ont été proposés pour vérifier ou infirmer ces modèles (A.Rzhetsky et M.Nei ; Mol. Biol. Evol. 12 pp131-51 (1995)).

Tests statistiques

- Test de l'invariant unique: sous le modèle de JC les paires AG et TC (transitions=P) sont 2 fois moins observables que les autres (transversions=Q). On attend donc : 2P-Q=0. On va donc estimer l'écart de JC à sa valeur théorique 0 avec

$JC = \sum_{i<j}^{n} (2P_{ij} - Q_{ij})c$ .Pour cela, on calcule la variance ($V = \sum_{i=1}^{i=N} \frac{(x_i - x)^2}{N}$) de JC (V(JC)) et on compare $\frac{|JC|}{\sqrt{V(JC)}} \rangle z_{\alpha/2}$ où α représente le degré de signification souhaité

et z la valeur seuil au-delà de laquelle la courbe de la probabilité a la surface $\alpha/2$. Dans le modèle Kimura cela revient à tester si α et β sont égaux. Dans le modèle Kimura au sein des transversions on peut distinguer les paires AT+GC=T et AC+GT= U qui doivent être égales. Le test va donc mesurer la probabilité que K soit significativement différent de 0 (rejet de l'hypothèse) $K = \sum_{i\langle j}^{m} T_{ij} - U_{ij}$

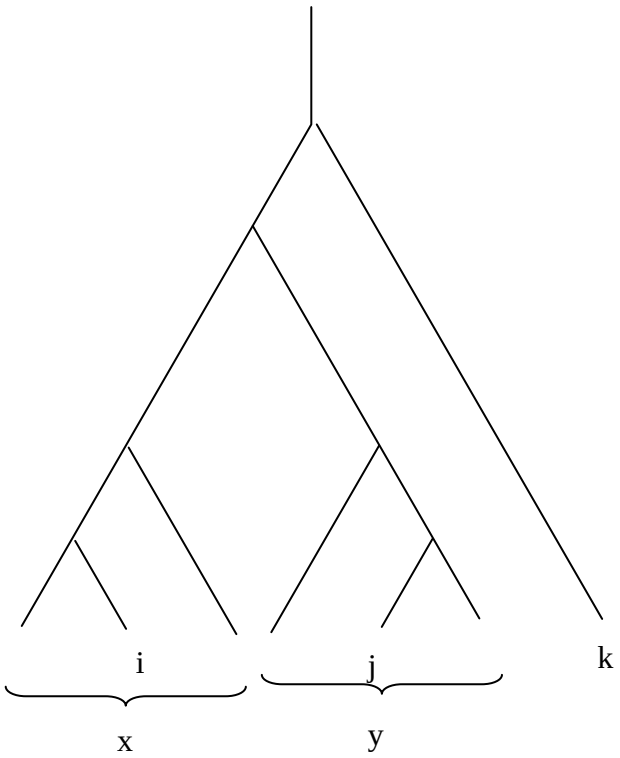
- Test de stationnarité Dans les autres modèles, à l'équilibre la probabilité g du nucléotide x dans la séquence 1,2, … ou m est la même : $g_{x1} = g_{x2} = ... = g_{xm}$. C'est ce que l'on va tester.
- Test des invariants multiples. Si le test précédent a établi que la fréquence des nucléotides remplissait bien la condition de stationnarité, on va chercher quel est le modèle le plus simple qui rende compte des données. On va considérer 10 couples de changements possibles (les changements réciproques étant de même probabilité) : AA, AT, AC, AG, TT, TC, TG, CC, CG, GG avec $AA = X_1, AT = X_2, ...$ Pour chaque modèle il est possible d'écrire pour X_i une équation de la forme $(\sum_s a_s X_s) + b = 0$ ou s indique le sème nucléotide, a et b les paramètres de chaque modèle. Le modèle de Kimura revient alors à $a_2 = a_7 = 1$, $a_3 = a_9 = -1$ et tous les autres a et b sont nuls soit $X_{AT} - X_{AC} - X_{GT} + X_{GC} = 0$

Procédures

Les méthodes décrites ici sont des méthodes agglomératives : après construction du tableau de distance entre tous les taxons pris deux à deux, on commence à regrouper deux Unités Evolutives en une Unité Evolutive Hypothétique. On reconstruit un tableau de distances en remplaçant ces deux UE par l'UEH et on agglomère de nouveau deux UE(ou UEH). Cette étape est recommencée jusqu'à ce que tous les taxons soient inclus dans l'arbre.

UPGMA

Ce qui signifie Unweighted Pair-Group Method of Arithmetic average.
Dans cette méthode le critère de regroupement de deux UE est la plus grande proximité : après le regroupement des deux UE les plus proches, on les remplace par une UE Hypothétique et on recommence à chercher les deux UE (ou UEH les plus proches) en calculant les distances entre UE et UEH comme une moyenne entre toutes les UE que comprend l'UEH.



$$d_{xy} = \frac{1}{rs} \sum_{i=1}^{r} \sum_{j=1}^{s} d_{ij}$$

avec r et s étant le nombre de UE comprises respectivement dans les UEH x et y.

WPGMA non rencontré dans les logiciels usuels calcule la distance entre deux UEH de façon un peu différente :

$d_{xy} = \sum_{i=1}^{r} \sum_{j=1}^{s} \frac{1}{2^{ci}} \frac{1}{2^{cj}} d_{ij}$ où ci et cj représentent le nombre d'étapes précédant l'étape d'agglomération de x et y

Les arbres obtenus par cette méthode sont obligatoirement racinés puisque la distance est répartie de façon uniforme sur chaque branche. Pour que cette méthode soit applicable l'horloge moléculaire doit être respectée.

	Tetrahy	Ginkgo	Epinard	Sureau	Poireau	Mouche	Bonite	Lapin	Rat	Cheval
Tetrahy	0									
Ginkgo	68	0								
Epinard	72	19	0							
Sureau	66	15	17	0						
Poireau	61	15	12	9	0					
Mouche	69	44	46	50	42	0				
Bonite	68	45	48	51	42	23	0			
Lapin	68	40	45	48	40	21	17	0		
Rat	69	39	44	47	39	20	16	2	0	
Cheval	68	43	48	50	42	22	18	6	6	0

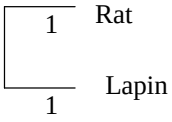
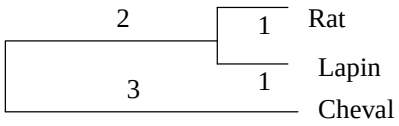


Figure II- 6. La première étape d'une procédure UPGMA. Choix des UE les plus proches et début du processus agglomératif. La distance est également répartie sur les deux branches.

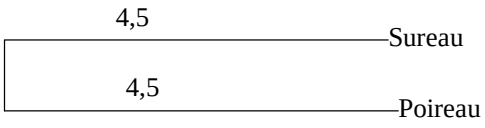
Etape 2

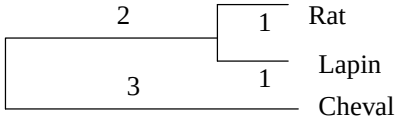
	Tetrahym.	Ginkgo	Epinard	Sureau	Poireau	Mouche	Bonite	L + R	Cheval
Tetrahym.	0								
Ginkgo	68	0							
Epinard	72	19	0						
Sureau	66	15	17	0					
Poireau	61	15	12	9	0				
Mouche	69	44	46	50	42	0			
Bonite	68	45	48	51	42	23	0		
L + R	68,5	39,5	44,5	47,5	39,5	20,5	16,5	0	
Cheval	68	43	48	50	42	22	18	6	0



Etape 3

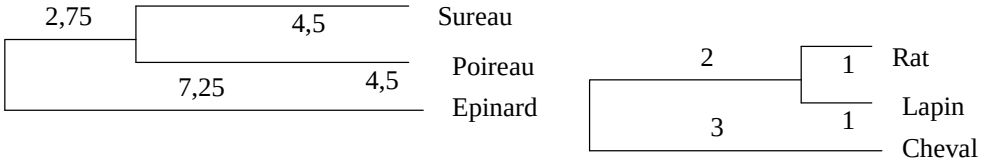
	Tetrahym.	Ginkgo	Epinard	Sureau	Poireau	Mouche	Bonite	L+R+C
Tetrahym.	0							
Ginkgo	68	0						
Epinard	72	19	0					
Sureau	66	15	17	0				
Poireau	61	15	12	9	0			
Mouche	69	44	46	50	42	0		
Bonite	68	45	48	51	42	23	0	
L+R+C	68,33	40,67	45,67	48,33	40,33	21	17	0





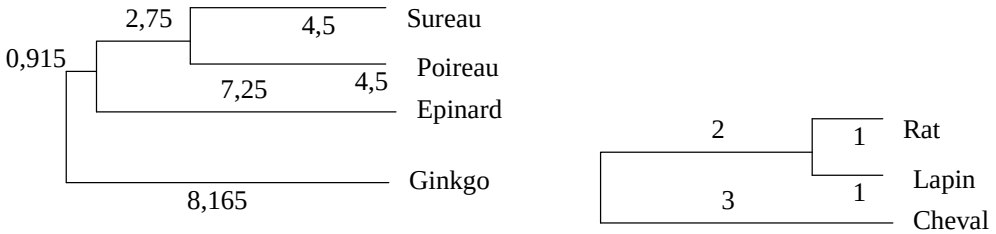
Etape 4

	Tetrahymena	Ginkgo	Epinard	S+P	Mouche	Bonite	L+R+C
Tetrahymena	0						
Ginkgo	68	0					
Epinard	72	19	0				
S+P	68	15	14,5	0			
Mouche	69	44	46	50	42	0	
Bonite	68	45	48	51	42	23	
L+R+C	68,33	40,67	45,67	44,33	21	17	0



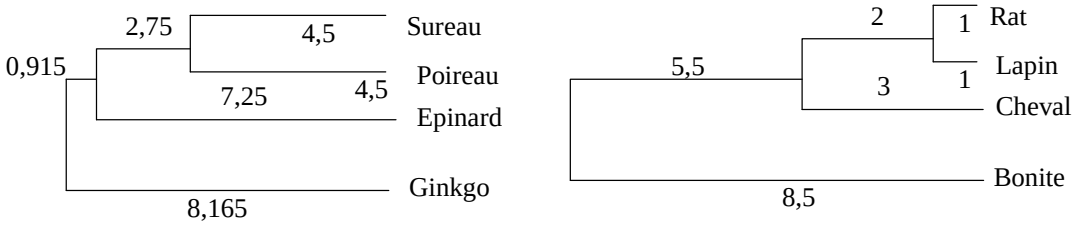
Etape 5

	Tetrahymena	Ginkgo	S+P+E	Mouche	Bonite	L+R+C
Tetrahymena	0					
Ginkgo	68	0				
S+P+E	69,33	16,33	0			
Mouche	69	44	46	42	0	
Bonite	68	45	47	42	23	
L+R+C	68,33	40,67	44,78	21	17	0



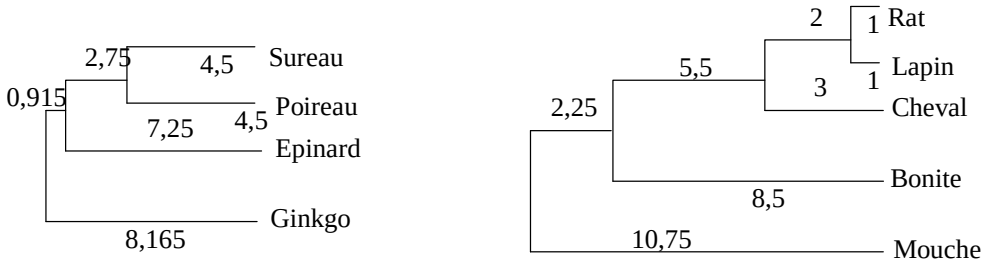
Etape 6

	Tetrahymena	S+P+E+G	Mouche	Bonite	L+R+C
Tetrahymena	0				
S+P+E+G	69	0			
Mouche	69	45,5	0		
Bonite	68	46,5	42	0	
L+R+C	68,33	43,75	21	17	0



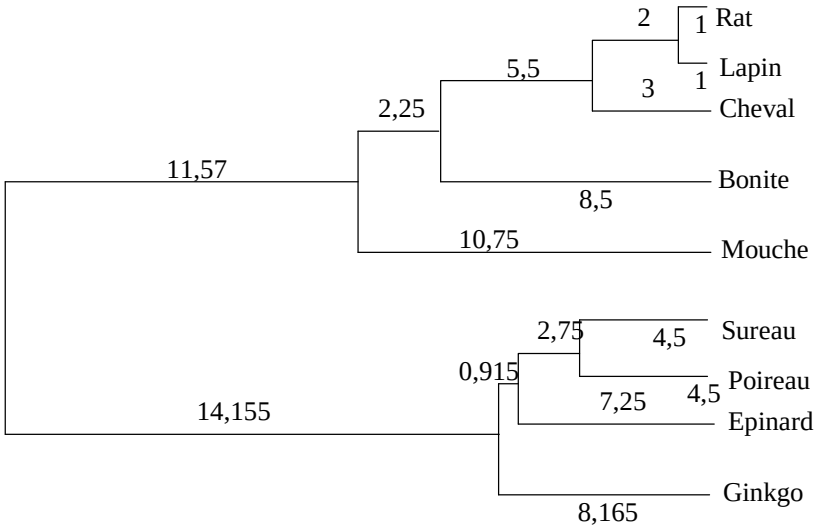
Etape 7

	Tetrahymena	S+P+E+G	Mouche	R+L+C+B
Tetrahymena	0			
S+P+E+G	69	0		
Mouche	61	45,5	0	
R+L+C+B	68,5	44,43	21,5	0



Etape 8

	Tetrahymena	S+P+E+G	R+L+C+B+M
Tetrahymena	0		
S+P+E+G	69	0	
R+L+C+B+M	67	44,64	0



Etape 9

	Tetrahymena	S+P+E+G+R+L+C+B+M
Tetrahymena	0	
S+P+E+G+R+L+C+B+M	67,89	0

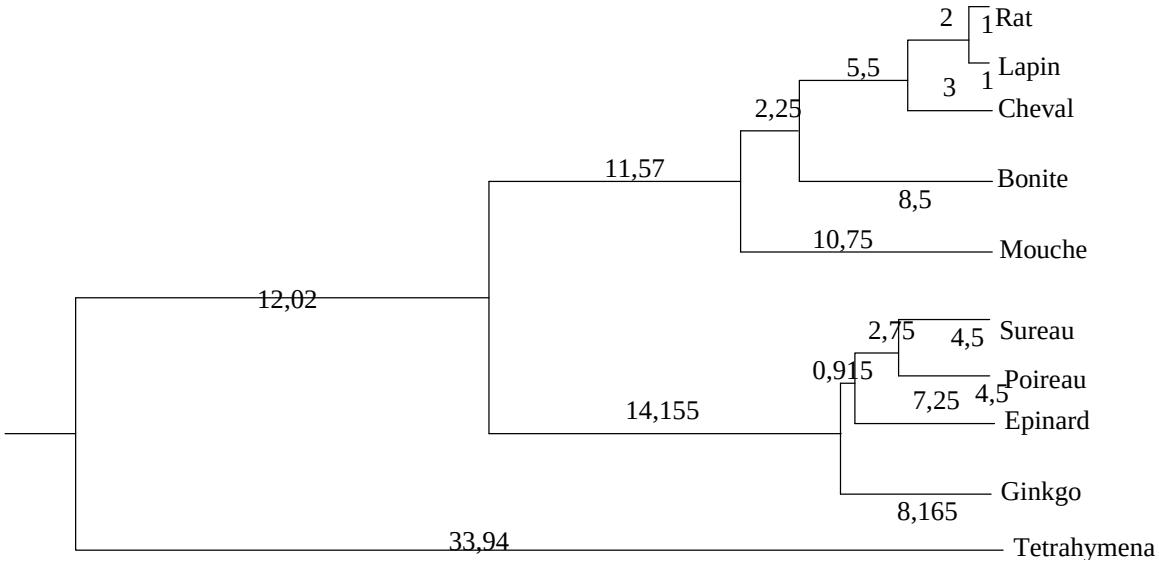


Figure II- 7. Les étapes successive pour construire l'arbre de tous les taxons par la méthode UPGMA.

NJ pour Neighbor Joining

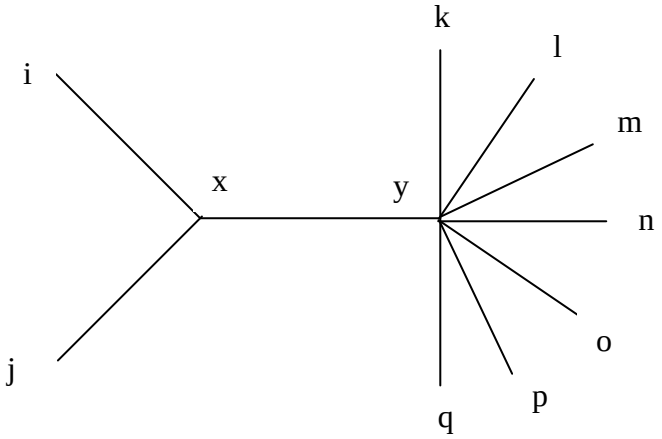
Etablissement de la formule qui permet la programmation du calcul.

$$Q = \sum_{i<j} D_{ij}$$

Dans l'arbre le moins résolu, l'arbre étoile la somme de toutes les branches est égale à

$$S_0 = \frac{Q}{(n-1)}$$

Dans la pratique, il peut exister des arbres plus courts en agglomérant deux taxons terminaux selon un schéma du type ci-dessous.



D représente une distance observée
B représente une distance estimée

La longueur de l'arbre correspondant s'exprime :

$$S_{ij} = B_{ix} + B_{jx} + B_{xy} + \sum_{k \neq i \neq j} B_{yk} \qquad 1$$

B représente la longueur estimée des différentes branches de l'arbre. Les distances observées D s'expriment de la façon suivante :

$$D_{ij} = B_{ix} + B_{jx} \qquad 2$$

$$D_{ik} = B_{ix} + B_{xy} + B_{yk}$$

et il y a n-2 distances de ce type donc

$$\sum_{k \neq i \neq j} D_{ik} = (n-2)(B_{ix} + B_{xy}) + \sum_{k \neq i \neq j} B_{yk} \qquad 3$$

De la même manière pour les distances de j à tous les autres différents de i

$$D_{jk} = B_{jix} + B_{xy} + B_{yk}$$

et il y a n-2 distances de ce type donc

$$\sum_{k\neq i\neq j} D_{jk} = (n-2)(B_{jx} + B_{xy}) + \sum_{k\neq i\neq j} B_{yk} \qquad 4$$

$$D_{kl} = B_{yk} + B_{yl}$$

soit pour toutes les distances entre les n-2 otus de l'étoile :

$$\sum_{k,l\neq i,j} D_{kl} = (n-3)\sum B_{yk} \qquad 5$$

Pour résumer on somme 1, 3, 4 et 5:

$$D_{ij} = B_{ix} + B_{jx}$$

$$\sum_{k\neq i\neq j} D_{ik} = (n-2)(B_{ix} + B_{xy}) + \sum_{k\neq i\neq j} B_{yk}$$

$$\sum_{k\neq i\neq j} D_{jk} = (n-2)(B_{jx} + B_{xy}) + \sum_{k\neq i\neq j} B_{yk}$$

$$\sum_{k,l\neq i,j} D_{kl} = (n-3)\sum B_{yk}$$

$$Q = (B_{ix} + B_{jx})(1+n-2) + 2(n-2)B_{xy} + (2+n-3)\sum B_{yk}$$

$$Q = (n-1)D_{ij} + 2(n-2)B_{xy} + (n-1)\sum B_{yk} \qquad 6$$

D'autre part, d'après (5) :

$$\sum_{k\neq i\neq j} B_{yk} = \frac{\sum_{k,l\neq i,j} D_{kl}}{(n-3)} \qquad 7$$

et

$$\sum_{k,l\neq i,j} D_{kl} = Q - R_i - R_j + D_{ij} \qquad 8$$

avec

$R_i = \sum_{j\neq i}^{n} D_{ij}$ distances de l'OTU i à toutes les autres

$R_j = \sum_{i\neq j}^{n} D_{ij}$ distances de l'OTU j à toutes les autres

On tire de 6

$$B_{xy} = \frac{Q - (n-1)D_{ij} - (n-1)\sum_{k\neq i\neq j}^{n} B_{yk}}{2(n-2)} \qquad 9$$

En substituant dans 1 les valeurs données par 2 et 9

$$S_{ij} = B_{ix} + B_{jx} + B_{xy} + \sum Byk \qquad (1)$$

$$S_{ij} = D_{ij} + \frac{Q-(n-1)D_{ij} - (n-1)\sum B_{yk}}{2(n-2)} + \sum B_{yk}$$

$$S_{ij} = \frac{\left[2(n-2)-(n-1)\right] D_{ij} + Q + \left[2(n-2)-(n-1)\right] \sum B_{yk}}{2}$$

$$S_{ij} = \frac{(2n-4-n+1)\ D_{ij} + Q + (n-3)\sum B_{yk}}{2(n-2)}$$

Puis en utilisant la valeur de $\sum B_{yk}$ exprimée dans 7 et en remplaçant ensuite $\sum D_{kl}$ par sa valeur donnée en 8

$$S_{ij} = \frac{(n-3)\ D_{ij} + Q + \sum D_{kl}}{2(n-2)} = \frac{(n-3)\ D_{ij} + Q + Q - R_i - R_j + D_{ij}}{2(n-2)}$$

$$S_{ij} = \frac{(n-2)\ D_{ij} + 2Q - R_i - R_j}{2(n-2)}$$

$$\boxed{S_{ij} = \frac{D_{ij}}{2} + \frac{2Q - R_i - R_j}{2(n-2)}}$$

Le calcul des longueurs de branches se fait selon la formule suivante :

$$D_{ix} = \frac{D_{ij}}{2} + \frac{R_i - R_j}{2(n-2)}$$

où x représente l'ancêtre hypothétique commun à i et j.

il faut donc noter que cette méthode est basée sur l'existence d'une horloge moléculaire, mais qu'elle tente d'en corriger les irrégularités : si un taxon est sur une longue branche, sa distance à tous les autres sera augmentée ; c'est le principe de la correction

UPGMA	NJ
•Distances ultramétriques •Vitesse constante sur toutes les branches •Arbre raciné	•Distances métriques et additives •Voir la formule •Arbre non raciné

Tableau II- 1. Comparaison des caractéristiques des deux processus de construction de phénogramme les plus utilisés.

Ces deux méthodes donnent des résultats qui peuvent être un peu différents.

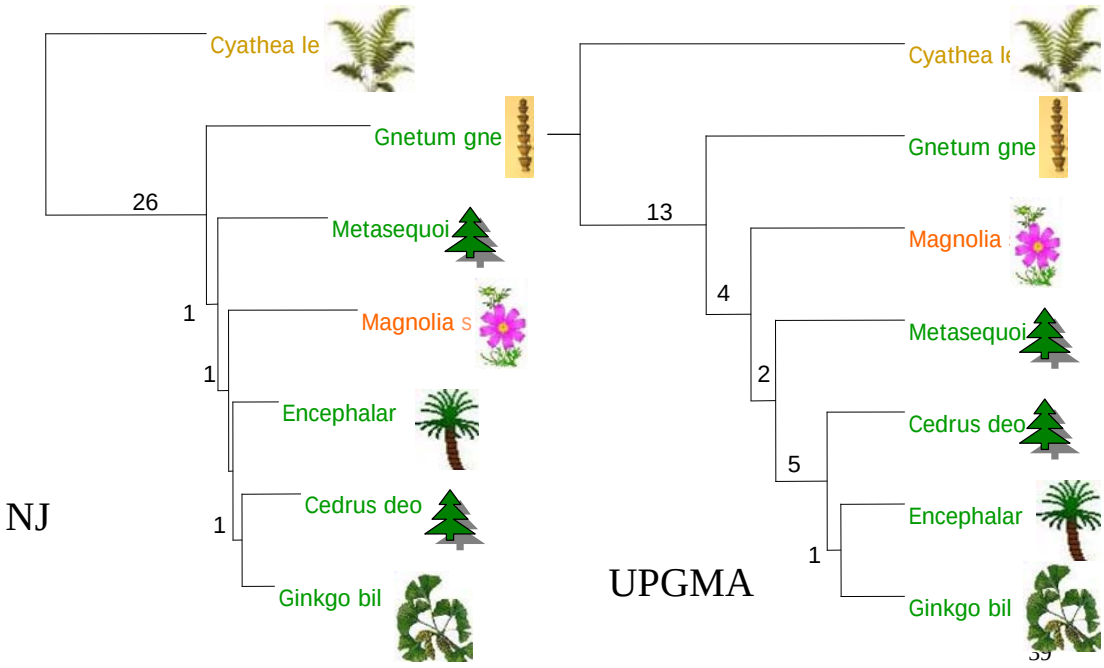


Figure II- 8. Arbres à 7 taxons obtenus avec NJ ou UPGMA. Les deux arbres sont orientés de la même façon pour les comparer commodément, cependant celui de gauche n'est pas raciné, alors que celui de droite l'est.