Méthodes probabilistes

Maximum de vraisemblance

Le principe de la méthode est dû à R.A. Fisher en 1922. Elle dépend de la complétude des données, et de la probabilité qu'a le modèle choisi de décrire les données. la probabilité d'observer les données expérimentales sous le modèle choisi dépend de la valeur des paramètres de ce modèle. la méthode du maximum de vraisemblance détermine ces paramètres pour que cette probabilité soit la plus grande possible.

L'application de ces méthodes probabilistes à la phylogénie fut suggérée tout d'abord par Edwards et Cavalli-Sforza en 1964. Mais à cette époque les problèmes de calcul posés étaient insolubles et on s'en tenait à des approximations simplifiées (la parcimonie appelée alors le minimum d'évolution et les méthodes d'estimation des phylogénies par les moindres carrées). Au fur et à mesure que les capacités de calcul croissaient, les modèles devenaient plus proches de la réalité biologique. En 1971 Neyman applique le ML à des séquences (AA ou nt) en utilisant un modèle de changements symétriques posant que les changements se faisaient au hasard et sans influence les uns sur les autres. En 1981 Felsenstein est le premier a proposer des implémentations qui mettent en jeu un vrai modèle biologique (cf. les modèles ci-après). ML inclut de façon explicite un modèle de substitution dans la procédure d'estimation comme le font les méthodes de distance alors que la parcimonie le fait implicitement.

Le problème des longues branches

Les méthodes cladistes peuvent présenter des biais dans certains cas.

Exemple

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
1	A	A	A	A	A	C	C	C	C	G	G	G	G	T	T	T	T
2	C	A	C	G	T	A	C	G	T	A	C	G	T	A	C	G	T
3	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A
4	C	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A

Tableau IV- 1. Jeu de données.

Dans le jeu de données présenté au tableau 5, les caractères 7, 12 et 17 sont des synapomorphies pour les taxons 1 et 2. Par contre le caractère 1 est contradictoire. Une méthode cladiste donne l'arbre le plus parcimonieux :

Etant donné que les branches 1 et 2 sont beaucoup plus longues que les deux autres il est très probable que tous les caractères sur ces deux branches ont évolué plus rapidement et que les caractères 7, 12 et 17 ne soient pas des synapomorphies.

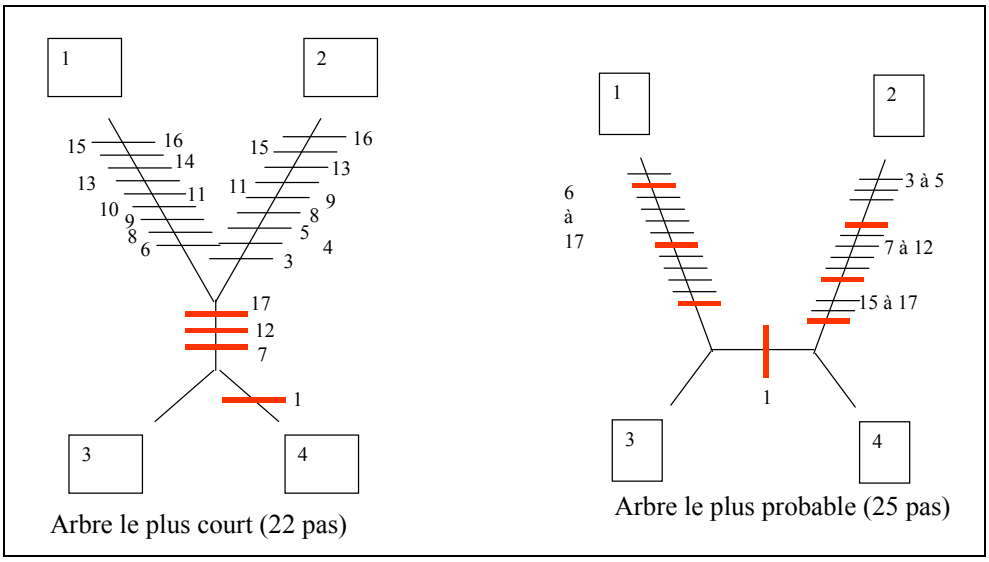


Figure IV- 1. A gauche, l'arbre le plus parcimonieux de 22 pas, a droite l'arbre le plus probable de 25 pas.

Cet exemple montre le problème lorsque tous les états de caractères sont limités à 4. Les caractères 2 à 17 représentent tous les états possibles pour les taxons 1 et 2

Avec un grand nombre de caractères on peut espérer que dans l'ensemble cela reflète ce que l'on peut décrire avec ces états 2 à 17.

En considérant un seul de ces caractères si l'un des taxons présente un C alors que tous les autres ont A à la place, la solution la plus parcimonieuse est un seul changement.

Mais, ce caractère a pu devenir C à la suite de multiples changements (5, 10 ou n)

Le même argument peut être appliqué à une autre branche qui présente l'état C.

Dans ce cas, les deux taxons ont tous les deux C mais uniquement par hasard et non en vertu d'une histoire commune. La différence avec l'arbre le plus court (le plus parcimonieux) est dans la vitesse variable de substitutions le long de différentes branches.

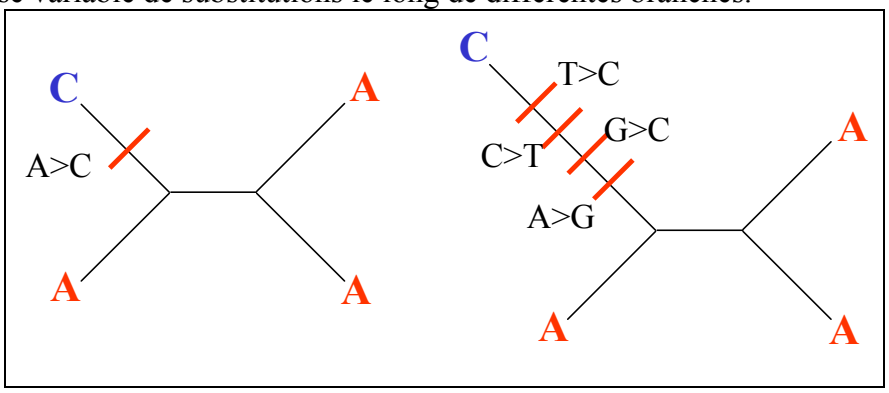


Figure IV- 2. Différentes interprétations pour une même distribution des états d'un caractère.

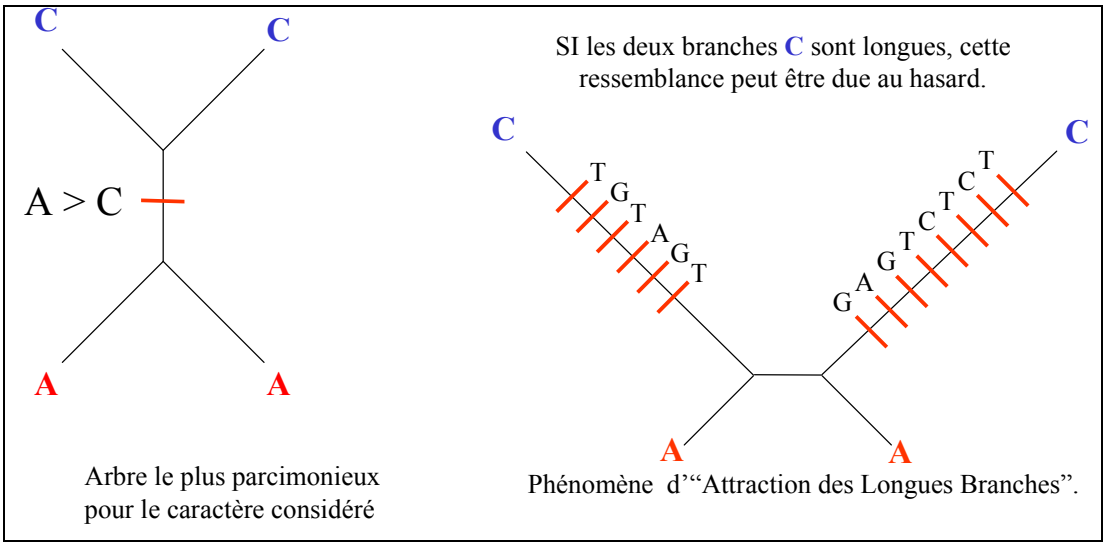


Figure IV- 3.Interprétation du phénomène d'attraction des branches longues.

C'est ainsi que l'on rend compte de l'attraction des branches longues.

Avec la méthode du maximum de vraisemblance, on va tenir compte de la longueur des branches. Comment le faire sans avoir une horloge vous demandez-vous ? Impossible, au moins directement ; néanmoins avec quelques hypothèses supplémentaires on peut utiliser d'autres informations pour en inférer ce qu'il a pu se passer à un site donné. Sur l'ensemble d'une séquence assez longue, même la parcimonie va mettre les taxons 1 et 2 au bout de longues branches. Ce sont soit des autapomorphies soit des homoplasies qui sont requises par l'arbre le plus parcimonieux. Autrement dit, la longueur des branches et donc l'espérance pour chaque caractère est basée sur l'information collectée sur l'ensemble des caractères.

Principe

Calcul de la probabilité d'un arbre

Sur un arbre le plus parcimonieux on obtient les longueurs de branche corrigées par la technique de Jukes et Cantor .

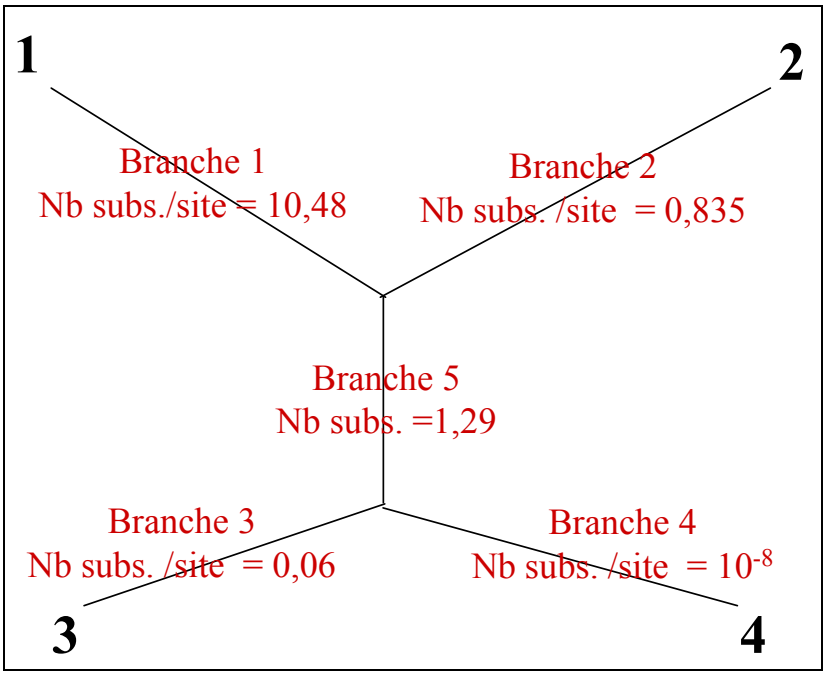


Figure IV- 4. Arbre obtenu par parcimonie avec des longueurs de branches calculées sur le nombre moyen de changements obtenus sur l'ensemble de tous les sites constants, variables et informatifs puis corrigés (JC69).

Ce qui permet de calculer la probabilité de changement par site sur chaque branche de cet arbre, en admettant que ces changements sont rares et que leur probabilité est donnée par une loi de Poisson

$$P_k = \frac{e^{-m} m^k}{k!}$$

qui a pour moyenne m. Il n'y a aucun changement pour un caractère si k=0 soit avec la probabilité

$$P_0 = \frac{e^{-m} m^0}{0!} = e^{-m}$$

Quant à la probabilité d'observer un changement c'est le reste des cas possibles

$$1\text{-}P_0$$

branche	Nb subs.	Pb 1 chgmt	Pb 0 chgmt
1	10,48	0,827	0,173
2	0,835	0,066	0,934
3	0,06	0,0047	0,9953
4	? 0	? 0	? 1
5	1,29	0,102	0,898

Tableau IV- 2. Probabilité de changement sur chacune des branches de l'arbre de la figure 80.

Pour un caractère donné on a la distribution suivante :

1	C
2	C
3	A
4	A

Si l'on utilise les longueurs de branches pour estimer la probabilité de changement sur les branches, pour ce caractère, l'arbre (1,2)(3,4) a la probabilité :
p(c,c | 1)* p(c,c | 2)* p(a,a | 3)* p(a,a | 4)*p(c,a | 5) voir le premier cas de changements d'états, figure suivante)
mais si l'on prend le second cas (1er,ligne2) de changements d'états, l'arbre (1,2(3,4)) cette probabilité devient
p(c,a | 1)* p(c,a | 2)* p(a,a | 3)* p(a,a | 4)* p(a,a | 5)

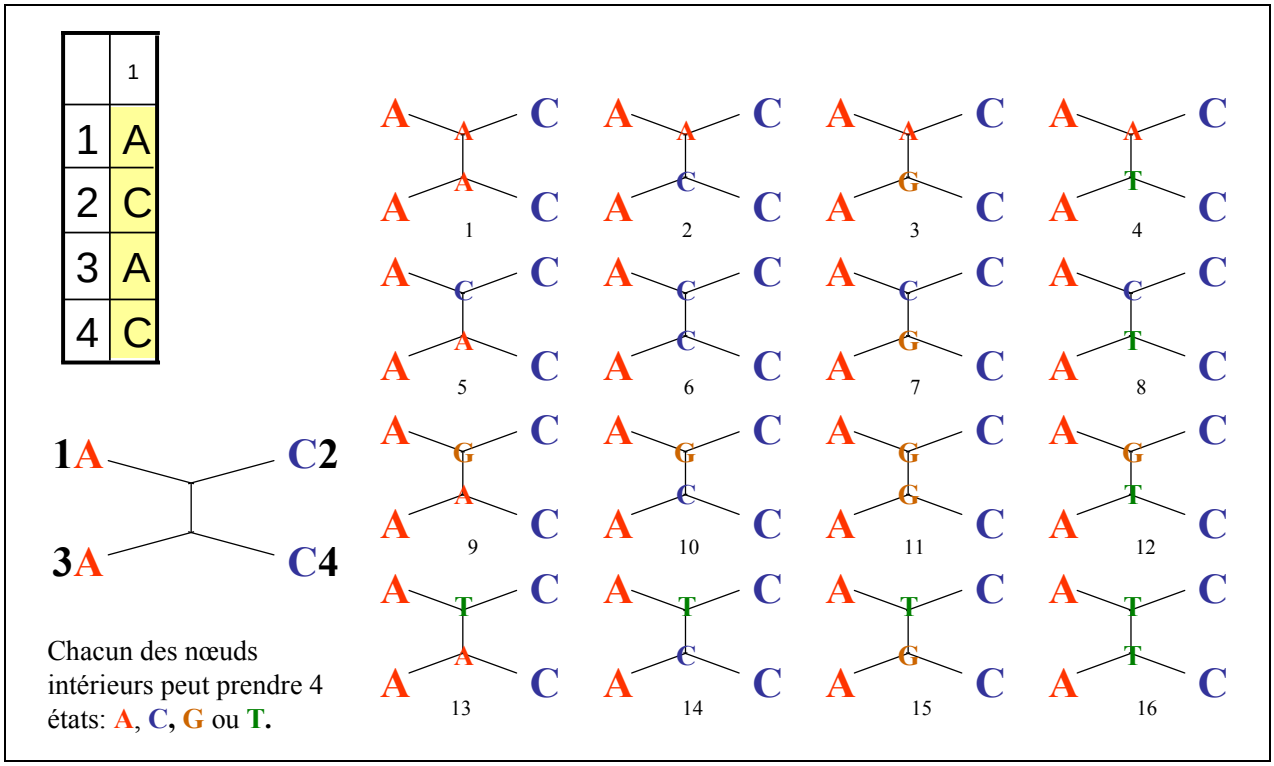


Figure IV- 5. Les 16 arbres possibles pour cette distribution de ces états de caractère.

Il y a ainsi 16 possibilités exclusives (OR) de rendre compte de cette répartition des états pour ce caractère parmi les 4 taxons, on doit donc sommer les probabilités pour avoir la probabilité totale de cette répartition des états. Cette somme donne donc la probabilité de cette distribution finale des états de caractères pour le caractère 1.
Certains de ces arbres présentent une différence sur la branche 4 et leur probabilité est négligeable (2, 3, 4, 6, 7, 8, 10, 11, 12, 14, 15 et 16). La probabilité des autres se calcule :

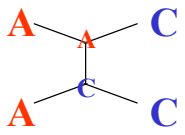
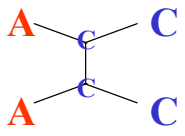
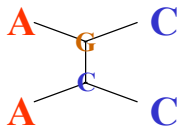
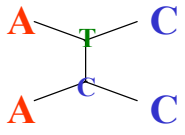
	Br 1	Br 2	Br 3	Br 4	Br 5	Pb globale
	0,173	0,066	0,0047	1	0,102	0,00000547
	0,827	0,934	0,0047	1	0,898	0,00328
	0,827	0,066	0,0047	1	0,102	0,0000262
	0,827	0,066	0,0047	1	0,102	0,0000262

Tableau IV- 3. Probabilité des arbres 2, 6, 10 et 14 pour le caractère 1.

Les seize arbres étant mutuellement exclusifs, il faut sommer leurs probabilités.

Pb = (0 *12) + 0,00000547 + 0,00328 + 0,0000262 + 0,0000262

Pb_1 = **0,00334**

Il faut refaire le même calcul pour chacun des 16 autre caractères (il y en a qui ont la même distribution et dont la probabilité sera la même : 1, 2, 3-4-5,6-10-14, 7-12-17, 8-9-11-13-15-16).

Il faut faire le même calcul pour chaque position.

La position 2 est constante

Pb_2 = **0,1441**

Les positions 7, 12 et 17 sont comparables.

$Pb_7 = Pb_{12} = Pb_{17}$ = **0,0735**

Les positions 3, 4, 5, 6, 8, 9, 10, 11, 13, 14, 15 et 16 sont comparables.

$Pb_3 = Pb_4 = Pb_5 = Pb_6 = Pb_8 = Pb_9 = Pb_{10} =$
$Pb_{11} = Pb_{13} = Pb_{14} = Pb_{15} = Pb_{16}$ = **0,09955**

Figure IV- 6. Probabilité pour chaque caractère de la matrice (tableau 5).

On obtient ainsi la probabilité de l'arbre le plus parcimonieux compte tenu de la longueur des branches estimées sur tous les caractères. Cette probabilité étant en général très faible (es : $0.2345*10^{-427}$ on préfère l'exprimer en log :log L=-427,6299 ou –lnL=427,6299. Si un autre arbre a une probabilité meilleure la valeur du logL sera plus grande (donc en valeur absolue, plus petite).

L'arbre préféré par ce type de méthode est celui qui présente la probabilité la meilleure.

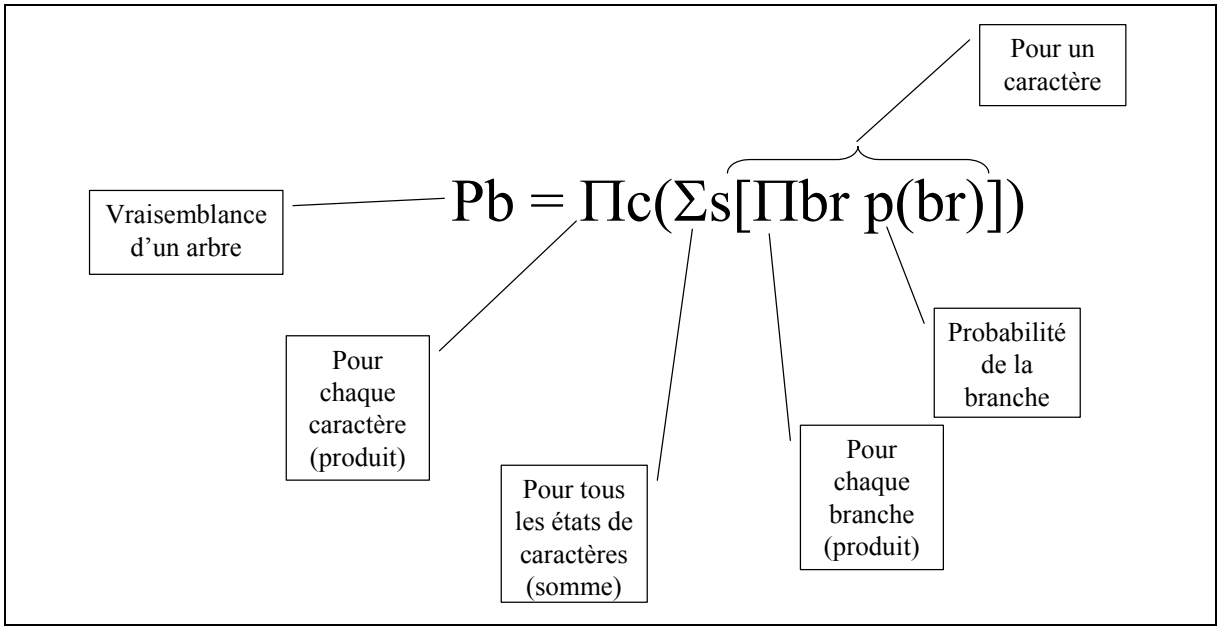


Figure IV- 7. Signification des différents paramètres intervenant dans le calcul de la probabilité d'un arbre.

Modèles et paramètres

La longueur des branches n'est pas le seul facteur qui peut être estimé. Les modèles à 4 états supposent l'indépendance des sites qui est indispensable dans les calculs d'arbres phylogénétiques. On sait cependant que ce n'est pas toujours le cas (exemple des ARNr). Il existe donc des modèles ou au lieu de considérer chaque site seul on prend les sites deux par deux afin de tenir compte de la succession et de l'interdépendance des sites. Certains auteurs (Dixon et Hill,1993) tiennent compte des structures tiges de l'ARNr pour pondérer les changements qui sont complémentaires. D'autres utilisent des matrices de substitutions 16*16 au lieu de 4*4 pour tenir compte de la non indépendance de la succession des sites (1995, Muse, Rzhetsky, Sonninger et von Haseler). D'autres modèles (Sharp, 1997) tiennent compte de la structure en codon et des propriétés différentes de substitution en fonction de la place du site dans ce codon ainsi que du résultat de cette substitution (changement synonyme ou pas). Dans de tels modèles on utilise fréquemment une matrice de substitution 61*61. On peut également tenir compte des propriétés physico-chimiques des acides aminés (Goldman et Yang 1993).

Fréquence des bases

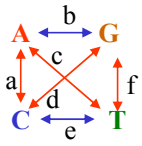
Dans les cas simplistes elle est considérée égale à 25% pour tous les organismes. Cependant on sait que cela peut varier. Si tous les organismes présentent une composition différente de 25% mais homogène entre eux, l'analyse par les différentes méthodes n'est pas perturbée. Par contre si un ou plusieurs organismes ont une composition globale significativement différente des autres cela peut provoquer une position aberrante dans l'arbre obtenu. Avec ML on peut tenir compte de ces différences et essayer de corriger les artéfacts qu'ils entraînent en choisissant des paramètres adaptés.

Rapport transitions / transversions

On a vu que les transitions possibles étaient au nombre de 4 alors qu'il y avait 8 transversions possibles. Si tous les changements sont équiprobables le rapport Si/Ve doit être

de 0.5. Cependant on admet souvent que les transitions sont plus faciles que les transversions (cf structure de l'ADN, voir également la notion de saturation) et l'on peut choisir de donner une valeur différente à ce rapport (qui signifie la probabilité relative d'un type de changement par rapport à l'autre). Sous cette appellation on peut entendre plusieurs choses, ce qui entretient la confusion.

- Si/Ve rate ratio c'est le rapport de la probabilité en un instant donné d'une Si à une Ve. on le nomme K. Si K est plus petit que 1 cela implique qu'à un instant la vitesse des Tv est supérieure à celle des Si.
- Si/Ve ratio représente en une unité de temps la probabilité d'une Si divisée par la probabilité d'une Ve. Si on a 2 fois plus de Ke que de Si, Si/Ve ratio = 0.5 quand K=1 à condition que A=C=G=T=25%.
- Enfin on peut distinguer Si/Ve pour l'ensemble des sites ou au contraire distinguer Si/Ve pour chaque site et dans ce cas il n'est pas forcément le même pour tous.



A = C = G = T = 0,25	A ≠C ≠ G ≠ T	transitions	transversions
JC (0)	F81 (3)	b=e=a=c=d=f	
K80 (1)	HKY85 (4)	b=e	a=c=d=f
TNef (2)	TN (5)	b, e	a=c=d=f
K81 (2)	K81uf (5)	b=e	a=f, c=d
TIMef (3)	TIM (6)	b, e	a=f, c=d
TVMef (4)	TVM (7)	b=e	a, f, c, d
SYM (5)	GTR (8)	b, e	a, f, c, d

Tableau IV- 4. Quelques modèles évolutifs pour l'ADN et leurs caractéristiques.

Deux modèles sont assez proches celui proposé en 1984 par Felsenstein (F84) et celui proposé en 1985 par Hasegawa, Kishino et Yano (HKY85).

HKY85					F84				
vers / de	A	G	C	T	vers / de	A	G	C	T
A	-	απG +βπG/πR	απC	απT	A	-	απG +βπG	απC	απT
G	απA +βπA/πR	-	απC	απT	G	απA +βπA	-	απC	απT
C	απA	απG	-	απT +βπT/πY	C	απA	απG	-	απT +βπT
T	απA	απG	απC +βπC/πY	-	T	απA	απG	απC +βπC	-

Tableau IV- 5. Caractéristiques des modèles HKY85 et F84.

Fréquence des sites variables

Jusqu'ici on a considéré que tous les sites variaient à la même vitesse, ce qui n'est pas réaliste. On peut corriger ce modèle en introduisant une certaine homogénéité parmi les sites. Plusieurs modèles sont disponibles. Ils peuvent être distribués selon une loi normale, ou une fraction des sites peut être invariant, ou encore la variation de vitesse peut suivre une loi Γ. La distribution Γ est une fonction de densité de probabilité continue (les distributions de χ2 etla distribution exponentielle en sont des cas particuliers). La forme de la distribution dépend de deux paramètres α (forme) et λ (échelle). Si l'on fixe α = 1 et λ = a, la variance est 1/a et la moyenne 1. Les vitesses sont déterminées comme des variables aléatoires d'une distribution Γ. Avec a égal à ∞ les sites sont à vitesse constante, si a<1 les différentes vitesses sont très hétérogènes. Expérimentalement on prend souvent des valeurs comprises entre 0.5 et 0.1.

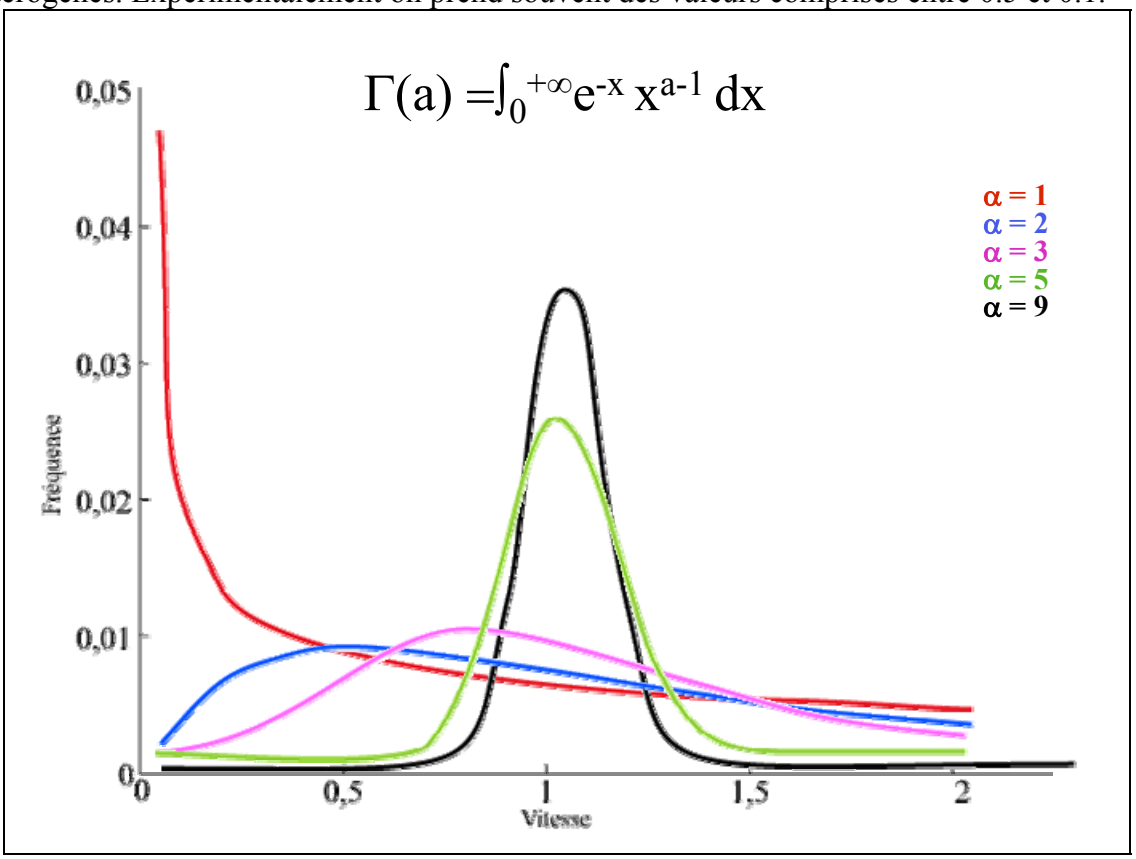


Figure IV- 8. La valeur du paramètre a modifie la forme de la courbe de distribution.

Beaucoup des modèles utilisés avec le maximum de vraisemblance sont impossibles à traiter algébriquement sous les autres méthodes. C'est grâce aux développements récents des capacités de calcul programmé que l'on peut appliquer ces modèles. Dans ce qui précède on a pu voir la déclaration explicite des hypothèses de la méthode du maximum de vraisemblance

- indépendance des sites les uns par rapport aux autres
- processus de substitution décrit à l'aide de paramètres divers
- longueur des branches de l'arbre phylogénétique
- hétérogénéité de la vitesse de transformation parmi les sites.

Avantages et inconvénients

Le maximum de vraisemblance, la parcimonie et certaines méthodes de distance sont robustes à la violation de nombreuses hypothèses, modèles de substitution inclus, variation de la vitesse entre site et non indépendance des sites. L'intérêt des hypothèses explicites est l'étude des mécanismes de l'évolution, en comparant différents modèles alternatifs.

Si au contraire, on ne s'intéresse qu'à la phylogénie des taxons, les paramètres supplémentaires générés par l'affinement du modèle sont alors considérés comme nuisibles.

Procédures

Tests d'hypothèses

On va comparer la probabilité d'un arbre sous diverses hypothèses évolutives.

Théorie

Si deux hypothèses sont à comparer , le rapport de leur probabilité est une bonne mesure.

$$\Delta = \frac{\max L_0(\mathrm{mod}\,\grave{e}le \cdot H0 \cdot appliqué \cdot aux \cdot données)}{\max L_1(\mathrm{mod}\,\grave{e}le \cdot H1 \cdot appliqué \cdot aux \cdot données)}$$

H0=général

H1=particulier

ou encore

$$\log \Delta = \ln L_0 - \ln L_1$$

Si $\Lambda<1$ c'est H1 qui gagne
Si $\Lambda>1$ c'est H0 qui gagne.
Ce second cas ($\Lambda>1$) ne peut être atteint que si H1 n'est pas un cas particulier de H0 (non nested models). Si H1 est un cas particulier de H0, $\Lambda<1$,puisqu'il y a plus de contraintes, et $-2\log\Lambda$ suit une distribution de $\chi 2$ avec un degré de probabilité q égal au nombre de paramètres en plus sous l'hypothèse H1.

On passe d'un modèle à l'autre en changeant un ou plusieurs paramètres.

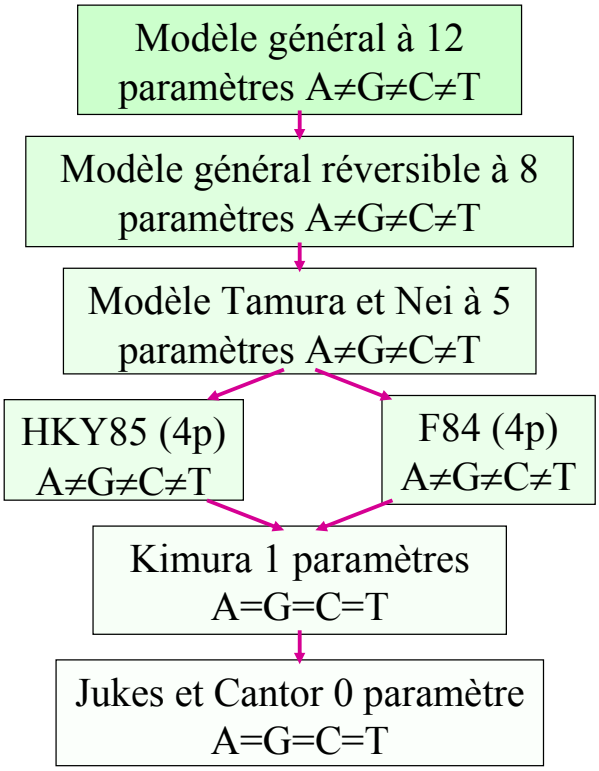


Figure IV- 9. Exemples de modèles dérivés (nested models).

On a pu montrer que cette distribution n'était pas toujours de type χ^2 aussi faut-il alors estimer le poids statistique de l'écart entre les deux probabilités à l'aide de simulations. On construit des jeux de données par simulation sous les hypothèses du modèle H0 et on calcule pour chaque la probabilité de l'arbre et la valeur $-2\log \Lambda$. On compare le nombre de fois où cette valeur par simulation a été plus forte que la valeur vraie. Si cette valeur est dépassée dans moins de 5% des simulations, on considère que la différence n'est pas significative et H1 est préféré (on ne choisit pas le cas particulier qui n'est pas forcément le meilleur).

Le logiciel ModelTest est fréquemment utilisé pour choisir un modèle d'évolution bien adapté aux séquences examinées. Les 14 modèles de la figure 8 sont essayés pour un arbre donné avec plusieurs options

- Tous les sites peuvent évoluer à la même vitesse ou à des vitesses différentes échantillonnées d'après une loi Γ (Γ)de distribution ou encore avec une proportion des sites invariables (I) ou la combinaison des deux (Γ + I)

Pour un modèle on dispose donc de 4 cas de figure :

1. JC
2. JC + G
3. JC + I
4. JC + G + I

Soit 56 modèles different tous derives les uns des autres.
La comparaison de ces modèles est faite selon deux méthodes.
Le test LRT (Likelihood Ratio Test)
Le critère d'information d'Akaïke (figure 86)

C'est une mesure de l'information perdue lorsqu'on utilise un modèle pour approcher une réalité.

$$AIC_i = -2\ln L_i + 2K_i$$

K = nombre de paramètres libres du modèle i

Pour comparer plusieurs modèles (dérivés ou non) on calcule la différence de leurs AIC

$$\Delta_i = AIC_i - \min AIC$$

Si Δ = 1 à 2 les modèles sont à prendre en considération

Si Δ = 3 à 7 les modèles ont beaucoup moins de support

Si Δ > 10 les modèles ne sont pas soutenus.

Dans le cas d'un échantillonnage petit (n/K<40, n étant le nombre de caractères de l'alignement) il convient d'utiliser l'AIC de second degré.

$$AIC_c = AIC + \frac{2K^*(K+1)}{n-K-1}$$

AIC = Akaike Information Criterion

Figure IV- 10. Le test d'Akaïke

Exemple

données sur la sérum albumine d'un poisson (Salso salar), d'un amphibien(Xenopus laevis), d'un oiseau(Gallus gallus), d'un rongeur (Rattus norvegicus) et d'un primate (Homo sapiens). On cherche à tester différents modèles concernant

1. la fréquence des bases JC69 contre F81
2. la vitesse de transition et de transversion F81 contre HKY85
3. la vitesse au sein des différents sites HKY85 contre HKY85+Γ
4. l'existence d'une horloge moléculaire. HKY85+Γ contre HKY85+Γ+horloge

Le choix des lignes 2, 3, etc est fonction du résultat de la ligne précédente.

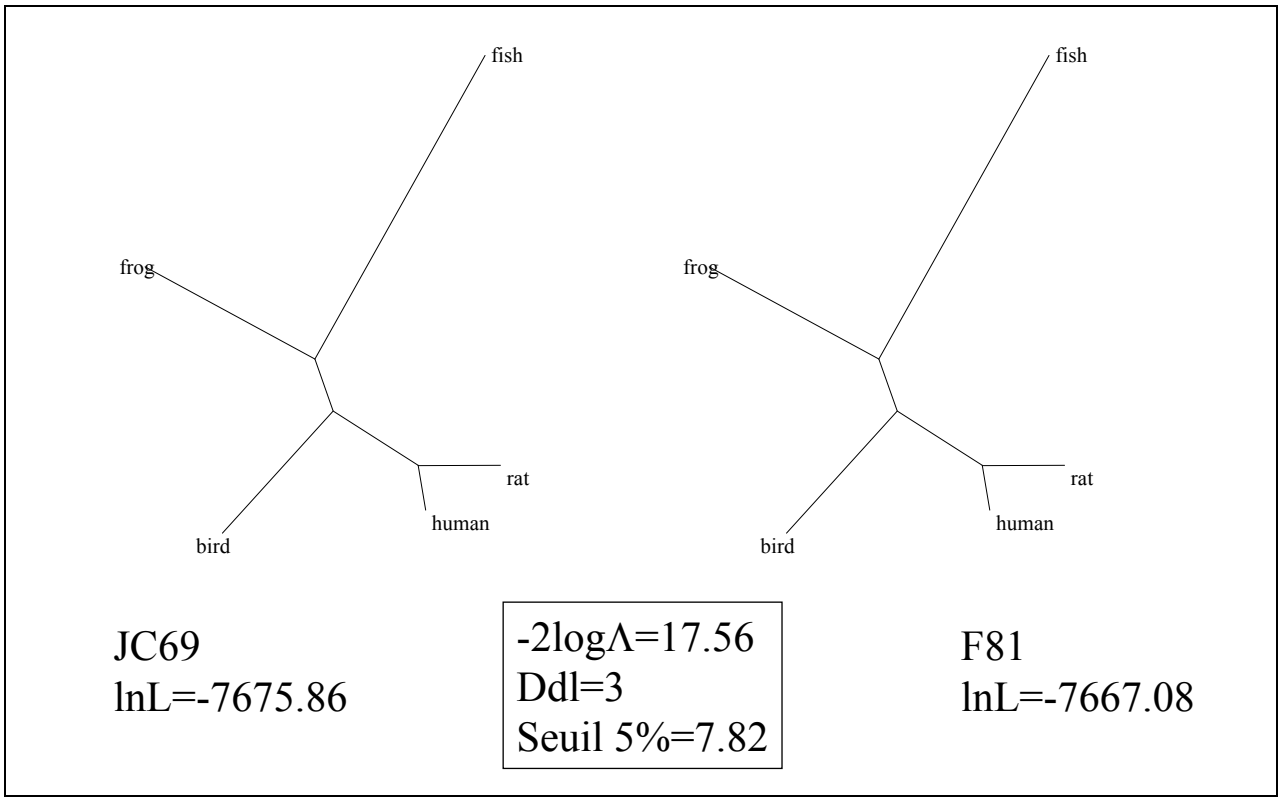


Figure IV- 11. La fréquence des bases est-elle égale ou non?

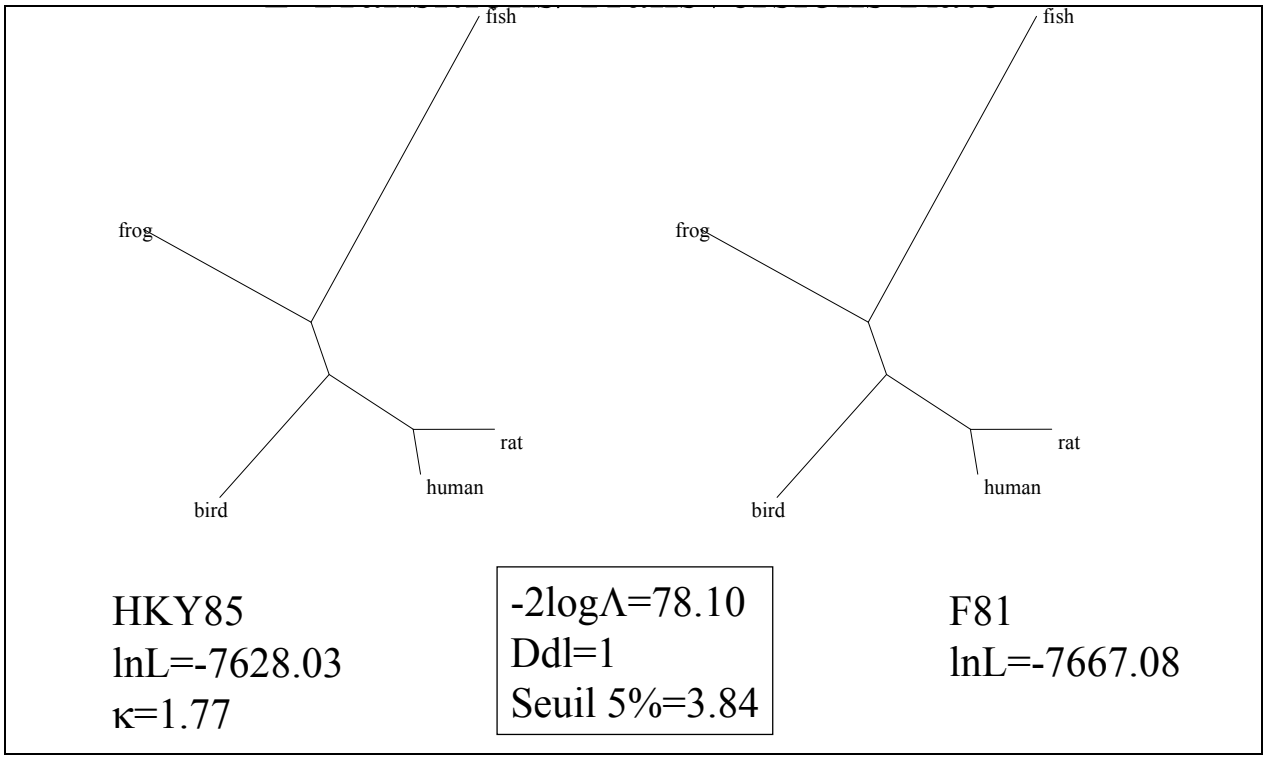


Figure IV- 12. Choix du rapport transitions/transversions.

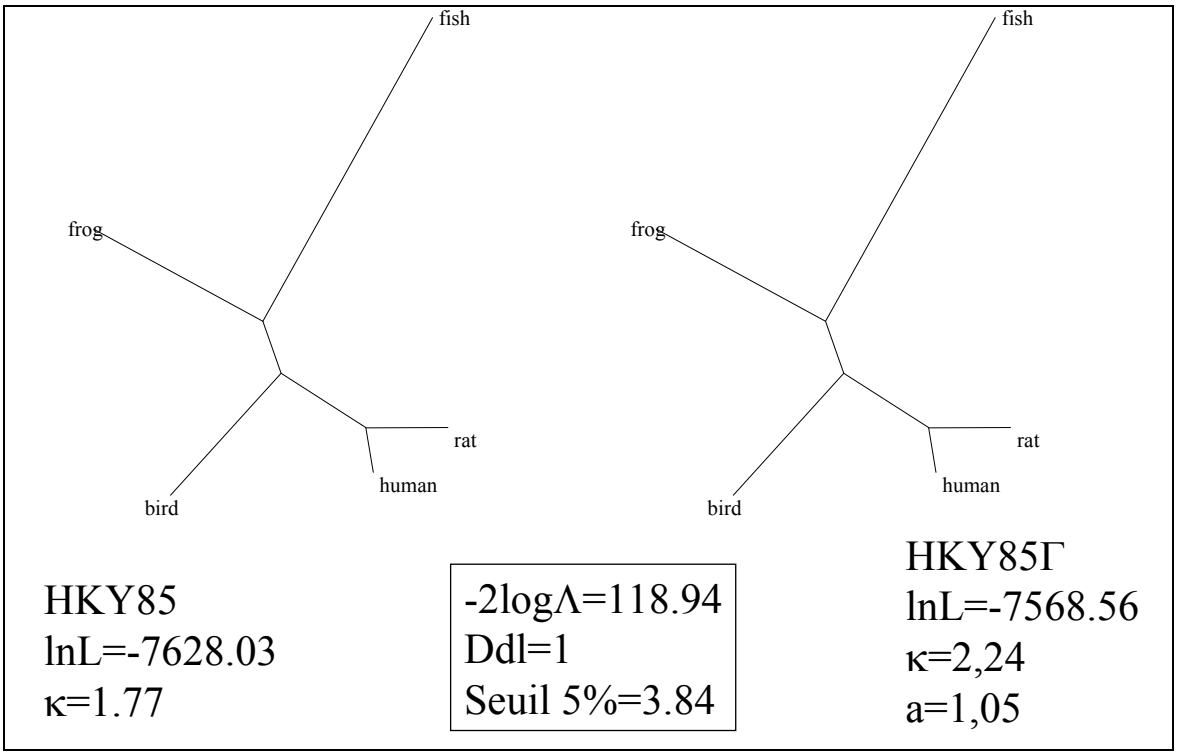


Figure IV- 13.Les différents sites évoluent-ils à la même vitesse ?

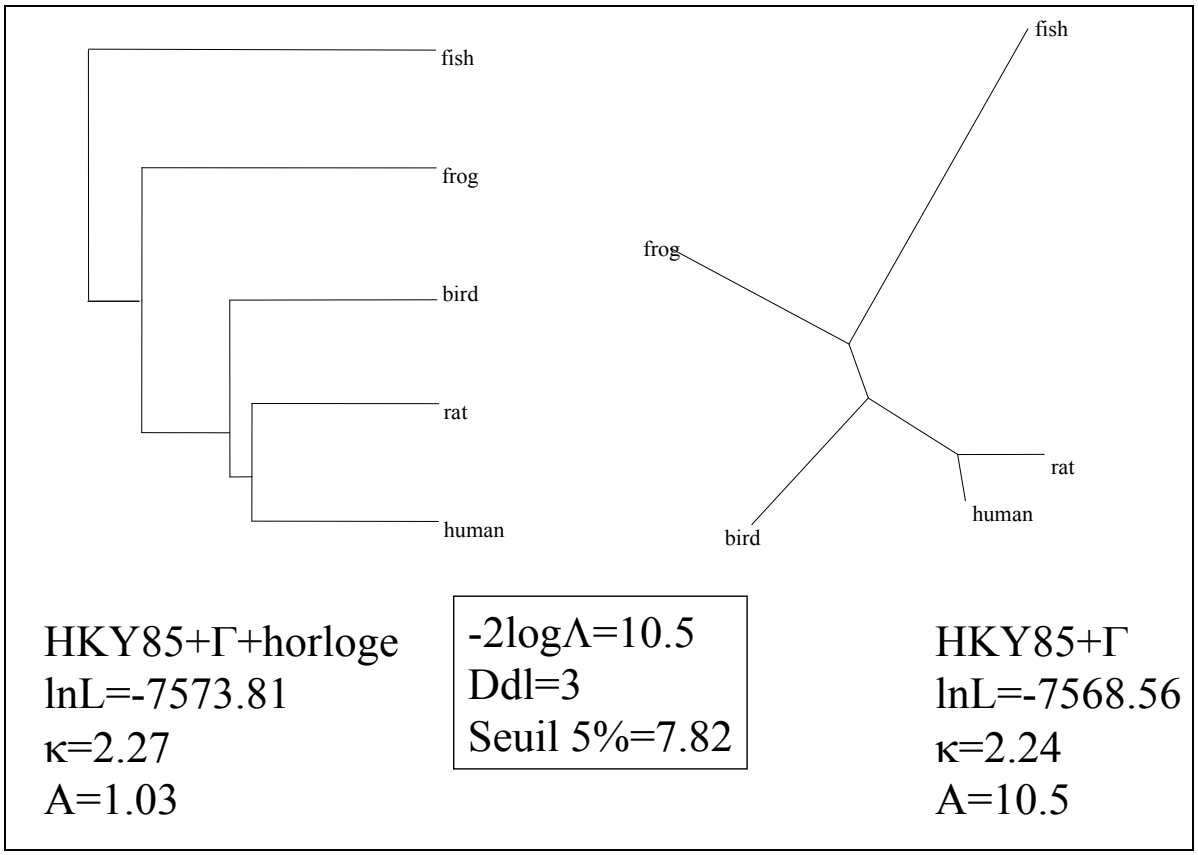


Figure IV- 14. L'horloge moléculaire est-elle respectée?

Comparaison de deux arbres

Ayant déterminé un arbre avec certaines méthodes, on peut vouloir le comparer à un autre obtenu avec les mêmes taxons , mais avec d'autres données.

cas des Chiroptrères (micro et macro) qui selon Pettigrow étudiant les caractères neurologiques trouve les mégachiroptères (le renard volant) plus proches des primates que des microchiroptères, ce qui implique soit que les chiroptères ne sont pas monophylétique et que la capacité de vol ait été acquise 2 fois indépendamment soit que les primates aient perdu secondairement cette aptitude. Le gène codant pour la protéine Interphotoreceptor Retinoid Binding Protein a été utilisé dans un programme de ML soit sans contrainte, soit avec une contrainte correspondant à l'hypothèse de Pettigrew.

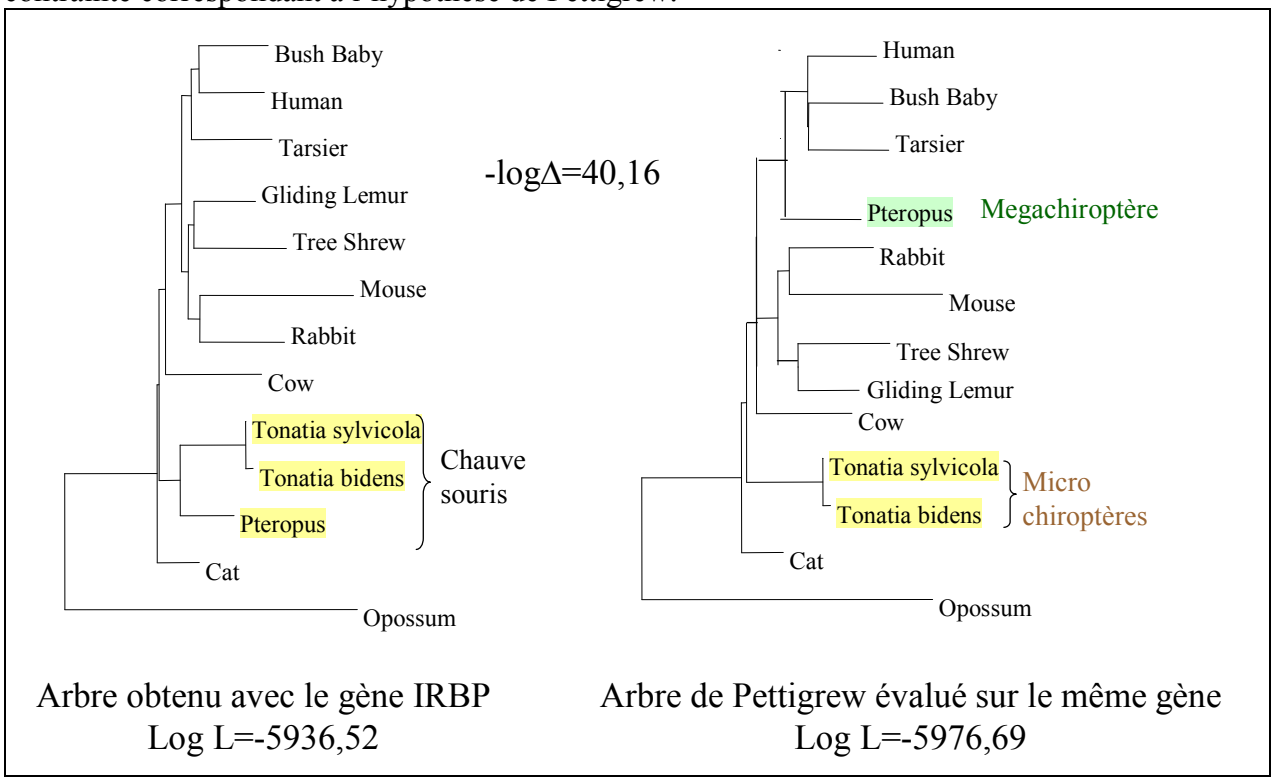


Figure IV- 15. Les deux arbres concurrents.

Les deux arbres n'étant pas produits par deux hypothèses dérivées l'une de l'autre le test du χ^2 n'est pas utilisable. Pour avoir une distribution de la différence des maximums de vraisemblance, on doit simuler de nombreux arbres et faire l'histogramme des différences. Ensuite sur cet histogramme on place la valeur réelle de la différence : est-elle dans les 95% les plus probables ?

La simulation est effectuée par tirages de type bootstrap sur le jeu de données établi avec la molécule IRBP. Pour chaque tirage, le maximum de vraisemblance des deux arbres est évalué ainsi que la valeur de la différence. On fait l'histogramme de toutes ces valeurs de différence La valeur réelle de la différence est placée sur cet histogramme est largement en dehors des 95% les plus probables. De cette façon, l'arbre de Pettigrew est rejeté.

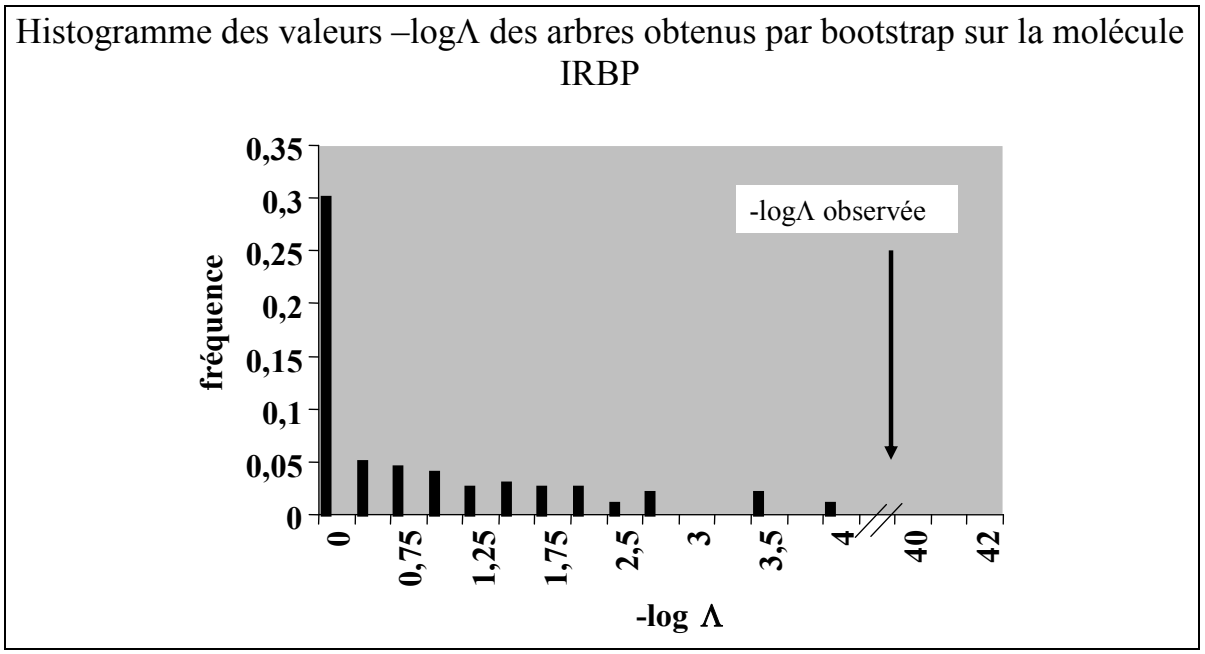


Figure IV- 16. Résultat de la simulation de nombreux arbres et comparaison avec la différence observée.

Conclusion

Les modèles que l'on peut définir restent simplistes au regard de la complexité de l'évolution. L'arbre obtenu n'est pas forcément le meilleur mais cette méthode permet de tester des hypothèses évolutives.

TREEPUZZLE ou la méthode des quartets

Méthode de construction dite des quartets

Le traitement des séquences préalablement alignées se fait par paquets de 4. Lorsque l'on ne prend que 4 taxons, l'information ne permet pas toujours de résoudre l'arbre. Ici, la combinaison des quartets donne beaucoup plus de résolution.

La procédure comporte trois étapes.

- construction de tous les quartets possibles. A chacun correspondent 3 arbres qui ,ont une certaine probabilité. Le plus probable est retenu.(L=max.{L1 ;L2 ;L3}).
- De façon répétitive, on combine les quartets pour construire un arbre global.
- L'arbre de consensus majoritaire est calculé à partir de la collection des arbres complets obtenus en 2.

Etape 1

Pour un quartet il existe 3 arbres qui ont chacun une probabilité Li. On choisit l'arbre qui a la L max. Si deux arbres sont équiprobables, l'un est tiré au hasard. Dans chaque quartet est ainsi définie une relation de voisinage.

❖

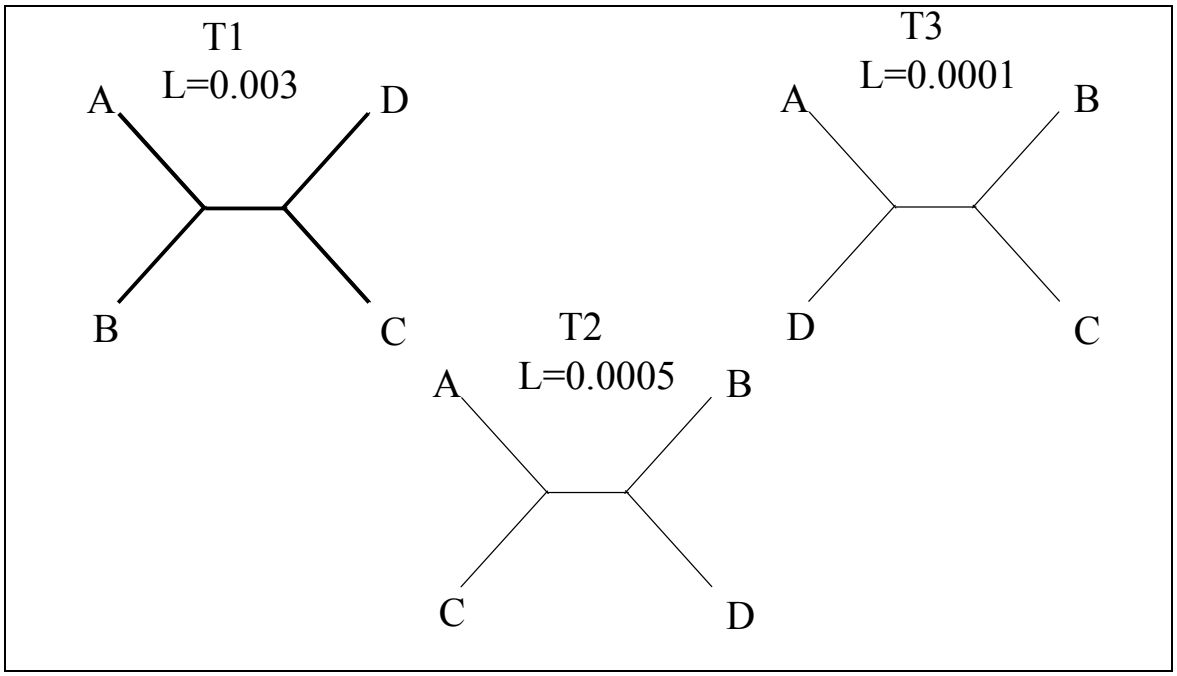


Figure IV- 17. Parmi les trois arbres possibles le plus probable est retenu. Il définit la relation de voisinage AB ‖ CD.

Etape 2

En général l'ensemble des relations de voisinage ci-dessus ne sont pas compatibles en un arbre. Afin de rechercher le ou les arbres possibles les étapes suivantes sont effectuées

- L'ordre des n taxa complet est retiré au, hasard (randomization)
- Pour ajouter le taxon suivant, si l'ordre tiré est ABCDE, on part de l'arbre obtenu avec le quartet ABCD et on va y ajouter E en tenant compte des relations de voisinage dans les différents quartets qui contiennent E et 3 des 4 autres (quartet i,j,k,E). Si l'on a la relation de voisinage ij ‖ kE, cela signifie que E ne peut être entre i et j. On peut ainsi repérer les segments sur lesquels ne doit pas être E dans les différents quartets. Ensuite chaque branche reçoit un score ; et celle qui a le score le plus bas est celle sur laquelle E est ajouté. Si deux segments ont le même score le plus bas l'un des deux est tiré au hasard. On continue jusqu'à ce que tous les taxons aient été ajoutés. La procédure est sensible à l'ordre d'entrée des taxons (plus il y a de taxa plus il faut de tirages). Il est donc indispensable de recommencer le plus de fois possible pour explorer l'ensemble des possibilités. Plusieurs arbres différents sont obtenus.

- Ces différents arbres permettent de calculer le consensus majoritaire dont la valeur à chaque nœud en indique la fiabilité. De cette manière les arbres qui servent pour le consensus sont indépendants les uns des autres, ce qui n'est pas le cas dans la procédure de TBR ou SPR ou NNI (cf. PAUP) où l'on part d'un arbre que l'on modifie. (Dans ces méthodes il y a également une procédure de tirage au hasard de l'ordre des taxons qui permet d'explorer les différents îlots). Dans cet arbre de consensus seuls les nœuds >= 50% sont acceptables, sinon on a des polyfurcations. Il ne faut pas oublier que l'on n'explore pas tous les arbres possibles.

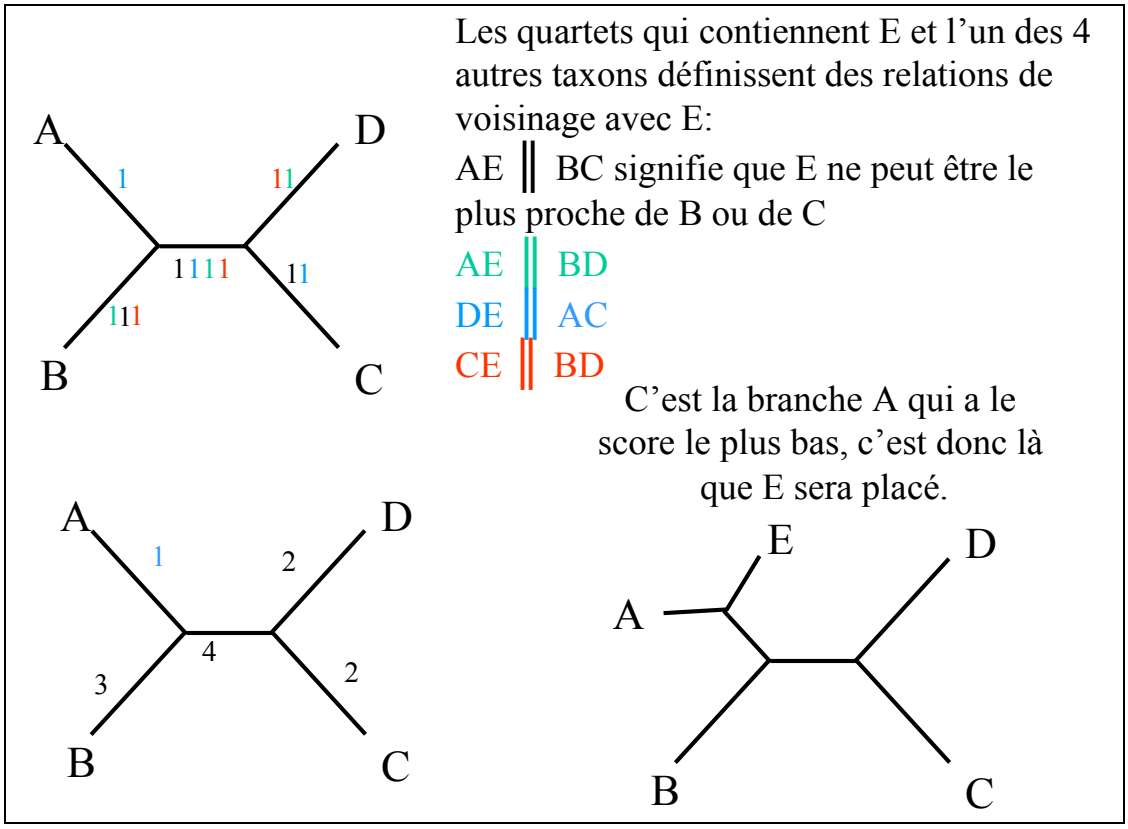


Figure IV- 18. Ajout d'un taxon.

Méthode graphique de visualisation de l'information

Quelle quantité d'information contiennent les données ?
Soient 4 taxons. Il existe 3 arbres et 3 seulement entièrement résolus qui sous un modèle donné d'évolution ont une certaine probabilité. la somme L1+L2+L3 est largement inférieure à1. Le théorème de Bayes permet de normaliser ces probabilités. Pour 1 compris de 1 à 3 :

$$p_i = \frac{L_i}{L_1 + L_2 + L_3}$$

Les probabilités p1, p2et p3 peuvent être interprétées comme les coordonnées barycentriques du point P appartenant au plan qui contient un triangle dont les trois sommets représentent chacun des 3 arbres résolus possibles. pi est tout simplement la longueur de la perpendiculaire de P à chaque côté du triangle. Si P est près d'un sommet, les probabilités pi sont clairement en faveur d'un arbre Ti, pi= max.{p1,p2,p3}. La représentation de P dans le triangle montre immédiatement quel arbre est préférable. Dans les données expérimentales, les séquences contiennent trop souvent du bruit qui brouille le signal phylogénétique(séquences courtes par ex., ou bien l'arbre vrai est une étoile). Au centre du triangle toutes les valeurs p1=p2=p3=1/3 de telle sorte que les 3 arbres sont équiprobables. Si P est près du centre, la relation

phylogénétique n'est pas résolue, si P se rapproche d'un côté un arbre peut être exclu sans que l'on soit capable de choisir entre les deux restants. A l'aide de ces 7 points d'attraction on détermine les bassins d'attraction. Voir la figure ci-dessous : information d'un quartet.

Avec n séquences alignées on a ($C_n^4 = \frac{n!}{4!(n-4)!}$) quartets différents possibles. Dans un même triangle, tous les vecteurs pi sont considérés et représentés par leur point P (1000 pris au hasard sont suffisants pour une bonne représentation). Suivant la répartition de ces points on a une bonne impression du niveau de résolution des quartets. Si la majorité tombe dans la zone A* la résolution n'est pas grande. Par contre si la résolution est grande, cela ne signifie pas que tous ces quartets sont compatibles et que l'arbre sera bien résolu.

Les quartets peuvent également contenir des lots disjoints de séquences. Dans ce cas les quartets sont constitués d'une séquence de chaque lot et on place les points P. C'est une façon de mesurer le soutien d'un nœud interne. On peut également considérer le % de quartets non résolus.

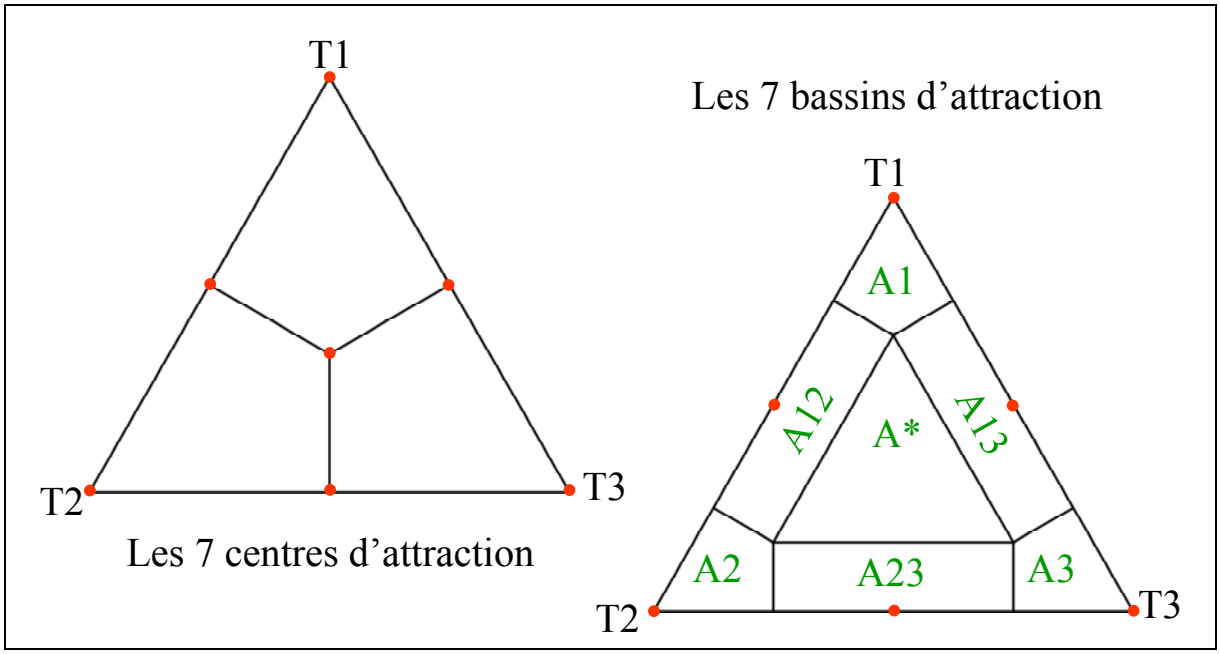


Figure IV- 19. Représentation graphique de l'information d'un quartet.

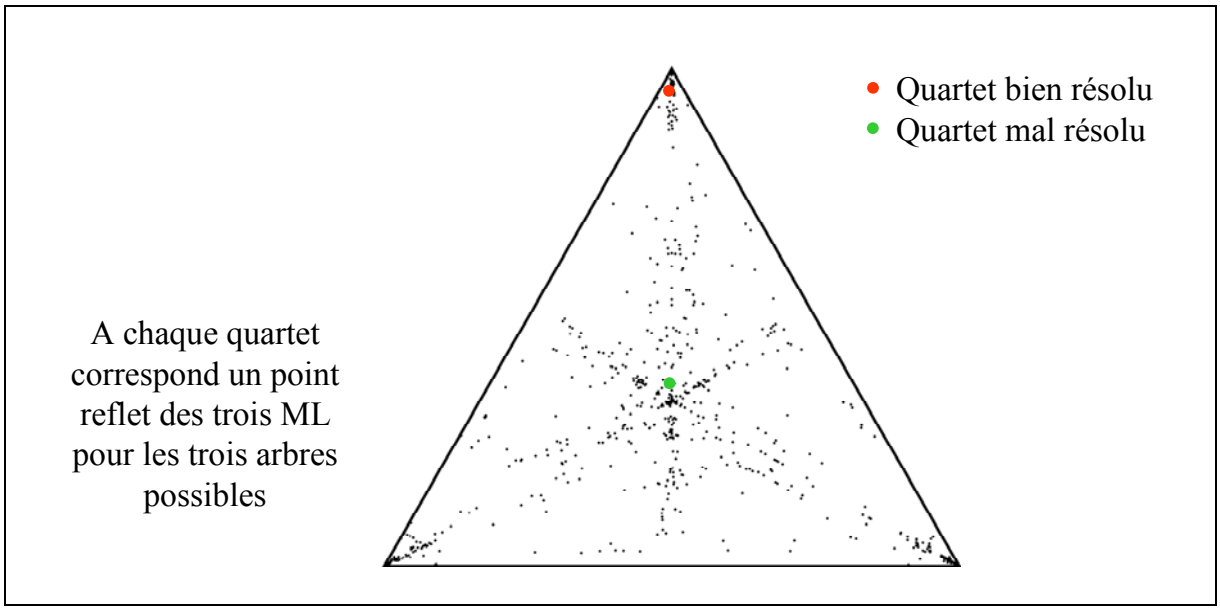


Figure IV- 20. Fichier de sortie de TreePuzzle "outlm.eps » (après une analyse de type « likelihood mapping ») représentant l'information de l'ensemble des quartets.

Méthode bayesienne

C'est une méthode probabiliste comme le maximum de vraisemblance mais à la différence de cette méthode qui calcule la probabilité des données en fonction du modèle, la méthode bayésienne calcule la probabilité du modèle en fonction des données.

Le théorème de Bayes

Un événement A dépend de n causes Ci toutes incompatibles (mutuellement exclusives).La loi des probabilités composées donne:

$$P(C_i / A) = \frac{P(C_i) * P(A/C_i)}{\sum_{k=1}^{N} P(C_k) * P(A/C_k)}$$

En théorie des probabilités, le théorème de Bayes énonce des probabilités conditionnelles : soit *A* et *B* deux *événements*, le théorème de Bayes permet de déterminer la probabilité de *A sachant B*, si l'on connaît les probabilités

- de *A*,
- de *B*
- de *B sachant A*.

Ce théorème élémentaire (originellement nommé de *probabilité des causes*) a des applications considérables.
Pour aboutir au théorème de Bayes, on part d'une des définitions de la probabilité conditionnelle :

$$P(A \mid B)P(B) = P(A,B) = P(B \mid A)P(A)$$

en notant $P(A,B)$ la probabilité que *A* et *B* aient tous les deux lieu, P(A |B) probabilité de A si B est réalisé. En divisant de part et d'autre par $P(B)$, on obtient

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

soit le théorème de Bayes.

Conventions d'écriture

P(A) = probabilité que A soit réalisé

P(A,B) = probabilité que A et B soient réalisés

P(A|B) =probabilité que A soit réalisé si B est vrai

Définition de la probabilité conditionnelle

P(A|B) P(B) = P(A,B) = P(B|A) P(A)

Théorème de Bayes (1702-1761)

$$\mathbf{P(A|B) = \frac{P(B|A)\ P(A)}{P(B)}}$$

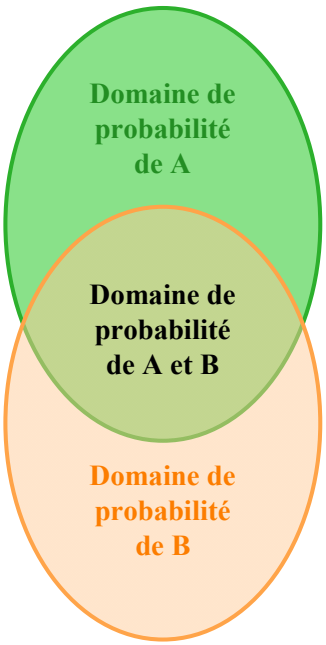


Figure IV- 21. Le théorème de Bayes.

Exemple

Exemple 1

D'où vient ce biscuit ?
Imaginons deux boites de biscuits.
L'une, A, comporte 30 biscuits au chocolat et 10 ordinaires.
L'autre, B, en comporte 20 de chaque.
On choisit les yeux fermés une boite au hasard, puis dans cette boite un biscuit au hasard. Il se trouve être au chocolat. De quelle boîte a-t-il le plus de chances d'être issu, et avec quelle probabilité ? Intuitivement, on se doute que la boite A a plus de chances d'être la bonne, mais de combien ?
La réponse exacte est donnée par le théorème de Bayes :
Notons H_A la proposition « le gâteau vient de la boite A » et H_B la proposition « le gâteau vient de la boite B ».
Si lorsqu'on a les yeux bandés les boites ne se distinguent que par leur nom, nous avons $P(H_A) = P(H_B)$, et la somme fait 1, puisque nous avons bien choisi une boite, soit une probabilité de 0,5. pour chaque proposition.
Notons D la phrase « le gâteau est au chocolat ». Connaissant le contenu des boîtes, nous savons que :
On peut donc écrire

$$P(H_A / D) * P(D) = P(D / H_A) * P(H_A)$$

$$P(H_A / D) = \frac{P(D / H_A) * P(H_A)}{P(D)}$$

Or
$P(D) = P(H_A) * P(D / H_A) + P(H_B) * P(D / H_B)$.

P(D | H_A) = 30/40 = 0,75
P(D | H_B) = 20/40 = 0,5.
Résolution utilisant la notation des probabilités
La formule de Bayes nous donne donc :

$$P(H_A|D) = \frac{P(H_A)\cdot P(D|H_A)}{P(H_A)\cdot P(D|H_A)+P(H_B)\cdot P(D|H_B)}$$

$$= \frac{0,5\times 0,75}{0,5\times 0,75+0,5\times 0,5}$$

$$= 0.6$$

Avant de regarder le gâteau, notre probabilité d'avoir choisi la boite A était $P(H_A)$, soit 0,5. Après l'avoir regardé, nous révisons cette probabilité à $P(H_A|D)$, qui est 0.6. Ajouter une donnée (ici gâteau au chocolat) modifie la probabilité d'un certain événement (ici gâteau tiré de la boite A).

Exemple 2

Prenons un exemple simple avec un lancé de dés. On dispose de 100 dés dont 90 sont corrects et 10 sont pipés.

valeur	Pb si dé correct	Pb si dé pipé
1	1/6	1/21
2	1/6	2/21
3	1/6	3/21
4	1/6	4/21
5	1/6	5/21
6	1/6	6/21

Choisissant un dé au hasard , on obtient un 4 et un 6. Ce dé est-il correct ou pipe?
Quelle est la probabilité de tirer un 4 et un 6 (deux lancés indépendants avec chacun des deux types de dés?

Avec un dé correct:

$$Pb = \frac{1}{6} x \frac{1}{6} = \frac{1}{36}$$

Avec un dé pipe

$$Pb = \frac{4}{21} x \frac{6}{21} = \frac{24}{441}$$ soit 1,96 fois plus

Intuitivement on pense que le dé est plutôt pipé.
A priori il y a une probabilité de 0,1 de choisir un dé pipé. Mais sachant que l'on a obtenu 4 et 6, la probabilité à posteriori est donnée par le théorème de Bayes :

$$Pb \cdot biaisé \cdot avec \cdot 4 \cdot et \cdot 6 =$$

$$= \frac{Pb \cdot 4 \cdot et \cdot 6 \cdot si \cdot biaisé \cdot x \cdot Pb \cdot biaisé}{(Pb \cdot 4 \cdot et \cdot 6 \cdot si \cdot biaisé \cdot x \cdot Pb \cdot biaisé) + Pb \cdot 4 \cdot et \cdot 6 \cdot si \cdot correct \cdot x \cdot Pb \cdot correct}$$

Ce qui donne

$$Pb \cdot biaisé \cdot si \cdot 4 \cdot et \cdot 6 = \frac{\frac{24}{441} * 0{,}10}{\left(\frac{24}{441} * 0{,}10\right) + \left(\frac{1}{36} * 0{,}90\right)} = 0{,}179$$

Prior = 0,10
Posterior = 0,179

Procédure

Les inferences bayésiennes phylogéniques sont basées sur la probabilité postérieure d'un arbre τ. Cette probabilité est donnée pour l'arbre numéro *i* conditionné à la matrice X des séquences alignées par la formule de Bayes:

$$F(\tau_i \mid X) = \frac{f(X \mid \tau_i) * f(\tau_i)}{\sum_{j=1}^{B_s} f(X \mid \tau_i) * f(\tau_i)}$$

F(ti | X) est la probabilité postérieuredu ième arbre et représente la probabilité que l'arbre soit correct étant donné l'alignement.
Au dénominateur on trouve la somme de toutes les probabilités pour tous les arbres (il y en a Bs) avec

- Arbres racinés Bs=(2s-5) !/2^{s-2}(s-2) !
- Arbres non racinés Bs = (2s-3) !/2^{s-3}(s-3) !

En général on utilise une probabilité à priori non informative
$F(\tau_i)=1/Bs$

Etape 1

On a une matrice de données (alignement) et un modèle phylétique composé

1. D'un arbre avec chaque taxon en bout de branche et des longueurs de branche déterminées.
2. Un modèle de substitution de l'ADN (JC, K2 HKY,etc.)

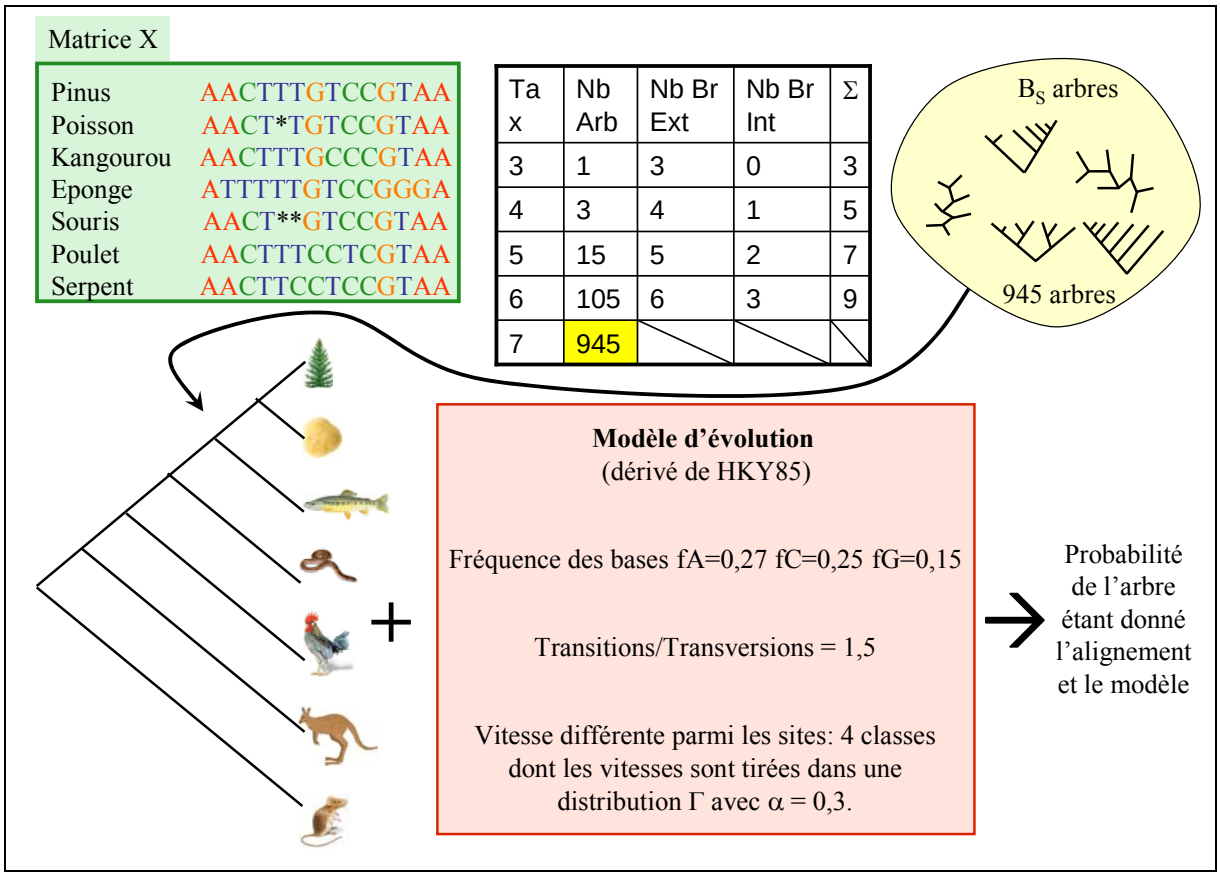


Figure IV- 22.Première étape de la méthode bayesienne: la matrice de caractères a permis de construire un arbre. En appliquant un modèle on calcule la probabilité de cet arbre.

On calcule la vraisemblance d'un arbre avec aux nœuds internes toutes les formes possibles

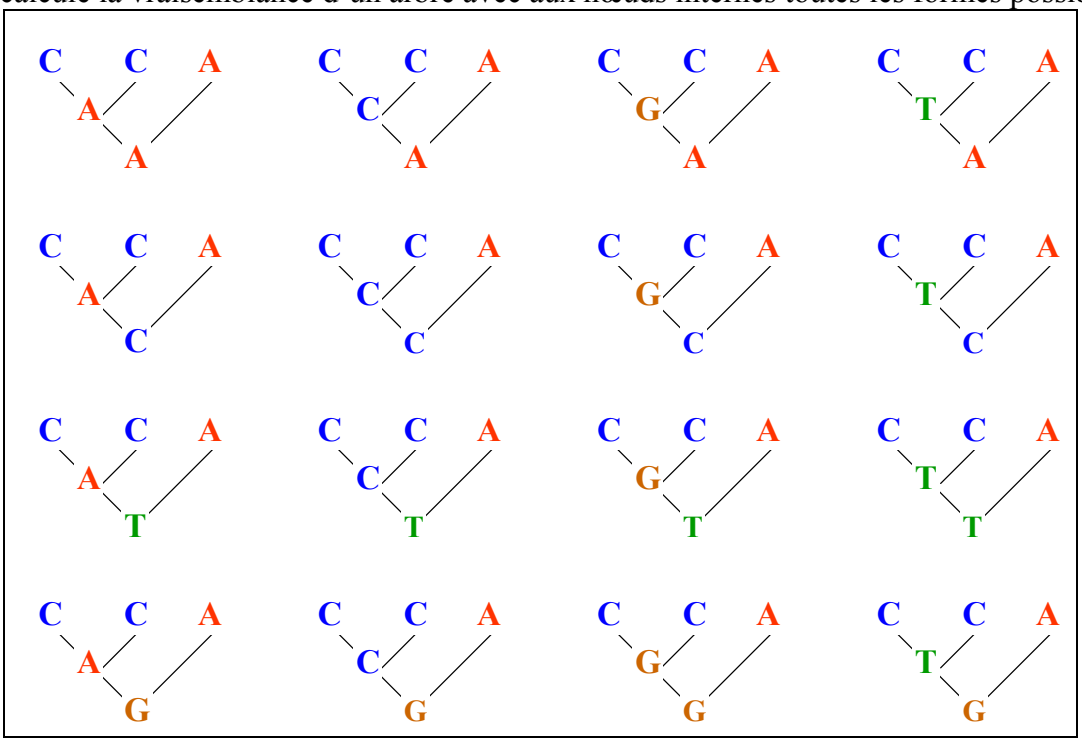


Figure IV- 23. Les 16 possibilités pour les nœuds internes d'un arbre donné à 3 taxa.

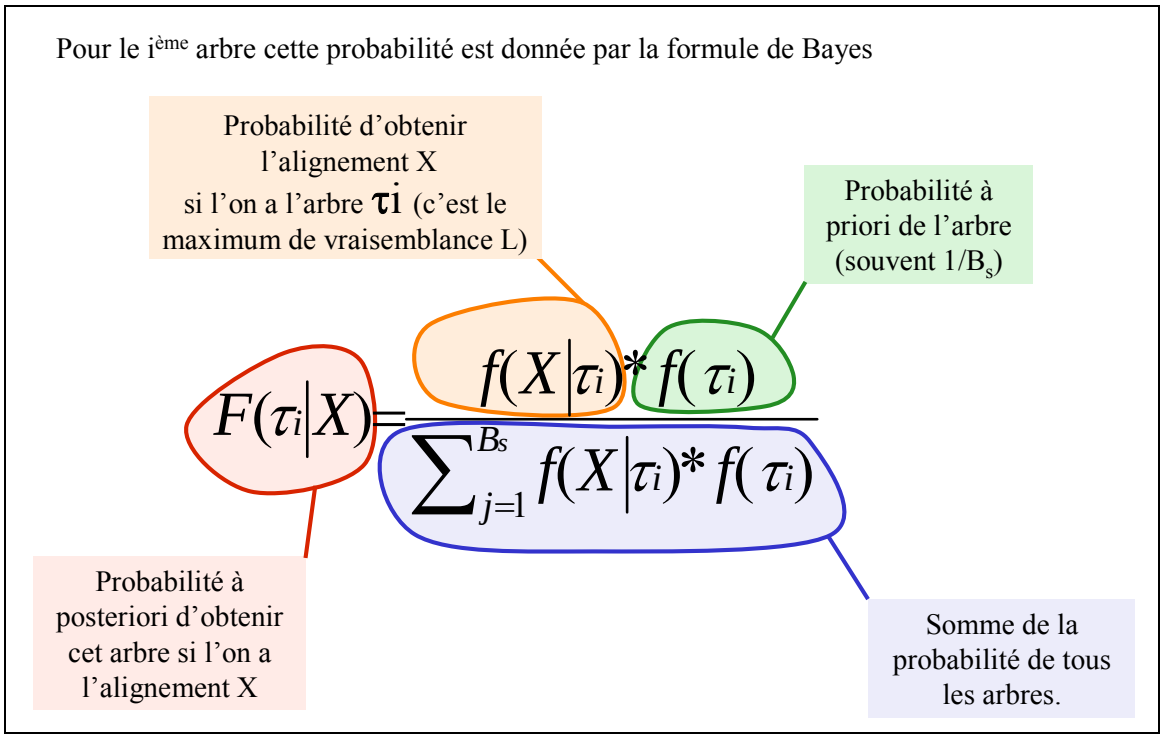


Figure IV- 24. Calcul de la probabilité à posteriori d'un arbre.

Exploration du paysage d'arbres par MCMC (Markov Chain Monte Carlo Algorithme)

1°) Un second arbre est propose pour lequel est également calculée la vraisemblance avec

- ψ le premier arbre avec ses longueurs de branches ses paramètres de substitution
- ψ' le second

2°) On calcule la probabilité d'accepter le second état.

Le nouvel état ψ' est accepté avec la probabilité

$$R = \min\left(1, \frac{f(\Psi'|X)}{f(\Psi|X)} \times \frac{f(\Psi|\Psi')}{f(\Psi'|\Psi)}\right)$$

$$= \min\left(1, \frac{f(X|\Psi')f(\Psi')/f(X)}{f(X|\Psi)f(\Psi)/f(X)} \times \frac{f(\Psi|\Psi')}{f(\Psi'|\Psi)}\right)$$

$$= \min\left(1, \underbrace{\frac{f(X|\Psi')}{f(X|\Psi)}}_{\text{Likelihood Ratio}} \times \underbrace{\frac{f(\Psi')}{f(\Psi)}}_{\text{Prior Ratio}} \times \underbrace{\frac{f(\Psi|\Psi')}{f(\Psi'|\Psi)}}_{\text{Proposal Ratio}}\right)$$

La Pb de tirer le second état en partant du premier est f(ψ' | ψ) et la Pb de faire l'inverse (qui n'est pas fait) est f(ψ | ψ')

3°) Un nombre au hasard est tiré, compris entre 0 et 1. Si R>à ce nombre, ψ' remplace ψ etc. C'est le principe de l'algorithme de Metropolis Hasting Green.

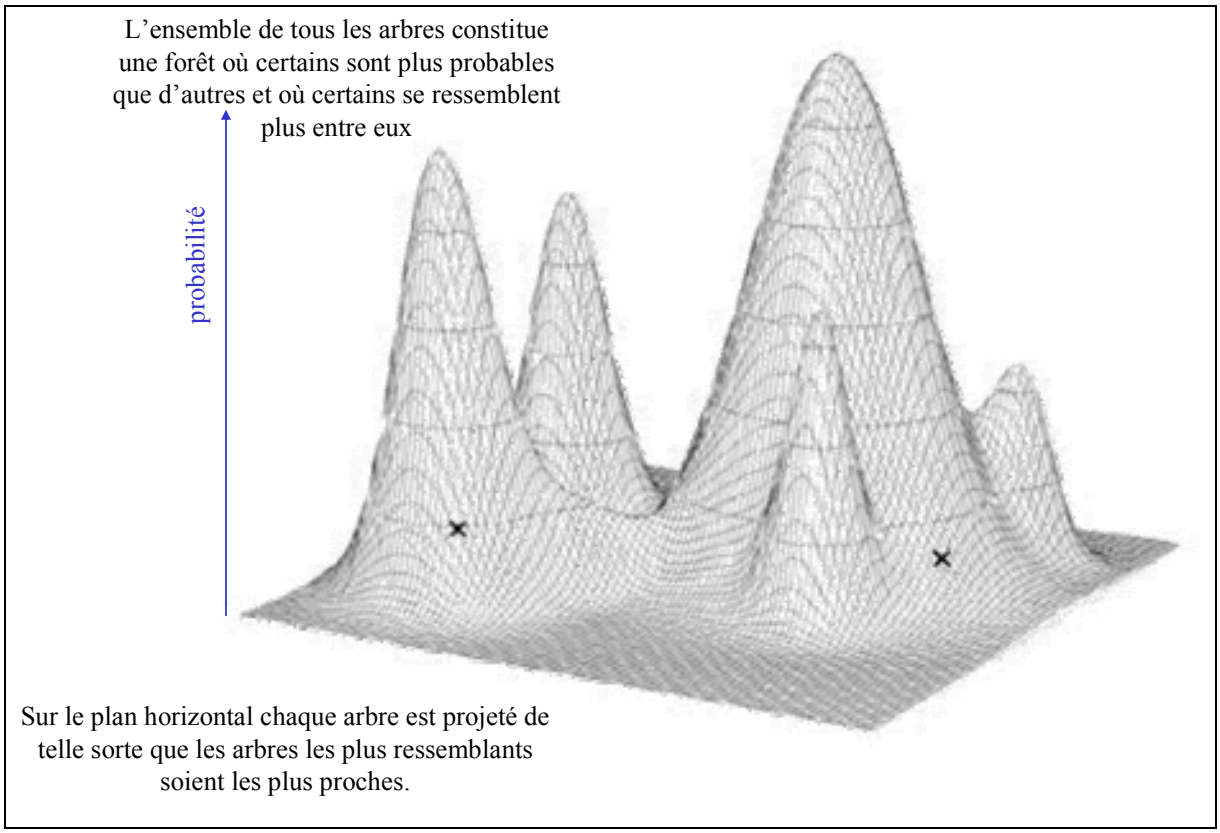


Figure IV- 25. Représentation du paysage de tous les arbres.

Chaque arbre retenu est conserve en mémoire et une façon d'apprécier sa Pb postérieure est de regarder le nombre de fois où le programme l'a visité.

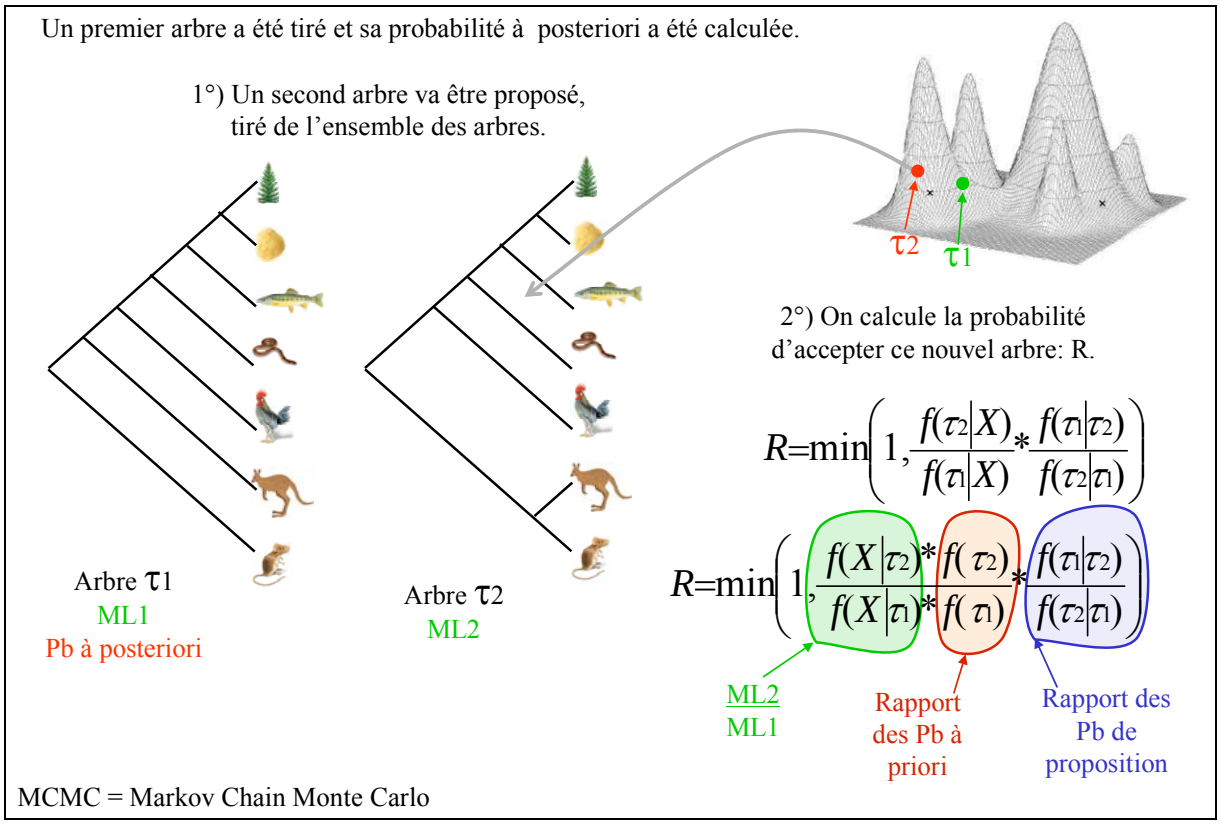


Figure IV- 26. Exploration du paysage d'arbres: algorithme MCMC (Markov Chain Monte Carlo).

Une variante est MCMCMC (Metropolis Coupled Markov Chain Monte Carlo). Plusieurs chaînes fonctionnent simultanément. Toutes sauf une vont être "chauffées" par un facteur

$$\beta = \frac{1}{1+T(i-1)}$$

Au fur et à mesure que l'on accumule les pas *I* augmente et β diminue. Cela a pour effet de permettre de traverser plus facilement de profondes vallées dans le paysage d'arbres (cela atténue le relief).

Lorsque toutes les chaînes ont fait un pas, un échange entre chauffée et froide est proposé. Les inférences sont faites uniquement sur la chaîne froide (β=1).

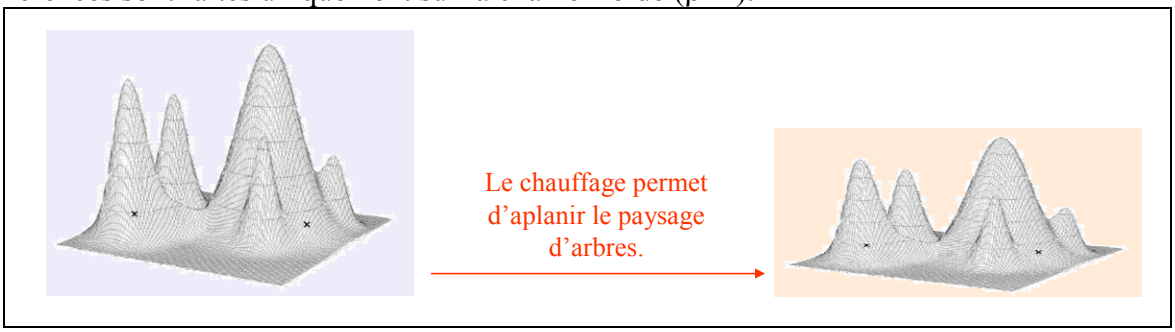


Figure IV- 27. Effet de l'algorithme MCMCMC (Metropolis Coupled Markov Chain Monte Carlo).

Analyse des résultats

Entrée des commandes à l'aide du fichier de données

On peut également le faire à la ligne de commande.
Sinon on ajoute à la fin du fichier un nouveau bloc

BEGIN mrbayes ;

qui se termine par

END ;

Comme pour le DATA block

Exemple pour un fichier de séquences codantes

```
begin mrbayes;
charset 1stpos = 1-720\3;
charset 2ndpos = 2-720\3;
charset 3rdpos = 3-720\3;
```

Indique la partition en codons

```
partition bycodon = 3:1stpos,2ndpos,3rdpos;
```

il y a une partition qui s'appelle bycodon en 3 lots pos 1, 2 et 3

```
lset nst=6 rates=sitespec sitepartition=bycodon;
```

Paramètres du modèle

lset nst=6 Modèle GTR (general time reversible)
rates=sitespec Les vitesses d'évolution sont site spécifiques
sitepartition=bycodon indique la partition à utiliser pour la variation de vitesse

```
usertree = (((FR,MS2),GA),((SP,NL95),((M11,MX1),QB)),PP7);
```

spécifie l'arbre initial au format Newick. Il doit être strictement dichotomique. Si l'arbre initial n'est pas défini, mrbayes en tire un au hasard. Ce qui dans certains cas est préférable, en particulier lorsqu'on examine la convergence de plusieurs tours indépendants

```
mcmc ngen=1000 printfreq=100 samplefreq=10 nchains=4 savebrlens=yes;
end;
```

mcmc ngen=1000 indique au programme qu'il doit effectuer 1000 générations
printfreq=100 et qu'il imprime à l'écran l'état toutes les 100 générations
samplefreq=10 alors qu'il sauve l'arbre en cours toutes les 10 générations dans le fichier d'arbres
nchains=4 4 chaines sont démarrées simultanément
savebrlens=yes;. Indique qu'il faut sauver les longueurs de branche des arbres.

execute <**filename**> *(*par exemple. execute replicase.nex), ce qui produit le résultat en figure 103.

Running Markov chain
Starting likelihoods
1---6429.147 -6429.147
2---6388.254 -6388.254
3---6678.446 -6678.446
4---6597.156 -6597.156

En faisant de plus en plus de tours on constate que la somme des ML sur les quatre chaînes augmente. On a donc des arbres avec une probabilité plus élevée. Seules les valeurs des chaînes froides peuvent être comparées (ML non pondérées).

Nb géné	Σ ML_n f(θ)	(ML_1)	ML_2	(ML_3	(ML_4)	Temps	Echange
100	-19215.03	(-6045.46)	[-5929.31]	(-6088.02)	(-6238.81)	2s	
200	-18522.17	(-5878.95)	[-5819.39]	(-5833.47)	(-5834.11)	4s	(4~2)
300	-18292.23	[-5732.10]	(-5725.53)	(-5821.09)	(-5810.95)	6s	
400	-18139.18	(-5712.11)	(-5703.27)	(-5780.78)	[-5693.39]	8s	
500	-18092.13	(-5702.02)	(-5684.19)	(-5762.62)	[-5680.81]	9s	
600	-18036.97	(-5696.36)	(-5680.07)	(-5701.56)	[-5671.27]	11s	
700	-18018.36	(-5682.72)	(-5679.98)	(-5700.30)	[-5662.94]	13s	(4~1)
800	-17981.18	(-5679.95)	(-5666.39)	(-5684.34)	[-5649.37]	15s	
900	-17970.38	(-5684.04)	(-5663.26)	(-5669.62)	[-5647.99]	17s	
1000	-17959.28	(-5684.56)	[-5657.12]	(-5657.73)	(-5648.36)	19s	(4~2)

( ) = chaîne chaude
[ ] = chaîne froide

Figure IV- 28. Exemple de sortie du programme MrBayes.

Le programme demande alors s'il doit continuer
Répondre « yes » pour continuer ou n'importe quoi d'autre pour terminer.

Faut-il continuer et sur combien de générations ?

- Tant que ML n'est pas stabilisé oui
- Combien ?? L'auteur du pgm suggère de dépasser largement le début du plateau apparent.
- Pour être sûr que les valeurs de maximum de vraisemblance ont convergé.

La somme des valeurs absolues des ML diminue de génération en génération et tend à se stabiliser. A partir de ce point, tous les arbres obtenus sont équivalents enprobabilité et on n'a aucun argument pour en préférer certains. MrBayes échantillonne les arbres suivant leur probabilité postérieure. Les arbres obtenus à partir du plateau peuvent être utilisés pour calculer un arbre consensus.

Nbgén	$\Sigma ML_n f_{(\theta)}$	ML_1	ML_2	ML_3	ML_4	Tps	Echange
500	17993.91	[-5651.94]	(-5704.28)	(-5686.40)	(-5658.09)	9s	
1000	17938.56	(-5653.61)	(-5652.86)	(-5671.55)	[-5644.75]	18s	
1500	17911.89	(-5642.16)	(-5657.28)	(-5642.82)	[-5643.71]	27s	(4~3)
2000	17900.22	[-5641.23]	(-5642.17)	(-5641.69)	(-5643.72)	35s	(4~2)
2500	17905.10	(-5651.80)	[-5642.50]	(-5642.14)	(-5640.15)	44s	
3000	17898.56	(-5650.13)	(-5641.28)	[-5635.48]	(-5641.77)	53s	
3500	17899.50	(-5645.09)	(-5645.03)	[-5637.74]	(-5640.57)	62s	(1~4)
4000	17888.31	[-5635.58]	(-5637.56)	(-5638.49)	(-5643.47)	71s	
4500	17901.53	[-5638.01]	(-5640.68)	(-5647.19)	(-5646.79)	80s	(1~3)
5000	17889.99	(-5638.23)	(-5640.35)	[-5636.40]	(-5641.58)	89s	(3~2)
5500	17906.17	(-5642.21)	(-5646.19)	[-5648.80]	(-5637.32)	98s	(1~4)
6000	17910.50	(-5642.70)	(-5648.94)	(-5656.92)	[-5637.01]	106s	(3~1)
6500	17906.43	(-5643.42)	[-5638.26]	(-5647.00)	(-5650.79)	116s	
7000	17895.80	(-5639.54)	[-5634.80]	(-5648.19)	(-5641.47)	125s	
7500	17890.95	(-5642.34)	(-5639.56)	[-5632.56]	(-5645.62)	134s	
8000	17897.81	(-5645.97)	(-5641.58)	(-5642.21)	[-5637.28]	143s	(2~3)
8500	17903.34	(-5642.63)	[-5639.20]	(-5644.76)	(-5647.57)	152s	(1~2)
9000	17892.54	(-5644.11)	(-5638.13)	(-5644.03)	[-5634.35]	161s	(1~3)
9500	17897.95	[-5641.69]	(-5637.69)	(-5640.14)	(-5646.34)	169s	(2~1)
10000	17893.30	(-5636.77)	[-5641.33]	(-5647.43)	(-5635.27)	178s	(4~1)
10500	17891.36	(-5642.92)	(-5640.35)	(-5649.60)	[-5629.41]	187s	
11000	17896.02	(-5648.41)	[-5635.74]	(-5640.64)	(-5641.29)	196s	
11500	17907.58	(-5642.24)	[-5634.30]	(-5659.54)	(-5646.23)	205s	(3~4)
12000	17879.13	(-5641.49)	(-5635.24)	(-5636.55)	[-5630.35]	214s	(2~3)
12500	17902.92	(-5656.38)	[-5636.68]	(-5648.66)	(-5633.30)	223s	
13000	17898.09	(-5645.50)	(-5640.01)	(-5645.86)	[-5636.92]	232s	(2~3)
13500	17889.75	(-5647.46)	[-5634.11]	(-5642.97)	(-5634.34)	241s	
14000	17889.64	(-5641.87)	(-5640.95)	(-5640.87)	[-5633.40]	250s	(1~3)
14500	17900.36	(-5646.60)	(-5643.01)	[-5638.79]	(-5642.06)	258s	(3~4)
15000	17909.86	(-5641.49)	(-5651.88)	(-5654.19)	[-5637.49]	267s	
15500	17905.15	(-5654.59)	(-5646.07)	(-5643.89)	[-5634.62]	276s	
16000	17894.56	(-5661.10)	(-5634.41)	(-5644.16)	[-5629.49]	285s	
16500	17895.79	[-5635.66]	(-5634.98)	(-5647.15)	(-5645.50)	294s	(3~1)
17000	17887.96	(-5635.99)	[-5634.48]	(-5647.32)	(-5636.63)	303s	(1~3)
17500	17889.21	(-5637.32)	(-5636.37)	(-5646.93)	[-5636.13]	311s	
18000	17898.34	(-5646.61)	(-5639.04)	(-5648.28)	[-5635.52]	320s	(1~3)
18500	17891.88	(-5635.24)	[-5638.67]	(-5648.28)	(-5637.44)	329s	
19000	17904.45	(-5637.69)	[-5637.90]	(-5647.41)	(-5655.37)	338s	(2~1)
19500	17902.31	(-5646.84)	(-5649.32)	(-5641.51)	[-5636.55]	346s	
20000	17891.69	[-5635.71]	(-5648.45)	(-5638.32)	(-5637.95)	355s	

Tableau IV- 6. Exemple de sortie de MrBayes avec 20000 générations (la dernière colonne indique les éventuels échanges entre chaîne froide et chaîne chaudes). De la génération 1 à 3000, les valeurs de ML ne sont pas encore stabilisées : c'est la période de rodage.

On constate que le plateau est obtenu à la génération 3000. tous les états qui précèdent sont rejetés c'est ce qui correspond à la période de rodage de la chaine. Les inférences phylogénétiques ne sont faites que sur les arbres obtenus après le rodage.

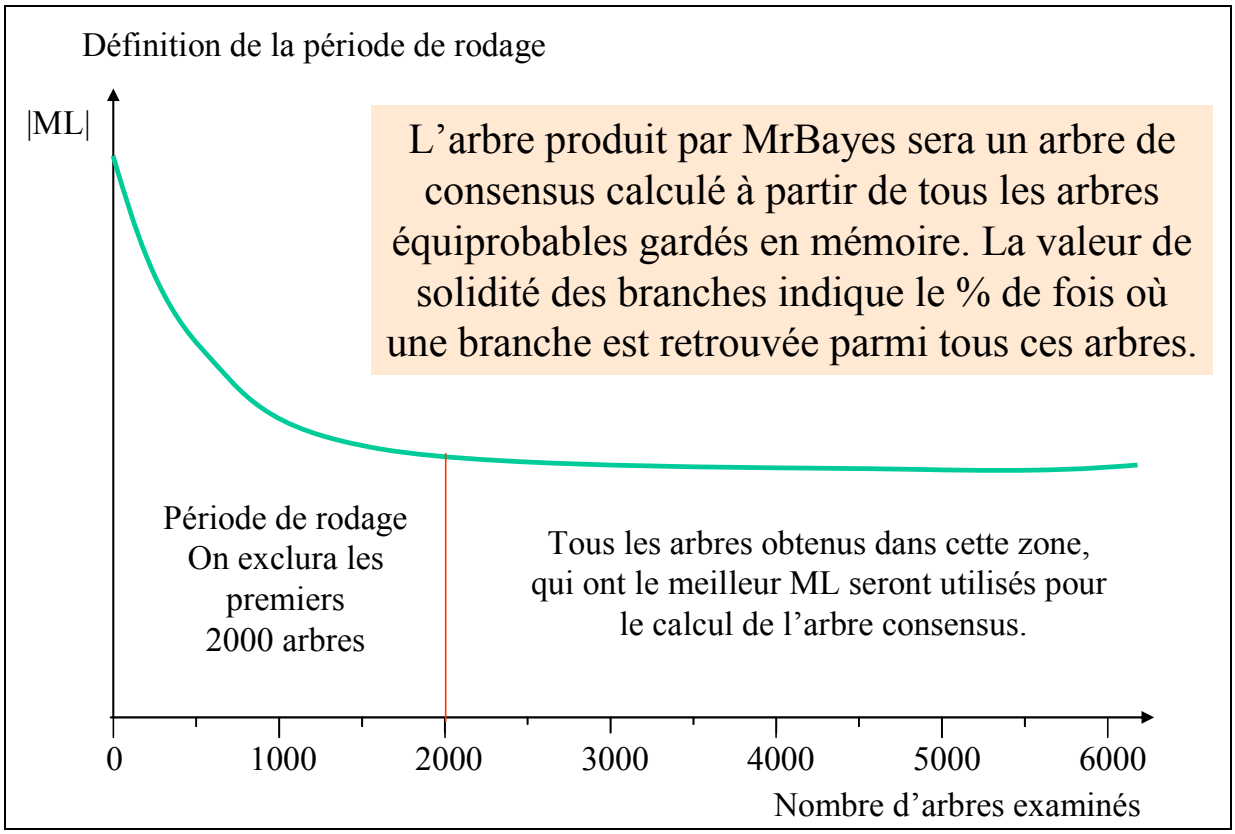


Figure IV- 29. Détermination graphique de la période de rodage.

Fichiers de sortie

Lorsqu'il est lance, le pgm ouvre 3 fichiers

Mbout.t c'est le fichier des arbres. Pour le calcul d'un consensus, penser à ôter tous ceux de la période de rodage

Mbout.p on y trouve les valeurs des différents parameters (lisible par un tableur cf Excel)

Mbout.bp (noms par défaut, on peut aussi les spécifier) au format txt on trouve tous les paramètres après une analyse

Combien faut-il d'arbres pour avoir un bon consensus? Cela dépend des données . En général plus il y en a mieux c'est. On peut décider de choisir ngen de façon à ce que ça tourne sur une nuit (ou n'importe quel temps qui vous paraisse supportable)
Par exemple avec le fichier adh.nex en 16h on obtient 480 000 générations soit 4 800 arbres. Il faut 15 000 générations pour atteindre le plateau, on va donc rejeter les 150 premiers arbres et calculer le consensus sur les 4 650 restants

Analyse après le run

Quand la recherché est terminée on peut récupérer les résultats par la commande

Sumt avec la syntaxe

sumt <filename> burnin = <nombre d'arbres à ignorer> contype = <allcompat or halfcompat>

- allcompat = consensus majoritaire 50% avec une table de tous les groupes trouvés et leur fréquence
- halfcompat = consensus majoritaire 50%

Dans cette sortie on a toutes les bipartitions (allcompat) avec leur fréquence et leur probabilité ainsi que la moyenne et la variance de la longueur de chaque branche dans un fichier *.parts .

Sumt ecrit les bipartitions leurs fréquences te leurs probabilités dans un fichier nommé *.parts. Si les longueurs de branche sont enregistrées, il écrit également un phylogramme consensus basé sur la longueur moyenne des branches. 3Enfin un arbre de consensus avec la probabilité de chaque clade est présenté à l'écran. Il est enregistré dans *.con qui peut être imprimé avec PAUP. On peut également construire le consensus avec PAUP en laissant de côté les arbres de la période de rodage.

Sump écrit dans un fichier *.bp (par défaut mbout.bp) la moyenne, la variance et l'intervalle de confiance de 95% pour les différents paramètres.

Syntaxe :
sump filename=<filename>.bp burnin=<number of trees to be ignored>.

Lorsqu'on a obtenu un arbre de consensus on possède

Une bonne estimation de la phylogénie
Les parties de l'arbre bien soutenues
Les modèles de substitution qui sont les plus adaptés

C'est en gros l'équivalent d'un calcul de maximum de vraisemblance suivi d'un bootstrap

Conclusion

Avec ces différentes méthodes nous avons vu plusieurs moyens pour tracer des dendrogrammes qui peuvent être interprétés en arbres phylogénétiques. Au cours de la recherche de ces arbres il a fallu faire des choix.
Le premier dont il n'a que peu été question ici est l'alignement des séquences. IL est cependant primordial. De lui dépendent les résultats finaux puisqu'il établit l'orthologie des caractères au sein des séquences. Et pourtant il n'y a pas souvent de « meilleur » alignement. Il y en a plusieurs possibles. Aussi est-il fréquent de trouver des travaux de phylogénie où plusieurs « bons » alignements sont utilisés afin de mettre en évidence les différences qu'ils entraînent au niveau des arbres.' Ces alignements peuvent avoir été générés par des logiciels (Clustal, Multalin, Malign, Dialign, BlockMaker, etc…) puis corrigés à la main ou faits directement « à l'œil ». En règle générale, pour des gènes codant des protéines, l'alignement est plus facile (code à 20 mots) lorsque la séquence est traduite en acides aminés. Par contre ce qui a évolué et qui peut donner le meilleur reflet de la phylogénie c'est l'ADN. Il est donc préférable d'utiliser les séquences nucléotidiques.

L'utilisation des différentes méthodes de construction d'arbres présente des avantages et inconvénients (voir tableau 11) dont il faut tenir compte. L'utilisation de chacune de ces méthodes implique certaines hypothèses implicites ou explicites. Différentes options sont proposées concernant le modèle d'évolution des séquences. Un choix erroné n'entraîne pas forcément de gros biais dans les relations phylogénétiques (certaines méthodes sont assez peu sensibles à un choix approximatif concernant le modèle évolutif). Cependant aucun arbre n'est sûr. La recherche des relations phylogénétiques réclame une grande prudence. La convergence des diverses méthodes vers un arbre commun est un signe favorable. Même une branche robuste peut par le jeu de multiples artéfacts être purement illusoire.

Avec le développement des techniques de clonage et de séquençage les banques proposent actuellement la séquence de nombreux gènes orthologues chez des organismes variés. Chaque gène n'étant pas soumis aux mêmes contraintes évolutives, leurs vitesses

respectives d'évolution sont variables et plus ou moins adaptées à un niveau de la classification phylogénétique. De plus, tous les gènes n'ont pas la même histoire évolutive. On sait que le génome nucléaire est hérité pour moitié de chaque parent alors que le génome mitochondrial est considéré comme le reflet d'un héritage en général maternel. Il est donc primordial, avant d'entamer une recherche phylogénétique d'acquérir une connaissance approfondie de la biologie des organismes que l'on veut y inclure. Lorsque des arbres sont obtenus il est toujours bon d'accumuler des arguments extérieurs pour les confirmer (ou infirmer).

Enfin tous les modèles utilisés ne représentent que des simplifications de l'évolution biologique réelle. En particulier, l' »arbre » est incapable de représenter des événements que de nombreuses observations ont validées (la symbiose entre cellule procaryote et cellule eucaryote pour donner des eucaryotes possédant une mitochondrie, les transferts horizontaux dont il est très probable qu'on observe actuellement des exemples et qui ont pu contribuer à l'évolution passée). Aussi ces méthodes sont-elles un outil qui est actuellement fort utile bien que faillible. L'accroissement des capacités de calcul permet d'espérer un affinement de ces méthodes et l'apparitions de méthodes nouvelles plus aptes à résoudre certaines questions.

L'évolution passée a laissé des traces obscurcies par sa complexité et aucune phylogénie ne peut être validée avec certitude.

Méthodes de distance	Méthodes de parcimonie	Méthodes probabilistes
*Calcul d'une distance globale	*Examen des caractères les uns après les autres (sans const.)	*Modèle explicite d'évolution *Examen de tous les caractères les uns après les autres
*Un seul arbre retourné par le programme	*La méthode peut retourner plusieurs arbres également parcimonieux	*La méthode peut retourner plusieurs arbres avec leurs probabilités respectives
*Pas de test de robustesse de l'arbre unique (excepté le bootstrap)	*Il y a un test de robustesse des nœuds (mesure de l'homoplasie dans l'arbre par le calcul du rapport de la longueur minimale de l'arbre à sa longueur réelle.	* La longueur de chaque branche est testée : on estime sa probabilité d'être supérieure à 0
*Pas de retour aux caractères pour pouvoir les reconsidérer	*Retour aux caractères pour éventuellement réévaluer ceux qui donnent des aberrations	*Retour aux caractères pour éventuellement réévaluer ceux qui donnent des aberrations
*Rapide, même avec un grand nombre de taxa	*Vitesse moyenne. Sur de grosses machines on peut en plusieurs jours traiter des données jusqu'à 500 taxa	*Lent :suivant les machines et le modèle évolutif défini.

Tableau IV- 7. Comparaison des trois types de méthodes de construction d'arbres.

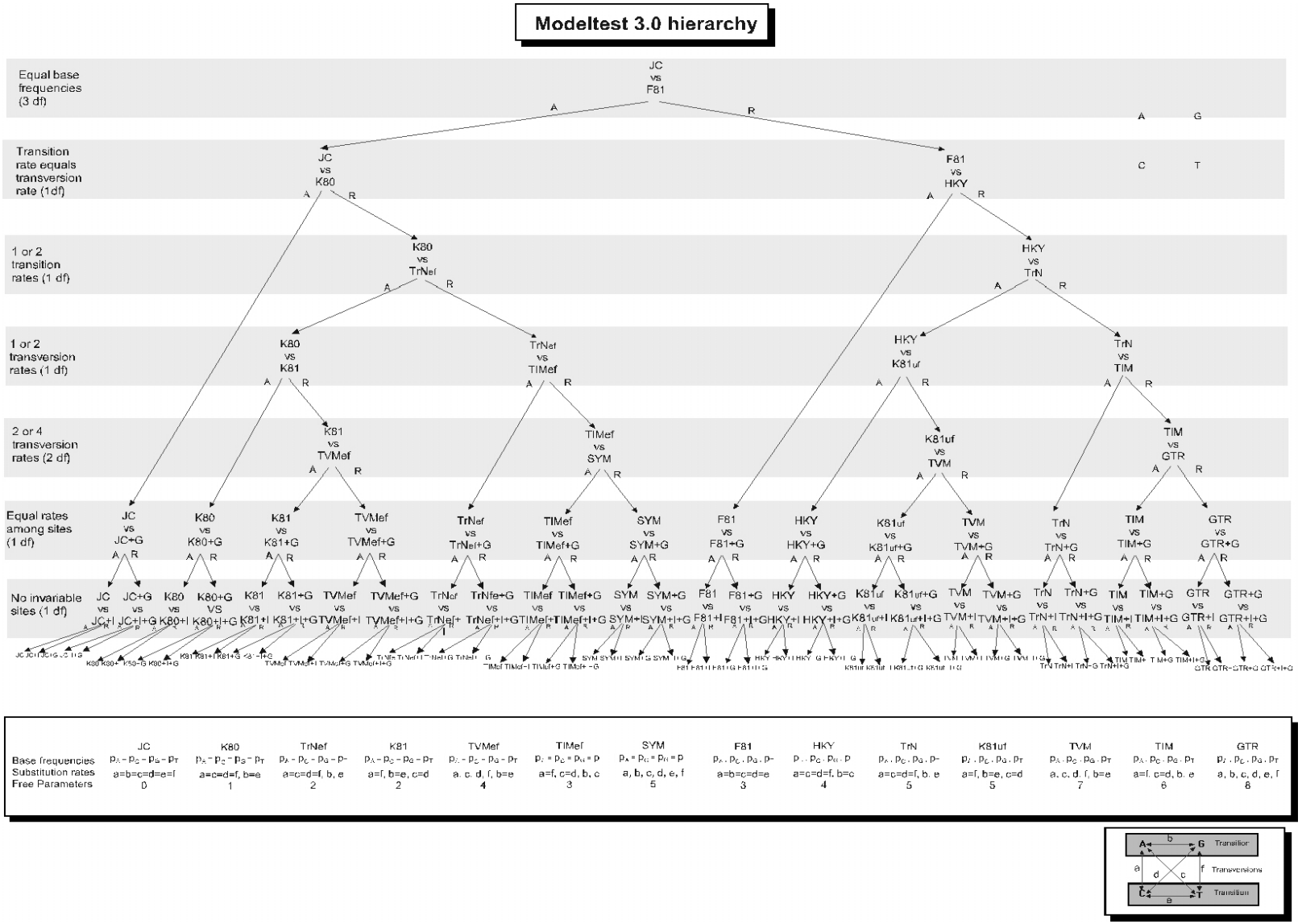


Figure IV- 30. Hiérarchie des modèles traités dans Modeltest