Université de Montréal
Faculté des études supérieures

Estimation des longueurs de branche et artefact sur la datation moléculaire

Par
Wafae El Alaoui

Programme de Bio-informatique
Département de Biochimie

Mémoire présenté à la faculté des études supérieures
en vue de l'obtention du grade de
Maître ès sciences (M. Sc)
en Bioinformatique

31 Août, 2008

Université de Montréal
Faculté des études supérieures

Mémoire intitulé :

ESTIMATION DES LONGUEURS DE BRANCHE ET ARTEFACT SUR LA DATATION MOLÉCULAIRE

Présenté par :
Wafae El Alaoui

A été évalué par un jury composé de :

Sylvie Hamel, Président-rapporteur

Franz Lang, membre du jury

Hervé Philippe, directeur de recherche

RESUMÉ

La phylogénie moléculaire fournit un outil complémentaire aux études paléontologiques et géologiques en permettant la construction des relations phylogénétiques entre espèces ainsi que l'estimation du temps de leur divergence. Cependant lorsqu'un arbre phylogénétique est inféré, les chercheurs se focalisent surtout sur la topologie, c'est-à-dire l'ordre de branchement relatif des différents nœuds. Les longueurs des branches de cette phylogénie sont souvent considérées comme des sous-produits, des paramètres de nuisances apportant peu d'information. Elles constituent cependant l'information primaire pour réaliser des datations moléculaires. Or la saturation, la présence de substitutions multiples à une même position, est un artefact qui conduit à une sous-estimation systématique des longueurs de branche. Nous avons décidé d'estimer l'influence de la saturation et son impact sur l'estimation de l'âge de divergence.
Nous avons choisi d'étudier le génome mitochondrial des mammifères qui est supposé avoir un niveau élevé de saturation et qui est disponible pour de nombreuses espèces. De plus, les relations phylogénétiques des mammifères sont connues, ce qui nous a permis de fixer la topologie, contrôlant ainsi un des paramètres influant la longueur des branches. Nous avons utilisé principalement deux méthodes pour améliorer la détection des substitutions multiples : (i) l'augmentation du nombre d'espèces afin de briser les plus longues branches de l'arbre et (ii) des modèles d'évolution des séquences plus ou moins réalistes.
Les résultats montrèrent que la sous-estimation des longueurs de branche était très importante (jusqu'à un facteur de 3) et que l'utilisation d'un grand nombre d'espèces est un facteur qui influence beaucoup plus la détection de substitutions multiples que l'amélioration des modèles d'évolutions de séquences. Cela suggère que même les modèles d'évolution les plus complexes disponibles actuellement, (exemple: modèle CAT+Covarion, qui prend en compte l'hétérogénéité des processus de substitution entre positions et des vitesses d'évolution au cours du temps) sont encore loin de capter toute la complexité des processus biologiques.
Malgré l'importance de la sous-estimation des longueurs de branche, l'impact sur les datations est apparu être relativement faible, car la sous-estimation est plus ou moins homothétique. Cela est particulièrement vrai pour les modèles d'évolution. Cependant, comme les substitutions multiples sont le plus efficacement détectées en brisant les branches en fragments les plus courts possibles *via* l'ajout d'espèces, se pose le problème du biais dans l'échantillonnage taxonomique, biais dû à l'extinction pendant l'histoire de la vie sur terre. Comme ce biais entraine une sous-estimation non-homothétique, nous considérons qu'il est indispensable d'améliorer les modèles d'évolution des séquences et proposons que le protocole élaboré dans ce travail permettra d'évaluer leur efficacité *vis-à-vis* de la saturation.

Mots clés : phylogénie, datation moléculaire, longueurs de branche, modèles d'évolution.

ABSTRACT

Molecular phylogeny provides an additional tool complementary to paleontological and geological studies, allowing the reconstruction of phylogenetic relationships between species and the estimate of their divergence time. Researchers are mainly focusing on the topology of a phylogenetic tree; i.e. the relative connection between different nodes. Whereas, the branch lengths of this phylogeny are often considered as secondary, i.e. as additional parameters containing little information. However, the branch lengths are the primary information for molecular dating. Importantly, saturation, the presence of multiple substitutions at the same position, is an artifact that leads to an underestimation of the branch length. We are therefore interested in estimating the magnitude of this phenomenon and its impact on divergence time.

We chose to study the mammalian mitochondrial genome, which is available for many species and displays a high level of saturation. Furthermore, the phylogenetic relationships of mammalians are known, thus allowing us to fix the topology, thus eliminating one of the parameters influencing the branch lengths. We used two main approaches to improve the detection of multiple substitutions: (i) an increase in the number of species breaks the longest branches of the tree, (ii) more realistic models of sequence evolution. The results demonstrate that there is a very pronounced underestimation of branch lengths (up to a factor of 3). Furthermore, the use of a large number of species is the factor that influences most the detection of multiple substitutions, not the improvement of the model of sequence evolution. This suggests that even the most complex evolutionary models currently available, like the CAT+ Covarion model, which takes into account the heterogeneity of the substitution process between sites and the rates of evolution over time, are still far from taking the entire complexity of biological processes into account.

Despite the important underestimation of branch lengths, the impact on dating appeared to be relatively limited, because the underestimation is more or less homothetic. This is obviously true for the complex evolutionary models. Since multiple substitutions are most effectively detected when breaking the long internal branches via the addition of species. This raises the problem of bias in the taxonomic sampling, due to the impact of extinction on the history of life on earth. Because this kind of bias leads to a non-homothetic underestimation, we consider it essential to improve models of sequence evolution and suggest that the protocol developed in this work will allow to evaluate their effectiveness towards saturation.

Keywords: Phylogeny, molecular dating, branch lengths, evolutionary models.

TABLE DES MATIÈRES

LISTE DES TABLEAUX

LISTE DES FIGURES

REMERCIEMENTS

Je voudrais tout particulièrement remercier mon directeur de recherche **Hervé Philippe** sans qui je n'aurais pas pu réaliser ce travail. Je lui serais toujours reconnaissante pour son aide, sa patience, sa compréhension, son enthousiasme, sa disponibilité (la liste est longue…) et surtout sa grande capacité à véhiculer ses connaissances. Sa passion pour la recherche lui confère les grandes qualités d'un professeur digne de ce nom. Et je ne peux être que fière d'avoir pu travailler avec lui durant ces dernières années. En espérant que des collaborations futures me permettront à nouveau de faire partie de son équipe.

J'aimerais remercier également **Nicolas Lartillot** qui a du subir toutes mes questions concernant Phylobayes ces dernières années. J'ai eu du plaisir à travailler avec lui à Montpellier durant mon stage au LIRMM et j'ai eu la chance de continuer cette collaboration à Montréal.

Je ne pourrais oublier de remercier Beatrice Roure, Henner Brinkmann et Claudia Kleinman pour leur soutien ces dernières années, que ça soit pour les maintes présentations orales qu'ils m'ont aidé à corriger ou pour la relecture de ce mémoire, je leur suis reconnaissante d'avoir pu m'aider à avancer dans mon travail.

Je remercie également tous les membres de l'équipe (actuels ou passés) :
Fabrice Baro, Dorothée Coste, Frederic Delsuc, Emmanuel Douzery, Jean-Christophe Grenier, Olivier Jeoffroy, Guy Larochelle, Nicolas Rodrigue, Yan Zhou

I. INTRODUCTION

Les études phylogénétiques des données moléculaires permettent d'établir les relations de parentés entre espèces actuelles. Lorsqu'elles sont associées aux données paléontologiques, elles sont aussi capables de dater des événements de spéciation. Nous allons tout d'abord décrire les principales méthodes de reconstruction phylogénétiques utilisées de nos jours, pour par la suite nous attarder sur les méthodes de datation moléculaire et les problématiques qui y sont associées.

Nous allons principalement discuter les différents facteurs qui peuvent affecter l'estimation des longueurs de branche par les méthodes de reconstruction phylogénétique actuelles. Les longueurs de branche sont un élément important pour estimer les âges de divergence. Les lacunes associées à leur mauvaise estimation peuvent avoir des conséquences sur la datation moléculaire. Nous allons évaluer différents moyens de contourner ces problématiques en essayant d'améliorer l'estimation de ces longueurs de branche.

1.1. Phylogénie

Définition d'une phylogénie :

Les phénomènes de spéciation et de duplication des gènes peuvent se schématiser sous la forme d'un arbre ; celui-ci est constitué de nœuds et de branches. Les nœuds externes sont les séquences actuelles, tandis que les nœuds internes sont des séquences hypothétiques représentant un ancêtre commun de deux espèces actuelles. Les branches de l'arbre sont caractérisées par leurs longueurs qui représentent la quantité de changements évolutifs inférés sur ses branches, généralement exprimée en nombre de substitutions par site. Une phylogénie permet en premier lieu de déterminer les relations de parenté entre différentes espèces (Figure 1).

La plupart des méthodes de reconstruction phylogénétiques vont inférer des arbres non enracinés, car elles détectent des différences entre séquences mais n'ont aucun moyen d'orienter temporellement ces différences. Il est donc nécessaire d'inclure un groupe de séquences connues *a priori* comme étant externe aux taxons d'intérêt, pour enraciner l'arbre sur la branche qui relie ce groupe aux autres séquences.

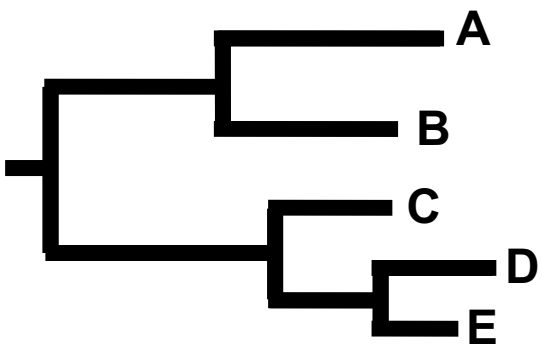

Figure 1 : Exemple d'une phylogénie de 5 espèces (A à E) caractérisée par des longueurs de branche

1.1.1. Calcul de la distance évolutive

L'idée principale des méthodes de distance est de convertir les données de séquence (alignement de séquences protéiques ou nucléotidiques) en une matrice de distances qui permet ensuite la construction d'une phylogénie. Cette matrice donne les distances évolutives entre toutes les paires de gènes pour l'ensemble des données. Une distance représente le nombre de substitutions s'étant produites le long du chemin séparant deux taxons. En pratique, on observe le nombre de différences entre deux séquences, qu'idéalement on espère proche du nombre de substitutions. Mais on est amené à corriger cette valeur pour prendre en compte la possibilité qu'un site ait subi plusieurs substitutions. Si les distances sont additives, c'est-à-dire que la distance entre deux paires de feuilles est égale à la somme des longueurs des branches qui les relient, alors il est possible de reconstruire un arbre. En pratique, les distances n'étant jamais additives, on est amené à minimiser l'écart entre matrice et longueurs de branche. Des méthodes ont été développées qui permettent d'ajuster la matrice à tous les arbres possibles et de calculer les longueurs de branche de l'arbre en choisissant celui avec l'écart minimal (Darlu et al. 1993). Parmi ces méthodes, on retrouve UPGMA (Unweighted Pair Group *Method* with Arithmetic mean) (Sneath et al. 1973) et NJ (Neighbor joining)(Saitou et al. 1987).

La distance évolutive que l'on appellera *d* est égale au nombre total de substitutions qui ont eu lieu entre 2 lignées depuis leur divergence. Celle-ci est exprimée en nombre de substitutions par site et est rapportée à la longueur des deux séquences. On peut estimer la distance entre deux séquences en considérant leur pourcentage de similarité (S = M/L). Pour des séquences nucléotidiques codant pour des protéines, la similarité (S) entre deux séquences peut être égale par exemple au nombre de sites synonymes identiques (M) divisé par la longueur de la séquence (L). La distance observée entre deux séquences (D) est donnée par (D = 1 – S) (Darlu et al. 1993).

Si le nombre de substitutions qui ont eu lieu depuis la divergence de deux séquences est petit, alors la plupart des substitutions sont probablement des substitutions simples (Figure 2a), puisque la probabilité que le même site ait subi plus d'une substitution est faible, ce qui fait que la distance observée (D) sera très proche de la vraie distance évolutive (d). Par contre lorsque le nombre de substitutions augmente, la probabilité qu'un site subisse plus d'une substitution augmente. Les substitutions multiples (parallèles ou inverses) (Figure 2b-f) peuvent donc fausser l'inférence du nombre de substitutions entre deux séquences en sous-estimant le nombre réel de changements évolutifs. Différents moyens statistiques et mathématiques sont utilisés afin d'estimer au mieux la vraie distance évolutive. Des hypothèses sur la nature du processus évolutif sont nécessaires afin d'inférer la distance évolutive à partir des différences observées.

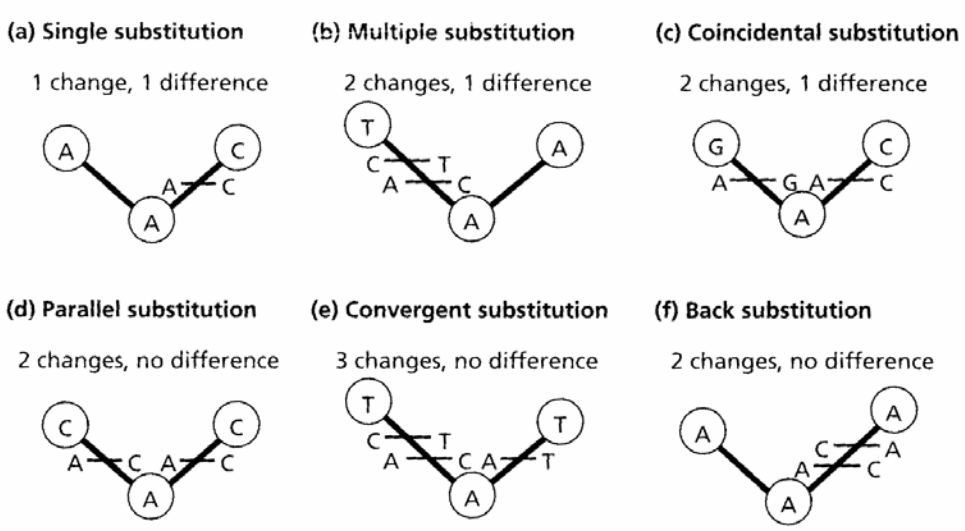

Figure 2 : Figure représentant les 6 types de substitutions possibles. Les séquences descendantes sont identiques alors qu'il a pu y avoir plus que 2 ou 3 substitutions, les différences observées sont donc plus petites que les changements qui ont vraiment eu lieu. L'homoplasie peut donc mener à une mauvaise inférence des relations évolutives entre les séquences (Page et al. 1998).

Lorsqu'on a plusieurs substitutions par site, la vraie distance évolutive (d) est supérieure au nombre de différences observées (D), on a donc :

$d = D$ + changements cachés

Ceci va affecter lors d'une inférence phylogénétique les longueurs de branche de l'arbre. Le nombre de substitutions inféré le long des branches va être inférieur au nombre réel de substitutions qui ont eu lieu, par conséquent nous aurons des branches trop courtes. Ce phénomène est appelé saturation[1] et diminue le caractère informatif des séquences étudiées. Si le taux de substitutions est élevé, on observe effectivement que le signal phylogénétique récent (substitution récente) va en quelque sorte effacer le signal phylogénétique ancien, de sorte que l'on ne le détectera pas, ce qui entraîne des longueurs pour les branches internes beaucoup trop courtes (Bromham & Penny, 2003).

[1] On parle souvent de manière erronée de saturation mutationnelle, lorsqu'on devrait parler de saturation substitutionnelle.

Saturation

Avec le temps, le nombre de différences entre deux séquences devient un estimateur de plus en plus mauvais du nombre de substitutions depuis la divergence des deux séquences de leur ancêtre commun. La Figure 3 (courbe de saturation) montre la relation entre le nombre de différences observées entre deux séquences et leur temps de divergence. Cette relation n'est pas linéaire, mais atteint un plateau dû à plusieurs changements étant intervenus sur les mêmes sites. Pour un site dans une séquence, après plusieurs substitutions successives, on peut retrouver un état ancestral (homoplasie). Ceux sont des positions qui sont saturées. En résultat, les substitutions les plus récentes ou les plus anciennes ont peu ou pas d'impact sur le nombre total de différences observées entre les séquences. Que deux séquences aient divergé il y a 20 millions d'années (M.A.) ou 15 M.A., le nombre de changements observés est le même (Page et al. 1998).

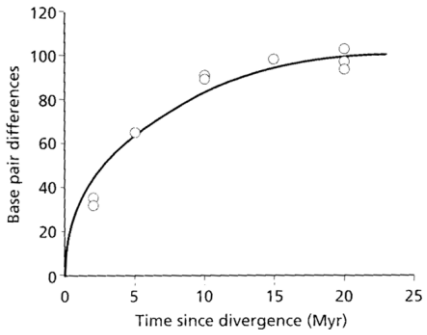

Figure 3 : Nombre de substitutions nucléotidiques entre des paires de séquences mitochondriales de mammifères bovins (684 paires de bases provenant du gène COII) en fonction du temps de divergence estimé (Janecek et al. 1996) (Figure de (Page et al. 1998)).

Exemple :

Pour un marqueur donné, les différents sites ne saturent pas à la même vitesse. Une claire illustration de ce phénomène est celle affectant la troisième position d'un codon codant pour une séquence protéique, puisque la troisième base sature beaucoup plus vite (beaucoup plus variable à cause de la dégénérescence du code génétique) que la première et deuxième base (position la plus conservée). La saturation mutationnelle peut donc affecter la séquence complète dépendant des contraintes évolutives sur cette séquence. En effet, moins il y a de contraintes sélectives, plus vite vont s'accumuler les substitutions et plus vite l'ancien signal phylogénétique va disparaître (saturation). Ce phénomène interviendra d'autant plus que le nombre d'états possibles sera faible et que les séquences appartiendront à des taxons très éloignés. (Hassanin et al. 1998).

1.1.2. Méthodes de corrections pour les substitutions cachées

Des techniques ont été développées pour corriger la distance à partir de la distance observée en estimant la quantité de changement évolutif qui a eu lieu (Felsenstein 2004). Les modèles d'évolution des séquences permettent de calculer les probabilités des différents changements entre les séquences en faisant des hypothèses concernant leur processus d'évolution. Ces modèles incorporent ainsi généralement des paramètres prenant en compte les fréquences stationnaires des nucléotides, la présence d'une fraction de sites invariables et l'hétérogénéité des taux de substitution entre sites (Figure 4). Pour tous ces modèles, une des hypothèses est que les processus de substitutions correspondent à un processus de Markov d'ordre O où la probabilité de passer de l'état *i* à l'état *j* dépend uniquement de l'état *i* et non pas des événements précédant cet état (Delsuc et al. 2004). On suppose également (en général) que pour ce processus les probabilités de substitutions ne changent pas le long de chacune des branches de l'arbre. Sous ces conditions, ceci peut être représenté de façon mathématique (pour les séquences d'ADN) sous la forme d'une matrice 4x4 de taux instantanés décrivant la vitesse à laquelle chaque nucléotide est remplacé par un nucléotide alternatif. Cette matrice notée Q correspondant au produit de la matrice des taux instantanées r par la matrice des fréquences stationnaires π. Celle-ci possède des composantes Q_{ij} qui correspondent aux taux r_{ij} de changement du nucléotide *i* en nucléotide *j* dans un intervalle de temps infinitésimal (Swofford et al. 1996; Delsuc et al. 2004).

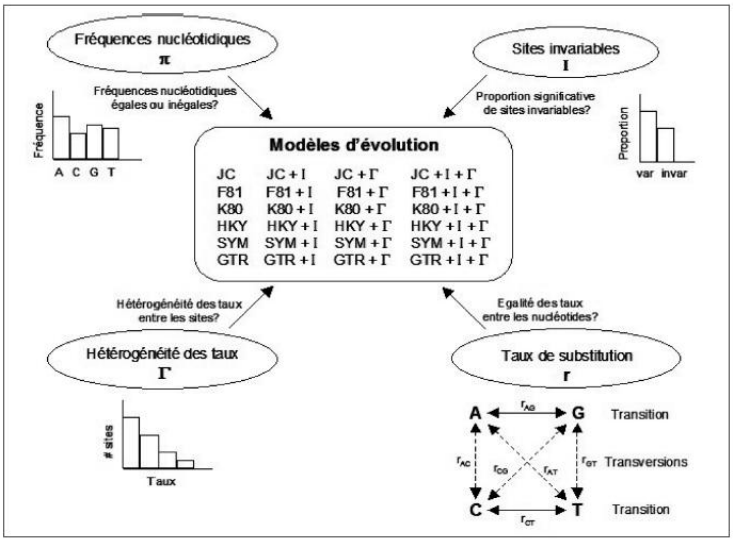

Figure 4: Principaux modèles d'évolution des séquences nucléotidiques (Delsuc et al. 2004)

Le premier modèle proposé est celui de Jukes et Cantor (Jukes et al. 1969). Ceux-ci font les hypothèses que tous les sites de deux séquences évoluent indépendamment et à la même vitesse, que toutes les substitutions sont équiprobables et que le processus ne varie pas au cours du temps (ce qui implique une fréquence stationnaire de 0.25 pour tous les nucléotides). On peut alors calculer la distance évolutive D_{jc}, en nombre de substitutions par site, à partir de la fréquence de différences observée D entre deux séquences (Équation 1) :

$$Djc = -\left(\frac{3}{4}\right)\ln\left(1-\left(\frac{4}{3}\right)D\right)$$

Équation 1 : Distance évolutive de Jukes et Cantor (1969).

La Figure 5 montre que la mesure de la distance augmente linéairement avec le temps (c'est une propriété recherchée pour la mesure de la distance quand il y a horloge moléculaire (voir section 1.2.1)). Cette correction indique clairement que la divergence est une fonction logarithmique en fonction du temps. On y observe l'augmentation de la variance en fonction du temps. C'est une indication que la mesure de la distance devient moins fiable lorsque le nombre de changements augmente.

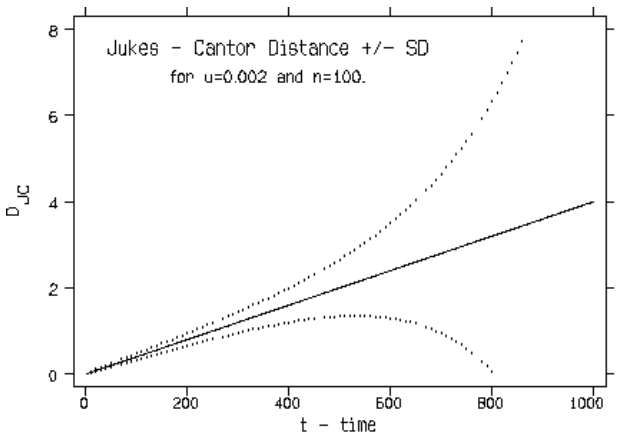

Figure 5: La courbe montre la correction de la distance Jukes et Cantor en fonction du temps t plus ou moins l'écart-type. On a "n" qui correspond à la longueur de la séquence, "μ" le taux de substitution et "t" le temps depuis le dernier ancêtre en commun de ces séquences. Cette courbe indique que la divergence est une fonction logarithmique du temps (http://helix.mcmaster.ca/721/distance/node3.html).

Dans ce modèle, le taux de substitution des nucléotides est le même pour toutes les paires de bases. Cela ne corrige donc pas les taux élevés de substitution de type transition (substitutions nucléotidiques entre les pyrimidines, C<=>T, ou les purines, A<=>G) comparés aux substitutions de type transversion (substitutions nucléotidiques entre pyrimidines et purines C<=>G, A<=>T, A<=>C ou G<=>T).

De nombreuses versions (Figure 6) successives de la matrice générale Q incorporant un nombre sans cesse croissant de paramètres ont été proposées depuis le modèle de Jukes et Cantor (Jukes et al. 1969). Chacune de ces matrices correspond à différentes hypothèses sur la nature du processus d'évolution moléculaire.
La plupart de ces modèles sont emboités entre eux, différant seulement dans le nombre de paramètres que ces derniers essayent d'inclure. La Figure 6 montre les relations et les différences entre quelques modèles d'évolution des séquences pour les nucléotides (Page et al. 1998). Le modèle GTR correspond à un des modèles les plus complexes, celui-ci assume que les différentes substitutions peuvent avoir lieu à des vitesses différentes, lesquelles dépendent de la fréquence à l'équilibre de chaque nucléotide. Celle-ci a été décrite en premier par Simon Tavaré en 1986. Les paramètres de cette matrice consistent en un vecteur de fréquence d'équilibre $\Pi = (\pi_1\pi_2\pi_3\pi_4)$ donnant la fréquence à laquelle chaque base apparait à chaque site ($\pi A \neq \pi T \neq \pi G \neq \pi C$), ainsi

que la matrice du taux. Celle-ci nécessite 6 paramètres pour les taux de substitutions en plus des 4 paramètres pour les fréquences d'équilibres (Tavare 1986).

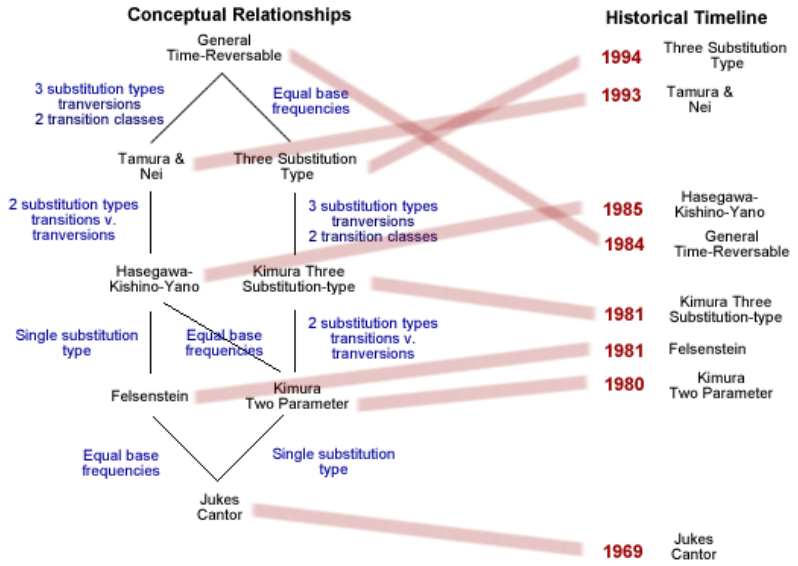

Figure 6 : Corrections apportées aux séquences nucléotidiques (Page et al. 1998). Les modèles intègrent un nombre plus ou moins élevé de paramètres.
– Correction de Jukes et Cantor (1969) – 0 paramètre
– Correction de Kimura (1980) – 1 paramètre (Kimura 1980)
– HKY (1985), F84 (1984) – 5 paramètres (Felsenstein 1984; Hasegawa et al. 1985)
– Tamura et Nei (1993) – 6 paramètres (Tamura et al. 1993)

De nombreuses méthodes de distances ont été développées et elles sont souvent utilisées, car elles sont très rapides. Des simulations (Lin et al. 1991; Pollock et al. 1995; Hastad et al. 1998) ont également montré qu'elles avaient des performances acceptables. Mais l'approche de distances qui ne permet de comparer que des séquences deux à deux empêche par exemple d'estimer le paramètre α de la distribution gamma (section 1.3.3.2). De plus, le fait de réduire la matrice de caractères à une matrice de distances induit la perte d'une certaine quantité d'information (Felsenstein 2004). Il existe des approches hybrides (distance calculée par maximum de vraisemblance) qui permettent de limiter ces problèmes. Mais nous allons plutôt nous concentrer sur les méthodes phylogénétiques qui permettent l'analyse de *n* séquences simultanément. Ces méthodes basées sur le principe du maximum de parcimonie ou du maximum de vraisemblance prennent en compte leur habilité à corriger pour les substitutions multiples et ainsi estimer leur nombre approximatif. (Moreira et al. 2000).

1.1.3. Méthodes comparant toutes les séquences simultanément

La méthode du maximum de parcimonie (MP) postule que, pour un groupe d'espèces, la meilleure phylogénie est celle qui nécessite le plus petit nombre de changements évolutifs. L'arbre phylogénétique des espèces est inféré de manière à impliquer le minimum d'événements évolutifs (Fitch 1971). L'avantage du MP est qu'il tient compte des types de caractères et ne réduit pas l'information à un nombre unique (une distance), car il compare plus que deux séquences à la fois. Même si l'arbre est connu avec certitude, on ne peut calculer le nombre minimal de changements pour un caractère, car il existe souvent plusieurs manières différentes mais équi-parcimonieuses de poser les changements sur l'arbre. Plusieurs stratégies en MP existent pour positionner les substitutions le long des branches. ACCTRAN (Farris 1970; Swofford et al. 1987) assume, lorsque les résultats sont ambigus, que la transition de caractères se produit le plus près possible de la racine de l'arbre. Elle renforce donc les synapormophies[2] sur les branches internes et favorise les réversions[3]. L'option DELTRAN (Farris 1970; Swofford et al. 1987) quant à elle a une approche opposée, car elle considère que les changements de caractères ont lieu le plus loin possible de la racine de l'arbre, et va donc repousser les transformations sur les branches terminales, ce qui favorise les convergences[4]. La méthode MP comporte des désavantages puisqu'elle ne corrige pas les substitutions multiples ayant eu lieu le long d'une branche donnée (ce qui a des conséquences beaucoup plus importantes quand peu d'espèces sont considérées) et calcule les longueurs de branche de manière ambigüe (indétermination à positionner les substitutions le long des branches). Celle-ci peut également être incohérente, c'est-à-dire que la phylogénie estimée peut converger vers une phylogénie incorrecte plus on augmente le nombre de caractères indépendants dans l'analyse (Hendy et al. 1989).

Puisque la méthode de MP ne permet pas de régler de façon efficace le problème de détection des substitutions multiples, deux méthodes probabilistes basées sur le

[2] Caractéristique nouvelle et distinctive partagée par un groupe d'organismes et qui en définit la lignée ou le clade.
[3] Apparition d'un caractère ayant l'apparence de la morphologie ancestrale.
[4] Apparition indépendante d'états de caractère identiques.

concept de vraisemblance ont été développées et appliquées au problème statistique de l'estimation des phylogénies : la méthode du maximum de vraisemblance et plus récemment l'approche bayésienne. Pour être statistiquement robuste et performante, les méthodes probabilistes doivent incorporer un modèle d'évolution de séquence et fournir un cadre statistique explicite pour permettre d'évaluer les hypothèses évolutives (Whelan et al. 2001). Elles semblent particulièrement adaptées au traitement des données moléculaires.

1.1.3.1. Maximum de vraisemblance (ML)

La vraisemblance est la probabilité d'observer les données X sachant l'hypothèse T (Pr (X|TM)). Dans le cas des analyses phylogénétiques, les données X sont l'alignement des séquences comparées, l'hypothèse T est l'arbre phylogénétique et le modèle d'évolution des séquences M (ainsi que les paramètres associés comme les longueurs de branche). Nous cherchons à trouver l'arbre dont la vraisemblance, étant donné les séquences observées et le modèle d'évolution choisi, est maximale. Il faut noter que cette vraisemblance de l'arbre n'est pas la probabilité que l'arbre soit « vrai », mais celle que l'arbre soit le plus vraisemblable (Felsenstein 1981).

Dans la méthode ML, les sites de l'alignement sont considérés comme étant indépendants. Il suffit donc de calculer la probabilité d'obtenir chaque site de l'alignement. La probabilité d'observer les états de caractères pour un site donné dépend de la topologie de l'arbre, de ses longueurs de branche et du modèle décrivant l'évolution des séquences le long de ses branches. Le calcul de la probabilité d'observer les données au site considéré est la somme des probabilités d'observer les différents états nucléotidiques possibles à ce site et à chaque nœud interne de l'arbre (Delsuc et al. 2004).

L'incorporation de modèles explicites d'évolution des séquences, dont les paramètres peuvent être estimés au cours de l'analyse ou donnés *a priori*, confère à cette méthode non seulement la capacité d'estimer la phylogénie mais également le mode d'évolution des séquences, ce qui permet de mieux estimer le nombre de substitutions qui a eu lieu au cours du temps. La méthode de ML est souvent considérée comme la plus fiable de toutes les méthodes phylogénétiques, celle qui conduit au

résultat le plus proche de l'arbre évolutif réel (Felsenstein 2004). Cette méthode permet d'appliquer les différents modèles d'évolution de séquences (par exemple les modèles cités précédemment) et d'estimer les longueurs de branche en fonction de changement évolutif, même si elle est coûteuse au niveau du temps de calcul**.**

1.1.3.2. Inférence bayésienne

Au cours du développement des méthodes probabilistes en phylogénie, l'approche bayésienne n'a été que très récemment appliquée (Par exemple, l'introduction du programme Mr. Bayes (Huelsenbeck et al. 2001)).

En théorie des probabilités, le théorème de Bayes estime les probabilités conditionnelles et marginales de deux événements aléatoires. Il est souvent utilisé pour déterminer la distribution de la probabilité *a posteriori* étant données certaines observations (Bayes 1763). Appliqué à la phylogénie, celui-ci a permis un important effort de modélisation de l'évolution des molécules (ex. modèle CAT (Lartillot et al. 2004)), alors que cette méthode est utilisée depuis fort longtemps en statistiques et mathématiques. Celui-ci a été formalisée par le calcul des probabilités postérieures – c'est-à-dire calculées *a posteriori* – d'arbres phylogénétiques à partir de probabilités définies *a priori* (Huelsenbeck et al. 2001). La notion de probabilité postérieure considérée ici, est la probabilité de l'hypothèse sachant les données X (Pr (T|X)).

L'inférence bayésienne (Équation 2) de la phylogénie combine la probabilité *a priori* Pr(T) d'un arbre T avec la vraisemblance Pr (X| T) des données X sachant cet arbre T pour produire une distribution de probabilité postérieure Pr(T | X) sur les arbres en utilisant la formule de Bayes :

$$\Pr(T \mid X) = \frac{\Pr(X \mid T).\Pr(T)}{\Pr(X)}$$

Équation 2 : Inférence bayésienne définie par la distribution de la probabilité postérieure.

Des priors sont définis sur tous les paramètres des modèles utilisés en bayésien. Les méthodes bayésiennes incorporent l'incertitude de certains paramètres (non connu

d'avance comme celui des longueurs de branche) en intégrant à travers une large fourchette de valeur plausible étant donné une distribution de probabilité *a priori* pour chaque paramètre. En revanche, des expériences réalisées par Yang et Rannala (Yang et al. 2005) ont établi que le choix des priors sur les longueurs de branche peuvent affecter les probabilités *a posteriori* et que la correspondance entre la probabilité *a posteriori* et la proportion correcte n'est pas toujours robuste avec l'utilisation de certains priors. Certains auteurs (Kolaczkowski et al. 2007) ont également montré que cela peut affecter les probabilités *a posteriori*, pouvant les faire dévier des valeurs qui auraient pu être inférées si les longueurs de branche étaient connues à l'avance.

Même si les probabilités postérieures ne peuvent pas être calculées analytiquement pour les jeux de données qui nous intéressent, les méthodes de "Markov Chain Monte Carlo" (MCMC) peuvent être implémentées pour trouver et examiner les « distributions d'équilibre » des arbres, ce qui permet de faire des probabilités d'hypothèse sur l'arbre "vrai" (Yang et al. 1997; Larget et al. 1999; Mau et al. 1999; Li et al. 2000). L'inférence bayésienne en phylogénie permet l'utilisation de modèles d'évolution sophistiqués, tandis que les MCMC retrouvent à partir de la probabilité de distribution *a posteriori* un échantillon de topologies où la fréquence relative empirique d'une topologie donnée converge vers sa probabilité marginale *a posteriori* (Tierney 1994).

Nous avons survolé dans cette section les principales méthodes d'analyses phylogénétiques, allant des méthodes de distances jusqu'aux méthodes d'inférences bayésiennes. Les méthodes phylogénétiques à proprement parler ne fournissent pas de datations absolues (en M.A), mais des quantités de divergence relatives (en substitutions\sites). Des méthodes de "datation moléculaire" vont permettre de coupler un arbre phylogénétique avec des points de calibration paléontologiques afin d'estimer des âges absolus de divergence entre taxons. Nous allons dans la section suivante approfondir un peu plus les méthodes de datations moléculaires et le contexte de leur utilisation.

1.2. Datation moléculaire

Quand les sédiments s'accumulent, les organismes vivants s'ensevelissent et, dans certaines conditions, forment des fossiles (empreinte ou trace) qui permettent aujourd'hui d'avoir un accès direct à l'histoire du passé. La paléontologie est la science qui permet d'étudier ces fossiles, et qui renseigne sur la vie passé en aidant à comprendre la nature des anciens organismes. Malheureusement, l'incomplétude du registre fossile ne permet pas de déterminer l'âge d'apparition de toutes les espèces, ce qui amène souvent à faire des extrapolations pour fournir des dates de divergence entre espèces ou à ne pas avoir de dates pour certains microorganismes. Or, les biologistes sont intéressés par toutes les dates de divergence pour toutes les espèces, en particulier à des fins comparatives (ex. cospeciation hôte-parasite).

En 1965, Zuckerkandl et Pauling (Zuckerkandl et al. 1965) ont eu l'idée d'utiliser des caractères moléculaires afin d'inférer un arbre phylogénétique. Grâce aux séquences primaires d'une protéine, la chaîne β de l'hémoglobine, ceux-ci ont proposé la première phylogénie moléculaire des Vertébrés, qui est plutôt en accord avec les phylogénies obtenues avec des données morphologiques et paléontologiques. Ils formulèrent aussi l'hypothèse de l'horloge moléculaire, qui stipule que les substitutions s'accumulent dans un gène à une vitesse proportionnelle au temps géologique, c'est-à-dire qu'un gène évolue à la même vitesse chez toutes les espèces. En effet, les datations fossiles fournissaient le seul moyen pour l'estimation de l'âge de divergence entre deux espèces, mais l'arrivée des données moléculaires (Zuckerkandl et al. 1965) ainsi que de nouvelles méthodes informatiques permettent d'estimer autrement le moment de divergence entre les lignées. En effet, lorsque les données moléculaires sont associées aux données paléontologiques, elles sont capables de dater des événements de spéciation en associant à un arbre phylogénétique une calibration fossile, ce qui offre la possibilité de dater tous les nœuds de l'arbre phylogénétique (Figure 7).

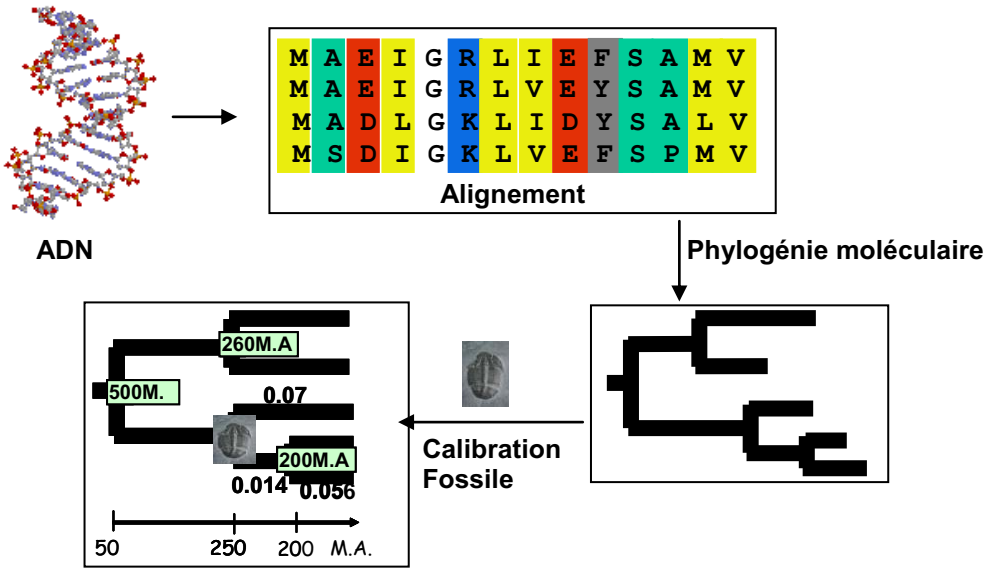

Figure 7 : Les différentes étapes pour réaliser la datation moléculaire à partir d'un alignement de séquences.

Comme on a vu précédemment, les longueurs de branche (LB= λ * Δt) lors d'une inférence phylogénétique sont fonction du temps Δt et du taux de substitution λ. Il nous est actuellement possible de mesurer LB, mais on n'a aucun élément pour séparer le temps Δt du taux λ, qui sont tous les deux *a priori* inconnus. Nous allons voir maintenant comment les différentes méthodes utilisées pour la datation moléculaire permettent de contourner ce problème.

1.2.1. La théorie neutraliste et hypothèse de l'horloge moléculaire

L'observation de la constance du taux d'évolution dans la formulation de l'hypothèse de l'horloge moléculaire fut surprenante, à cause de la variation de plusieurs facteurs pouvant entrer en jeu, tels que la variation du temps de génération ou la variation de la pression de sélection. Bien que la théorie de l'horloge moléculaire fût rapidement très contestée (Fitch et al. 1967; Kimura et al. 1971; Uzzell et al. 1971), il apparut que l'horloge moléculaire fonctionnait assez bien sur de longues périodes évolutives pour plusieurs gènes (Figure 8), mais que chaque région du génome accumulait des substitutions à un rythme spécifique dicté par la pression de sélection à

laquelle elle était soumise (c'est-à-dire que plus la sélection négative[5] est importante, moins le gène évolue vite).

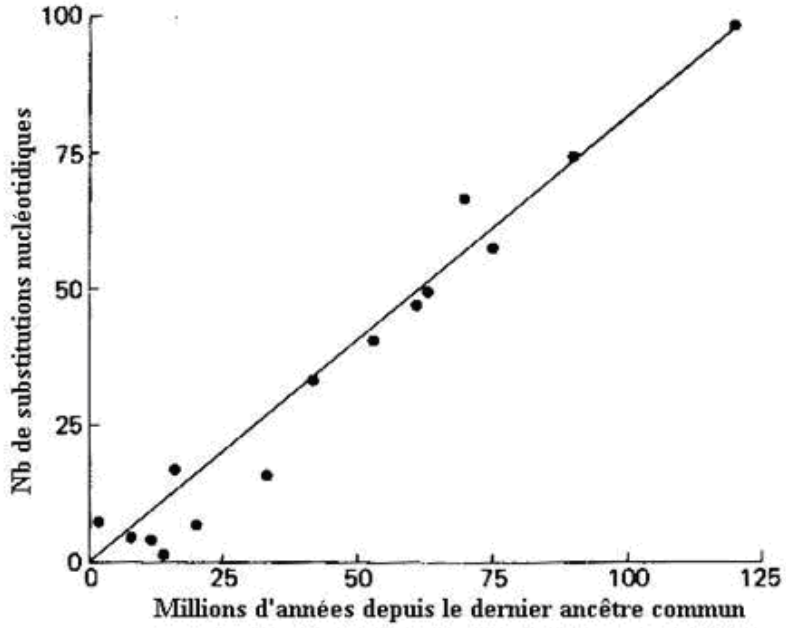

Figure 8 : Comparaison des séquences nucléotidiques de diverses protéines (Cytochrome C, Hémoglobine, Myoglobine, Insuline) et proportionnalité entre la divergence nucléotidique et le temps de divergence (Fitch et al. 1968)

Une des grandes forces de l'horloge moléculaire est qu'elle permet de définir le même taux λ pour tout l'arbre lorsqu'on a une calibration fossile, ce qui rend alors possible le calcul du temps de divergence Δt. Le fait d'attribuer le même taux d'évolution dans tout l'arbre fournit donc une solution assez simple aux problèmes de datation. Mais ceci a aussi soulevé un problème important, puisque l'horloge moléculaire semblait être en contradiction avec la théorie de la sélection naturelle de Darwin[6]. Cependant, la théorie neutraliste de l'évolution de Kimura, prédit l'existence d'une horloge moléculaire (pour les substitutions neutres), tout en étant en accord avec la sélection darwinienne. En effet, le taux d'évolution constant des protéines de l'hémoglobine ainsi que d'autres

[5] Mécanisme de sélection naturelle par lequel les gènes désavantageux sont éliminés (Darwin, C. (1859). The origin of species by means of natural selection. London, Murray.).
[6] L'idée centrale de la théorie darwinienne est la sélection naturelle (Ibid.).

protéines était la clé pour le développement de la théorie neutraliste de Kimura (Kimura M. 1968), puisque l'article de Kimura commence par une discussion sur le taux de substitution des acides aminés obtenus par Zuckerkandl et Pauling (Zuckerkandl et al. 1965). Dans cette théorie neutraliste, les changements dans les acides aminés sont très majoritairement limités aux changements qui n'affectent pas la fonction (changements neutres). Cette théorie est pleinement compatible avec le darwinisme, car elle suppose non seulement que les changements génétiques sont aléatoires, mais aussi que la sélection darwinienne élimine la plupart des mutations, la fixation des changements neutres se faisant par dérive stochastique. La théorie neutraliste propose surtout que la sélection positive est rare, question aujourd'hui encore débattue (Graur et al. 2000). Elle s'applique plus aux substitutions synonymes, pour lesquelles on s'attend à l'existence d'une l'horloge (peu ou pas de sélection positive). Mais puisque ces dernières saturent très vite, on doit également utiliser les substitutions non-synonymes quand on a une datation ancienne à réaliser (souvent supérieure à 50 M.A.). Ainsi, la découverte de l'horloge moléculaire et le concept d'évolution neutre au niveau moléculaire se sont renforcé mutuellement.

Cependant, plusieurs arguments ont malgré tout contribué à remettre en cause l'hypothèse de l'horloge moléculaire. Tout d'abord le taux d'évolution ne serait pas constant au cours du temps (Goodman M. 1962; Goodman M. 1963) : lors de la formation de nouvelles espèces, les mutations avantageuses seraient fixés plus rapidement. Certains travaux (Fitch et al. 1967; Kimura et al. 1971; Uzzell et al. 1971) ont révélé que la variance du nombre de substitutions d'acides aminés est généralement plus grande que celle attendue par la théorie neutraliste.

À partir de cette observation, d'autres auteurs (Uzzell et al. 1971) ont postulé que différents sites dans une même protéine peuvent évoluer à des taux différents. Une prédiction majeure, testable, de la théorie neutraliste est que le processus de substitution devrait suivre un processus de Poisson[7], avec une moyenne du nombre de substitutions par unité de temps égale à la variance. Si le paramètre du taux dans le processus de Poisson varie à travers les sites suivant une distribution gamma (section

[7] Modèle mathématique modélisant des événements aléatoires qui se reproduisent au cours du temps

2.4.3), une distribution binomiale négative en résulte, distribution qui a été effectivement observée pour plusieurs protéines (Uzzell et al. 1971).
Quelques années plus tard, Gillespie (Gillespie 1991) en utilisant le ratio (R) de la variance sur la moyenne du nombre de substitutions, a trouvé que l'horloge protéique est sur-dispersé ($R > 1$) et que l'évolution des protéines est épisodique, c'est-à-dire que les substitutions n'ont pas lieu de façon indépendante au cours de l'évolution, et qu'il y a des épisodes d'accumulation suivis d'arrêts évolutifs. Selon lui, les accumulations de substitutions se feraient en réponse à des changements environnementaux. Les protéines évoluant lentement et à un rythme à peu près constant ne seraient pas soumises à de grandes contraintes fonctionnelles, mais se trouveraient plutôt dans un environnement constant.

Malgré cet intense débat qui se poursuit encore, l'horloge moléculaire est devenue un outil important pour l'étude de l'évolution moléculaire. La plus simple interprétation est d'assumer que la théorie neutraliste de Kimura s'applique à de nombreux marqueurs moléculaires. Il est également possible d'assumer, « bien que la neutralité soit l'hypothèse conventionnelle, que des processus du type de ceux envisagés par Gillespie peuvent se moyenner sur de longues périodes et le taux apparaître à peu près constant (traduction de : (Takahata 2007)) ». À ce jour, l'horloge moléculaire a malgré tout été utilisée par de nombreux chercheurs afin de dater des événements de spéciations. Doolittle et al. (1996) et Feng et al. (1997) ont respectivement reporté l'âge de divergence des eucaryotes et procaryotes; Hedges et al. (1996) et Kumar et Hedges (1998) ont proposé une échelle d'évolution pour les vertébrés; Wray et al. (1996) ont daté la diversification des métazoaires; (Korber et al. 2000) ont étudié l'ancêtre commun le plus récent des principales souches de VIH.
Mais lorsqu'on souhaite appliquer l'hypothèse de l'horloge moléculaire, il est essentiel de préalablement tester son existence sur le jeu de données en question. Dans cette perspective, un certain nombre de tests ont été développés.

1.2.2. Les Tests du taux relatif

Puisqu'on n'a pas de connaissance directe de la séquence ancestrale des deux séquences modernes et qu'on a peu d'information sur leur temps de divergence, le taux d'évolution ne peut être comparé directement. Le plus simple, le test de taux relatif (Sarich et al. 1973) permet de comparer ce taux entre deux espèces. L'idée de base du test du taux relatif est la suivante (Figure 9) : si on considère par exemple trois séquences (A, B et C) avec C étant le groupe externe, soit n_{ijk} le nombre de sites observés où les séquences A, B et C ont les nucléotides i, j et k. Sous l'hypothèse de l'horloge moléculaire, $E(n_{ijk}) = E(n_{jik})$ peu importe le modèle d'évolution et que le taux de substitution varie avec le site. Si cette égalité est rejetée, alors l'hypothèse de l'horloge moléculaire peut être rejetée pour ce lot de séquences. Par exemple, Wu et Li (WU et al. 1985) ont appliqué le test du taux relatif pour comparer des séquences nucléotidiques de Rongeurs et d'Homme avec un groupe externe. Ils ont découvert que les Rongeurs évoluaient approximativement deux fois plus rapidement que l'Homme. Cela suggère que l'effet du temps de génération est plus fort pour des substitutions nucléotidiques silencieuses que pour des substitutions d'acides aminés.

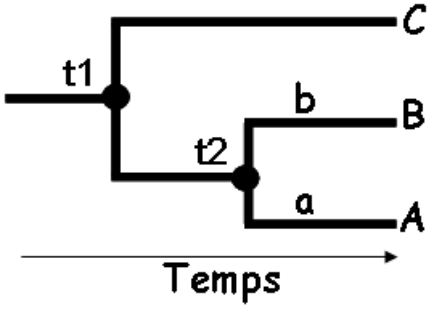

Figure 9 : Relation phylogénétique pour trois séquences hypothétiques d'espèces A, B et C. L'ancêtre commun le plus récent de A et B se trouve au temps t2, et l'ancêtre commun le plus récent pour les trois espèces se trouve au temps t1 ; a et b sont la quantité de changements dans les séquences (mesurée par le nombre de substitutions qui ont eu lieu) le long des branches des espèces A et B à partir du temps t2 jusqu'au temps actuel. La différence expérimentale d (A, C)-d (B, C) est assumée être égale à a-b. On peut donc tester si a est égal à b en regardant si d(A,C)-d(B,C) est significativement différent de zéro (Sarich et al. 1973).

Une autre approche consiste à implémenter un test de taux relatif non paramétrique (Gu et al. 1992; Tajima 1993). Ce test est non paramétrique puisqu'il ne nécessite pas d'information sur le modèle d'évolution. Il est basé sur des triplets de séquences, et stipule que le nombre de sites où un même nucléotide (ou un même

acide aminé) est partagé entre un des deux groupes internes et le groupe externe est égal pour les deux groupes internes lorsque le taux de substitution est constant. Lorsque les deux gènes évoluent sous l'hypothèse de l'horloge moléculaire, on s'attend à ce qu'ils accumulent un nombre similaire de substitutions « uniques ». « D'un autre coté, lorsqu'une des deux copies a accumulé un nombre plus grand de substitutions, l'horloge moléculaire ne s'applique pas et un des paralogues (gènes ayant évolué à partir de la duplication d'un même gène de départ) a alors subi une augmentation dans son taux d'évolution » (traduction de (Van de Peer et al. 2001)). Puisque cette approche ne prend pas en compte un modèle explicite, elle ignore les événements de substitutions multiples qui peuvent mener à la saturation ou biaiser le nombre de changements inférés. Ça ne permet pas non plus une généralisation simple pour plus de trois séquences. Mais son avantage est que cette méthode peut être utilisée pour comparer des taux pour lesquels il n'y a pas de modèle disponible, comme les insertions ou les délétions (Gu et al. 1992).

Le test $X2$ de l'horloge moléculaire utilisant le maximum de vraisemblance (voir section 1.5) a également été proposé (Felsenstein 1981) et semble être valide si la taille de l'échantillon observé est suffisamment importante comparée avec le nombre d'états de caractères possibles aux différents sites de l'alignement, ce nombre qui augmentant assez rapidement avec le nombre de séquences (Goldman 1993). Ce test peut être appliqué pour trois séquences également, ce qui permet de revenir à un test de taux relatif si une séquence est contrainte à être le groupe externe par rapport aux deux autres dans un arbre phylogénétique enraciné.

Plusieurs auteurs ont proposé par la suite un test du taux relatif pour plus de trois séquences (Li et al. 1992; Takezaki et al. 1995). Il est donc possible dans ce cas de comparer le taux de substitutions de plusieurs lignées, en considérant pour chacune d'elles plusieurs séquences et en utilisant une lignée composée de plusieurs séquences comme groupe externe.

1.2.3. Les limites de l'horloge moléculaire pour les datations

La théorie neutraliste est maintenant connue comme étant une explication incomplète de l'hypothèse de l'horloge moléculaire. Par exemple, Ayala a noté : « la fondation théorique proposée originalement par l'horloge, nommé la théorie neutraliste de l'évolution moléculaire, est intenable. La variation du taux d'évolution moléculaire a contribué largement à invalider la théorie » (Ayala 1999). Pulquerio et Nichols ont également noté : « la théorie neutre n'est pas une explication complète. Par exemple, elle prédit un taux de substitution constant par génération, alors que des évidences empiriques suggèrent quelque chose de plus proche d'un taux constant par année. »(Pulquerio et al. 2007)

Dans les dernières années, plusieurs études ont donc démontré que l'hypothèse de l'horloge moléculaire était violée (Ayala 1999; Ho et al. 2006; Pulquerio et al. 2007) et cela même pour les changements synonymes (Zeng et al. 1998). Dans de nombreux cas en plus, l'absence de l'horloge moléculaire peut ne pas être perçue à cause de la saturation (crée une variation du taux artefactuelle et par conséquent une illusion d'horloge) (Scherer 1989; Springer 1995) et/ou du faible pouvoir statistique des tests utilisés (Fitch 1976; Gingerich 1986; Springer 1995; Robinson et al. 1998). Face aux nombreuses violations de l'horloge moléculaire, certains auteurs ont également proposé pour permettre de réaliser des datations, d'enlever des taxons ou des gènes, quand leurs taux évolutifs varient de manière statistiquement significative. Une méthode permettant d'éliminer les gènes ou les taxons évoluant trop vite ou trop lentement par rapport à un taux de mutation moyen a ainsi été proposée (Takezaki et al. 1995), mais vite critiquée (Rambaut et al. 1998). Lorsque l'absence de l'horloge moléculaire est visible, il est effectivement possible d'éliminer les séquences qui gênent (Hedges et al. 2004) en appliquant les tests de détection des écarts par rapport à l'hypothèse d'horloge moléculaire. Mais le retrait de taxons ayant un taux d'évolution différent, peut conduire à la perte de diversité taxonomique et peut avoir comme conséquence la perte de points de calibration paléontologique (Rambaut et al. 1998).

Quand il y a alors absence d'horloge moléculaire, cela conduit de façon certaine à des erreurs lors de la datation. Pour contrer ce problème, plusieurs modèles « d'évolution des taux » ont été proposés afin d'assouplir l'hypothèse d'horloge moléculaire stricte par des modèles statistiques plus sophistiqués, dite d'horloge moléculaire relâchée.

1.2.4. Comment améliorer l'horloge moléculaire ?

1.2.4.1. Horloge locale

Le principe de l'horloge locale (Figure 10) est que même si le taux de substitution varie sur l'ensemble de l'arbre, il existe certaines parties de l'arbre où il est constant. Une des difficulté de cette horloge est d'identifier correctement les branches ou régions de l'arbre dans lesquelles les taux de substitutions différent significativement les uns des autres ; cette difficulté explique pourquoi il existe alors plusieurs méthodes du type « horloge locale » (Kumar 2005).

La méthode de datation par « quartets de taxon » (Rambaut et al. 1998) est une des implémentations les plus simple de la méthode d'horloge moléculaire locale. Cette dernière fonctionne avec un quartet d'espèces qui combine deux paires d'espèces, chacune d'entre elles ayant un temps de divergence connue. Pour chaque paire, un taux peut être estimé, ce qui permet d'estimer le temps de divergence entre les paires (l'âge du quartet). Cette méthode permet d'éviter les problèmes reliés à l'incertitude de la topologie de l'arbre, mais il est difficile de combiner les estimations de plusieurs quartets d'une façon significative (Bromham et al. 1998).
Une autre méthode, probablement la plus connue d'horloges moléculaires locales, basée sur une approche de maximum de vraisemblance a été proposée par Yoder et Yang (2000). Elle permet de résoudre de manière simple le problème de la séparation du taux λ et du temps Δt en continuant à garder le taux λ fixe (dans certaines parties de l'arbre). Ces derniers ont implémenté une horloge moléculaire locale en maximum de vraisemblance qui permet de considérer différents taux d'évolution pour certaines lignées tandis que d'autres sont assumées être constantes (Figure 10). Dans un tel

modèle d'horloge moléculaire locale, on assume que chaque branche dans la phylogénie peut prendre un des k taux possibles. On considère que $\lambda_0 = 1$ est le taux par défaut et on utilise simplement k-1 facteur de multiplication du taux comme paramètres additionnels. Lorsque k=1, toutes les branches ont le même taux, et le modèle est alors réduit à une horloge moléculaire globale. Le modèle a donc *n*+k-2 paramètres (*n* étant le nombre d'espèces). On peut noter que le nombre de branches dans une topologie d'arbre non-enracinée est 2*n*-3, donc *n*+k-2 ne devrait pas dépasser 2n-3. Par contre certaines spécifications du taux pour les branches peuvent rendre impossible l'identification de tous les paramètres du modèle, et ces problèmes d'identifiabilité doivent être évités (Yoder et al. 2000). Yoder et Yang ont appliqué ce principe à des gènes du génome mitochondrial codant pour des protéines en utilisant plusieurs points de calibration (voir section 1.3.1). Les résultats de ce modèle statistique sur l'estimation des dates a permis de donner une estimation de la date de divergence des primates qui est bien supportée par les données et qui est généralement congruente avec les informations paléontologiques (Yoder et al. 2000).

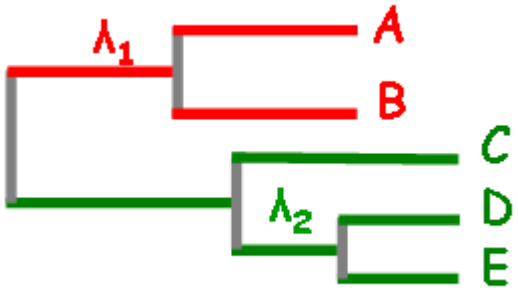

Figure 10 : Illustration d'une horloge moléculaire locale pour une phylogénie de 5 espèces (A à D) avec différents taux λ.

Il reste le problème de la définition objective des horloges locales. De plus, pour de grandes phylogénies, il y a de nombreuses manières de spécifier différentes horloges locales dans différents sous-arbres. Une méthode basée sur le principe d'horloge moléculaire locale a été proposée en 2004 (Yang 2004). Celle-ci dérive d'une autre méthode appelée « ad hoc rate smoothing » (AHRS) qui permet de contourner une des difficultés de l'horloge locale, soit celle de placer les horloges moléculaires locales sur l'arbre phylogénétique, puisque cette dernière permet de le faire automatiquement.

Cette approche (AHRS), implique quatre étapes: 1) les longueurs de branche sont estimées pour chaque gène dans un arbre pré-spécifié sans restriction d'horloge moléculaire; 2) les taux d'évolution initiaux spécifiques aux branches sont approximés en se basant sur une approche paramétrique de lissage qui minimise la variation du taux ; 3) par la suite, une algorithme de « clustering » ad hoc est utilisé pour grouper ces taux de branche spécifiques dans des groupes qui constituent les horloges locales ; 4) les temps de divergence sont finalement estimés par maximum de vraisemblance d'après (Yang et al. 2003). Même si l'algorithme AHRS rend possible de placer les différentes horloges locales sur la phylogénie, le nombre de ces horloges reste à déterminer par l'utilisateur.

Il existe également une extension de l'algorithme AHRS (Aris-Brosou 2007), qui modifie surtout l'étape 2 et 3 de l'algorithme original AHRS. Celle-ci permet d'utiliser une approche hybride qui semble mieux performer que les méthodes de vraisemblance pénalisée ou que la plupart des modèles bayésiens (Section 1.2.4.2.).

1.2.4.2. Horloge relâchée

Plusieurs récentes approches indépendantes ont toutes pour point commun d'essayer de contourner les problèmes survenant lors de l'utilisation d'horloges globales ou locales dans la réalisation de datations moléculaires en modélisant, implicitement ou explicitement, le taux. Ces méthodes incluent des approches non-paramétriques (Sanderson 1997; Sanderson 2002), et des modèles paramétriques bayésiens (Thorne et al. 1998; Huelsenbeck et al. 2000; Kishino et al. 2001; Aris-Brosou et al. 2002; Aris-Brosou et al. 2003; Drummond et al. 2006; Lepage et al. 2007).

Parmi toutes ces méthodes, l'approche paramétrique offre l'opportunité d'explorer une grande diversité de modèles alternatifs, chacun ayant des hypothèses spécifiques concernant la forme de l'arbre et la manière dont les taux de substitutions vont varier dans le temps. Ces modèles diffèrent donc dans plusieurs aspects, certains décrivent le taux comme un processus continu (Kishino et al. 2001), d'autres comme une fonction constante à travers le temps (Huelsenbeck et al. 2000). D'autres modèles diffèrent dans leur dynamique : les taux sont autocorrélés (Kishino et al. 2001; Aris-Brosou et al.

2003), ou non (Drummond et al. 2006). Elles diffèrent aussi dans la stratégie d'incorporer les contraintes d'âges (les points de calibrations) dans l'analyse. Finalement, différents priors ont été proposés pour la forme de l'arbre (qui dépend de l'histoire des espèces et de l'échantillonnage), incluant une distribution uniforme (Kishino et al. 2001; Drummond et al. 2006) ; une distribution de Dirichlet (Thorne et al. 2002), et une distribution découlant d'un processus de naissance/mort (Aris-Brosou et al. 2003).

Il existe encore à travers cette variété de modèles de relaxation des controverses sur la possibilité de biais (Aris-Brosou et al. 2003). Ces biais seraient dû au modèle de relaxation lui-même dans certains cas, ou sur le prior sur le temps de divergence (Blair et al. 2005; Welch et al. 2005).

1.2.4.2.1 Autocorrélation des taux

L'autocorrélation des taux (Figure 11) proposé par Thorne (1998) est basée sur le principe que, lorsque deux taxons se séparent par spéciation, leurs taux d'évolution respectifs à un locus donné sont identiques (ou du moins très proches). Ensuite, les différences de taux d'évolution peuvent se propager indépendamment le long des deux branches descendantes. En pratique, chaque branche descendant d'un nœud se voit attribuer un nouveau taux, qui est échantillonné dans une distribution log-normale dont la moyenne est égale au taux de la branche directement ascendante et dont la variance est le produit du coefficient d'autocorrélation υ par la durée cumulée des deux branches. Par convention bayesienne, υ est appelé un hyperparamètre car il gouverne une distribution probabiliste définie *a priori*.

Les variations graduelles de taux présentées ci-dessous (Figure 11) sont modélisées de la manière suivante. « Chaque branche est caractérisée par un unique taux, qui est la moyenne de son taux initial et de son taux final. Le taux r_1 d'une branche descendante est tiré (disque vert) dans une distribution normale centrée sur le taux r_0 de la branche ascendante. Le taux r_1 est ici plus rapide que r_0. Les autres relations d'ordre illustrées sont : $r_2 > r_1 > r_0$, $r_3 > r_4$, et $r_5 > r_4$. Une variance de cette distribution normale des taux, respectivement importante ou faible, entraîne un écart important ou faible par rapport à l'hypothèse d'horloge moléculaire » (Douzery et al. 2006).

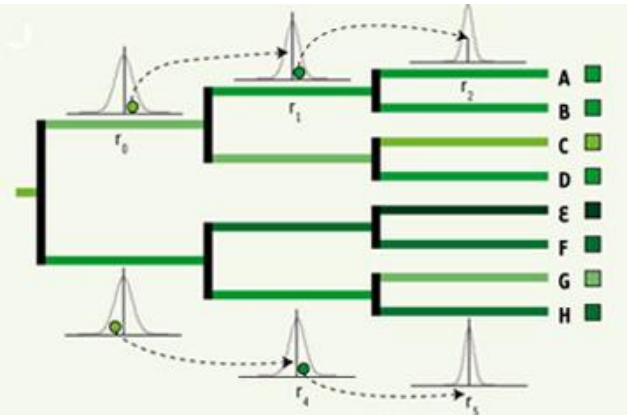

Figure 11 : Principe de l'autocorrélation des taux.

Une approche proposée par Sanderson (méthode non-paramétrique s'appuyant sur le critère des moindres carrés) permet de considérer que les taux d'évolutions évoluent à partir d'un taux ancestral, ce qui implique que les taux d'évolution sont autocorrélés (Sanderson 1997). Cette autocorrélation impose des limites par rapport à la vitesse avec laquelle un taux d'évolution peut changer d'une lignée ancestrale à une lignée descendante : plus le coefficient d'autocorrélation des taux est élevé, moins il y a de contraste entre le taux de substitution des branches descendantes vis-à-vis de leur branche ascendante, et plus proche nous sommes de l'hypothèse d'horloge globale. « Cette méthode pénalise les paramètres estimés du taux en les comparant directement à leur voisin immédiat au niveau phylogénétique. Elle utilise un estimateur naïf du taux local en tant que nombre de substitutions inférées sur une branche divisé par le temps inféré » (traduction de (Sanderson et al. 2004)).

Plus récemment, Sanderson (Sanderson 2002) change son approche non-paramétrique d'utilisation du critère des moindres-carrés, par une approche semi-paramétrique qui combine la puissance des méthodes non-paramétriques et les avantages des méthodes paramétriques. Cette dernière approche s'appuie sur un modèle contenant plusieurs paramètres, et définit un paramètre de lissage qui en limite les fluctuations. « Ce paramètre contrôle le lissage des taux d'évolution en empêchant de trop grandes variations d'une branche à l'autre et permet de contrôler également la qualité d'ajustement des données au modèle. Lorsque ce paramètre de lissage est égal à zéro,

il n'y a pas de pénalisation des différences de taux, et chaque branche de l'arbre évolue selon son propre taux. Lorsque celui-ci tend vers l'infini, la moindre variation de taux est sanctionnée, et le modèle converge vers l'horloge globale. Un large éventail de possibilités offrant différents niveaux de lissage des taux d'évolution peut donc être étudié » (Douzery 2002).

1.2.3.2.2. Distribution *a priori*

Les approches bayésiennes qui ont été proposées pour estimer le temps de divergence ne sont pas seulement caractérisées par le modèle d'évolution des taux, mais aussi par tout un ensemble de priors. En particulier, il y a les priors du modèle pour la spéciation et sur les changements autocorrélés des taux. Ainsi, à partir de l'alignement de séquences, des intervalles de contraintes paléontologiques et des distributions *a priori* des différents paramètres constituant "l'hypothèse" vont permettre de calculer les valeurs *a posteriori* de ces paramètres. Dans le modèle de (Thorne et al. 1998), il est effectivement possible d'incorporer l'information du prior sur l'âge minimum et maximum d'un nœud, sur la base des enregistrements fossiles mais aussi sur le taux d'évolution moléculaire sur différentes branches étant donné l'âge d'un nœud interne (Kumar 2005). L'implémentation proposée offre la possibilité de paramétrer la moyenne et la déviation standard des distributions *a priori,* du taux à la racine, et de l'âge de la racine (Thorne et al. 1998). Les priors de ce modèle sont en effet équivalents à une fonction pénalisée particulière de l'approche de Sanderson (Sanderson 1997). Des études de simulation montrent que ce raffinement du modèle incorporant les contraintes de dates imposées par le registre fossile conduit à augmenter la précision des estimations d'âges de divergence (Kishino et al. 2001). En effet, l'adjonction de contraintes paléontologiques au modèle a pour effet immédiat de réduire les intervalles de confiance sur les âges estimés *a posteriori* dans le voisinage immédiat du point de calibration qui a été contraint (Kishino et al. 2001).

Même si les distributions *a priori* et les détails de l'implémentation varient, l'approche bayésienne et la vraisemblance pénalisée ont en commun d'adoucir ou de minimiser la variation du taux à travers le temps avec le processus d'autocorrélation. Puisque ceci

peut mener à des estimations de dates inappropriée lorsque les taux varient grandement (Ho et al. 2005), d'autres approches bayésiennes ont été proposées. Le modèle de Huelsenbeck et collaborateurs (Huelsenbeck et al. 2000) permet aux taux de changer n'importe où dans l'arbre. En effet, Huelsenbeck *et al.* (2000) n'autorisent pas uniquement les changements de taux d'évolution aux nœuds, mais n'importe où le long des branches de l'arbre. Leur modèle n'introduit que deux paramètres (processus de Poisson avec paramètre ρ et le paramètre α qui contrôle la forme de la distribution gamma) par rapport à une horloge globale stricte. Ce modèle correspond à un "modèle de Poisson composé". Les substitutions nucléotidiques ont lieu le long des branches de l'arbre selon un premier processus de Poisson (Gillespie 1991). Les positions sur l'arbre où sont introduits les changements de taux de substitution sont définis selon un second processus de Poisson. Ainsi, contrairement au modèle précédent, le taux d'évolution peut varier à l'intérieur même d'une branche. Un des avantages de traiter la variation du taux comme un processus de Poisson est que le modèle introduit la variation du taux à n'importe quel point de l'importe phylogénétique, tandis que la plupart des autres méthodes introduisent le changement du taux seulement à des points correspondant à des événements de spéciations (Rutschmann 2006).

D'autres approches ont été introduites plus récemment. Ces dernières assignent directement un paramètre pour la moyenne du taux à chaque branche (Aris-Brosou et al. 2003; Drummond et al. 2006). L'approche de Aris-Brosou utilise trois types de distribution *a priori* des taux d'évolution dans un modèle bayésien : une distribution gamma, une distribution exponentielle et une distribution normale équivalente à celle utilisée par Thorne et *al*. (1998).

Un nouveau modèle basé également sur l'autocorrélation a été proposé en 2007 (Lepage et al. 2007). Ce dernier suit un processus CIR (Cox et al. 1985) dans lequel des propriétés stationnaires sont bien définies. Le modèle CIR est défini comme un mélange de processus « Ornstein-Uhlenbeck carré » (Aris-Brosou et al. 2003). Ce processus préserve la positivité du processus du taux, et évite les biais systématique (les taux tendent à être plus élevés proche de la racine de l'arbre)(Lepage et al. 2007).

Il est également possible de relâcher l'hypothèse de l'horloge moléculaire sans autocorrélation. Drummond et *al* (2006) ont étudié trois jeux de données et ont

découvert qu'il n'existe pas d'autocorrélation significative des taux le long des branches, ce qui suggère que pour certains jeux de données les modèles d'autocorrélation ne soient pas nécessaires. Dans cette approche sans corrélation les longueurs de branche sont également distribuées d'après une distribution gamma ou une distribution exponentielle (Drummond et al. 2006).

Une étude récente a permis de comparer certains modèles d'horloge relâchée (Lepage et al. 2007). En appliquant plusieurs modèles à trois jeux de données protéiques différents (Eucaryotes, vertébrés et mammifères), ils ont évalué l'ajustement relatif de ces modèles et des priors (grâce au facteur de bayes en utilisant une méthode numérique nommé « intégration thermodynamique »). Ils ont pu démontrer que les modèles CIR et log-normal, ont un ajustement semblable et meilleur par rapport au modèle non-corrélé (Drummond et al. 2006) et cela sur les trois jeux de données. Puisque jusqu'à ce jour, c'est la seule étude comparative, il faudrait certainement d'autres études afin de comparer plus de modèles différents.

En résumé, l'introduction des horloges moléculaires relâchées permet de résoudre une partie des problèmes posés par l'horloge moléculaire globale, en prenant en compte l'existence d'hétérogénéité du taux d'évolution entre espèces. Un exemple de datation réalisée grâce aux horloges relâchées : « Adoptant une approche bayésienne basée sur l'autocorrélation des taux pour calibrer l'horloge moléculaire relâchée, Springer *et al*. (Springer et al. 2003), ont estimé les âges de divergence des principaux groupes de mammifères placentaires. Les résultats suggèrent une diversification qui serait survenue au cours du Crétacé, il y a environ 100 Ma, et indiquent qu'au moment où les dinosaures s'éteignaient, il y a environ 65 Ma, la plupart des ordres de placentaires étaient déjà apparus, sinon diversifiés. L'occupation de niches écologiques laissées vacantes par l'extinction des dinosaures n'aurait donc pas été le facteur déclenchant à l'origine de la radiation évolutive des ordres modernes de mammifères placentaires (Bromham et al. 1999)» (Douzery et al. 2006).

1.2.4. Principales controverses entre paléontologie et datation moléculaire

L'exactitude de l'horloge moléculaire, même avec une reconnaissance précoce de sa nature stochastique, a longtemps été sujet à controverse. En supposant une divergence de temps assez précoce des Hominoïdes (23-25 M.A.) d'après les paléontologues (Steiper et al. 2004), Goodman (1962, 1963) a stipulé un ralentissement du taux dans la lignée menant vers l'Homme. Le ralentissement du taux pour les Hominoïdes est attribué à l'effet du temps de génération qui assume que les espèces avec un temps de génération plus petit vont évoluer plus rapidement que les espèces avec un temps de génération plus long. Cette corrélation entre le temps de génération et le taux de substitution apparaîtrait si les erreurs dépendant de la réplication de l'ADN étaient la source majeure des mutations, ce qui ferait que les organismes avec un temps de génération plus court subiraient plus de réplications par unité de temps que ceux qui auraient un temps de génération plus long. À l'opposé, (Sarich et al. 1967) et (Wilson et al. 1969) soutiennent qu'il n'y a pas eu de ralentissement, postulant un récent ancêtre entre l'Homme et les autres Anthropoïdes au Pliocène[8]. Le test du taux relatif, proposé par ces derniers, leur a permis de montrer que l'horloge était respectée, et qu'il n'y avait donc pas de ralentissement. Sarich et Wilson (1967) ont donc conclu que la divergence entre *Gorilla*, *Pan* et *Homo* a eu lieu il y a environ 5 millions d'années (M.A.); ils ont en effet observé que la distance moléculaire dans l'arbre se trouvait être 1/6 de la distance entre Cercopithecoidea (ex. Baboin, Macaque) et Hominoidea. Puisque le temps de divergence entre Cercopithecoidea et Hominoidea est estimé à 30 M.A. grâce aux données paléontologiques (malgré des enregistrements fossiles incomplets), les auteurs ont conclu que la divergence entre *Gorilla*, *Pan* et *Homo* a eu lieu il y'a 30/6 = 5 M.A. L'estimation moléculaire de la divergence entre *Pan* et *Homo* à 5 M.A. (Sarich et al.

[8] Ère pliocène : Terrains tertiaires les plus récents, la faune y présente des ressemblances avec le monde vivant actuel et l'apparition, chez les Mammifères des premiers Hominidés (Australopithèques). (http://www.universalis.fr/encyclopedie/O142281/PLIOCENE.htm)

1967) venait contredire l'âge estimé par les paléontologues, environ 30 M.A. Cette estimation a été confirmée en 1984 par Charles Sibley et John Ahlquist (Sibley et al. 1984), en utilisant la méthode expérimentale connue sous le nom d'hybridation de l'ADN. La date actuellement acceptée par tous les chercheurs est de 5-7 M.A. (Brunet et al. 2002), les paléontologues ayant drastiquement révisé leurs estimations, en particulier grâce à la découverte de nouveaux fossiles.

Un autre exemple célèbre de désaccord entre la paléontologie et la datation moléculaire est le cas des Métazoaires. D'après les données morphologiques, les Métazoaires sont le groupe-frère des Choanoflagellés, des organismes unicellulaires ressemblant aux choanocytes d'Éponges. Dans les sédiments du Néoprotérozoïque, on retrouve des Éponges qui sont les organismes Métazoaires les plus simples. Les Métazoaires diploblastiques (ex. les Porifères), possèdent, comme eux, 11 types de cellules spécialisées tandis que les vers qui sont des organismes triploblastiques plus complexes, en ont à peu près 55. La majeure partie des organismes fossiles édiacariens étaient des organismes à corps mou; cette faune semble avoir été pratiquement éteinte à la frontière entre le Précambrien et le Cambrien, il y a 543 M.A. et il n'apparaît pas de continuité claire entre les organismes du Néoprotérozoïque et ceux du Précambrien (Erwin et al. 1997). L'examen du registre fossile semble indiquer que la diversification des Bilateria (triploblastiques) s'est produite rapidement à la base du Cambrien, phénomène appelé explosion cambrienne (Conway 2000), soit il y a environ 540 M.A. La datation moléculaire quant à elle propose des dates très variables pour la divergence des Métazoaires bilatériens (Protostomes et Deutérostomes). Ces estimations diffèrent largement, allant de 582 M.A. (Aris-Brosou et al. 2002), 573-656 M.A. (Peterson et al. 2004), 670-736 M.A. (Ayala et al. 1998), 830 M.A. (Gu 1998), à 976 M.A. (Hedges et al. 2004), ce qui est toujours plus ancien que l'âge proposé par la paléontologie de 543 M.A. Même si des découvertes récentes de Métazoaires dans l'Ediacarien[9] ont peut-être étendu les enregistrements d'Éponges et d'animaux bilatéraux à 570 Ma. (Li et al. 1998; Xiao et al. 1998), les affinités biologiques de plusieurs organismes édiacariens restent controversées, et les preuves paléontologiques de la vie métazoaire ne dépassent pas 600 M.A. L'absence précoce de fossiles métazoaires peut être causée par des biais

[9] Désigne un biozone dans l'Eocambrien, daté de 700 millions d'années

systématiques dûs à une absence de préservation qui a empêché de les retrouver auparavant dans les enregistrements fossiles (Bowring et al. 1993). La diversification des Métazoaires ainsi que les désaccords entre paléontologie et datation moléculaire reste encore un sujet d'actualité.

Il est attendu que, par leur nature, la génétique et les données fossiles soient destinées à donner des estimations différentes de l'âge de divergence, puisque la forme morphologique des fossiles peut seulement identifier les morphoclades, tandis que les données génétiques peuvent seulement identifier les premières étapes de spéciation bien avant que les morphoclades définis par des synapomorphies morphologiques soient établis (Archibald et al. 1999). Pour certains généticiens, les arguments pour justifier ces différences sont plutôt tournés vers les lacunes dans l'enregistrement fossile (Hedges et al. 1996; Kumar et al. 1998). Les paléontologues désapprouvent en répliquant avec des preuves du contraire (Alroy 1999; Foote et al. 1999). Mais il a été argumenté que les estimations moléculaires de l'âge des clades vont toujours donner des âges plus anciens que celles proposées avec les fossiles (Archibald 1999). Cependant tout cela ne suffit probablement pas à expliquer toutes les différences observées. Un nombre de facteurs peuvent expliquer ce désagrément. Nous allons donc voir dans la section suivante, les différentes problématiques associées aux datations moléculaires et proposer différentes solutions afin d'améliorer les estimations des âges de divergences à partir des données moléculaires.

Les modèles d'horloges relâchées ont donc pour avantage de donner des temps de divergence entre espèces plus en accord avec les données paléontologiques (Sanderson 1997; Huelsenbeck et al. 2000; Kishino et al. 2001; Aris-Brosou et al. 2002; Douzery et al. 2004; Welch et al. 2005). L'avenir réside donc dans ces méthodes d'horloge relâchée, « notamment celles qui proposent d'intégrer plusieurs fourchettes d'âges de divergence plutôt qu'un unique point en guise de calibration, et qui développent un modèle explicite de variation des taux de substitution le long des différentes branches des phylogénies. Dans un avenir proche, nous devrions donc obtenir des datations moléculaires suffisamment fiables et assorties d'écart-types suffisamment réduits pour permettre d'identifier les âges de divergence entre taxons

compatibles avec les données paléontologiques, et de repérer les points de désaccord entre données nucléotidiques ou protéiques et fossiles. La résolution de ces désaccords nécessitera alors un nouveau retour analytique et critique sur les caractères moléculaires et paléontologiques. »(Douzery 2002)

1.3. Problématiques associées aux datations moléculaires et solutions proposées

1.3.1. Calibrations paléontologiques

Dans le but d'estimer le temps de divergence absolue entre espèces, il est nécessaire de calibrer l'horloge moléculaire avec des données paléontologiques. L'approche traditionnelle (Wray et al. 1996) utilise les fossiles pour assigner des dates à certains nœuds de l'arbre phylogénétique. À partir de là, il est possible grâce aux différentes méthodes de datations décrites précédemment d'estimer le taux de substitution pour différents taxons, ce qui permet d'estimer l'âge de divergence pour les autres nœuds de l'arbre phylogénétique. Malheureusement, l'utilisation de fossiles pour assigner les dates sur des nœuds de la phylogénie peut être problématique.

Si la succession des couches géologiques contenait un bon échantillon de toutes les plantes et animaux vivants à cette période, il serait facile de lire à travers le temps pour déterminer exactement l'apparition et l'extinction de toutes les espèces. Malheureusement les fossiles ont des modalités de conservation très différentes ; cette hétérogénéité est due à plusieurs facteurs dont les changements dans les paramètres physico-chimiques de l'environnement (conditions redox, pH, pression, température, bactéries), le niveau des mers ainsi que le temps d'exposition à l'érosion des sédiments, etc (Smith 1994). L'incomplétude du registre fossile nous empêche de déterminer facilement l'âge exact d'apparition de toutes les espèces. En effet, non seulement il existe des lacunes dans le registre fossile, mais aucun paléontologue n'est jamais assez chanceux pour découvrir un fossile datant exactement du moment précis auquel deux taxons divergent l'un de l'autre.

Pour illustrer ce genre de problème, disons qu'un fossile est identifié dans une couche stratigraphique qui détermine son âge dans un intervalle entre (T1-T2) M.A. Ce dernier possède des caractères qui impliquent une relation avec trois espèces (X, Y et Z). Deux d'entre elles (Y et Z), sont supposés être plus proches entre elles que l'autre espèce (X), sur la base de leur comparaison morphologique. L'existence du fossile seule ne nous apprend absolument rien sur le temps de divergence des trois espèces de Y à partir de Z ou de Y et Z à partir de X. En effet, l'estimation de temps de divergence obtenus à partir de fossiles dépend de l'interprétation phylogénétique du fossile (Figure 12). (Easteal 1999)

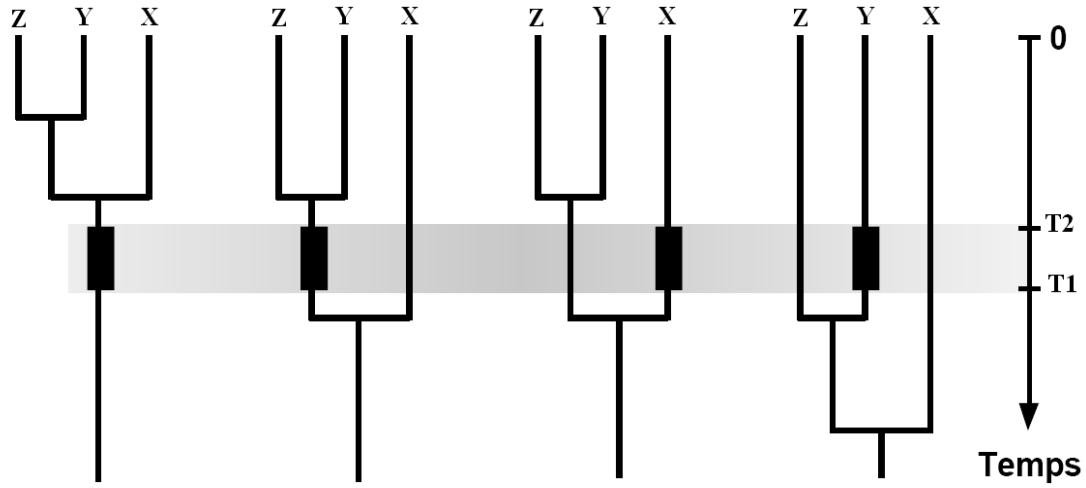

Figure 12: Quatre temps de divergences différents obtenus à partir de différentes interprétations phylogénétiques d'un fossile en relation avec 3 taxa, X, Y et Z. Estimation de temps de divergence obtenus à partir d'évidences fossiles dépend de l'interprétation phylogénétique du fossile. Le rectangle noir représente la région de l'arbre où se situe l'intervalle de confiance de la date du fossile (Easteal 1999).

Les fossiles ne fournissent donc pas une estimation du temps de divergence, mais plutôt un intervalle pour ce temps de divergence (Sanderson 1997). La meilleure situation possible serait de trouver un fossile qui se positionne sur la branche avant le nœud d'intérêt, et d'en trouver un autre qui se positionne sur la branche en dessous du nœud (Figure 13). N'importe quel fossile qui se positionne sur une branche avant un nœud fournit un intervalle supérieur sur le temps de divergence de ce nœud, alors que

celui qui se trouve en dessous du nœud fournit un intervalle inférieur sur le temps de divergence de ce même nœud (Cutler 2000).

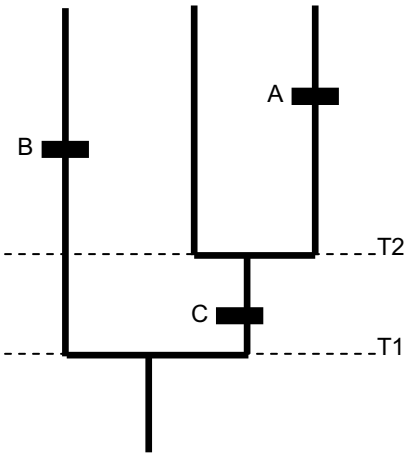

Figure 13 : Le fossile A fournit l'intervalle inferieur pour les temps T1 et T2. Le fossile B fournit l'intervalle inférieur pour le temps T1. Le fossile C fournit un intervalle supérieur pour le temps T2 et un intervalle inférieur pour le temps T1 (Cutler 2000).

Les fossiles fournissent donc plutôt des intervalles d'âges de divergence pour les taxons étudiés que des âges exacts, superposables aux nœuds des arbres notamment moléculaires. Si, par suite d'une erreur d'identification du fossile ou de datation de la strate à laquelle il appartient, les calibrations sont incorrectes, tous les âges de divergence vont être affectés, potentiellement de façon considérable. Dans de nombreuses études, les auteurs utilisent une calibration fossile unique (primaire, secondaire (Shaul et al. 2002) ou tertiaire (Heckman et al. 2001)), car ils considèrent souvent peu d'espèces, pour qu'elles soient disponibles pour beaucoup de gènes**.** Une calibration primaire se fait directement à partir de l'âge estimé d'un fossile. Une calibration secondaire dérive de l'âge estimé sur un nœud de l'arbre à partir d'une calibration primaire, (c-à-d. qu'une calibration primaire permet d'inférer un âge de divergence entre espèces à l'aide d'une méthode de datation moléculaire et que cette date inférée est à nouveau utilisée dans une autre étude comme calibration initiale sur un nœud de l'arbre pour inférer d'autres dates de divergence sur d'autres nœuds). Une

calibration tertiaire quant à elle dérive de l'âge estimé sur un nœud à l'aide d'une calibration secondaire (Shaul et al. 2002). Plusieurs équipes d'évolutionnistes moléculaires ont donc inféré des dates précises pour la divergence entre des lignées d'espèces à partir de calibrations secondaire et tertiaire. Ils les ont utilisées en tant que calibrations uniques, sans y associer l'incertitude liée à ces données, ce qui affecte considérablement les âges mesurés (Hedges et al. 1996; Kumar et al. 1998; Wang et al. 1999).

Graur et Martin (2004) montrent que le point de calibration choisi dans ces études n'est pas bien référencé, et qu'il contient des erreurs pouvant être associées à plusieurs facteurs, alors que la calibration fossile de 310 M.A +/- 0 (Homme-Poulet) est considéré comme une date dépourvue d'incertitude par ces auteurs (Heckman et al. 2001; Shaul et al. 2002). Il existe effectivement trois sources d'erreurs possibles par lesquelles ce point de calibration peut être affecté : 1. erreur dans la topologie de l'arbre phylogénétique, 2. erreur dans l'identification taxonomique du matériel fossile, ainsi que 3. erreur dans la détermination chronologique des strates géologiques. Il est donc important de considérer l'incertitude reliée aux différentes calibrations paléontologiques pour éviter ce genre de problème et d'utiliser des intervalles de temps plutôt que des points fixes dépourvus d'incertitude (Sanderson 1997; Kishino et al. 2001).

Des méthodes ont été proposées pour modéliser l'incertitude reliée aux données paléontologiques en assumant que les âges des calibrations sont distribués uniformément entre deux intervalles avec une probabilité nulle de se retrouver à l'extérieur de cet intervalle (Thorne et al. 1998). Ces intervalles « rigides » (Figure 14) surestiment souvent la confiance dans les données fossiles (Yang et al. 2006) puisque la paléontologie se trompe moins souvent sur la borne inférieure (car un fossile est présent) que sur la borne supérieure (où en général il y a absence de fossiles). Ceci permet donc à la méthode d'intégrer une bonne borne inférieure (âge minimum du nœud) mais pas une bonne borne supérieure (âge maximum du nœud). En pratique, les chercheurs peuvent être forcés d'utiliser une borne supérieure non réaliste pour éviter d'assigner un âge trop ancien. Cette stratégie est problématique, puisque l'intervalle imposé sur la prior peut influencer l'estimation du temps *a posteriori* (Yang et al. 2006).

Il est donc préférable d'utiliser des intervalles « souples » (Figure 14), avec des distributions plus flexibles, c'est-à-dire avec une probabilité non nulle de se retrouver à l'extérieur de l'intervalle de calibration. Des études ont été effectuées pour montrer comment des analyses bayésiennes sur des données moléculaires avec des calibration "rigides" ou "souples" peuvent être utilisées pour évaluer l'exactitude des calibrations proposés (Sanders et al. 2007). Les résultats qui en ressortent sur des données réelles ainsi que des données simulées ont permis de démontrer potentiellement la large différence pour l'estimation du temps de divergence obtenue avec des calibrations " rigides" ou "souples", avec généralement la supériorité des résultats obtenus avec une calibration "souple"(Yang et al. 2006).

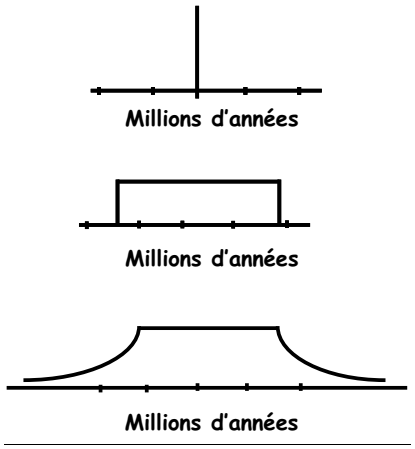

Figure 14 : Schématisation de calibrations possibles. En haut une calibration stricte. Au milieu une calibration de type « rigide ». En bas une calibration de type « souple ».

Le couplage fossiles et molécules a donc été amélioré non seulement par des méthodes de datation pouvant inclure plusieurs calibrations paléontologiques en même temps (section horloge relâchée), mais également par des méthodes permettant de considérer ces dernières comme des intervalles de temps plutôt que comme des points fixes dépourvus d'incertitude (Sanderson 1997; Kishino et al. 2001)

1.3.2. Problèmes dûs aux erreurs stochastiques

En dehors du problème de calibration paléontologique, Il existe encore plusieurs problèmes qui peuvent affecter les âges estimés. Le choix du gène peut être une source d'erreur stochastique (le choix d'un locus donné plutôt qu'un autre dans le génome) et peut avoir des répercussions importantes sur les estimations de l'âge des nœuds. Pour ce qui est de l'échantillonnage génomique, de multiples gènes ou protéines doivent être considérés afin de ne pas rendre les estimations d'âges de divergence trop dépendantes du choix d'un seul locus et de l'importante erreur stochastique qui lui est associée (Douzery et al. 2006). Des auteurs ont choisi d'utiliser un grand nombre de gènes afin de réduire l'effet de ce genre d'erreur stochastique (Hedges et al. 2004). Outre les problèmes reliés au choix du locus ou du genre, des erreurs systématiques souvent dû au choix du modèle peuvent affecter l'âge de divergence estimé.

1.3.3. Problèmes dûs aux erreurs systématiques

Les erreurs systématiques sont des artefacts des modèles qui se traduisent par un positionnement statistiquement significatif sur les arbres, bien qu'erroné, d'espèces évoluant plus rapidement que la moyenne. Le fait que les gènes n'évoluent pas à la même vitesse, va en premier lieu accentuer les inégalités des longueurs de branche, au-delà de la seule distance évolutive, dans les arbres phylogénétiques ce qui mène à une mauvaise reconstruction de l'arbre phylogénétique. Typiquement, on retrouve l'artefact du phénomène appelé l'attraction des longues branches (« long-branch attraction », LBA) (Figure 15) (Felsenstein 1978).

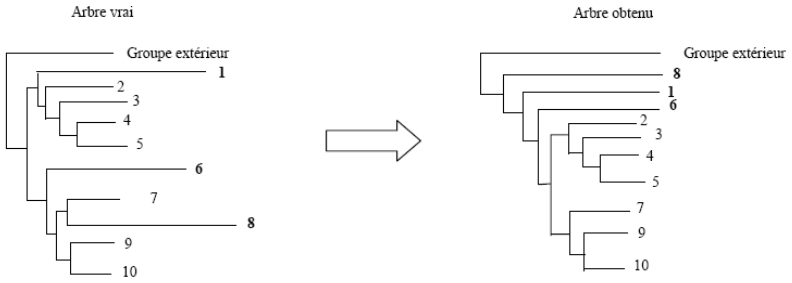

Figure 15 : Phénomène d'attraction des longues branches. Trois taxons (1, 6 et 8) évoluent plus vite que les autres taxons (branche beaucoup plus longue). Si on choisit un groupe extérieur qui est distant des autres espèces, ces taxons qui évoluent rapidement vont se retrouver groupés sur l'arbre non raciné, ou d'émergence basale sur l'arbre raciné (Philippe et al. 1998).

Quand certaines séquences évoluent beaucoup plus vite que d'autres, celles qui évoluent moins vite gardent plus de caractères ancestraux et finissent par se ressembler plus entre elles qu'aux autres. Avec la plupart des méthodes « classiques » de phylogénie, les séquences évoluant rapidement se regroupent ainsi artificiellement ensemble (Philippe et al. 1998). De façon similaire, lorsqu'un groupe extérieur distant est utilisé, ce qui est souvent le cas, on peut obtenir des arbres à base asymétrique où la longue branche du groupe extérieur « attire » les autres taxons à longue branche (Philippe et al. 1998). L'utilisation d'un groupe extérieur à branche courte est un bon moyen pour limiter les effets de l'attraction des longues branches.

Les différences observées entre les arbres utilisant des marqueurs génétiques différents peuvent aussi s'expliquer par la saturation mutationnelle des séquences. (Philippe et al. 1998). Comme dit précédemment, dans le cas de datations moléculaire, lorsque l'absence de l'horloge moléculaire est visible, il est possible d'éliminer les séquences qui gênent (Hedges et al. 2004) en appliquant les tests de détection des écarts par rapport à l'hypothèse d'horloge moléculaire. Dans de nombreux cas, l'absence de l'horloge moléculaire peut ne pas être perçue à cause de la saturation mutationnelle et/ou du faible pouvoir statistique des tests utilisés. La saturation fait effectivement croire à une horloge moléculaire et donc les différences de vitesses d'évolution des gènes surtout lorsque l'attraction des longues branches est présente ne sont pas apparentes, ce qui mène systématiquement à des erreurs lors de la datation.

1.3.3. Problème de Saturation : Moyens pour améliorer la détection des substitutions multiples

Plusieurs approches ont été appliquées pour résoudre ce genre de problème d'erreur systématique comme l'augmentation de l'échantillonnage taxonomique ainsi que l'amélioration des modèles d'évolution des séquences, permettant ainsi une détection plus efficace des substitutions multiples. Il a également été démontré que le retrait d'espèces évoluant rapidement (Aguinaldo et al. 1997), de gènes (Brinkmann et al. 2005; Philippe et al. 2005) ou des positions dans les séquences (Brinkmann et al. 1999; Hirt et al. 1999; Ruiz-Trillo et al. 1999; Burleigh et al. 2004) permettait de diminuer les erreurs systématiques reliées à ce phénomène. Ces méthodes qui ont été proposées pour éviter les problèmes d'erreurs systématiques devraient aussi être efficaces pour améliorer l'estimation des longueurs de branche.

1.3.3.1. Augmenter le nombre d'espèces

Il est assez facile de comprendre pourquoi l'augmentation du nombre d'espèces permet une meilleure estimation des longueurs de branche. En effet, nous avons vu précédemment que, pour un site donné, après plusieurs substitutions successives, on peut retrouver un état ancestral et donc que ce site est saturé. Les longues branches (en particulier celles des espèces évoluant rapidement) peuvent partager par chance des états de caractères plus fréquemment que les espèces plus proches qui partagent des caractères dérivés à partir d'un ancêtre commun (Felsenstein 1978). Lorsqu'on infère la phylogénie avec un grand nombre d'espèces, il est possible de briser les plus longues branches et révéler par conséquent les substitutions multiples jusque-là non inférées (Fitch et al. 1987; Fitch et al. 1990). Puisqu'une espèce intermédiaire peut contenir à une position donnée un état de caractère jusque-là non connu, cela peut effectivement permettre de détecter une nouvelle substitution.

On voit dans l'exemple de la Figure 16 qu'avec un ancêtre commun ayant une Adénine (A), celle-ci va être substituée deux fois pour finalement obtenir une Cytosine (C) chez le descendant. Lorsqu'on a peu d'espèces, on n'infère donc qu'une substitution de A vers C. Par contre, si on ajoute des espèces intermédiaires, celles-ci vont venir se placer de

façon à briser la branche qui mène au C, et donc à détecter qu'il y a eu une autre substitution vers une Thymine (T) avant de substituer en C.

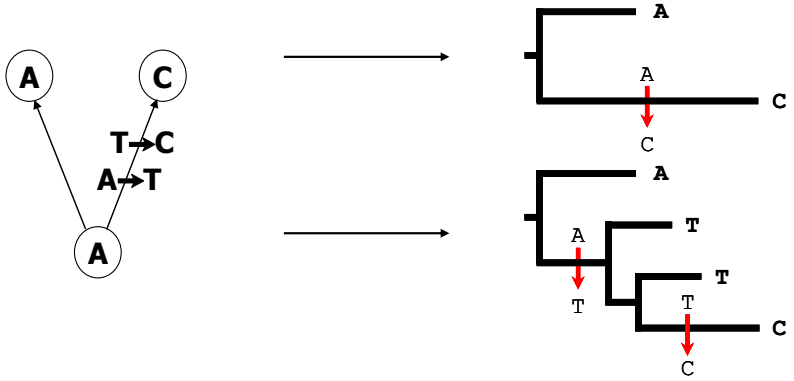

Figure 16 : Détection de substitutions supplémentaires par l'ajout d'espèces intermédiaires. On retrouve à gauche la schématisation d'une position d'un alignement contenant une adénine (A) chez l'ancêtre commun, et l'adénine et cytosine (C) pour les deux fils. La position contenant C peut être due à 2 substitutions qu'on ne peut pas détecter si on avait peu d'espèces. À droite, on voit que l'ajout d'espèces permet de détecter une substitution supplémentaire soit la thymine (T).

Plusieurs débats ont eu lieu sur l'importance de l'augmentation du nombre d'espèces et son impact sur l'inférence des arbres, mais très peu sur l'estimation des longueurs de branche. Hendy and Penny (1989) ont examiné les inférences phylogénétiques par la méthode de MP pour le cas où l'horloge moléculaire est respectée. Ils ont trouvé que les plus longues branches sont attirées les unes par les autres et que la méthode MP peut donc être incohérente, même avec une horloge. Ils ont donc suggéré que l'addition judicieuse de taxons peut briser les plus longues branches et aider la méthode de MP à mieux converger vers la « vraie » phylogénie (Hendy et al. 1989).

L'addition prudente de taxons est donc particulièrement importante lorsqu'on effectue des analyses de parcimonie puisque cette dernière est plus sujette au problème de l'attraction des longues branches. Puisque les méthodes paramétriques comme le ML, incorporent des modèles qui peuvent prendre en compte les substitutions cachées, ces méthodes sont moins sujettes à l'effet de l'attraction des longues branches (Swofford et al. 2001), tant que le modèle d'évolution utilisé est adéquat (Huelsenbeck et al. 1993; Kuhner et al. 1994; Huelsenbeck 1995; Swofford et al. 2001). Malgré tout, un nombre suffisant de taxons doit être échantillonné pour paramétriser ces modèles efficacement

(Pollock et al. 2002). De plus, les plus longues branches requièrent des modèles d'évolution de séquences plus proche de la réalité (puisque plus de changements cachés doivent être inférés), ce qui fait que l'augmentation du nombre d'espèces (qui brise les plus longues branches) est grandement bénéfique pour les méthodes paramétriques ainsi que les méthodes non paramétriques (Heath et al. 2008).

Néamoins, Kim (Kim 1996) montre que le problème d'inférence phylogénétique devient important pour la méthode de MP lorsqu'il y a un rajout d'espèces, et que non seulement les longues branches sont des mauvais indicateurs des mauvaises conditions de l'inférence phylogénétique, mais que même un arbre avec des branches « égales » peut produire des estimations incohérentes. L'addition de taxons supplémentaires atténue le problème d'incohérence seulement si la moyenne des taux de changements est basse. Si le taux est élevé, le problème peut devenir pire. L'auteur suggère que si l'on veut ajouter des taxons pour éviter ce problème, les espèces qui sont ajoutées doivent avoir un taux d'évolution peu élevé et être proche de l'ancêtre en commun des clades (qui peut être mesuré en comparaison avec le groupe extérieur). Les longues branches nécessitent un grand nombre d'espèces supplémentaires pour la reconstruction de la phylogénie, mais l'addition d'espèces doit être faite seulement pour les branches d'intérêt que l'on désire briser bien que les autres branches de l'arbre (où on ne fait pas d'ajout d'espèces) peuvent encore avoir des estimations incorrectes (Kim 1996).

Le débat sur l'efficacité de l'utilisation d'un grand nombre d'espèces pour éviter l'attraction des longues branches a été très actif. Hillis (1996) a étudié l'effet de l'échantillonnage taxonomique sur l'inférence phylogénétique directement en essayant d'évaluer l'exactitude de la reconstruction phylogénétique. L'auteur a analysé des données simulées sur une phylogénie de 228 espèces d'angiospermes, et a démontré que la phylogénie peut être correctement reconstruite avec des séquences longues de 5000 nucléotides. alors que plus de nucléotides sont nécessaire pour construire un arbre avec quatre espèces (Hillis et al. 1994). Ce résultat est surprenant puisqu'il existe seulement trois solutions possibles pour le problème des quatre taxons alors qu'il existe $1{,}2 \times 10^{502}$ solutions possibles pour le problème des 228 taxons. Même des méthodes simples comme le MP qui incorporent peu de complexité d'évolution dans le modèle

performent de façon remarquable à retrouver la topologie d'arbre de départ. Ils ont en conclut qu'un grand nombre de séquences peut fournir assez d'informations pour estimer de façon efficace une phylogénie (Hillis 1996), ce qui est en contraste avec d'autres études qui affirmaient le contraire (Hillis et al. 1994; Swofford et al. 1996).

Outre le fait d'augmenter le nombre d'espèces, l'augmentation du nombre de caractères a été discutée longuement. Certains auteurs ont suggéré que l'augmentation du nombre total de caractères peut augmenter la résolution et le support pour une phylogénie (Graybeal 1994; Hillis et al. 1994; Rannala et al. 1998). En 1999 Poe et Swofford (Poe et al. 1999) démontrent, grâce à des données simulées, que l'ajout de caractères peut être une meilleure stratégie que l'ajout d'espèces, même pour des arbres contenant des longues branches, et que le fait d'ajouter des espèces évoluant lentement, qui brisent les longues branches peut réduire la précision des résultats. Ceci a été contesté par d'autres études plus récentes (Pollock et al. 2002; Zwickl et al. 2002) qui ont permis d'établir que l'ajout de taxons a un effet positif sur l'exactitude de la reconstruction phylogénétique, et ont démontré dans plusieurs cas que l'augmentation du nombre d'espèces a beaucoup plus d'effets bénéfiques sur l'inférence phylogénétique que l'augmentation du nombre de caractères. (Phillips et al. 2004) (Delsuc et al. 2005).

Malgré ce débat qui se poursuit encore (Heath et al. 2008), plusieurs études sur l'importance d'utiliser un grand nombre d'espèces pour les inférences phylogénétiques ont démontré que l'ajout d'espèces supplémentaires dans les analyses phylogénétiques résulte en moyenne en de meilleurs estimations des relations d'évolution entre espèces (Lecointre et al. 1993; Hillis 1996; Rannala et al. 1998; Pollock et al. 2002; Zwickl et al. 2002; Poe 2003; DeBry 2005; Hedtke et al. 2006). Ces études représentent une large gamme d'approches incluant des simulations, l'analyse de groupes biologiques bien connus, et la comparaison à des phylogénies connues. Chacune de ses approches a des avantages et désavantages distincts (Hillis 1995), mais ensemble elles fournissent un message consistant sur l'importance d'utiliser un grand nombre d'espèces.

1.3.3.2. Améliorer les modèles d'évolution des séquences

Il est également important d'améliorer les modèles d'évolution des séquences afin de mieux décrire le processus évolutif. Ceci permet alors de mieux estimer le nombre de substitutions qui ont eu lieu le long des branches de l'arbre en ajustant le modèle d'évolution de manière plus efficace aux données. Pour plusieurs analyses, particulièrement pour des grandes distances évolutives, l'évolution est modélisée au niveau protéique. Puisque ce ne sont pas toutes les substitutions de l'ADN qui changent les acides aminés (a.a), de l'information est perdue lorsqu'on regarde au niveau des a.a au lieu de l'ADN. Malgré tout, plusieurs avantages sont en faveur de l'utilisation des a.a, puisque l'ADN est plus sujet à contenir des biais de compositions (dégénérescence du code génétique), puisque ce ne sont pas toutes les positions dans l'ADN qui évoluent à la même vitesse (les mutations synonymes sont plus fréquemment fixées dans une population que les mutations non-synonymes), et le peu de possibilités de caractères (quatre états de caractères possibles A, C, G et T). Tout cela fait en sorte que l'ADN subit plus de substitutions multiples (voir section 1.1.1). Ceci rend plus difficile d'estimer des longues distances évolutives que pour les a.a qui possèdent 20 états de caractères possibles.

Traditionnellement, les modèles d'acides aminés sont empiriques. La première matrice proposée en 1978 par Dayhoff (Dayhoff et al. 1978) estime le taux de substitutions à partir d'un alignement de plusieurs familles de protéines pour lesquelles les séquences homologues ont au moins 85% d'identité. Ceci minimise les chances d'avoir des substitutions multiples à un site. À partir de la matrice des taux estimés, une série de matrices de probabilité de substitutions fut dérivée, connu sous le nom de PAM. Les matrices Dayhoff PAM ont été basées sur relativement peu d'alignements de séquences (d'après la disponibilité de l'époque), mais dans les années 1990, de nouvelles matrices ont été estimées en utilisant la même méthodologie, mais basées sur des plus grandes bases de données protéiques (matrice JTT (Jones et al. 1992)).

Les différents modèles d'évolution de séquences ont souvent été utilisés dans un contexte de maximum de vraisemblance. Dans les analyses phylogénétiques, les paramètres sont estimés en utilisant le ML pour chaque jeu de données, ou sont prédéfinis à des valeurs pré-estimées auparavant à partir d'un large jeu de données représentatif (paramètres empiriques). Les paramètres empiriques sont utiles lorsqu'on se retrouve avec un grand nombre de paramètres ou lorsque les facteurs spécifiques des processus d'évolution entre les données sont similaires. Par exemple, dans les modèles d'évolution des protéines, les 190 paramètres d'échange entre les a.a sont empiriques pour la matrice de substitution Wag, puisque du coté pratique il est difficile d'estimer ce nombre de paramètres pour la majorité des jeux de données protéiques utilisés en inférence phylogénétique (Whelan et al. 2001).

Le dilemme qui reste par contre avec les modèles probabilistes est le contrôle de l'expression du modèle. Des modèles avec trop de paramètres peuvent «sur-ajuster » les observations, tandis que ceux ayant trop peu de paramètres peuvent être non réalistes, résultant en des conclusions erronées. Un exemple classique de la simplification est l'hypothèse d'un taux d'évolution égal entre tous les sites d'une protéine (Felsenstein 2001). De nombreux programmes de reconstruction phylogénétique permettent de modéliser cette hétérogénéité des taux entre les sites, puisqu'on sait que les différents sites d'une protéine ou d'un gène n'évoluent pas à la même vitesse. Une distribution gamma est l'approche la plus souvent utilisée à cause de sa grande flexibilité (Yang 1996). La moyenne de cette distribution est contrainte à 1, et un paramètre unique, appelé paramètre alpha, contrôle la forme (la variance) de cette distribution (Figure 17). Lorsqu'alpha a une valeur inférieure à 1, il existe une grande variation de taux à travers les sites présents dans l'alignement. Plus grande est la valeur d'alpha, plus faible est l'hétérogénéité.

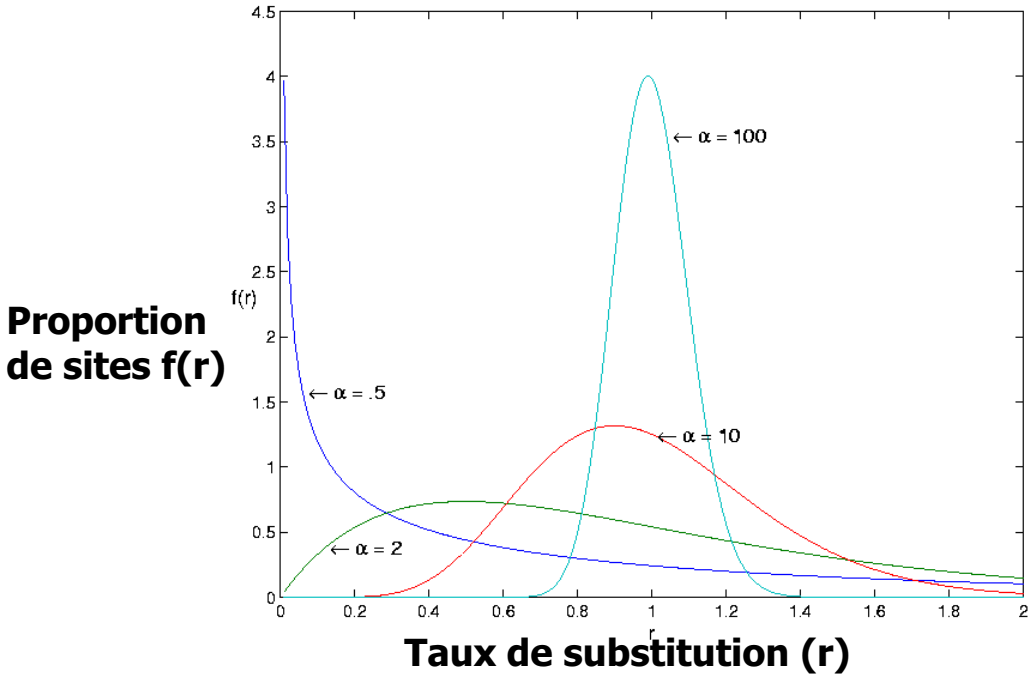

Figure 17 : Distribution gamma avec différents paramètres α qui contrôlent la forme de sa courbe.

Mis à part l'utilisation d'une distribution gamma pour modéliser le taux d'évolution, d'autres méthodes ont été introduites. En effet, Tuffley and Steel ont proposé en 1998 (Tuffley et al. 1998) d'utiliser l'idée de covarion de (Fitch et al. 1970). Dans le modèle de covarion, les positions sont autorisées à changer entre état variable (c'est-à-dire que la position peut accepter des substitutions) et état invariable. Ceci fournit des contraintes additionnelles et plus réalistes sur le taux d'évolution comparé aux techniques plus simples qui permettent au temps d'évolution sur chaque branche d'être sélectionnée aléatoirement à partir d'une probabilité de distribution comme la gamma distribution (Zhou et al. 2007). Plus récemment, Huelsenbeck et al. (Huelsenbeck et al. 2006) ont appliqué des priors du processus de Dirichlet dans une approche bayésienne pour modéliser les variations des taux entre sites pour permettre de faire varier la force de la sélection à travers les sites. La prior sur le processus de Dirichlet a deux composantes : (1) un paramètre, appelé α, influence la probabilité d'un élément des données à se retrouver dans une même catégorie, et (2) la probabilité de distribution qui décrit la probabilité du paramètre assigné à chaque catégorie.

En 2004, (Lartillot et al. 2004) ont proposé le modèle CAT (en inférence bayésienne) pour décrire le remplacement des a.a en autorisant un processus de substitution distinct pour différents sites de l'alignement. Ce modèle assume l'existence de classes distinctes différant par la fréquence d'équilibre pour les 20 acides aminés. La matrice d'échange peut être décomposée en deux jeux de paramètres : d'une part, des taux relatifs d'échange (190 paramètres), et d'autre part, des probabilités stationnaires (ou fréquences d'équilibre) pour les 20 acides aminés. L'utilisation de la prior sur le processus de Dirichlet permet également au nombre total de catégories et à leur profil respectif de fréquences d'a.a, ainsi que l'affiliation de chaque site à une certaine classe, d'être des variables libres du modèle. Ceci permet au modèle CAT de s'adapter à la complexité des données analysées. Les résultats obtenus avec ce modèle permet de montrer la supériorité de celui-ci par rapport aux modèles standard (Lartillot et al. 2004; Lartillot et al. 2008).

1.3.3.3. Retrait de sites rapides

Outre l'amélioration des modèles d'évolution des séquences ainsi que l'augmentation du nombre d'espèces, il existe d'autres moyens de contourner les problèmes de saturation mutationnelle des branches ainsi que les problèmes de l'attraction des longues branches. En effet, le retrait de sites à évolution rapide permet de ne garder que la fraction des sites qui ont le moins de substitutions multiples, donc ceux qui ont le moins besoin du modèle pour prédire les changements cachés. Ainsi, on peut limiter l'impact des imperfections des modèles actuels.

La variation des taux à travers les sites peut influencer grandement les méthodes de reconstructions d'arbres (Yang 1996). Il existe plusieurs méthodes qui permettent d'éliminer les positions qui évoluent rapidement. Ces méthodes varient dans la manière avec laquelle elles sélectionnent les sites. Par exemple la méthode SF ('Slow-Fast method') (Brinkmann et al. 1999) (Figure 18) permet de sélectionner les positions en fonction de leur taux d'évolution à l'intérieur de groupes d'espèces prédéfinis. Le nombre de changements pour chaque position est calculé avec des analyses de parcimonie à l'intérieur de groupes monophylétiques prédéfinis. Le taux d'évolution à une position est alors estimé comme la somme du nombre de substitutions à l'intérieur de chaque

groupe, et est donc indépendant des relations entre les groupes. Si un groupe évoluant rapidement est mal placé par des méthodes standards de reconstruction phylogénétique à cause de l'attraction des longues branches, sa bonne position a plus de chances d'être retrouvée avec les sites les plus lents, et va donc progressivement apparaître au fur et à mesure que les positions évoluant rapidement sont éliminées. La méthode SF permet d'être efficace pour détecter les groupes qui peuvent être mal placés à cause de l'attraction des longues branches (Gribaldo et al. 2002).

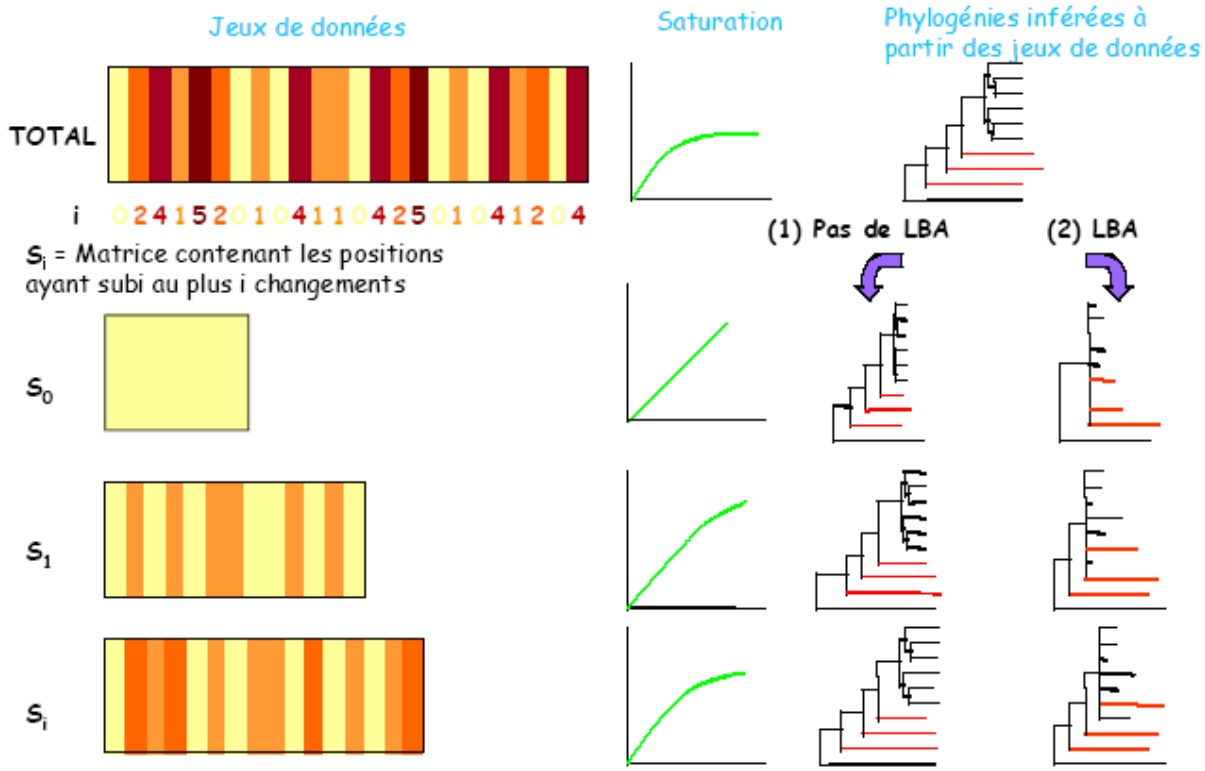

Figure 18 : Méthode SF (Brinkmann et al. 1999). « Le calcul des vitesses d'évolution de chaque position d'un jeu de données permet la construction de matrices S_i ou i est le nombre de changements maximums tolérés pour une position. L'augmentation du seuil i va permettre de créer des matrices contenant de plus en plus de positions variables, jusqu'à ce que l'ensemble des positions soit utilisé. » (Brochier 2002)

Dans une méthode proposée plus récemment par (Rodriguez-Ezpeleta et al. 2007), les sites évoluant rapidement sont identifiés en utilisant une modification d'une méthode datant de 1999 par Ruiz-Trillo et al. (Ruiz-Trillo et al. 1999). Au lieu d'éliminer les sites d'après la catégorie gamma à laquelle ils appartiennent, ils sont plutôt éliminés d'après la moyenne des taux pondérés sur toutes les catégories avec une pondération donnée par la probabilité *a posteriori* de chaque catégorie (Rodriguez-Ezpeleta et al. 2007). Peu

importe la méthode utilisée, ce sont des méthodes approximatives qui permettent d'éliminer une partie importante du bruit en éliminant les sites qui sont le plus aptes à contenir de l'homoplasie en se concentrant sur les positions qui contiennent le plus d'informations évolutives pour la reconstruction phylogénétique (Loughran et al. 2008).

Pour résumer ce chapitre, trois grandes classes de méthodes ont été proposées pour combattre la saturation : (i) l'échantillonnage taxonomique, (ii) le modèle d'évolution, et (iii) le retrait des sites rapides. Cependant, l'efficacité relative de ces trois méthodes, ainsi que l'impact de la saturation sur les datations moléculaires, n'ont pas été étudiés en détail.

II. Matériels & méthodes

2.1. Alignement protéique

Un jeu de données protéiques provenant de séquences mitochondriales de mammifères est utilisé pour l'ensemble de l'analyse (Genbank : (www.ncbi.nlm.nih.gov/Genbank/)). Les mammifères ont été choisis puisqu'ils possèdent un riche registre fossile nécessaire pour les calibrations paléontologiques et le choix des gènes est basé sur le fait que les marqueurs mitochondriaux possèdent une saturation relativement importante, ce qui nous intéresse particulièrement dans les analyses actuelles puisqu'on cherche à évaluer la qualité des méthodes d'estimation du nombre de substitutions le long des branches dans un arbre phylogénétique. Toutes les espèces dont le génome mitochondrial était disponible en novembre 2006 ont été retenues, sauf celles qui étaient trop similaires (en général, différents individus de la même espèce).

Chaque espèce est représentée par la concaténation de 12 gènes mitochondriaux codant les protéines (nd1, nd2, nd3, nd4, nd4L, nd5, co1, co2, co3, cyb, atp6, atp8) (Figure 19). Le nd6 étant codé sur le brin complémentaire n'a pas été concaténé dans notre super-alignement, car il est soumis à des biais compositionnels différents.

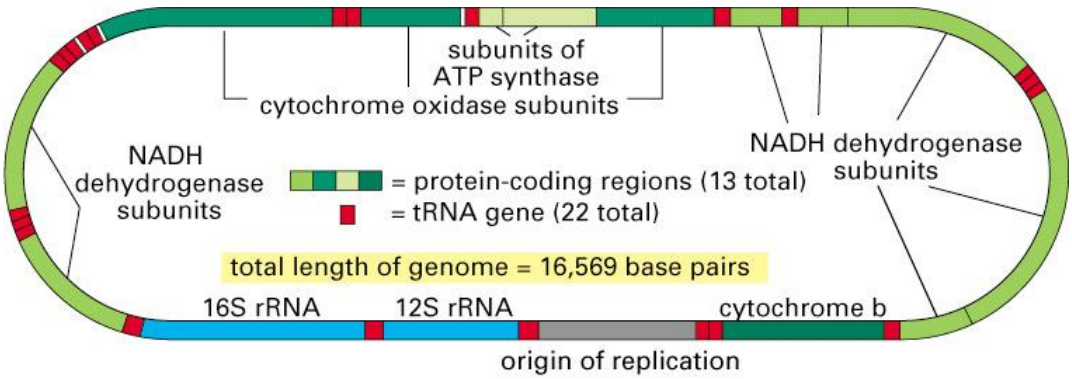

Figure 14–58. Molecular Biology of the Cell, 4th Edition.

Figure 19 : Organisation du génome mitochondrial.
Molécule d'ADN circulaire double brin de 16 569 paires de bases. Les gènes qui le composent sont 2 gènes d'ARN ribosomiques (12S et 16S), 22 gènes d'ARN de transfert nécessaires à l'expression de l'ADNmt et 13 gènes codant pour des protéines de la chaîne respiratoire (Alberts et al. 2002).

Un alignement protéique a été réalisé à l'aide de BLAST(Altschul et al. 1990), puis raffiné manuellement dans notre laboratoire par Hervé Philippe. Après élimination des régions mal alignées grâce à Gblocks (Castresana 2000), on obtient un alignement de 3540 acides aminés.

Une inférence phylogénétique par maximum de vraisemblance a été réalisée avec Treefinder (Jobb et al. 2004) (modèle MtREV+Γ_4), puisque ce dernier est un programme qui implémente des outils statistiques puissants pour inférer des arbres phylogénétiques et est facile d'utilisation. Comme nous voulions avoir un arbre de référence, la topologie obtenue a été raffinée par Emmanuel Douzery et Pierre-Henry Fabre (Institut des Sciences de l'Évolution de Montpellier) pour la rendre compatible avec les résultats les plus fiables obtenus à partir des marqueurs nucléaires et de la morphologie (Figure 20).

Nous avons choisi de travailler à topologie fixe, car nous cherchons à évaluer l'impact de différents modèles d'évolution des séquences ainsi que l'effet de l'échantillonnage taxonomique sur l'estimation des longueurs de branche (LB).

Figure 20 : Topologie de l'arbre phylogénétique à 196 espèces.

2.1.1. Alignement nucléotidique

Les séquences nucléotidiques correspondant au jeu protéique ont également été récupérées (GenBank:www.ncbi.nlm.nih.gov/Genbank/) pour les 196 espèces. L'alignement des nucléotides a été dicté par l'alignement des acides aminés. Ce jeu de données a été partitionné en trois selon la position du nucléotide dans le codon (NT1, NT2 et NT3). Ces trois alignements ont été considérés séparément dans l'analyse.
Les mêmes 196 espèces, les mêmes codons ainsi que la même topologie de l'arbre considérés pour les analyses protéiques ont été utilisés pour les analyses nucléotidiques.

2.2. Comparaison des arbres inférés lorsque l'échantillonnage taxonomique est différent

Sur quoi peut-on se baser pour comparer deux arbres phylogénétiques contenant des espèces différentes? Nous avons abordé le problème en choisissant d'avoir des espèces communes à tous les arbres inférés peu importe le nombre total d'espèces ou l'échantillonnage taxonomique. Cela permet de comparer les longueurs de branche des sous-arbres correspondant aux espèces en commun et cela en gardant toujours une topologie fixe de départ et en n'inférant que les longueurs de branche sur les arbres phylogénétiques.
Quatre espèces de référence (*Homo sapiens, Mus musculus, Dugong dugon, Zalophus californianus)* parmi les 196 ont été choisies puisqu'elles font partie des clades majeurs de Placentaires (Euarchontes, Glires, Afrotheria et Laurasiatheria) ainsi qu'une espèce faisant partie de l'outgroup (Metatheria) soit *Macropus robustus*. Ces groupes sont bien échantillonnés, ce qui permet de mieux étudier l'impact de l'échantillonnage taxonomique. Les Xenarthra, quatrième clade majeur des Placentaires n'ont pas été retenus ici, car seulement quatre génomes mitochondriaux sont disponibles.

Pour permettre la comparaison d'arbres contenant un nombre d'espèces et un échantillonnage taxonomique différents, nous avons mesuré la longueur du sous-arbre à 5 espèces provenant des différents arbres obtenus (Figure 21). Nous effectuons donc la somme de chacune des 7 branches du sous-arbre (Équation 3).

$$LB_5^n = \sum_{1}^{i=7} LB_i^n$$

Équation 3 : Formule générale pour la longueur totale du sous-arbre à 5 espèces de référence quand n espèces sont considérées.

Exemple de nomenclature utilisée : LB_5^{75} est la longueur du sous-arbre à 5 espèces à partir d'un arbre inféré avec 75 espèces, ce qui correspond à la somme des LB_i de 1 à 7 calculées avec 75 espèces.
$LB_5^{75} = \sum_{1}^{i=7} LB_i^{75}$

Le sous-arbre correspond au chemin qui relie les espèces de référence entre elles, donc peu importe les espèces présentes et leur nombre on peut mesurer la longueur de chacune des branches de la racine de l'arbre à chacune des branches externes correspondant à nos espèces de référence (Figure 22). Chacune des 7 branches (en considérant les deux branches internes) de l'arbre réduit peut également être comparée séparément.

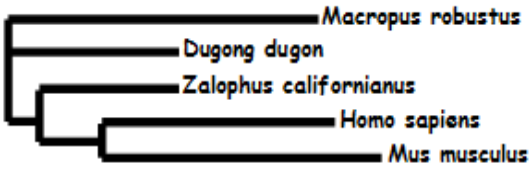

Figure 21 : Arbre à 5 espèces de référence.

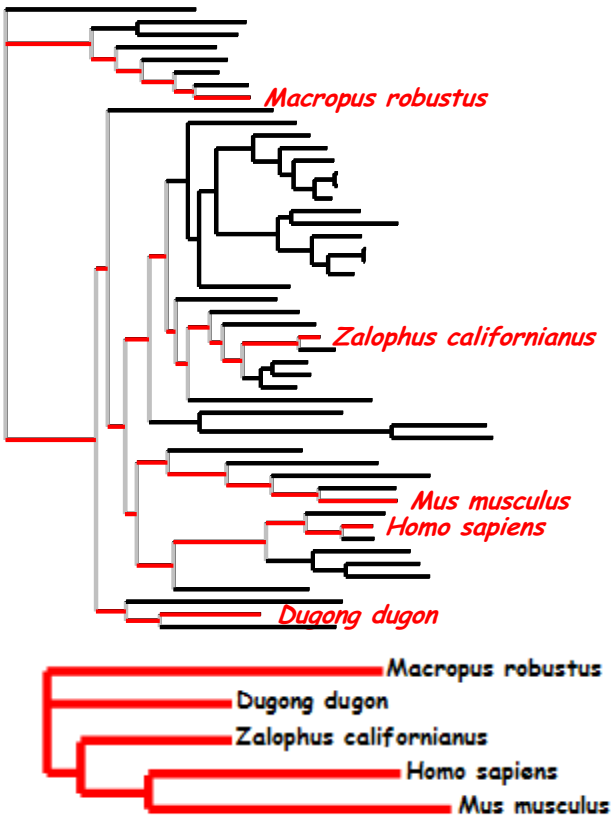

Figure 22: Extraction du sous-arbre à 5 espèces de référence d'un arbre contenant un échantillonnage de 50 espèces. L'arbre représente un arbre avec un échantillonnage au hasard; Les branches d'intérêt du sous-arbre sont en rouge et permettent l'extraction du sous-arbre à 5 espèces (*Macropus robustus; Dugong dugon, Zalophus californianus; Homo sapiens; Mus musculus*) en bas de la figure.

Pour chacun des arbres obtenus (échantillonnage et nombre d'espèces variables, modèle d'inférence phylogénétique variable), le sous-arbre à 5 espèces est extrait avec le programme de Statistique R 2.4.1 (R Development Core Team 2004) en utilisant le package spécialisé pour la phylogénie « APE » (Paradis et al. 2004) qui permet de calculer la somme des longueurs de branche. Nous avons choisi cet outil puissant pour nos analyses, parce que non seulement il intègre plusieurs options de statistiques nécessaires pour l'obtention de nos résultats, mais il permet également de faire de la programmation en utilisant des fonctions pour l'analyse phylogénétique déjà écrites, ce qui facilite largement les analyses.

2.3. Sous-échantillonnage taxonomique

En plus des 5 espèces de référence, un programme en JAVA a été écrit qui réalise un échantillonnage aléatoire parmi les 191 autres espèces de mammifères afin d'obtenir un nombre d'espèces croissant, totalisant soit 25, 50, 75, 100, ou 150 espèces. Pour chaque sous-échantillon de séquences on effectue:

1. Échantillonnage aléatoire de N espèces parmi les 191 espèces.
2. Réduction de l'arbre de départ à 196 espèces (topologie sans longueur de branche) aux N+5 espèces présentes dans le nouvel alignement (programme Puz2ceb (Bapteste et al. 2002)).
3. Répéter l'étape 1 et 2 pour obtenir 100 sous-échantillons à N+5 espèces.

On obtient 502 arbres (sans longueur de branche) et 502 alignements différents contenant un nombre croissant d'espèces (Tableau 1).

Nombre d'espèces	5	25	50	75	100	150	196
Nombre de jeu de données	1	100	100	100	100	100	1

Tableau 1 : Nombre d'espèces dans les échantillons aléatoires avec le nombre de jeux de données correspondant.

2.4. Modèles d'évolution des séquences utilisés pour l'évaluation des longueurs de branche

2.4.1. Séquences protéiques

On veut étudier l'impact du modèle et son interaction avec l'échantillonnage taxonomique sur les valeurs $\sum LB_5^n$. Nous avons utilisé différents modèles d'évolution des séquences qui sont consignés dans le Tableau 2 :

Modèles	Parcimonie non pondérée (Fitch 1971)	Poisson (Felsenstein 1981)	Wag (Whelan et al. 2001)	MtREV (Adachi et al. 1996)	GTR (Tavare 1986)	CAT (Lartillot et al. 2004)	CAT-GTR (Lartillot et al. 2004)
Programme	Paup(Swofford 2000)	Tree-puzzle 5.2 & Phylobayes 2.2	Tree-puzzle 5.2 (Schmidt et al. 2002)	Tree-puzzle & Phylobayes 2.2	Phylobayes 2.2(Lartillot et al. 2004)	Phylobayes 2.2	Phylobayes 2.2

Tableau 2 : Modèle d'évolution de séquences pour les séquences protéiques.

Mais on veut également savoir si les résultats en ML sont différents des résultats obtenus par inférence bayésienne, donc on va aussi les comparer entre eux quand les modèles le permettent.

2.4.2. Séquences nucléotidiques

Des matrices de substitutions nucléotidiques ont aussi été utilisées pour inférer les longueurs de branche sur nos arbres. Les différents modèles sont consignés dans le Tableau 3:

Modèles	Parcimonie non pondérée (Fitch 1971)	JC (Jukes et al. 1969)	HKY (Hasegawa et al. 1985)	GTR (Tavare 1986)
Programme	Paup	Tree-Puzzle 5.2	Tree-Puzzle 5.2	Tree-Puzzle 5.2 & Paup

Tableau 3 : Modèle d'évolution de séquences pour les séquences nucléotidiques.

2.4.3. Distribution Gamma

Pour prendre en compte l'hétérogénéité de vitesse entre sites, on a utilisé une distribution gamma avec tous les modèles décrits précédemment. Le taux continu de la distribution gamma est approximé avec une distribution discrète où les sites sont divisés en k catégories de taux avec une probabilité égale : 4 catégories gamma ont été utilisées avec phylobayes, et 8 catégories avec Tree-Puzzle.

2.4.4. Sites invariants

La contribution des sites invariants dans ce jeu de données a été examinée. Les sites invariants sont les positions dans l'alignement qui ne sont pas libres de varier entre des séquences. L'effet des modèles a donc été testé en prenant en compte les sites invariants avec le modèle Wag en plus de la modélisation de l'hétérogénéité de taux par la distribution gamma (Γ). Les résultats obtenus avec (Wag+I+Γ_8) pour les LB_5^n sont très comparables avec les résultats de LB_5^n avec le modèle « Wag+Γ_8 ». Nous n'avons donc pas pris en compte cette option de proportion de sites invariants pour les autres modèles utilisés dans notre analyse (Lanave et al. 1984);(Yang 1993); (Swofford et al. 1996).

2.5. Évaluation de l'ajustement des modèles aux données

Afin de connaître l'ajustement des modèles aux données, nous avons considéré une log-vraisemblance pénalisée pour ML, l'Akaike Information Criterion(AIC) (Akaike 1973) qui permet de comparer différents modèles entre eux en prenant en compte le nombre de paramètres de chaque modèle. Plus la valeur de l'AIC est petite, meilleur est l'ajustement du modèle aux données (Équation 4).

$$AIC = 2k - 2\ln(L)$$

Équation 4 : Critère AIC où *k* est le nombre de paramètres dans le modèle et L la vraisemblance maximisée (la vraisemblance estimée à partir de l'alignement étant donnés l'arbre et le modèle).

Avec ce critère, la déviance du modèle (-2ln(L)) est pénalisée par 2 fois le nombre de paramètres. L'*AIC* représente donc un compromis entre le biais (qui diminue avec le nombre de paramètres) et la parcimonie (nécessité de décrire les données avec le plus petit nombre de paramètres possible).

Quand le nombre de paramètres *k* est grand par rapport au nombre d'observations K, c'est-à-dire si K/*k* < 40, il est recommandé d'utiliser l'AIC corrigé (Équation 5)(Hurvich et al. 1995):

$$AIC_C = AIC + \frac{2k(k+1)}{K-k-1}$$

Équation 5 : Critère AIC corrigé.

En inférence bayésienne, il existe beaucoup plus de modèles intéressants à tester. Il aurait fallu utiliser le facteur de Bayes (Kass et al. 1995) ou la cross-validation pour évaluer l'ajustement de ses modèles à nos données, mais cela devient très vite couteux en temps de calcul. Nous nous sommes donc basés sur les données publiées (Lartillot et al. 2006; Lartillot et al. 2008) qui suggèrent cet ordre pour l'ajustement des modèles : Poisson << Wag/MtREV << GTR << CAT << CAT-GTR.

2.6. Méthodes de calcul des longueurs de branche

2.6.1. Maximum de parcimonie

Les reconstructions par maximum de parcimonie ont été réalisées avec le logiciel PAUP (Swofford 2000) et deux différentes stratégies d'optimisation ont été utilisées soit ACCTRAN et DELTRAN.

Les valeurs des BL pour la parcimonie sont divisées par le nombre de sites (3540) pour obtenir le nombre de substitutions par site (valeur typique donnée par les programmes ayant une approche probabiliste).

2.6.2. Maximum de vraisemblance

Puisqu'on travaille à topologie fixe, la méthode de ML va juste chercher les valeurs des paramètres (par exemple les longueurs de branche) qui maximisent le log L (le log de la vraisemblance). Les applications standards de la prior de la distribution gamma nécessite d'avoir une moyenne β de la prior égale à 1, sinon il faut réajuster les longueurs des branches de l'arbre originale (Mayrose et al. 2005).

Nous avons utilisé Tree-Puzzle (Schmidt et al. 2002) et Paup(Swofford 2000) pour nos inférences en ML. Nous avons alors remarqué qu'un problème avec la moyenne de la distribution gamma existait, bien que la valeur moyenne des taux relatifs pour toutes les catégories gamma devrait être contrainte à être égale à 1 (cette valeur représente la moyenne des taux relatifs entre sites *a priori*, mais n'implique pas automatiquement que la moyenne *a posteriori* soit égale à 1). En effet après vérification dans nos analyses, la valeur moyenne des taux relatifs est inférieure à 1 (Exemple Tableau 4). En conséquence, les longueurs de branche sont surestimées pour compenser cette valeur trop faible de la moyenne des taux relatifs. On a donc du calculer à partir de tous les fichiers de sortie de Tree-Puzzle la valeur de la moyenne de la distribution gamma pour ajuster les valeurs des longueurs de branche en les multipliant par la nouvelle moyenne calculée.

On calcule la valeur moyenne des taux, c'est-à-dire la moyenne ponderée (Équation 6):

$$\bar{r} = \frac{\sum_{1}^{n} TR_i \times NS_i}{\sum_{1}^{n} NS_i}$$

Équation 6 : Moyenne des taux relatifs, $\bar{r}$ La moyenne des taux relatifs, TR_i le taux relatif pour la catégorie i et NS_i le nombre de sites de la catégorie i, et n le nombre total de catégories.

Exemple

MtREV+ Γ_8 à 196 espèces (Tree-Puzzle)

Catégories	Taux relatifs	Nombre de sites	TR x NS	Distribution gamma
1	0.0001	1082	0.1082	
2	0.0073	0	0	
3	0.0453	297	13.4541	
4	0.1528	412	62.9636	
5	0.3897	476	185.4972	
6	0.8669	495	429.1155	
7	1.8605	556	1034.438	
8	4.6773	222	1038.3606	
TOTAL		3540	2763.9272	**0.7808**

Tableau 4 : Exemple de correction pour la Gamma pour le modèle MtREV+Γ_8 avec Tree-Puzzle.

On ajuste les valeurs des longueurs de branche en les multipliant par la moyenne de la distribution gamma calculée, soit : LBcorrigé = LBinferé * **0.7808**

2.6.3. Inférence bayésienne

La méthode des "Chaînes de Markov Monte Carlo" (MCMC) échantillonne des distributions de probabilité inconnues *a priori*. Le programme PhyloBayes (Lartillot et al. 2004) implémente cette technique en échantillonnant la distribution de probabilité *a posteriori* induite sur les paramètres d'un modèle d'évolution à partir des alignements de séquences. Comparée à d'autres méthodes d'échantillonnage MCMC (e.g MrBayes (Huelsenbeck et al. 2001)), la particularité de Phylobayes est que celui-ci utilise le modèle CAT qui permet de prendre en compte l'hétérogénéité entre les sites des processus de remplacement entre les acides aminés.

Nous avons vérifié visuellement la convergence des différentes chaînes générées par MCMC avec Phylobayes avant de les arrêter, basé sur le critère de la convergence du log-L principalement et du paramètre alpha (lorsqu'une distribution gamma est considérée). Nous avons également lancé à quelques reprises des chaines indépendantes (sur les mêmes données avec le même modèle) ainsi que des chaines plus longues pour confirmer que les chaînes étaient bien convergées. Le nombre de cycles ainsi que le temps de convergence sont différents dépendamment du modèle considéré et du nombre d'espèces. Les premiers points des chaînes ont été éliminés de l'analyse grâce a l'option « burnin » afin de ne garder pour notre analyse que la partie stationnaire des chaînes.

2.7. Analyses statistiques

En maximum de vraisemblance (Tree-Puzzle), lorsque les 100 échantillonnages aléatoires d'espèces étaient faits (pour 25, 50, 75, 100 et 150 espèces), la moyenne ainsi que l'écart-type ont été calculés grâce au package APE du programme de Statistique R 2.4.1 (R Development Core Team 2004).
En inférence bayésienne, on peut calculer l'écart-type inter-échantillon (pour les 100 échantillons aléatoires d'après la formule précédente) ainsi que l'écart-type intra-échantillon. Pour l'écart-type intra-échantillon, il est possible de récupérer directement les différents arbres de la distribution *a posteriori* du programme Phylobayes, pour ainsi mesurer la moyenne et l'écart-type de LB_5^n.

Il est plus difficile de calculer l'écart-type intra-échantillon pour le maximum de vraisemblance. Cette approche surestime la variance, car les branches ne sont pas indépendantes; nous n'avons donc calculé que l'écart-type pour n=5 et 196. Le programme Tree-Puzzle fournit l'écart-type pour les différentes branches de l'arbre. Nous calculons l'écart-type de la façon suivante:

$$\sigma_{LB_5^n} = \sqrt{\sum_{i=1}^{c} \sigma_{Li}^{2}}$$

Équation 7 : Èart-type intra-échantillon où *c* est le nombre de branches.

2.8. Asymétrie de la distance de la racine aux feuilles

La saturation mutationnelle (Section 1.6.2) va surtout raccourcir les branches longues et donc réduire le coefficient d'asymétrie de la distribution de la distance de la racine aux feuilles. Ce coefficient mesure l'asymétrie d'une probabilité de distribution d'une variable aléatoire réelle par rapport à une distribution normale comme décrit dans la Figure 23:

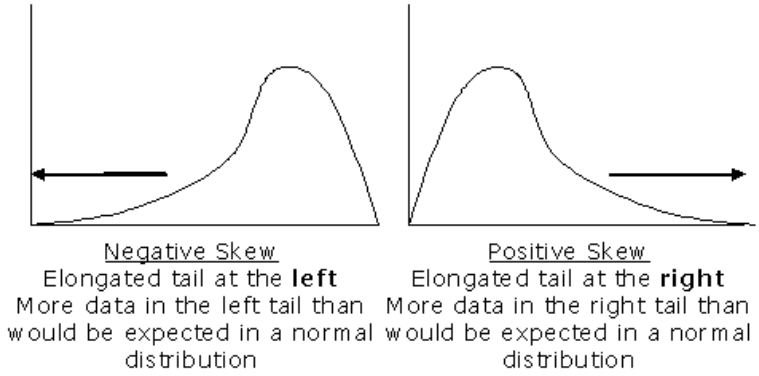

Figure 23 : Figure montrant la forme de la courbe pour un coefficient d'asymétrie positif et un coefficient d'asymétrie négatif. (http://en.wikipedia.org/wiki/Skewness)

Les coefficients d'asymétrie des différents arbres inférés avec les 196 espèces ont été comparés. Pour cette analyse nous avons enraciné les arbres avec les 3 espèces de monotrèmes présentes dans notre jeu de données. Puis, nous avons enlevé les branches correspondant à ce groupe externe. Le coefficient d'asymétrie est donc mesuré sur un arbre à 193 espèces en utilisant l'option « skew » du logiciel de statistique R 2.4.1 qui calcule le coefficient d'asymétrie (Équation 8) de la distance entre la racine de l'arbre et une des branches externes selon la formule suivante :

$$\text{Coefficient d'asymétrie} = \frac{\sum_{i=1}^{N}(Y_i - \bar{Y})^3}{(N-1)s^3}$$

Équation 8 : Coefficient d'asymétrie où Y_1, Y_2, ..., Y_N sont les longueurs de la racine aux feuilles correspondant aux N espèces, $\bar{Y}$ est la moyenne, s est l'écart-type et *N* est le nombre de points. Le coefficient d'asymétrie pour une distribution normale (données symétriques) est égal à zéro.

2.9. Analyse séparée des 12 gènes mitochondriaux

L'impact de l'augmentation du nombre d'espèces sur chacun des gènes est étudié en séparant notre alignement d'après les gènes que celui-ci contient. On retrouve donc 12 gènes qui correspondent à nd1, nd2, nd3, nd4, nd4L, nd5, co1, co2, co3, atp6, atp8. L'alignement a été séparé grâce aux numéros d'accessions correspondent au blast de *Macropus robustus* sur la base de donnée GenBank (www.ncbi.nlm.nih.gov/Genbank/) selon le Tableau 5:

Numéro d'accession	**Gènes**	**Longueurs du gène (a.a)**
NP_007404.1	ND5	570
NP_007396.1	COX1	511
NP_007403.1	ND4	457
NP_007406.1	CYTB	379
NP_007394.1	ND1	317
NP_007395.1	ND2	342
NP_007400.1	COX3	261
NP_007397.1	COX2	220
NP_007399.1	ATP6	226
NP_007401.1	ND3	110
NP_007402.1	ND4L	98
NP_007398.1	ATP8	49

Tableau 5 : Tableau présentant les numéros d'accession ainsi que les noms des 12 gènes mitochondriaux avec leur longueur correspondante en nombre d'acides aminés.

De façon logique, on s'attend à ce que les gènes qui évoluent lentement soient peu saturés (vont subir moins de substitution multiples). Avec le programme Tree-Puzzle, nous avons inféré les arbres phylogénétiques avec le modèle MtREV+Γ_8, respectivement pour 5 et 196 espèces, afin d'établir la corrélation entre le taux d'évolution (correspondant au LB du sous-arbre à 196 espèces (LB_5^{196})) et le niveau de saturation des gènes ($\frac{LB_{196}^5}{LB_5^5}$). Le niveau de saturation est mesuré en calculant le ratio des LB de 196 espèces sur les LB du sous-arbre à 5 espèces, ce qui permet de mesurer le nombre de substitutions supplémentaires qui sont détectées à 196 espèces et non pas à 5 espèces.

Cette corrélation a également été calculée avec le programme Phylobayes en utilisant le meilleur modèle disponible CAT-GTR+Γ_4.

2.10. Simulation de séquences protéiques

Nous avons simulé des séquences à partir de notre arbre phylogénétique de départ et reproduit la méthodologie décrite précédemment pour comparer les résultats obtenus et ainsi avoir une image plus complète de la performance des différents modèles.

Les alignements de séquences simulées ont des propriétés différentes et un signal phylogénétique plus fort que les données réelles (Brinkmann et al. 2005). Typiquement plus d'effort computationnel est requis pour trouver la "bonne" phylogénie pour les données réelles (Alexandros 2005).

On voulait donc tester l'hypothèse selon laquelle il y aurait moins de sous-estimation des longueurs de branche pour des séquences simulées que pour les données réelles. Nous avons donc utilisé le même protocole que précédemment, et nous avons procédé à la comparaison des LB_5^n d'arbres inférés à partir des séquences simulées et des séquences réelles

Méthodologie :

À partir de l'arbre à 196 espèces inféré grâce au modèle CAT avec hétérogénéité des taux entre les sites (Γ_4), 100 simulations de séquences ont été obtenues grâce à l'option « PPRED » qui permet d'effectuer des analyses postérieures prédictives, c'est-à-dire que, pour 100 points pris régulièrement dans la chaîne de Markov après le « burn-in » (le burn-in est une option qui permet d'éliminer les premiers points de la chaîne de Markov, avant sa convergence), les séquences sont simulées en prenant en compte les paramètres évolutifs de chaque position estimés par le modèle CAT+Γ_4.
Chaque ensemble de séquences simulées est utilisé pour effectuer les mêmes étapes que précédemment, soit un échantillonnage aléatoire d'espèces pour un nombre croissant de 25 à 150 espèces. Chaque échantillon réduit d'espèces (avec la topologie d'arbre correspondante) est utilisé pour inférer les longueurs de branche avec Tree-Puzzle pour les modèles Poisson, Wag et MtREV (avec/sans Γ).
Les longueurs de branche obtenues avec les séquences simulées sont alors comparées, à modèle d'inférence et nombre d'espèces égaux, avec celles obtenues à partir des données réelles.

2.11. Retrait de sites

Un site rapide peut s'ajuster aussi bien au modèle qu'un site lent. Le problème de l'ajustement au modèle réside dans le fait qu'un site rapide a besoin d'être plus « corrigé » par le modèle qu'un site lent (plus de substitutions sont prédites). Nous avons donc regardé l'effet du retrait des sites rapides sur l'inférence des longueurs de branche de l'arbre phylogénétique obtenu pour l'ensemble des espèces. Il n'est possible d'étudier que le coefficient d'asymétrie de l'arbre inféré, puisque les longueurs de branche inférées ne peuvent plus être comparées.
Le programme PAML (Yang 1997) (modèle Wag+Γ_8 (Whelan et al. 2001)) a été utilisé afin d'estimer pour chaque site de l'alignement le taux d'évolution en assumant une distribution gamma avec 8 catégories. Les sites évoluant les plus rapidement ont été

éliminés de l'analyse par pas de 250 sites par ordre décroissant de taux, pour un total de 2500 sites éliminés.
Les alignements des sites restants ont été utilisés pour inférer les nouvelles longueurs de branche avec le programme Tree-Puzzle (modèle MtREV+ Γ_8).

2.12. Datation

La datation moléculaire permet d'estimer l'âge de divergence entre espèces en prenant en compte le nombre de substitutions qui ont eu lieu (voir section 1.3). Si une méthode ne permet pas d'estimer de façon efficace les longueurs de branche de l'arbre phylogénétique en question, cela a pour conséquence de sous-estimer les dates des nœuds de l'arbre.

2.12.1. Jeux de données

Nous avons utilisé le même jeu de données mitochondriales de 196 espèces de mammifères pour estimer l'âge de divergence entre espèces. Par contre, nous avons adapté notre jeu de données de façon à permettre des points de calibration différents. Nous avons choisi pour chaque jeu de donnée des espèces de référence qui permettent d'y intégrer la calibration correspondante. Ainsi nous avons trois jeux de données qui permettent d'utiliser quatre calibrations. Nous avons alors calibré sur des nœuds récents correspondant à la divergence entre l'homme et les chimpanzés ainsi que la divergence entre le rat et la souris. Pour les calibrations plus anciennes, nous avons choisi de calibrer sur des nœuds correspondant aux carnivores (*Felis catus*/*Zalophus californianus)* et aux périssodactyles (*Equus caballus*/*Rhinoceros unicornis*).

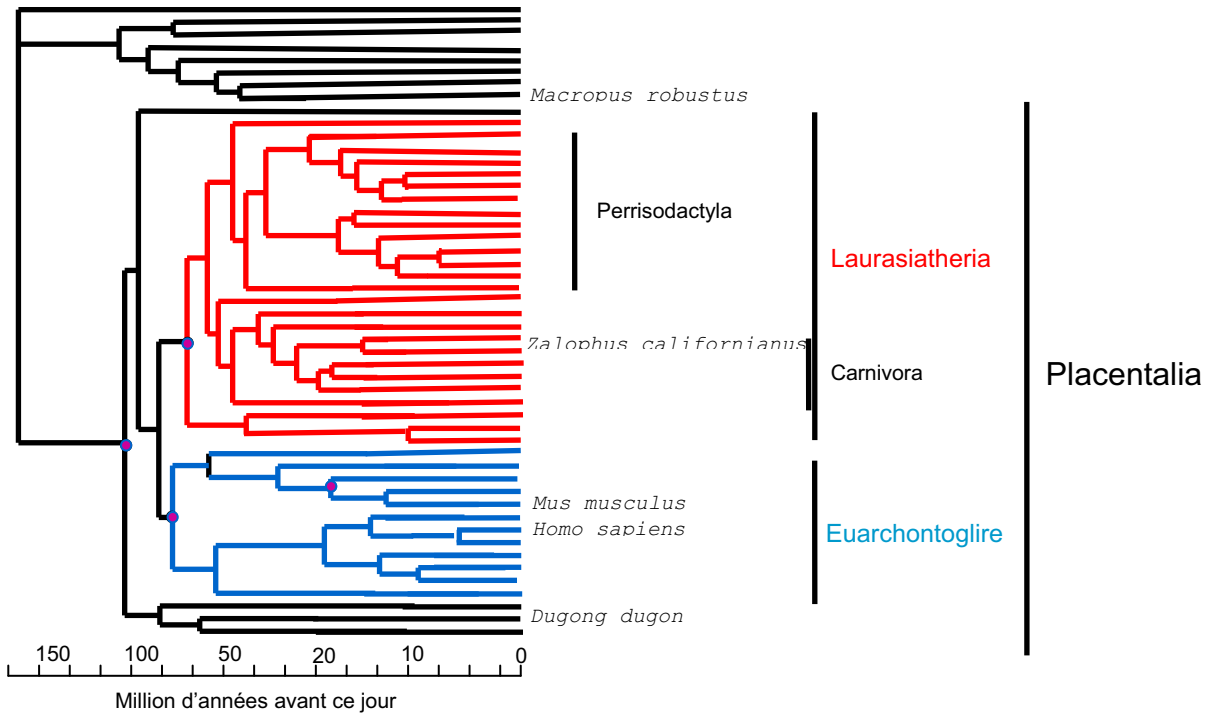

Figure 24 : Position des 4 nœuds (rond mauve) qu'on cherche à dater; soit un nœud à la base des Laurasiatheria, à la base des Euarontoglires, des placentaires et un nœud à la base de *Mus musculus*/*Rattus norvegicus*. Échelle de temps en M.A. d'après (Springer et al. 2003).

Le Tableau 6 et la Figure 24 montrent les différents jeux de données utilisées. Puisque nous regardons l'âge estimé du nœud *Mus musculus/ Rattus norvegicus*, ces deux espèces sont toujours présentes dans les trois jeux de données. Les espèces qui servent à la calibration sont obligatoirement présentes dans le groupe des espèces de référence.

Pour les points de calibration *Homo sapiens\Pan paniscus* et *Mus musculus\Rattus norgevicus*, le même jeu de données est utilisé, puisque celui-ci permet l'intégration de ces deux points de calibration et permet la datation des nœuds recherchés (Figure 25)

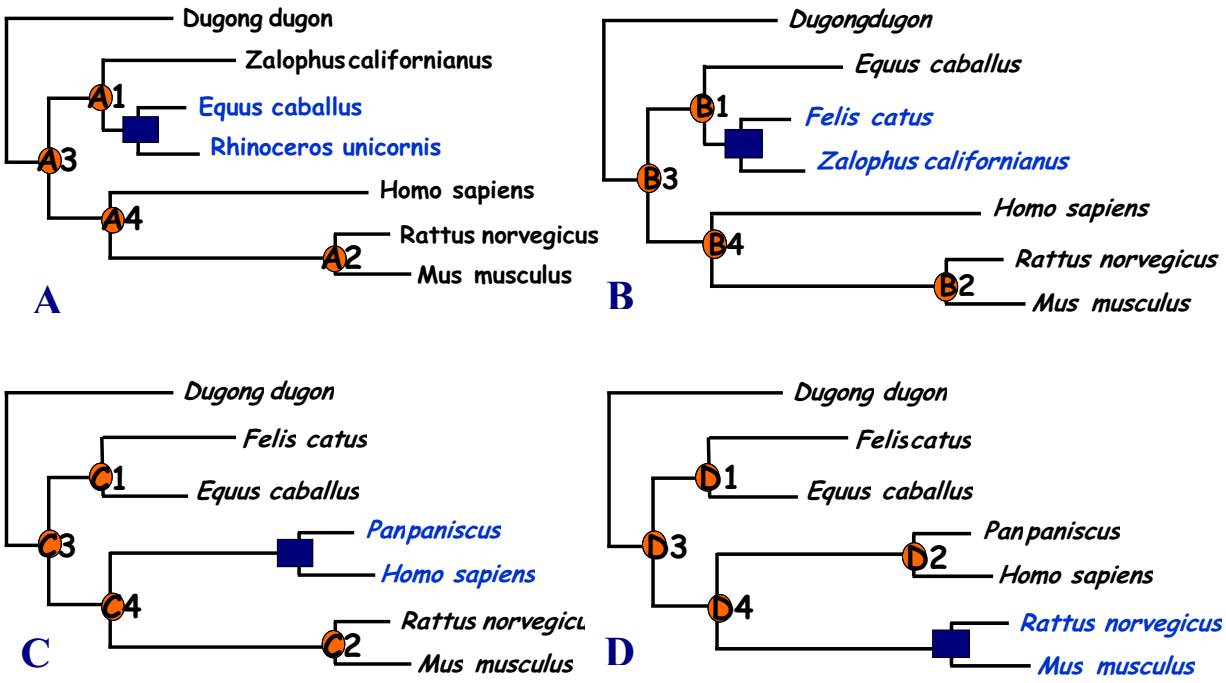

Figure 25 : Arbres phylogénétiques représentant les espèces de référence, les intervalles de calibrations (en bleu) ainsi que les différents nœuds à dater pour les différents jeux de données.

Les quatre nœuds dont on cherche l'estimation de l'âge sont les suivants:

Laurasia; Placentalia; Euarchontoglire; *Mus musculus/ Rattus Norvegicus* ou *Homo sapiens\Pan paniscus*

Jeu de données	**A**	**B**	**C**	**D**
Point de Calibration	*Equus caballus/ Rhinoceros unicornis*	*Felis catus/ Zalophus californianus*	*Homo sapiens/ Pan paniscus*	*Mus musculus /Rattus norvegicus*
Age du nœud (en M.A.)	54 à 58	50 à 63	5 à 7	7 à 16
Espèces de références*	*Zalophus californianus;Rhinoceros unicornis;*	*Felis catus; Zalophus californianus;*	*Felis catus; Pan paniscus*	*Felis catus; Pan paniscus*
Nœuds dont l'âge est estimé	1.Laurasia 2.Placentalia 3.Euarchontoglire 4.*Mus musculus/ Rattus norvegicus*	1.Laurasia 2.Placentalia 3.Euarchontoglire 4.*Mus musculus/ Rattus norvegicus*	1.Laurasia 2.Placentalia 3.Euarchontoglire 4.*Mus musculus/ Rattus norvegicus*	1.Laurasia 2.Placentalia 3.Euarchontoglire 4.*Homo sapiens/ Pan paniscus*

Tableau 6 : Jeux de données utilisés pour la datation moléculaire.
*Les espèces de références suivantes sont présentes dans tous les jeux de données : *Dugong dugon*; Macropus robustus ;*Equus caballus; Homo sapiens; Rattus norvegicus; Mus musculus*.

Nous avons étudié l'impact de l'échantillonnage taxonomique de la même façon que précédemment. Un échantillonnage aléatoire est donc effectué afin d'obtenir un nombre d'espèces croissant, totalisant soit 25, 50, 75. Pour chaque sous-échantillon de séquences :

1. Sélection aléatoire de 17, 42 ou 67 espèces
2. Réduction de l'arbre de départ à 196 espèces (topologie sans longueur de branche) aux espèces présentes dans l'alignement (programme Puz2ceb).
3. Répéter l'étape 1 et 2 pour obtenir 100 sous-échantillons.
4. Répéter les étapes 1 à 3 pour les 4 types de calibrations.

On obtient donc 302 arbres (sans longueur de branche) et 302 alignements différents contenant un nombre croissant d'espèces et cela pour chacun des points de calibrations, soit un total de 904 jeux de données (Tableau 7):

Nombre d'espèces	8	25	50	75	196
Nombre de jeu de données	3	300	300	300	1

Tableau 7 : Nombre d'espèces dans les échantillons aléatoires avec le nombre de jeux de données correspondant.

2.12.2. Multidistribute (http://statgen.ncsu.edu/thorne/multidivtime.html)

Le paquet Multidistribute emploie une méthode bayésienne de datation moléculaire qui utilise un arbre phylogénétique optimal conjointement avec l'estimation par maximum de vraisemblance des longueurs de branche et de la matrice de variance-covariance correspondante. Ces valeurs sont utilisées dans une analyse bayésienne MCMC pour approximer la distribution *a postériori* des temps de divergence et des taux de substitutions (Thorne et al. 1998) (Thorne et al. 2002) (Kishino et al. 2001).

2.12.2.1. Estbranches

Estbranches est un programme qui estime les longueurs de branche d'un arbre enraciné ainsi que la matrice de variance-covariance en approximant la surface de vraisemblance avec une distribution normale multivariée centrée sur l'estimation du maximum de vraisemblance des longueurs de branche. L'arbre optimal est celui qui a maximisé la vraisemblance de ses longueurs de branche (Thorne et al. 1998).
Nous avons enraciné nos arbres avec les espèces de marsupiaux (*Macropus robustus* faisant partie des espèces de références) que chaque jeu de données contenait et avons utilisé une matrice de substitution mtMAM (Yang et al. 1998) qui est optimisée pour les mitochondries de mammifères.
Afin d'utiliser Estbranches, pour chaque analyse un nouveau fichier de contrôle avec l'arbre enraciné correspondant à l'alignement était nécessaire, il a donc fallu automatiser

tous les processus. Comme Estbranches est très lent, on n'a pas analysé tous les modèles qui étaient disponibles et on a seulement fait les analyses pour 5, 25, 50, 75 et 196 espèces.

2.12.2.2. Multidivtime

Multidivtime utilise une méthode d'autocorrélation des taux. Cette méthode est modélisée de sorte à s'ajuster à une distribution log-normale, ce qui permet aux taux d'évolutions « d'évoluer » le long des branches de l'arbre. (Thorne et al. 1998; Kishino et al. 2001) (Thorne et al. 2002).

Multidivtime va donc lire non seulement les longueurs de branche de l'arbre généré par Estbranches mais également la matrice de variance-covariance fournie par ce dernier afin de dater les différents nœuds de l'arbre phylogénétique en question.

2.12.3. Protocole

Lorsque Multidivtime est lancé une première fois, il imprime à l'écran la topologie de l'arbre avec les numéros des nœuds assignés par Estbranches. On utilise donc le numéro de nœud correspondant à notre point de calibration dans le fichier de contrôle de Multidivtime (multicntrl.dat) avec l'âge de ce nœud.

Exemple (Extrait du fichier de contrôle multicntrl.dat):

number of constraints on node times
L 27 5 /* 01 Homo_Pan */
U 27 7 /* 02 Homo_Pan */

L (Lower bound) : borne inférieure de calibration à 5 M.A. pour le nœud 27 (*Homo sapiens/Pan paniscus*)
U (Upper bound) : borne supérieure de calibration à 7 M.A. pour le noeud 27 (*Homo sapiens/Pan paniscus*)

Les différentes étapes sont donc :

1. Enraciner notre arbre sur les marsupiaux présents dans notre jeu de données.
2. Utiliser Estbranches pour produire un arbre avec des longueurs de branche ainsi qu'une matrice de variance covariance.
3. Trouver le nœud correspondant à notre point de calibration et lui imposer une contrainte d'âge inférieur et supérieur.
4. Lancer Multidivtime avec les paramètres correspondant au jeu de données.
5. Lire l'âge estimé pour les 4 nœuds choisis (Laurasia, Placentalia, Eurchontoglires, *Mus musculus/Rattus norvegicus*) grâce à un programme maison qui cherche le nœud parent correspondant (à partir des espèces de référence).
6. Répéter l'opération pour les 4 points de calibration
7. Automatiser ces étapes pour nos 302 alignements de séquence (chacun ayant un fichier de contrôle différent)

Les valeurs des paramètres par défaut de Multidivtime ont été utilisées pour le MCMC, puisque ces derniers semblent être les meilleurs pour une convergence optimal de la chaine de Markov (Thorne et al. 1998) (Kishino et al. 2001). La chaine a alors été échantillonnée 1000 fois, avec 100 cycles entre chaque échantillon et un burnin de 100000 cycles avant le premier échantillon de la chaine de Markov.

La quantité d'évolution nécessaire pour calculer la moyenne et l'écart-type du prior sur le taux d'évolution à la racine est de 0,18 et 0,09 respectivement. Pour ce qui est de la valeur de la prior sur l'âge de la racine, c'est-à-dire l'intervalle de temps de la racine au temps présent, celle-ci est égale à 100 +/- 50 M.A..

III. Résultats & Discussion

À partir du jeu de données décrit dans la partie Matériels et Méthodes, nous avons comparé plusieurs méthodes de reconstruction phylogénétique ainsi que plusieurs modèles d'évolution des séquences. Plus précisément, on cherche à savoir comment l'estimation de la longueur d'un sous-arbre de cinq taxons est influencée par la méthode, le modèle et l'échantillonnage taxonomique. On se focalise sur la somme des longueurs de branche du sous-arbre à 5 espèces ($\sum LB_5^n$) provenant de phylogénies avec échantillonnage aléatoire d'un nombre croissant d'espèces, afin de comparer les différentes phylogénies à l'aide de repères constants, c'est-à-dire les 5 espèces toujours présentes.

3.1. Parcimonie

Les analyses de maximum de parcimonie (MP) ont été effectuées sur les séquences protéiques avec le programme Paup. Deux méthodes, ACCTRAN et DELTRAN, ont été utilisées pour mesurer le nombre de substitutions sur les branches de l'arbre (voir Section 1.5.3.). Elles diffèrent dans la répartition des substitutions dans l'arbre (pas dans le nombre total). La Figure 26 montre les résultats obtenus avec MP ainsi qu'avec la méthode de maximum de vraisemblance (Tree-Puzzle) pour deux modèles, soit Poisson et MtREV+Γ_8. Le modèle Poisson a été choisi car c'est le modèle le plus simple, considérant une probabilité égale de substitution entre les différents acides aminés et entre les différents sites (Felsenstein 1981). C'est le modèle le plus proche de la parcimonie, la plus grande différence étant qu'il prend en compte les longueurs des branches. Le choix du modèle MtREV+Γ_8 (Adachi et al. 1996) est justifié par le fait que c'est un modèle empirique basé sur l'observation des protéines mitochondriales, et par conséquent on s'attend à avoir une meilleure estimation avec ce dernier puisque notre jeu de données est constituée de séquences mitochondriales.

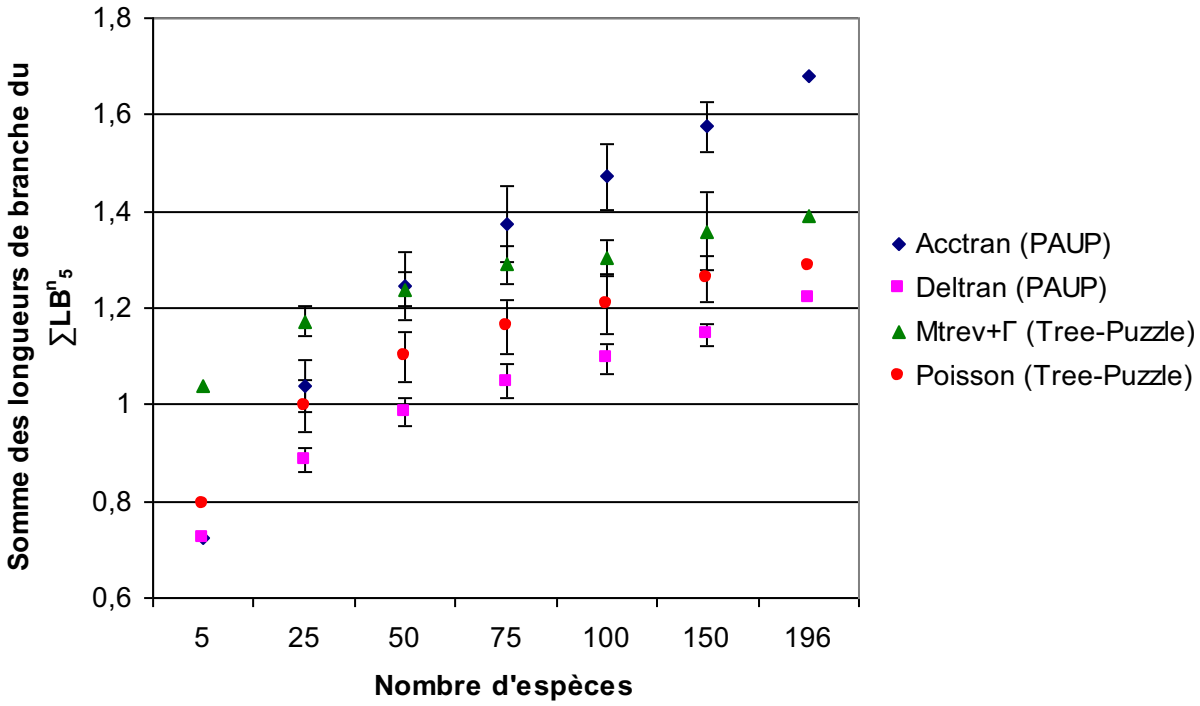

Figure 26 : Comparaison entre les longueurs de branche obtenues par parcimonie (PAUP), et par ML (PUZZLE) avec les modèles MtREV+Γ_8 et Poisson. Deux méthodes sont utilisées pour localiser les substitutions sur les branches de l'arbre, ACCTRAN (maximise les réversions) ou DELTRAN (maximise les convergences). Le graphique montre la longueur du sous-arbre à 5 espèces en fonction de l'augmentation du nombre d'espèces (de 5 à 196). Chaque point représente la moyenne de la longueur du sous-arbre avec l'écart-type (barre d'erreur) calculé sur 100 échantillons d'espèces.

Pour toutes les méthodes, la longueur du sous-arbre à 5 espèces augmente avec l'échantillonnage taxonomique. Ceci indique la présence de nombreuses substitutions multiples, que l'augmentation de l'échantillonnage permet de détecter peu à peu. Mais cela indique également que les deux méthodes sous-estiment beaucoup le nombre total de substitutions lorsque l'échantillonnage taxonomique est faible. Même à 196 espèces, l'estimation des LB ne semble pas très claire à cause de l'écart important entre les différentes méthodes. On voit également qu'il n'y a pas de plateau, ce qui suggère qu'il est encore possible de détecter des substitutions multiples le long des branches avec un ajout supplémentaire d'espèces (voir saturation section 1.3.2.2).

Avec la méthode ACCTRAN, la taille du sous-arbre à 5 espèces c'est-à-dire le nombre moyen de substitutions détectées par site passe de 0,72 pour 5 espèces

($\sum LB_5^5$) à 1,68 pour 196 espèces ($\sum LB_5^{196}$). Le rapport $\frac{\sum LB_5^{196}}{\sum LB_5^5}$ permet de calculer le gain de substitutions détectées en passant de 5 à 196 espèces; sa valeur est de 2,33 avec la méthode ACCTRAN. Par contre avec la méthode DELTRAN, ce rapport n'est que de 1,69 avec 0,72 et 1,22 substitutions détectées respectivement pour 5 et 196 espèces. Si on compare également les deux méthodes entre elles (ACCTRAN vs DELTRAN) à 196 espèces, on voit que le rapport $\frac{\sum LB_5^{196}acctran}{\sum LB_5^{196}deltran}$ est de 1,38 fois (1,68 vs 1,22). La différence d'estimation du nombre de substitutions par ACCTRAN et DELTRAN est donc notable, cette variation est due à une localisation différente des substitutions le long des branches de l'arbre.

Comme nous pouvons le voir sur le graphique, les courbes obtenues avec la méthode ML ont des valeurs inférieures à celles obtenues avec la méthode ACCTRAN et supérieures à la courbe de DELTRAN. Par ML, les courbes atteignent presque un plateau à partir de 100 espèces. Avec une approche probabiliste et un modèle simple (Poisson), on constate qu'il n'y a pas d'amélioration accrue de l'estimation des branches car toutes les branches sont courtes.

L'incertitude reliée à la détermination des substitutions par la méthode MP nous empêche d'utiliser cette méthode par la suite. Le très grand écart observé entre les deux courbes MP, même avec beaucoup d'espèces, est un signe de saturation, car cela implique la présence de nombreuses substitutions multiples que l'on peut positionner de façon différente le long de l'arbre. Cela indique que notre choix de marqueurs a bien été effectué puisqu'ils contiennent un niveau élevé de saturation, mais que la méthode MP est inadaptée. Nous allons donc abandonner la méthode MP pour se concentrer sur les résultats obtenues avec les méthodes probabilistes.

3.2. Analyse de vraisemblance

L'écart important observé avec la méthode MP entre ACCTRAN et DELTRAN implique de vérifier préalablement la valeur de la variance dans l'estimation des

longueurs de branche. Pour que les analyses qui vont suivre soient intéressantes, il faut que les différences entre modèles ou échantillonnages taxonomiques sont significatives. La Figure 27 montre l'écart-type dû à l'incertitude intrinsèque de la méthode ML (voir méthodologie section 2.7.). Elle présente la valeur moyenne de la taille des sous-arbres à 5 et 196 espèces ($\sum LB_5^5$ et $\sum LB_5^{196}$) ainsi que l'écart-type correspondant pour les différents modèles. Le plus petit écart-type observé est celui de l'arbre à 5 espèces avec le modèle Poisson avec une valeur de 0,017 tandis qu'avec le modèle MtREV+Γ_8 on observe un écart-type de 0,03. À 196 espèces, le même phénomène est obtenu, c'est-à-dire une plus grande variance pour le modèle MtREV+Γ_8 que pour le au modèle Poisson (0,10 vs 0,06). On a donc plus de variance quand le modèle est plus réaliste et qu'il contient plus de paramètres (ex. MtREV+Γ_8 vs Poisson) et lorsque les valeurs des LB sont plus grandes (effet plus ou moins mécanique).

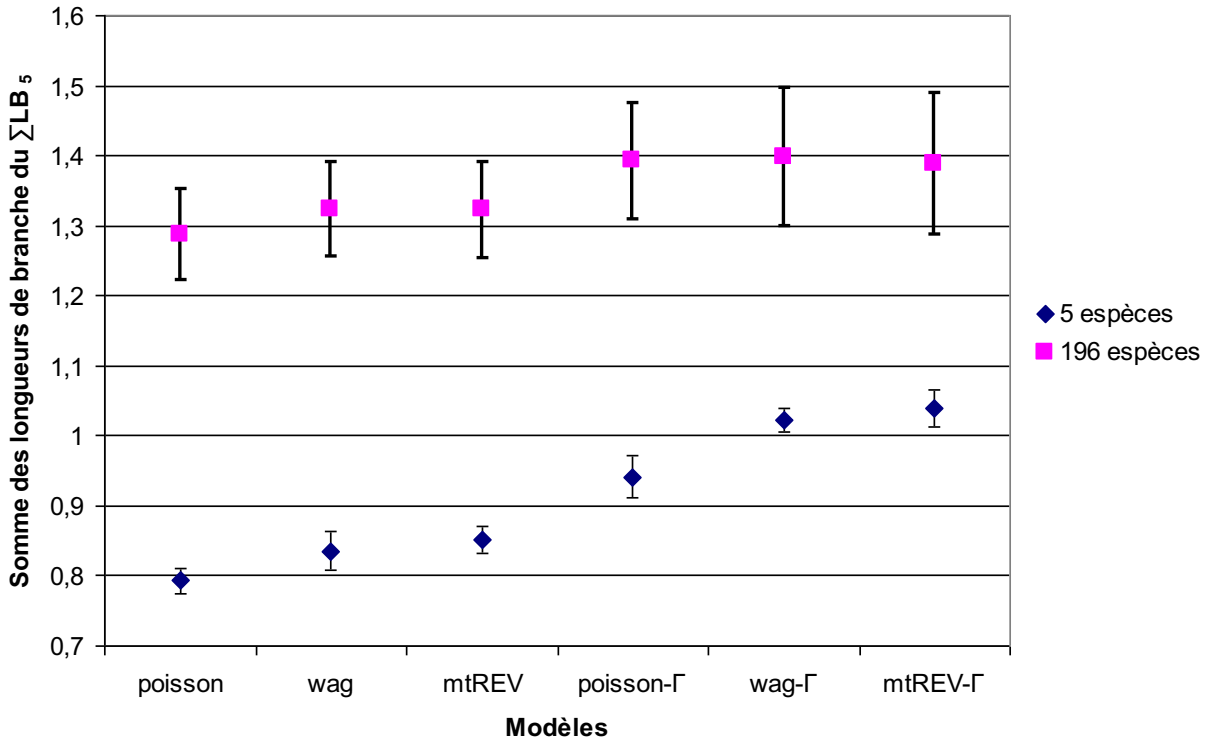

Figure 27 : Valeur de $\sum LB_5^5$ et $\sum LB_5^{196}$ pour différents modèles d'évolution de séquences avec l'écart-type (barre d'erreur). L'axe-X représente les modèles tandis qu'on retrouve sur l'axe-Y les valeurs du $\sum LB_5^n$.

La variance est relativement importante, mais elle n'empêche pas de comparer les modèles ou les échantillonnages taxonomiques. À 196 espèces, la comparaison des modèles est plus délicate, mais il est certain que l'on surestime la variance dans ce cas. En effet, tree-puzzle ne fournit que la variance pour chaque branche, et on ne peut pas calculer la variance du sous-arbre réduit, mais la somme des variances de toutes les sous-branches (cf. matériels et méthodes). En résumé, le problème posé par l'énorme variance entre ACCTRAN et DELTRAN avec la méthode MP est beaucoup moins important en ML, et ne nous empêchera pas de comparer différents modèles et différents nombres de taxons.

La comparaison des différents modèles en ML est présentée sur la Figure 28. On y voit l'effet de l'augmentation du nombre d'espèces sur la détection des substitutions multiples ainsi que l'influence du modèle choisi. Pour le modèle Poisson, on a une valeur de 0,79 substitutions par site pour 5 espèces $\sum LB_5^5$, alors que cette valeur atteint 1,29 pour 196 espèces $\sum LB_5^{196}$, ce qui donne un rapport $\frac{\sum LB_5^{196}Poisson}{\sum LB_5^5 Poisson}$ de 1,63 fois plus de substitutions par site à 196 espèces. Cela veut dire que plus de 30% des substitutions détectées avec 196 espèces ont été manquées à 5 espèces. À nombre d'espèces constant, la taille de l'arbre augmente avec la complexité du modèle. Le rapport de la somme des longueurs de branche à 196 et 5 espèces tend à diminuer avec la complexité du modèle, ainsi la moins forte augmentation est obtenue pour le modèle MtREV+Γ_8 avec un rapport égal à 1,34 (1,39/1,04). Cela indique que de meilleurs modèles comme MtREV détectent mieux les substitutions multiples à faible échantillonnage taxonomique.

Nous avons calculé l'écart-type inter-échantillon (voir section 2.7.) à partir des différents échantillons aléatoire de taxons. Comme on peut le voir, cet écart-type est assez important mais il devient de plus en plus petit au fur et à mesure que l'on augmente le nombre d'espèces dans notre échantillon. De plus, à faible échantillonnage taxonomique (25 en particulier), cette variance n'empêche pas de discriminer les modèles avec et sans distribution Γ, ce qui indique que prendre en compte l'hétérogénéité des taux entre les sites est dans ce cas particulièrement important.

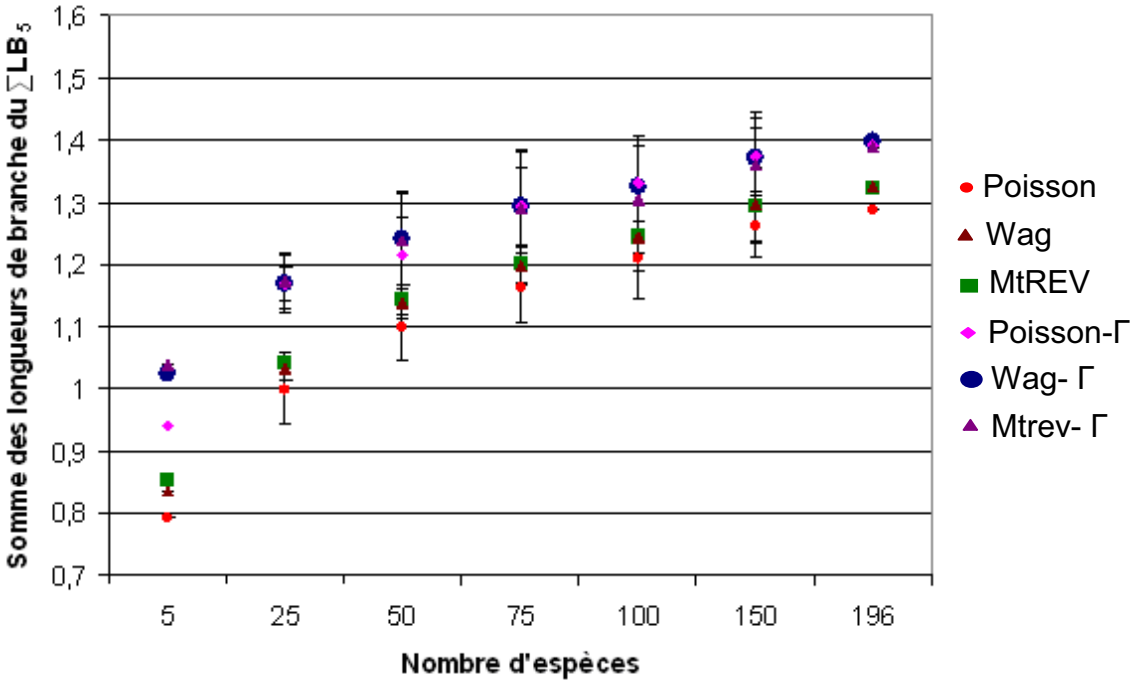

Figure 28 : Comparaison des 6 modèles (Wag, MtREV et Poisson) avec ou sans une distribution gamma à 8 catégories en ML (Tree-Puzzle). Chaque point représente la moyenne des longueurs de branche pour le sous-arbre à 5 espèces de référence avec l'écart-type (barre d'erreur) sur 100 arbres.

Entre le modèle le plus simple (Poisson) et le modèle le mieux adapté à nos données *a priori* (MtREV+Γ_8), on voit que le rapport $\frac{\sum LB_5^{196}}{\sum LB_5^5}$ de l'augmentation des substitutions détectées est nettement supérieur pour le modèle Poisson (1,63 vs 1,34). Comme l'écart entre les modèles Poisson et MtREV+Γ_8 est beaucoup plus grand à 5 espèces qu'à 196 espèces, on en déduit que l'effet de l'augmentation du nombre d'espèces est de rendre les estimations plus robustes (moins dépendantes du modèle). Étant donné leur impact sur le nombre de substitutions détectées, cela suggère qu'améliorer l'échantillonnage taxonomique est plus efficace que d'améliorer le modèle utilisé. De manière surprenante, les deux modèles empiriques (MtREV et WAG) diffèrent très peu, en particulier quand la distribution Γ est présente, alors que l'on s'attendrait à ce que MtREV, qui a été optimisé pour les données mitochondriales, soit plus performant.

Nous remarquons aussi que plus on augmente le nombre d'espèces, plus la pente de la courbe entre chaque échantillonnage successif devient plus petite. Ceci peut être d'abord expliqué par l'échantillonnage taxonomique aléatoire. En effet plus on augmente le nombre d'espèces, plus la probabilité d'avoir des espèces en commun augmente. Entre 150 et 196 espèces, la probabilité d'avoir des espèces en commun est plus grande qu'entre 25 et 50 espèces, l'échantillonnage aléatoire étant fait sur 191 espèces. Une autre justification de la diminution de cet écart-type serait que plus il y a d'espèces, plus les branches sont courtes, ce qui fait qu'elles sont moins sujettes à subir des substitutions multiples (les substitutions sont mieux résolues). Enfin, plus il y'a d'espèces, plus il y'a de branches qui ne sont pas situées sur le chemin connectant les 5 espèces de référence, donc moins il y'a de chance de briser les branches qui nous intéressent ici, et donc de détecter plus de substitutions.

On peut noter également qu'il n'y a toujours pas de plateau et qu'avec l'augmentation du nombre d'espèces, le nombre de substitutions détectées continue à augmenter. La même explication que pour MP serait valide, c'est-à-dire que cela suggère qu'il est encore possible de détecter des substitutions multiples le long des branches avec un ajout supplémentaire d'espèces.

Puisque les différents modèles utilisés ont un nombre de paramètres semblable (différant seulement avec l'ajout de Γ), le critère AIC (voir section 2.5.) permet de les comparer entre eux : plus ce critère est petit, mieux le modèle s'ajuste aux données. Pour 196 espèces, le modèle Poisson a une valeur de 560718,74, tandis que pour le modèle MtREV+Γ_8 cette valeur diminue à 435729,2 (Figure 29). Cela montre que le modèle qui s'ajuste le mieux à nos données est le modèle MtREV+Γ_8 suivi des modèles Wag+Γ_8 et MtREV ; ce résultat attendu est donc confirmé par le critère AIC. On voit aussi que même si Poisson+Γ_8 détecte plus de changements que MtREV, le critère AIC montre qu'il a un moins bon ajustement aux données. Donc non seulement le modèle MtREV+Γ_8 détecte plus de substitutions multiples à 196 espèces que d'autres modèles, mais celui-ci détient un meilleur ajustement du modèle par rapport aux données (modèle le plus vraisemblable). En résumé, les meilleurs modèles tendent à mieux détecter les substitutions multiples. Toutefois, cette relation n'est pas totalement systématique.

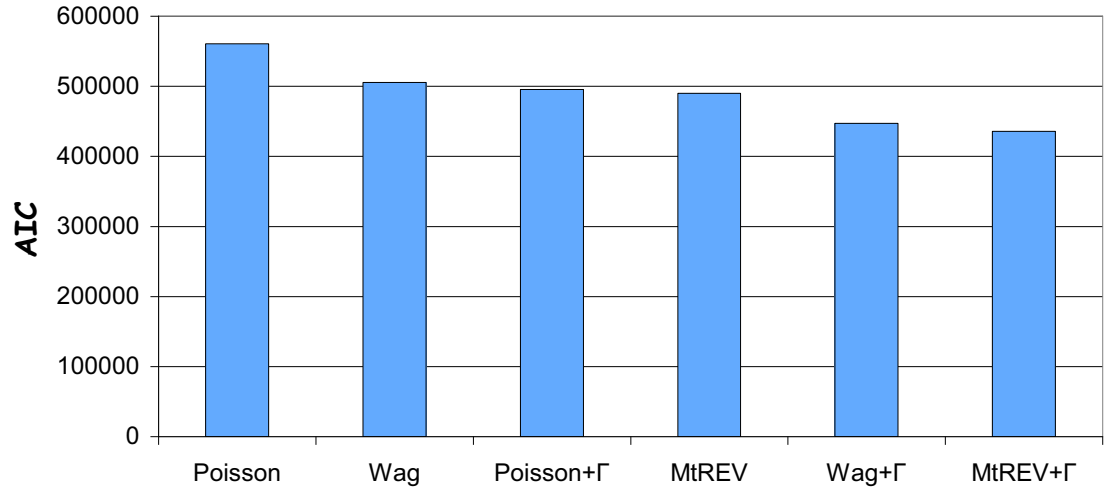

Figure 29: Le critère AIC permet la comparaison de la vraisemblance des 6 modèles d'évolution des séquences. La valeur AIC permet de corriger la vraisemblance des modèles en prenant en compte le nombre de paramètres du modèle

3.2.1. Analyse des différentes branches du sous-arbre à 5 espèces

Pour chaque branche du sous-arbre à 5 espèces, nous avons suivi son évolution avec l'augmentation du nombre d'espèces afin de s'assurer que chacune de ses branches est influencée par cette augmentation. La Figure 30A montre qu'avec le modèle MtREV+Γ_8, les valeurs des branches externes arrivant aux 5 espèces augmentent alors que les deux branches internes de l'arbre ne semblent pas varier de façon significative. Par contre, l'augmentation de la longueur varie de 5 à 196 espèces pour les différentes branches, avec un rapport $\frac{LB_5^{196}}{LB_5^5}$ compris entre 1,24 (*Homo sapiens* et *Dugong dugon*), et 1,73 (pour *Zalophus californianus*).

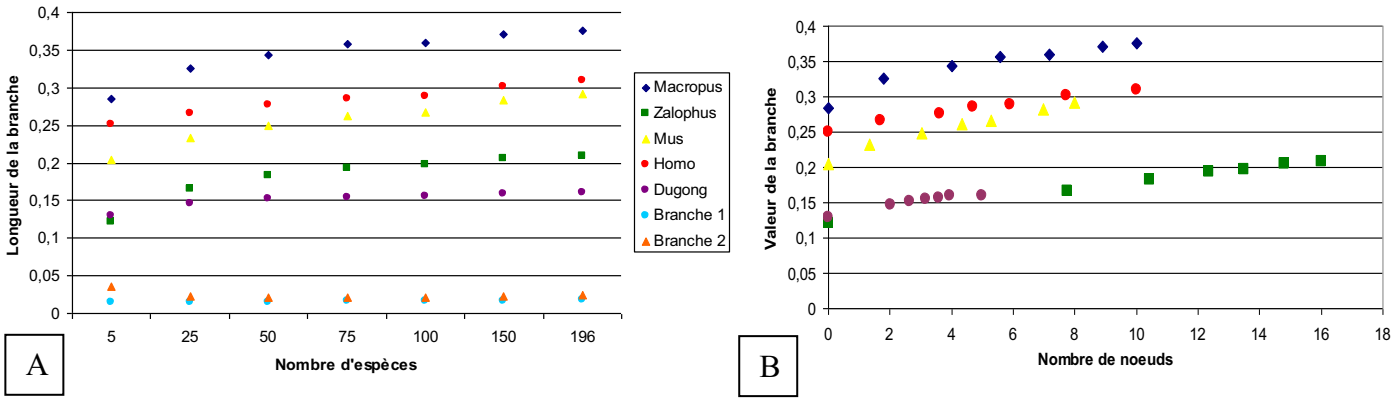

Figure 30 : A) Longueur de chaque branche du sous-arbre à 5 espèces en fonction de l'augmentation de l'échantillonnage taxonomique. Chaque point de 25 à 150 représente la moyenne du tirage aléatoire différent d'espèces sur 100 arbres. Les arbres ont été obtenus avec le modèle MtREV+Γ_8 en ML (Tree-Puzzle). Branche 1 et branche 2 représentent les branches internes de l'arbre réduit à 5 espèces. B) L'axe-X représente le nombre de nœuds (avant extraction du $\sum LB_5^n$) qui sont fonction du nombre d'espèces tandis que l'axe-Y représente la longueur de chaque branche du LB_5^n. Chaque point représente la moyenne du nombre de nœuds sur 100 tirages.

On sait que l'augmentation du nombre d'espèces dans un jeu de données permet de mieux estimer la longueur de certaines branches de l'arbre mais cela dépend également de l'échantillonnage taxonomique que l'on retrouve dans le jeu de données en question. Plus on va avoir d'espèces dans un groupe donné, plus les substitutions multiples dans ce groupe ont de chance d'être détectées; ceci est dû au fait que les espèces intermédiaires vont venir briser les branches les plus longues de l'arbre. Il est donc important lorsqu'on fait un ajout d'espèces dans un jeu de données quelconque, de prendre soin d'ajouter des espèces de façon « attentionnée ».

Dans la Figure 31, nous avons calculé la moyenne du nombre de nœuds brisant une branche de référence, c'est-à-dire le nombre de fois où la branche a été « cassée » par l'ajout d'espèces. Sur la Figure 30B, le nombre de nœuds (qui est corrélé à l'augmentation du nombre d'espèces) est présenté sur l'axe-X par rapport à la longueur de la branche correspondant à chacune des 5 espèces de référence (axe-Y). La valeur à 5 espèces du nombre de nœuds est 0 puisque celui-ci est notre arbre de référence. Les deux branches internes ne sont pas représentées sur cette figure, car ces dernières n'ont pas été « cassées » avec l'ajout d'espèces.

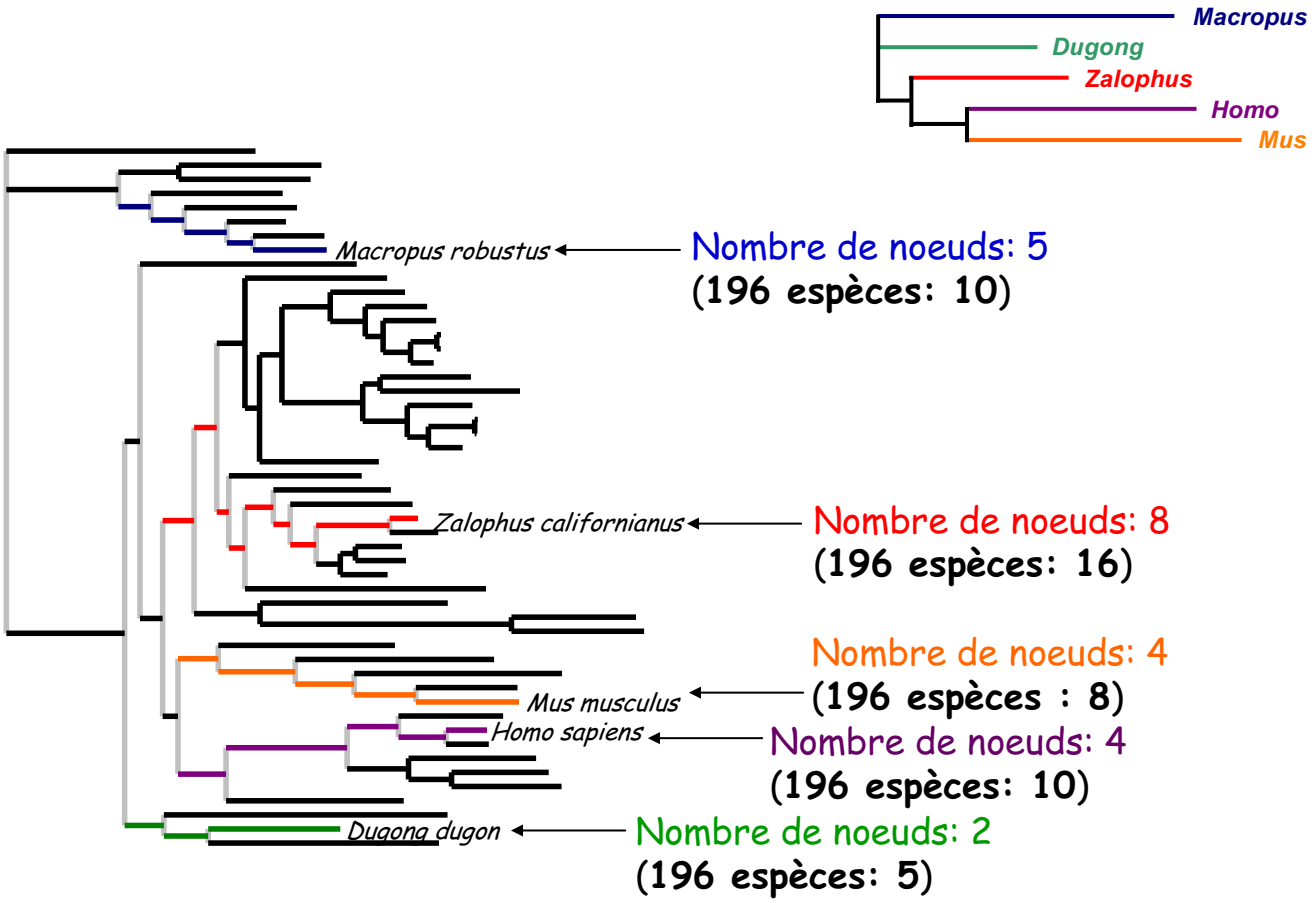

Figure 31: Exemple de calcul du nombre de nœuds dans un arbre phylogénétique à 50 espèces. Le nombre de nœuds correspond au nombre de fois qu'une branche a été cassée par l'ajout d'espèces dans le jeu de données. Le nombre de nœuds à 196 espèces est indiqué entre parenthèses.

On voit que la branche de l'espèce *Dugong dugon* n'est cassée que 5 fois à 196 espèces alors que celle de *Zalophus californianus* a un nombre de nœuds intermédiaires de 16. Cette différence est due à l'échantillonnage taxonomique car le groupe Laurasiatheria contient beaucoup plus d'espèces que les autres groupes. Ainsi, l'espèce *Dugong dugon* présente une courbe qui atteint un plateau car malgré l'augmentation du nombre d'espèces, cette branche ne peut pas être cassée un grand nombre de fois.

3.3. Analyse des 12 gènes mitochondriaux en ML

Les analyses précédentes étaient basées sur un alignement de 3540 acides aminés représentant la concaténation des 12 gènes mitochondriaux. Afin de connaître l'impact de l'augmentation du nombre d'espèces sur chacun de ses gènes, et de connaître leurs taux d'évolution, nous les avons analysés séparément. La Figure 32A présente les valeurs de $\sum LB_5^5$ et $\sum LB_5^{196}$ pour MtREV+Γ_8 en ML. Nous avons choisi ce modèle, car c'est celui qui s'ajuste le mieux à nos données jusqu'à maintenant. Les noms des différents gènes sont notés sur l'axe-X de la figure et la valeur de la longueur du sous-arbre est représentée sous forme d'histogramme. On voit que la taille du sous-arbre $\sum LB_5^5$ à 5 espèces a presque toujours une valeur inférieure à celle du sous-arbre $\sum LB_5^{196}$ à 196 espèces, sauf pour ATP8 pour lequel les deux valeurs sont pratiquement identiques. Cette variation intra-gène, bien que faible, était attendue vu que l'augmentation du nombre d'espèces permet de « casser » les plus longues branches de l'arbre. Le rapport $\frac{\sum LB_5^{196}}{\sum LB_5^5}$ pour le gène ATP8 vaut 0,98, tandis que cette valeur atteint 2,22 pour le gène COX3. De plus, à nombre d'espèces constant, les valeurs de LB varient de 0,30 pour le gène COX1 à 3,01 pour le gène ND2, ce qui donne une très grande différence de la vitesse moyenne d'évolution des 12 gènes mitochondriaux.

Existe-t-il une relation entre la vitesse d'évolution d'un gène et son niveau de saturation ? On s'attend à ce que plus un gène évolue vite, plus il soit saturé, car il a plus de chance d'être affecté par des substitutions multiples. Sur la Figure 32B (voir méthodologie section 2.8), l'axe-X représente la valeur $\sum LB_5^{196}$, qui peut être interprétée comme le taux d'évolution du gène tandis que l'axe-Y représente le niveau de saturation, c'est-à-dire le nombre de substitutions non détectées à 5 espèces mais détectées à 196 espèces. Cette valeur peut être calculée en faisant le rapport $\frac{\sum LB_5^{196}}{\sum LB_5^5}$.

Cette figure montre de façon surprenante une pente négative, avec un coefficient de corrélation d'une valeur de 0,61. Nos résultats suggèrent l'inverse de ce que l'on attendait, c'est-à-dire que les gènes qui évoluent le plus rapidement ont le plus bas taux de saturation. Cette contradiction peut être expliquée par le fait que $\sum LB_5^{196}$ est surtout dictée par le nombre de sites constants que notre alignement contient, tandis que le rapport $\frac{\sum LB_5^{196}}{\sum LB_5^5}$ est surtout dicté par la saturation des sites variables.

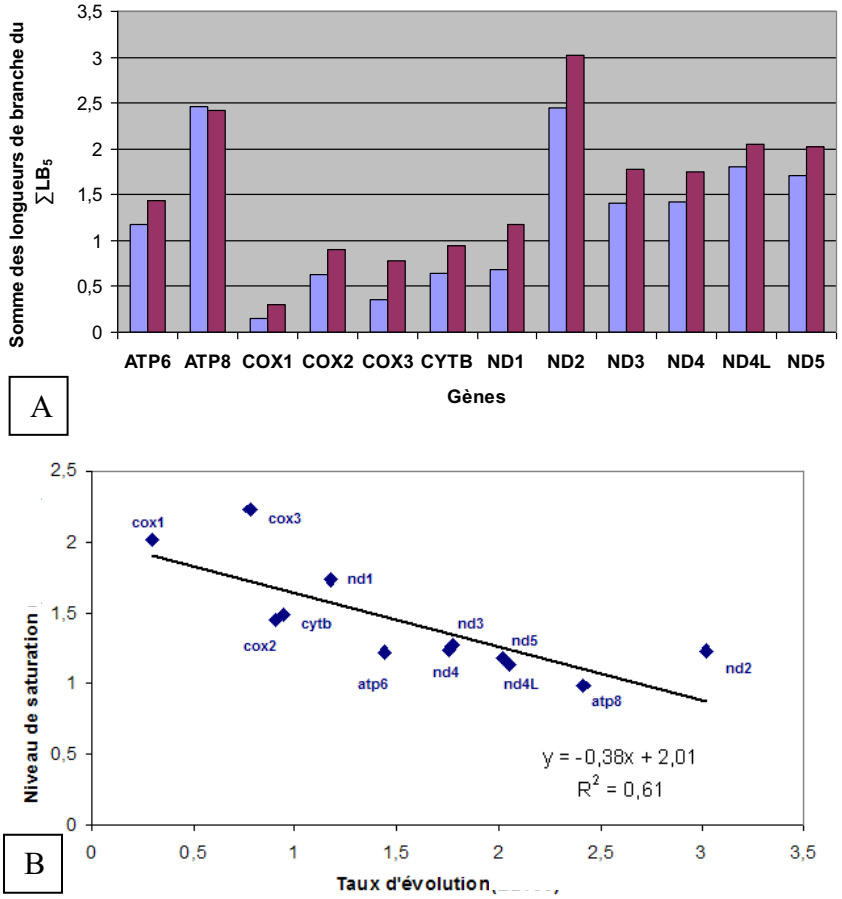

Figure 32 : Analyse des 12 gènes mitochondriaux séparément en vraisemblance avec MtREV+Γ (A). Chaque barre représente la valeur du sous-arbre à 5 espèces de référence pour 5 et 196 espèces. Le taux d'évolution ($\sum LB_5^{196}$) est corrélé avec le niveau de saturation ($\frac{\sum LB_5^{196}}{\sum LB_5^5}$) pour les 12 gènes (B).

Le Tableau 8 permet le classement croissant du taux d'évolution des 12 gènes mitochondriaux pour 5 et 196 espèces avec MtREV+Γ_8. On suppose que le classement à 196 espèces est plus exact que celui à 5 espèces, étant donné qu'on y détecte plus de substitutions multiples. Il est donc important d'utiliser un grand nombre d'espèces afin d'avoir un classement plus exact des taux d'évolution des différents gènes, puisque le tableau montre la très grande différence que l'on peut obtenir avec peu d'espèces.

On peut donc conclure de cette analyse qu'il est non seulement important d'utiliser un nombre important d'espèces pour les analyses phylogénétiques, mais que le choix du gène peut également influencer par son taux de saturation le niveau d'évolution (par conséquent le nombre de substitutions à prédire).

5 espèces	**196 espèces**
Cytochrome c oxidase subunit I	Cytochrome c oxidase subunit I
Cytochrome c oxidase subunit III	Cytochrome b
Cytochrome b	Cytochrome c oxidase subunit III
Cytochrome c oxidase subunit II	Cytochrome c oxidase subunit II
NADH dehydrogenase subunit 1	NADH dehydrogenase subunit 1
ATP synthase F0 subunit 6	NADH dehydrogenase subunit 4
NADH dehydrogenase subunit 3	ATP synthase F0 subunit 6
NADH dehydrogenase subunit 4	NADH dehydrogenase subunit 4L
NADH dehydrogenase subunit 5	NADH dehydrogenase subunit 5
NADH dehydrogenase subunit 4L	NADH dehydrogenase subunit 3
ATP synthase F0 subunit 8	ATP synthase F0 subunit 8
NADH dehydrogenase subunit 2	NADH dehydrogenase subunit 2

Tableau 8 : Tableau croissant des taux d'évolution des 12 gènes mitochondriaux pour 5 et 196 espèces avec le modèle MtREV+Γ_8 en ML.

3.4. Analyse des séquences nucléotidiques

On a pu voir qu'avec les séquences protéiques, l'augmentation du nombre d'espèces permet d'augmenter le nombre de substitutions détectées. Le phénomène est-il comparable avec les séquences nucléotidiques ? On sait que le code génétique est dégénéré, et qu'il peut exister plusieurs triplets pour un même acide aminé. On s'attend donc à ce que chacune des trois positions du codon soit influencée de façon différente, la 3ème position du codon (NT3) étant celle qui varie le plus souvent car ce

changement s'accompagne plus rarement d'un changement d'acide aminé. Nous avons donc pris les trois positions du codon séparément pour notre analyse.

Dans la Figure 33A, nous avons utilisé le modèle JC (Tree-Puzzle), qui assume des taux de transition ainsi que des fréquences d'équilibre égales pour toutes les bases (Jukes et al. 1969). On retrouve une corrélation entre l'augmentation du nombre d'espèces et l'augmentation de la taille du sous-arbre $\sum LB_5^n$ avec des écarts-types inter-échantillon très petits. On voit que pour NT1, la taille du sous-arbre passe de 0,70 à 1,07 de 5 ($\sum LB_5^5$) à 196 espèces ($\sum LB_5^{196}$). Pour NT2, ces valeurs passent de 0,30 à 0,46. Le rapport $\frac{\sum LB_5^{196}}{\sum LB_5^5}$ pour ces deux positions est égal à environ 1,52. De manière surprenante, le rapport est identique pour la troisième position du codon (NT3), avec des valeurs respectives de 2,73 et 4,14. Donc bien que JC soit le modèle ayant la matrice d'échange la plus simple, on est capable grâce à l'augmentation du nombre d'espèces de détecter des substitutions supplémentaires avec un rapport (1,52 fois plus de substitutions à 196 espèces vs 5 espèces) identique pour les 3 positions du codon. La différence étant qu'avec NT3 on détecte évidemment plus de substitutions par sites qu'avec NT1 ou NT2.

Lorsqu'on teste d'autres modèles d'évolution de séquences, comme le modèle HKY (Figure 33B), qui intègre un paramètre de plus que JC (rapport transition/transversion), on a une tendance de courbe peu corrélée à l'augmentation du nombre d'espèces pour NT3 et une influence presque identique aux analyses avec le modèle JC pour NT1 qui présente un rapport $\frac{\sum LB_5^{196}}{\sum LB_5^5}$ de 1,55 et NT2 où ce rapport est de 1,52. Les écarts-types pour NT3 sur les différents échantillons ont des valeurs trop grandes et font en sorte que les valeurs se chevauchent entre elles. On peut noter que cet écart-type tend à diminuer avec l'augmentation du nombre d'espèces, mais cela est expliqué par le fait qu'on retrouve plus d'espèces en commun entre les différents échantillons lorsque le nombre d'espèces augmente. Des problèmes numériques pourraient également expliquer les résultats obtenus pour NT3. Lorsque les longueurs de branche sont grandes (>1 substitution par site), leur valeur a très peu d'effet sur le

logarithme de la vraisemblance, c'est-à-dire que la surface de vraisemblance est presque plane et qu'il est difficile d'explorer efficacement la surface. On peut facilement rester bloqué dans un minimum local, ce qui empêche de trouver la position du maximum de la vraisemblance. De plus, la précision numérique des logiciels utilisés peut être insuffisante. Les valeurs que l'on infère pour des très longues branches peuvent donc facilement être erronées. Ce problème diminue par contre lorsqu'on augmente le nombre d'espèces, car les branches deviennent alors plus courtes (l'ajout d'espèces ayant permis de les briser à plusieurs reprises), ce qui fait en sorte que la surface de vraisemblance acquière plus de relief.

Dans la Figure 33C, nous avons utilisé le modèle GTR (Tavare 1986) avec le programme Paup (avec estimation des paramètres du modèle à partir des données) ; ce dernier modèle intègre plus de paramètres que précédemment et c'est le modèle le plus complexe pour les nucléotides à être couramment utilisé (section 1.5.3). Nous observons que pour NT2 le patron habituel de croissance de $\sum LB_5^n$ avec l'augmentation du nombre d'espèces est retrouvé, mais que pour NT1, après un plateau autour de 100 espèces, la valeur $\sum LB_5^n$ diminue. Nous n'avons pas du tout d'explications à ce phénomène unique dans toutes nos analyses. Pour la 3ème position du codon, les valeurs ne sont pas corrélées à l'augmentation du nombre d'espèces, avec une valeur plus élevée ($\sum LB_5^n$) à 5 espèces (12,08) qu'à 196 espèces (11,60) et des écarts-types très élevés entre 25 et 150 espèces et qui se chevauchent entre eux. Bien qu'un logiciel différent ait été utilisé (Paup au lieu de Tree-Puzzle), l'instabilité numérique est encore probablement la cause de ces résultats et permet de comprendre ces écarts-types.

Tous ces modèles étaient utilisés en maximum de vraisemblance, mais qu'en est-il maintenant de la méthode de maximum de parcimonie ? Est-ce qu'on aura le même problème qu'avec les acides aminés pour l'indétermination des substitutions le long des branches avec les deux méthodes ACCTRAN et DELTRAN ? La Figure 33D montre les résultats de la parcimonie où les tailles du sous-arbre à 5 espèces sont corrélées avec l'augmentation du nombre d'espèces. Mais le même problème se pose entre les deux méthodes (voir section 3.1.), car nous avons toujours une grande différence pour le nombre de substitutions détectées, en particulier pour NT3.

On aurait pu utiliser une concaténation des séquences correspondant aux trois positions nucléotidiques ou utiliser des modèles à codon, mais l'idée de départ était de voir l'effet des substitutions multiples sur l'inférence des longueurs de branche pour les différentes positions (NT1, NT2, et NT3) indépendamment. On s'attendait donc à avoir un plateau de saturation beaucoup plus marqué pour NT3 que NT1 et NT2. Mais à cause des différents problèmes computationnels évoqués ici, il a été impossible de le distinguer. Compte tenu de tous ces problèmes liés à la dégénérescence du code génétique et aux grandes différences observées entre les différentes positions du codon selon les différents modèles d'évolution des séquences, il est préférable de continuer à utiliser les séquences protéiques pour nos analyses.

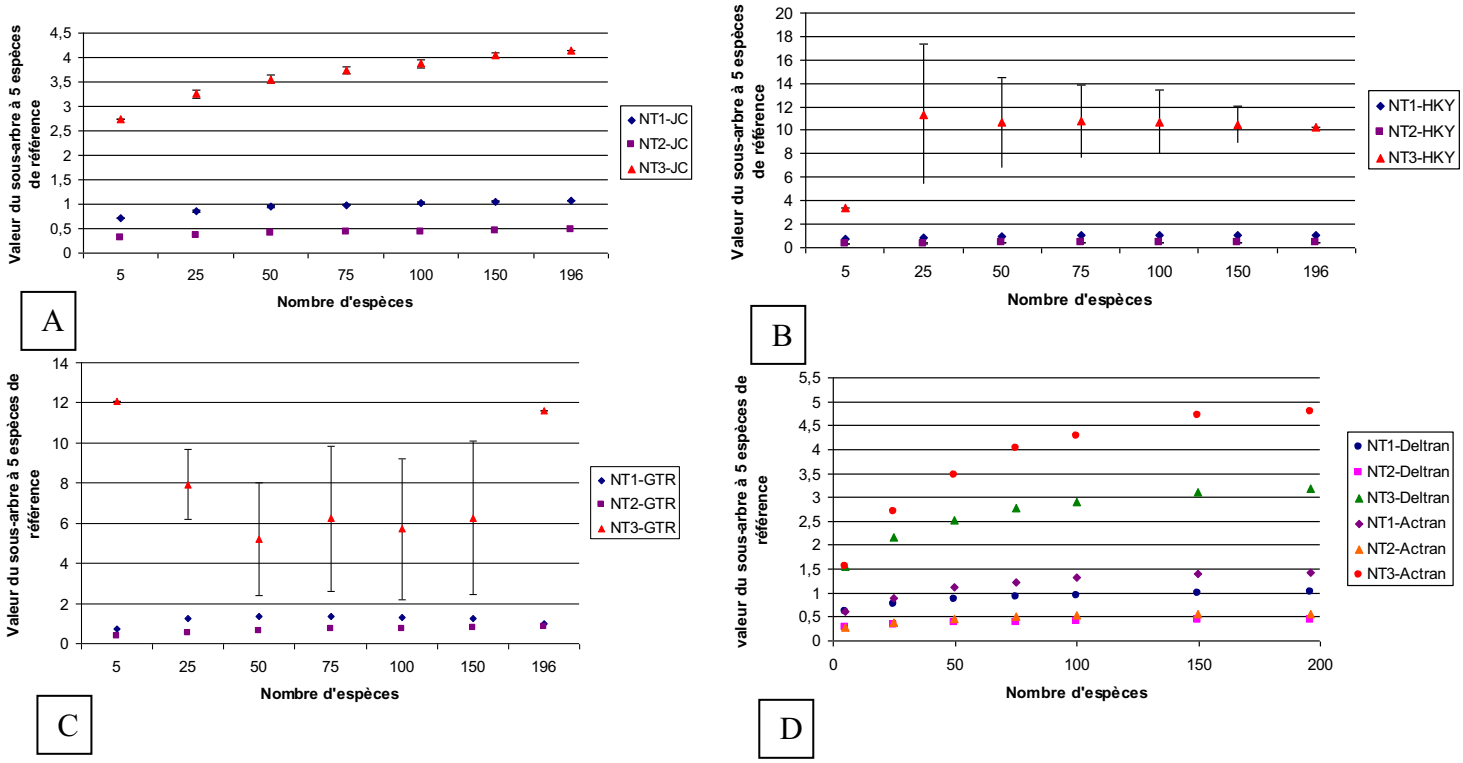

Figure 33 : Analyse des séquences nucléotidiques pour les trois positions du codon séparément soit NT1, NT2 et NT3 avec le modèle JC (Tree-Puzzle) (A), le modèle HKY (Tree-Puzzle) (B) modèle GTR (Paup) (C) et la parcimonie (Paup) (D). Deux méthodes sont utilisées pour définir les substitutions sur les branches de l'arbre pour la parcimonie, ACCTRAN (maximise les réversions) ou DELTRAN (maximise les convergences). Les graphiques montrent la longueur du sous-arbre à 5 espèces en fonction de l'augmentation du nombre d'espèces (de 5 à 196). Chaque point représente la moyenne de longueur de branche $\frac{\sum_{1}^{100} LB_5^n}{100}$ avec l'écart-type (barre d'erreur) sur 100 arbres.

3.5. Résultats des analyses bayésiennes avec différents modèles d'évolution de séquences

Plusieurs problèmes avec la méthode de maximum de vraisemblance nous ont incités à réaliser des analyses avec une méthode bayésienne. Tout d'abord le problème lié à l'utilisation de la distribution Gamma avec le programme Tree-Puzzle (voir section 2.6.2.), qui surestimait les longueurs de branche pour corriger la valeur trop faible de la moyenne des taux relatifs (≠ 1) ainsi que l'indisponibilité de tous les modèles d'évolution des séquences dans Tree-Puzzle.

Nous avons choisi d'utiliser le programme Phylobayes (voir section 2.6.3.). Sa particularité est qu'il utilise le modèle probabiliste CAT (Lartillot et al. 2004) qui prend en compte l'hétérogénéité entre sites du processus de remplacement des acides aminés. La supériorité du modèle CAT a été démontrée auparavant : il permet en particulier de prendre en compte le niveau de saturation générale, en ayant une meilleure estimation des préférences spécifiques en acides aminés pour chacun des sites (Lartillot et al. 2007). En plus du modèle CAT, quelques variations ont été étudiées : avec et sans distribution Γ, et avec différents jeux de taux relatifs d'échanges (MtREV, GTR).

3.5.1. Analyse de la concaténation

Le même protocole que précédemment pour l'échantillonnage aléatoire ainsi que la mesure de la taille du sous-arbre à 5 espèces a été utilisé. Le Tableau 9 contient les différentes valeurs obtenues avec Phylobayes pour plusieurs modèles d'évolution de séquences protéiques. Les valeurs pour 25 à 150 espèces représentent la moyenne des résultats sur 100 échantillons aléatoires. Nous avons utilisé quelques modèles (ex. CAT-MtREV) à titre de contrôle en analysant seulement la somme des longueurs de branche ($\sum LB_5^n$) pour 5 et 196 espèces. Afin de mieux visualiser ces résultats, quelques valeurs ont été schématisées dans la Figure 35 .

3.5.1.1. Rôle du logiciel dans l'estimation des longueurs de branche

Afin de vérifier l'influence éventuelle de la méthode statistique (maximum de vraisemblance versus Bayes) et du logiciel Phylobayes sur l'estimation des résultats, nous avons comparé la taille du sous-arbre estimé par Phylobayes et Tree-Puzzle avec les modèles Poisson et MtREV avec ou sans $\Gamma_{(4 \text{ ou } 8)}$ avec des jeux de données de 5 à 196 espèces (Figure 34). Le coefficient de corrélation pour les modèles comparés a une valeur presque égale à 1 (≈0.99) suggérant que les différences observées ne sont pas dues au logiciel ni à la méthode mais bien aux modèles utilisés. Cela confirme que les priors ont très peu d'influence sur les résultats bayésiens (à condition d'avoir suffisamment de données, ce qui est le cas ici). Comme attendu, les inférences en maximum de vraisemblance et en bayésien sont virtuellement identiques quand le modèle d'évolution est le même. Les résultats qui vont suivre sont donc représentatifs de l'approche probabiliste, et pas seulement de l'approche bayésienne.

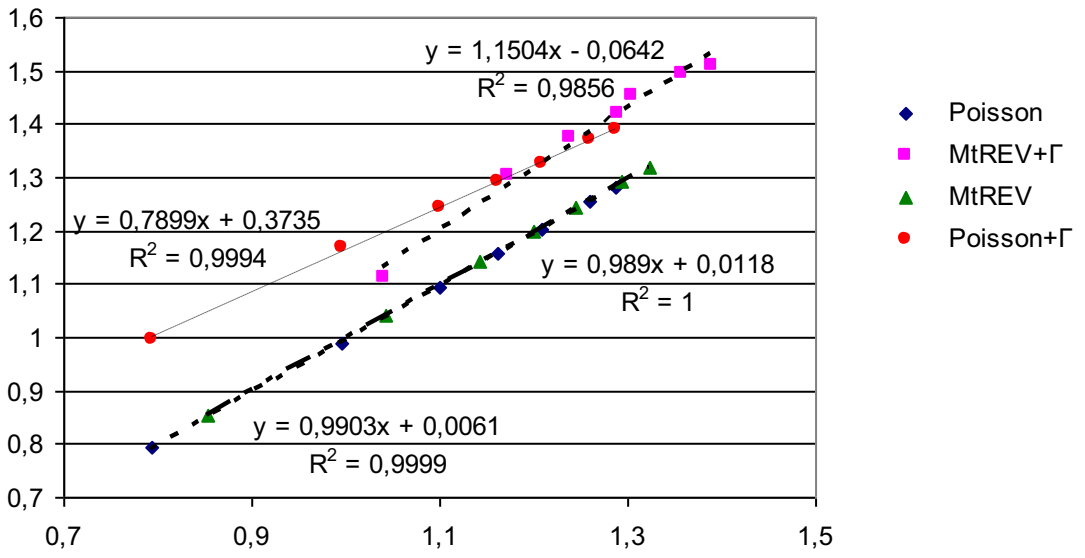

Figure 34: Corrélation de deux modèles (MtREV et Poisson) avec une inférence bayésienne (Phylobayes) et de maximum de vraisemblance (Tree-Puzzle), avec ou sans une distribution Γ. L'axe-X et l'axe-Y représentent respectivement les valeurs du $\sum LB_5^n$ avec Tree-Puzzle et Phylobayes pour un nombre croissant d'espèces (de 5 à 196 espèces).

3.5.1.2. Estimation avec CAT

À partir de la Figure 35, on voit qu'en général, les modèles de type CAT montrent la même tendance à croître en fonction de l'augmentation du nombre d'espèces, à l'exception du modèle CAT+Γ_4, où la valeur à 5 espèces (2,35) est plus élevée que toutes les autres alors que les valeurs diminuent jusqu'à 50 espèces pour ensuite suivre une tendance normale comparable aux autres modèles. Le meilleur ajustement de CAT aux données n'a été démontré auparavant que pour des jeux de données contenant un nombre d'espèces supérieur à 45 (Lartillot et al. 2004; Lartillot et al. 2008). Ces résultats à 5 espèces pourraient être expliqués par le fait que le modèle CAT n'est pas adapté aux analyses avec peu d'espèces, puisque celui-ci n'a pas assez d'information pour inférer des profils biologiquement pertinents, et surtout pour les attribuer aux différents sites de l'alignement. Cette hypothèse est illustrée par la grande variance observée avec peu d'espèces (écart-type important pour 25 espèces). Par contre, l'écart-type tend à diminuer avec l'augmentation du nombre d'espèces, c'est-à-dire à partir de 50 espèces, car le modèle est alors capable de bien attribuer les différents profils aux sites, ce qui diminue la variance observée entre les différents échantillons.

Nous avons également comparé la variance intra-chaîne avec ce modèle, c'est-à-dire la variance sur 100 arbres extrait directement de Phylobayes à partir de la chaine convergée de MCMC (Figure 36). À moins de 50 espèces, on a une dispersion très grande de la valeur moyenne de la taille des sous-arbres à 5 espèces ainsi que des écarts-types très larges. Ces écarts-types intra-chaînes sont comparables à ceux inter-échantillons calculés précédemment avec ce modèle, car on voit que ces derniers tendent également à diminuer avec l'augmentation du nombre d'espèces. Cela permet d'appuyer notre hypothèse qu'à très peu d'espèces le modèle CAT estime difficilement les longueurs de branche. On va donc pour ce modèle considérer seulement les résultats obtenus à partir de 50 espèces puisqu'avec moins d'espèces les résultats ne sont pas fiables.

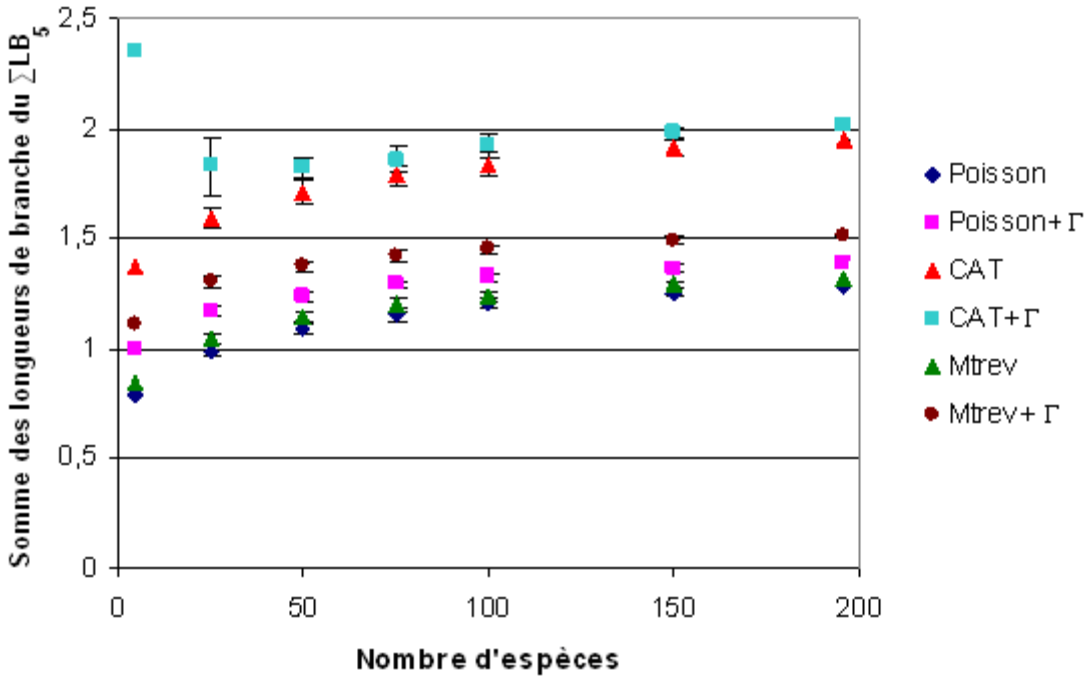

Figure 35 : Représentation graphique de quelques modèles du Tableau 9. Comparaison des 3 modèles (Wag, MtREV et Poisson) avec\sans une distribution gamma Γ_4 avec (Phylobayes). Chaque point représente la moyenne de longueur de branche pour le sous arbre à 5 espèces de référence avec l'écart-type (barre d'erreur) sur 100 arbres.

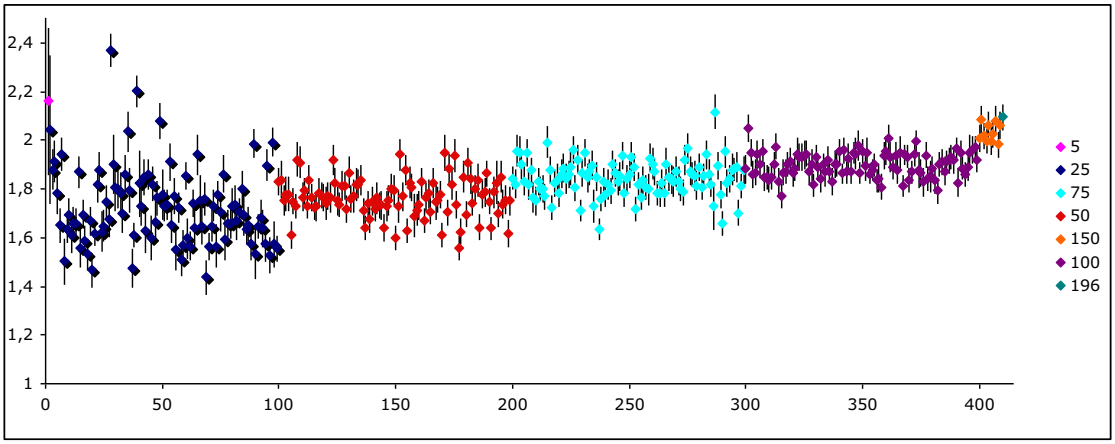

Figure 36 : Variance intra-chaine pour le modèle CAT+Γ_4. Cette variance est calculée sur 100 points de la chaine MCMC convergée. On retrouve sur le graphique la moyenne de 5 à 196 espèces avec l'écart-type correspondant sous forme de barre d'erreur.

3.5.1.3. Estimation avec les autres modèles

Les résultats des modèles Poisson, Wag, MtREV et GTR ont des valeurs très proches et comparables aux résultats obtenus avec la méthode ML (Figure 35 et Tableau 9). Par contre, on a une estimation des longueurs de branche nettement plus grande avec CAT par rapport aux autres modèles, avec une tendance de la courbe à augmenter proportionnellement au nombre d'espèces.

Modèle	5	25	50	75	100	150	196
Poisson	0,79	0,98	1,09	1,15	1,20	1,25	1,28
MtREV	0,85	1,04	1,14	1,19	1,24	1,29	1,31
Wag	0,83	1,03	1,14	1,20	1,24	1,29	1,32
GTR	0,84	1,04	1,15	1,21	1,25	1,30	1,33
CAT	1,37	1,59	1,71	1,78	1,84	1,91	1,95
Poisson+Γ_4	0,99	1,16	1,24	1,29	1,32	1,36	1,38
MtREV+ Γ_4	1,11	1,30	1,37	1,42	1,45	1,49	1,50
Wag+ Γ_4	1,09						1,51
GTR+ Γ_4	1,20	1,52	1,63	1,70	1,74	1,80	1,82
CAT+ Γ_4	2,34	1,82	1,82	1,86	1,92	1,97	2,01
CAT-MtREV	1,06						1,98
CAT-Wag	0,94						1,93
CAT-GTR	1,13						2,48
CAT-MtREV+ Γ_4	1,39						1,98
CAT-GTR+ Γ_4	1,49	2,45	2,33	2,43	2,53	2,59	2,65

Tableau 9: Tableau des valeurs moyennes du sous-arbre à 5 espèces de référence obtenues avec Phylobayes pour différents modèles d'évolution de séquences pour un nombre d'espèces croissant. De 25 à 150 espèces, la valeur est la moyenne sur 100 échantillons.

3.5.1.4. Impact de l'hétérogénéité de taux

Lorsqu'on considère une hétérogénéité entre les sites en introduisant le paramètre Γ dans nos modèles, on observe des différences entre les modèles Poisson, MtREV et GTR, avec une augmentation sensible de la taille du sous-arbre ($\sum LB_5^n$) de Poisson+Γ_4 à MTREV+ Γ_4 et de MTREV+ Γ_4 à GTR+ Γ_4 (le modèle WAG+ Γ_4 n'est pas considéré ici car seules les longueurs de branche à 5 et 196 espèces ont été calculées). Ceci peut être expliqué par le fait que la matrice d'échange du modèle MtREV a été

calculée sans considération d'une gamma distribution, ce qui fait que l'ajout de ce paramètre donne des estimations assez mauvaises (Le et al. 2008). Cette remarque peut aussi être faite pour la matrice Wag. La matrice GTR permet d'estimer à partir des données ses différents paramètres, ce qui pourrait expliquer qu'on obtient de meilleurs résultats. L'introduction du modèle GTR+Γ_4 permet donc d'avoir une estimation plus grande du nombre de substitutions.

3.5.1.5. Combinaison entre CAT et les autres modèles

Le modèle CAT suppose que la probabilité d'échange d'un a.a vers un autre est uniforme (Poisson), ce qui n'est peut etre pas réaliste. On peut, meme si c'est couteux d'un point de vue computationel, utiliser une matrice non homogène(MtREV, Wag ou GTR). Dans le Tableau 9, on voit que la combinaison du modèle CAT avec MtREV ou Wag nous donne des résultats inférieurs à ceux obtenus avec CAT. Et cela pour 5 et 196 espèces. Par contre, on peut noter la valeur élevée pour la combinaison du modèle CAT avec GTR, que ce soit avec Γ ou non, le nombre de substitutions détectées dépasse largement les estimations avec les autres modèles. La question reste à savoir s'il existe une surestimation des longueurs de branche avec ce modèle ou si on est en présence d'une sous-estimation des autres modèles ? Un contrôle a été effectué en introduisant la matrice d'échange GTR générée par les analyses avec CAT-GTR+Γ_4 sur une nouvelle analyse CAT-GTR+ Γ_4, la matrice GTR étant alors fixée pour l'analyse. Le calcul obtenu pour $\sum LB_5^{196}$ est de 2,62, cette valeur est très proche de celle dans le Tableau 9 et permet donc de confirmer la valeur de CAT-GTR+ Γ_4 mais ne répond toujours pas à la question de savoir si c'est dû à une surestimation. Des analyses sont en cours par Nicolas Lartillot (Département de biochimie, Faculté de Médecine, université de Montréal) afin de vérifier l'ajustement de ce modèle aux données et ainsi montrer sa supériorité par rapport aux autres modèles, puisqu'il permettrait alors de détecter beaucoup plus de substitutions.

3.5.1.6. Coût d'exécution des analyses et complexité théorique

Le Tableau 10 et le Tableau 11 sont des exemples de temps de calcul pour 2 modèles différents, soit Poisson et CAT-GTR. On peut y noter la grande différence de temps de calcul à 196 espèces par exemple pour Poisson (≈ 5h) par rapport à CAT-GTR (≈15 jours). Le temps de calcul est une sérieuse limitation à l'utilisation de CAT-GTR+Γ. Les différents résultats avec Phylobayes ont nécessité une longue période de temps afin d'avoir une bonne convergence des chaînes de MCMC.

CAT-GTR	**Nombre d'échantillon**	**Temps par échantillon**	**Convergence**
5	1	20h	500 points
10	100	16h	500 points
25	100	46h	500 points
50	100	92h	500 points
75	20	6,5 jours	500 points
100	20	8,5 jours	500 points
150	10	≈13 jours	500 points
196	1	≈15 jours	500 points

Tableau 10 : Exemple de temps de calcul (avec un processeur Xeon, 2.4 GHz) avec Phylobayes pour le modèle CAT-GTR.

PB-Poisson	**Nombre d'échantillon**	**Temps par Échantillon**	**Convergence**
25	100	45 min	100
50	100	1,25 h	100
75	100	2,5 h	100
100	100	2,5 h	100
150	100	≈ 4h	100
196	1	≈ 5h	100

Tableau 11: Exemple de temps de calcul (avec un processeur Xeon, 2.4 GHz) avec Phylobayes pour le modèle Poisson.

On peut conclure des résultats obtenus avec le modèle CAT dans Phylobayes que ces derniers permettent de donner une meilleure estimation des longueurs de branche et par conséquent du nombre de substitutions multiples qui ont pu avoir lieu. Des modèles complexes comme CAT ou CAT-GTR (avec/sans Gamma) ont permis de donner des valeurs largement supérieures aux autres modèles (même si la question de surestimation reste posée pour CAT-GTR).

La complexité théorique des algorithmes de Monte Carlo est assez difficile à calculer, à cause de l'aspect stochastique de la convergence. C'est vrai en particulier, si l'on a affaire a des algorithmes basés sur des principes de rééchantillonnages différents (Metropolis-Hastings (Metropolis et al. 1953; Hastings. 1970) versus échantillonnage de Gibbs (Geman et al. 1984; Casella et al. 1992), par exemple). Cela dit, un point particulier mérite d'être mentionné, concernant les différences algorithmiques entre les implémentations des modèles CAT et CAT-GTR, qui vont tous les deux dans le sens d'une plus grande efficacité sous le modèle CAT: Les calculs de vraisemblance sont effectués en utilisant l'algorithme du pruning (Felsenstein 1981), qui est un cas de programmation dynamique. Ils ont une complexité proportionnelle à S^2, ou S est le nombre d'états (20 pour les protéines). Cela dit, pour le modèle CAT, ou l'on a affaire a des processus de type Poisson, il est possible de recoder l'espace d'états du processus de substitution en considérant l'ensemble des états non observes en une colonne donnée de l'alignement comme un seul état 'X'. La fréquence d'équilibre de cet état virtuel est la somme des fréquences d'équilibres des états qu'il représente. La complexité du calcul sera alors de $(M_i+1)^2$, ou M_i est le nombre d'états observés en la colonne i. Sur l'ensemble de l'alignement le rapport de complexité est de 20x20 / $<(M_i+1)^2>$ ou <> est la moyenne prise sur l'ensemble de l'alignement. Ce recodage n'est exact que pour les processus de Poisson, et donc, n'est pas valide dans le cas du modèle CATGTR ou MtREV.

3.5.2. Analyse séparée des 12 gènes mitochondriaux avec Phylobayes

Nous avons effectué la même analyse que la section 3.3, mais en bayésien avec le modèle CAT-GTR+Γ_4 afin d'évaluer l'impact du modèle sur l'estimation des longueurs de branche à 5 et 196 espèces. Nous observons qu'on a toujours la même tendance à avoir une valeur supérieure, peu importe le gène, pour 196 espèces par rapport à 5 espèces (Figure 37A).

Les rapports $\frac{\sum LB_5^{196}}{\sum LB_5^5}$ entre la taille du sous-arbre à 196 espèces et à 5 espèces (Figure 37B) sont par contre différents de ceux observés avec le modèle MtREV+Γ_8 (Figure 32B). Par exemple le gène ND2 a un rapport de 1,23 pour MtREV+Γ_8 versus 1,95 pour CAT-GTR+Γ_4. On note qu'il n'a d'ailleurs pas été possible d'obtenir des valeurs de longueurs de branche pour le gène ND4L à 5 espèces avec le modèle CAT-GTR+Γ_4 compte tenu de la longueur trop petite du gène. Quand le taux d'évolution est représenté en fonction du niveau de saturation pour le modèle CAT-GTR+Γ_4 (Figure 37B), on observe que la pente est presque nulle et que le coefficient de corrélation n'est que de 0,0038. L'interprétation de ce graphique est ici différente du graphique avec le modèle MtREV+Γ_8 (Figure 32B).

On a pu voir précédemment que les résultats avec le modèle MtREV+Γ_8 peuvent être faussés. La corrélation négative obtenue dans la Figure 32B est probablement erronée à cause de l'absence de paramètre gamma (Γ) dans la détermination de la matrice MtREV (Le et al. 2008). Ceci peut être confirmé par l'absence de corrélation dans la Figure 37B (modèle CAT-GTR+Γ_4) et nous permet de conclure qu'il n'y a pas de corrélation entre le taux de saturation et la vitesse d'évolution d'un gène. Ce résultat semble plus logique car les positions constantes qui influencent la vitesse d'évolution n'ont pas d'impact sur les positions rapides qui elles comptent dans l'estimation de la saturation.

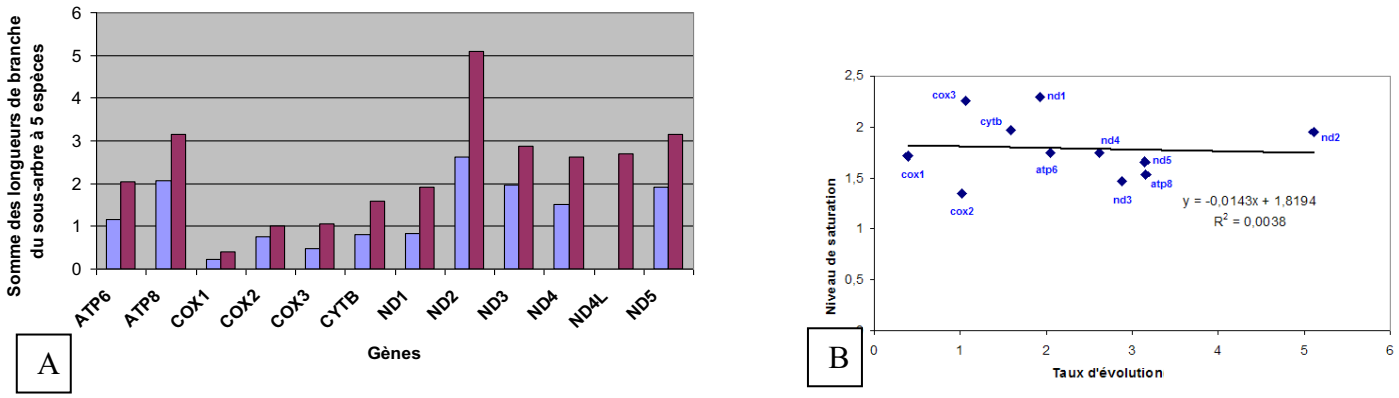

Figure 37: Analyse des 12 gènes mitochondriaux séparément en bayésien avec le modèle CAT-GTR+Γ$_4$. Chaque barre représente la valeur de $\sum LB_5^n$ pour 5 et 196 espèces. Le taux d'évolution ($\sum LB_5^{196}$) est corrélé avec le niveau de saturation ($\frac{\sum LB_5^{196}}{\sum LB_5^5}$) pour les 12 gènes (B).

3.5.3. Simulations de séquences avec Phylobayes

Après avoir obtenu plusieurs résultats avec les séquences réelles mitochondriales, on voulait tester l'hypothèse selon laquelle il n'y aurait moins de sous-estimation des longueurs de branche pour des séquences simulées que pour les données réelles. Nous avons donc effectué du « posterior prédictive » avec Phylobayes, cela consiste à simuler de nouvelles séquences en utilisant les valeurs de paramètres estimées à partir de données réelles. L'alignement de départ a permis d'obtenir 100 arbres avec le modèle CAT+Γ$_4$ à 196 espèces correspondant à différents points de la chaîne MCMC après convergence. Chacun de ses arbres a servi à simuler un jeu de séquences (voir section 2.10.).

Les séquences simulées sont utilisées pour inférer à nouveau des arbres avec longueurs de branche pour différents modèles, par exemple MtREV+Γ$_8$ en ML (Tree-Puzzle) représenté sur la Figure 38A. Sur l'axe-Y on retrouve la valeur ($\sum LB_5^{196}$) pour chacun des 100 arbres obtenus à partir des séquences simulées (l'axe-X). On voit que les valeurs varient entre 1,4 et 1,6 pour le modèle MtREV+Γ$_8$, tandis qu'elles ont une valeur nettement supérieure, comprise entre 2 et 2,2 en moyenne pour les 100 arbres obtenus avec le modèle CAT+Γ$_4$ correspondants aux 100 points extraits de la chaine MCMC. On a donc une estimation des longueurs de branche inférieure pour les

séquences simulées par rapport aux données réelles avec le modèle CAT+Γ_4. De plus, les $\sum LB_5^{196}$ des séquences simulées inférées avec MtREV+Γ_8 ont une valeur supérieure à la valeur obtenue avec les données réelles en utilisant le même modèle (1,39 substitutions par site contre 1,52 en moyenne pour les séquences simulées). Ceci permet d'appuyer l'hypothèse selon laquelle il existe moins de sous-estimation des longueurs de branche pour des séquences simulées que des séquences réelles. Ces résultats sont la confirmation qu'avec les données réelles la saturation nous empêche d'inférer les bonnes longueurs de branche.

Dans la Figure 38B, la valeur $\sum LB_5^n$ est représentée pour différents modèles à 5 et 196 espèces. Que ce soit avec 5 ou 196 espèces, on a des valeurs qui sont similaires peu importe le modèle, avec une valeur nettement plus grande à 196 espèces. La moyenne des tailles de $\sum LB_5^{196}$ extraits à partir de l'inférence à 196 espèces avec le modèle CAT+Γ_4 (données réelles) reste largement plus élevée, elle est représentée par le losange orange sur le graphe. Mais la conclusion reste la même que précédemment, c'est-à-dire une sous-estimation moindre des longueurs de branche avec des séquences simulées pour tous les modèles en ML, avec des valeurs plus élevées de $\sum LB_5^5$ et $\sum LB_5^{196}$ que pour les séquences réelles avec tous ces modèles (Figure 38B). On aurait pu tester plus de modèles pour voir l'effet des simulations sur les longueurs de branche inférées, mais nous nous sommes contentés des six modèles dans la Figure 38B pour des raisons computationnelles. À partir des 100 alignements simulés à 196 espèces, on aurait également pu utiliser le même modèle que celui qui a servi à les simuler, soit CAT+Γ_4 afin de vérifier qu'on arrive à retrouver les résultats de départ. Comme le temps computationnel est très long pour un aussi gros jeu de données, nous avons fait le test avec un seul alignement simulé tiré parmi les 100 afin de vérifier que l'inférence des longueurs de branche est la même. Comme les valeurs réelles et inférées sont très proches (2,10 vs 2,09), cela a permis de vérifier la cohérence du modèle CAT.

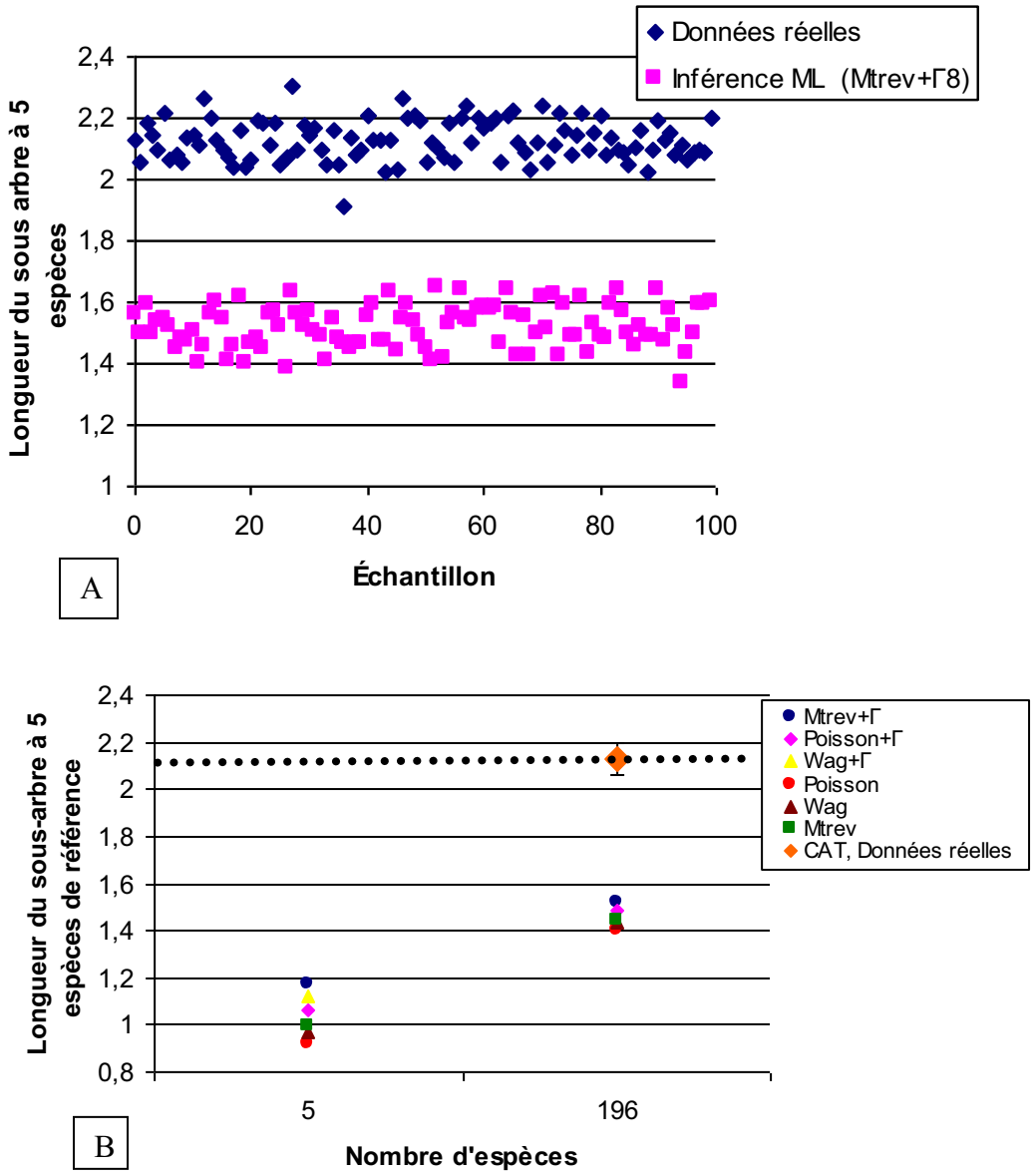

Figure 38 : À partir de 100 arbres obtenus avec Phylobayes et le modèle CAT (données réelles), 100 alignements de 196 espèces ont été simulés et utilisés pour une inférence avec 6 modèles (Wag, MtREV et Poisson) avec\sans une distribution Γ_8 en ML (Tree-Puzzle). A) représente la comparaison de $\sum LB_5^{196}$ pour les données réelles extraites de l'inférence par CAT+Γ_4 et pour les arbres inférés avec le modèle MtREV+Γ_8 à partir des séquences simulées. B) Représente la valeur moyenne des 100 sous-arbres inférés ($\frac{\sum_{1}^{100} LB_5^n}{100}$) pour 5 et 196 espèces à partir des séquences simulées avec les 6 modèles.

3.6. Retrait de sites et coefficient d'asymétrie

Les trois espèces de monotrèmes sont exclues de l'analyse du coefficient d'asymétrie, car elles servent de groupe extérieur. Le coefficient d'asymétrie est donc calculé sur 193 espèces d'après la méthodologie décrite dans la section 2.8. et les retraits de sites d'après la méthodologie dans la section 2.11. (taux des sites calculés avec le modèle Wag+Γ_8 et longueur des branches avec MtREV+ Γ_8).

Comme attendu, la taille du sous-arbre $\sum LB_5^{196}$ tend à diminuer avec le retrait progressif de sites rapides (Figure 40A); cette valeur sans retrait de sites est de 1,39 et elle passe d'une valeur de 0,95 (pour un retrait de 250 sites) à 0,05 (pour un retrait de 1750 sites). Ce qui nous intéresse est de connaître le coefficient d'asymétrie de l'arbre, puisque la saturation va surtout raccourcir les branches longues et donc réduire le coefficient d'asymétrie de la distribution des distances de la racine aux feuilles terminales. On peut voir sur la Figure 40 prise à titre de référence puisqu'aucun retrait de sites n'y a été effectué, que la valeur du coefficient d'asymétrie pour les six modèles d'évolution est inférieure à 1 (présence de saturation), et que cette valeur est nettement inférieure à celle obtenue du coefficient d'asymétrie lorsqu'on fait des retraits de sites. On peut voir sur la Figure 40B, que plus on retire des sites de notre alignement, plus le coefficient augmente, il passe d'une valeur de 0,68 (sans retrait) à 1,9 (1750 sites retirés). Cela permet de confirmer que le retrait de sites a permis de garder les sites qui ont le moins de substitutions multiples, donc ceux qui ont le moins besoin du modèle pour prédire les substitutions cachées.

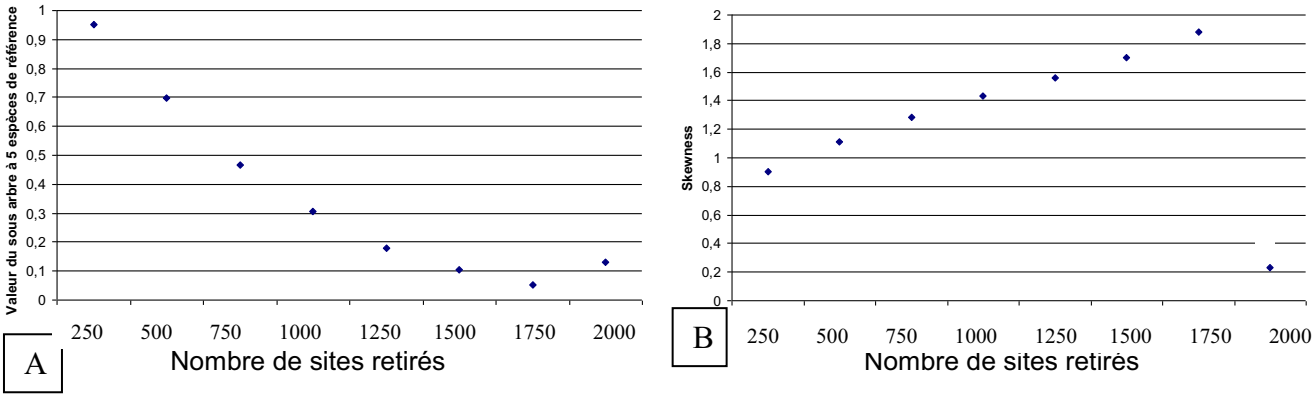

Figure 39 : Longueur du sous-arbre à 5 espèces de référence (A) et coefficient d'asymétrie (B) pour un retrait de 250 sites à la fois sur les 3540 positions départ. Chaque point représente la moyenne sur 100 arbres inférés pour 196 espèces avec le modèle MtREV+Γ (Tree-Puzzle). Le coefficient d'asymétrie est calculé sur les longueurs de branche de 193 espèces dans l'arbre, les 3 monotrèmes étant exclus.

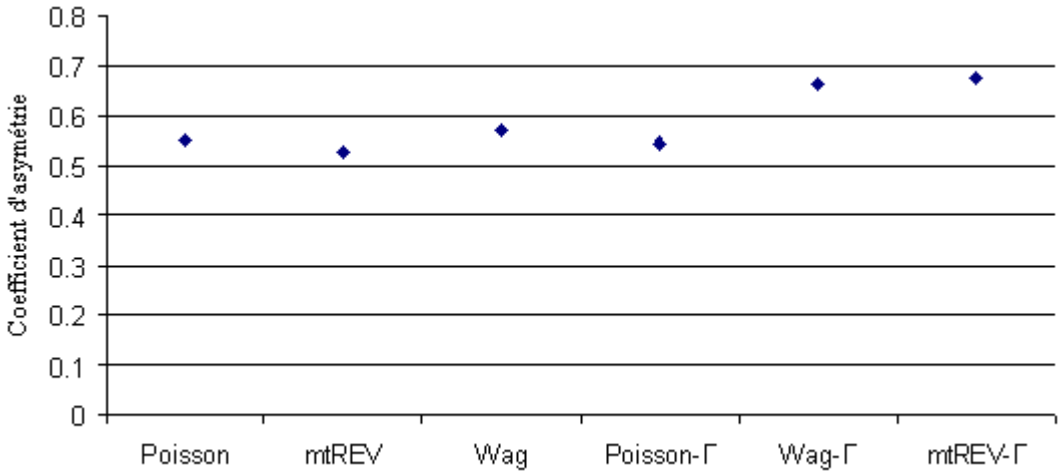

Figure 40 : Coefficient d'asymétrie pour 6 modèles d'évolution de séquences sans retrait de sites.

3.7. Estimation des longueurs de branche et datation moléculaire

Les longueurs des branches sont souvent très sous-estimées, de manière non proportionnelle comme on a pu le voir avec les résultats précédents, puisque les plus longues branches sont beaucoup plus sous-estimées que les plus courtes (les longues branches étant plus apte à contenir des substitutions multiples). Étant donné que cette sous-estimation est présente avec tous les modèles d'évolution de séquences, on

s'attend à avoir un effet important sur les datations. Rappelons que les longueurs de branche sont fonction du taux de substitutions et du temps (section 1.3.1).
Pour des raisons de temps de calcul, nous avons choisi de travailler avec les deux programmes Estbranche et Multidivtime; ces programmes sont les plus rapides, mais ne permettent de tester que peu de modèles. Par contre cela nous permet de regarder l'effet de l'augmentation du nombre d'espèces, puisqu'on a montré jusqu'à maintenant, grâce aux analyses sur les longueurs de branche, que son impact semble plus important que l'effet du modèle choisi. On voulait également vérifier si différentes calibrations (en particulier jeunes ou vieilles) ne biaisaient pas les résultats. Nous avons cependant toujours utilisé une seule calibration à la fois afin de pouvoir utiliser au départ des jeux de données avec peu d'espèces (8 espèces). Des analyses avec le programme Phylobayes qui permet de réaliser de la datation moléculaire avec d'autres modèles, sont en cours et n'ont pas pu être intégrées à ce mémoire.

La datation a été effectuée pour 8, 25, 50 ,75 et 196 espèces, avec 100 échantillons pour 25 à 75 espèces; le modèle d'évolution des séquences mtmMAM (Yang et al. 1998) et un modèle d'autocorrélation lognormal pour les taux (Kishino et al. 2001) ont été utilisés. On retrouve les résultats de la datation moléculaire dans les Figure 41-Figure 44 sur lesquelles on a inclus l'écart-type entre échantillons taxonomiques sous forme de barre d'erreur. L'écart-type pour 8 et 196 espèces extrait directement du programme Multidivtime est indiqué dans le Tableau 12. On retrouve dans ces figures différents points de calibration et l'estimation de l'âge de plusieurs nœuds d'après les jeux de données A, B, C et D (voir une description des points de calibration et des jeux de données à la section 2.12).

3.7.1. Calibration à partir du noeud *Equus caballus/Rhinoceros unicornis*

Pour évaluer l'effet de l'augmentation du nombre d'espèces sur l'estimation de l'âge des nœuds, nous avons étudié l'âge prédit pour quatre nœuds différents. Dans la Figure 41, avec une calibration ancienne (*Equus caballus*/*Rhinoceros unicornis* : 54 à 58 M.A.), on voit que l'âge inféré des nœuds varie grandement. Pour les Laurasia par

exemple, l'âge va de 78 M.A pour 8 espèces à 89 M.A. pour 196 espèces, en passant par des valeurs souvent supérieures pour un nombre intermédiaire de taxons. On remarque que l'écart-type inter-échantillon entre 25 et 75 espèces (valeur moyenne de 96 M.A. ± 8,5 M.A.) est très grand, ce qui fait que les âges du nœud Laurasiatheria se chevauchent entre eux et avec l'âge estimé avec 196 espèces (mais sont toujours plus grands que celui estimé avec 8 espèces). Cette grande variance s'explique très probablement par l'effet de l'échantillonnage taxonomique sur l'estimation des longueurs de branche qui, comme on l'a vu précédemment, est très grand avec ces faibles nombres d'espèces. En plus, l'échantillonnage taxonomique influe sur la modélisation du taux. En effet, dépendamment des espèces qui sont échantillonnées aléatoirement parmi le jeu de données, on peut avoir plus d'espèces dans un groupe que dans un autre ce qui peut influencer la manière dont le taux varie au cours du temps et donc peut influencer l'estimation de l'âge des nœuds. À 8 et 196 espèces, l'écart-type intra-échantillon est fourni directement par le programme Multidivtime. On peut noter qu'en tenant compte de la marge d'erreur, l'âge des nœuds à 8 et 196 espèces se chevauchent pour la plupart des jeux de données (Tableau 12) ; par exemple les valeurs pour le nœud Laurasiatheria sont de 84.9±8.5M.A à 8 espèces et de 88.2±7.6M.A à 196 espèces. Notre approche ne donne donc pas un âge assez précis pour pouvoir tirer des conclusions statistiquement significatives sur l'impact de l'échantillonnage taxonomique sur les datations, contrairement à ce que l'on avait pu faire pour les longueurs de branche.

On a également la même tendance de courbe et le même problème d'écart-type avec le nœud d'Euarchontoglires et avec le nœud Placentalia (excepté que l'âge à 196 espèces est plus récent que celui à 5 espèces). Le nœud le plus intéressant est celui de *Mus musculus*/*Rattus norvegicus*, car ce dernier a été évalué directement par les paléontologues, grâce aux données fossiles comme ayant divergé entre 7-16 M.A. (Jacobs et al. 1980), ce qui permet de vérifier l'estimation de l'âge. La Figure 41 (nœud 4) montre que l'ajout d'espèces fait tendre vers cette valeur. Partant de 24 M.A.(± 5,5 M.A) avec 8 espèces, on voit qu'à 196 espèces, l'âge estimé du nœud se rapproche de l'intervalle des paléontologues avec une valeur de 18 M.A. (± 4M.A.). Dans ce cas-ci,

l'augmentation du nombre d'espèces permet de donner un âge raisonnable à au moins un des nœuds.

3.7.2. Calibration à partir du noeud *Felis catus/Zalophus californianus*

La Figure 42 montre les résultats pour une autre calibration avec un nœud ancien (*Felis catus* et *Zalophus californianus* : 50 à 63 M.A.). La forme des courbes ressemble à celle obtenue avec la calibration *Equus caballus*/*Rhinoceros unicornis*, à l'exception de l'estimation de l'âge du nœud Euarchontoglires où l'âge estimé à 196 espèces est légèrement plus jeune que celui à 8 espèces (90 M.A. vs 95 M.A.). On peut également noter que les âges sont relativement plus anciens entre 8 et 75 espèces pour tous les nœuds mesurés, mais qu'à 196 espèces, l'âge estimé est presque identique à celui estimé auparavant avec une calibration *Equus caballus*/*Rhinoceros unicornis* avec des écarts-types très comparables (Tableau 12). Les mêmes conclusions que précédemment s'appliquent ici, puisqu'on ne peut vérifier que l'âge estimé pour *Mus musculus*/*Rattus norvegicus*, l'estimation à 196 espèces se rapproche également beaucoup plus de la paléontologie que celle effectuée avec peu d'espèces.

3.7.3. Calibrations à partir de nœuds plus récents

Pour les Figure 43 et Figure 44, les calibrations sont beaucoup plus récentes, puisqu'elles sont faites pour le jeu de données C sur le nœud *Homo sapiens*/*Pan paniscus* (5 à 7 M.A.) et *Mus musculus*/*Rattus norvegicus* (7 à 16 M.A.) pour le jeu de données D.

Sur la Figure 43, on voit que pour les nœuds 1 et 3, on a une tendance croissante de la courbe en fonction du nombre d'espèces, contrairement aux nœuds 2 et 4 où l'âge estimé tend à diminuer avec l'augmentation du nombre d'espèces. Comparé aux calibrations précédentes, l'âge des nœuds estimés est beaucoup plus jeune. Par exemple, avec 196 espèces, les placentaires sont estimés avoir divergé à (109.5 ± 10.4)

M.A. avec *Felis catus*/*Zalophus californianus* et à (89.4 ± 16.4) avec *Homo sapiens*/*Pan paniscus*. Cela confirme la grande sensibilité des datations moléculaires aux points de calibration. De manière intéressante, malgré ces différences, le nœud 4 (14.5 ± 3.4 M.A.) se rapproche encore de la valeur paléontologique (7 à 16 M.A.) de l'âge de divergence entre *Mus musculus* et *Rattus norvegicus*. Les écarts-types gardent des valeurs élevées, et cette incertitude due à l'échantillonnage taxonomique nous empêche de confirmer que l'augmentation du nombre d'espèces est un facteur qui influence significativement l'estimation de l'âge des nœuds.

Dans la dernière figure (Figure 44), l'âge estimé de tous les nœuds est encore plus jeune que précédemment, toujours avec des tendances de courbes très différentes dépendamment du nœud que l'on date. L'âge estimé dans ce cas-ci du nœud *Homo sapiens/Pan paniscus* à 196 espèces est de 4.3 ± 1.3 M.A. (valeur d'ailleurs identique à l'âge estimé avec 8 espèces), ce qui fait que cette valeur est inférieure à l'âge de divergence estimé par les données fossiles pour ces deux espèces, soit entre 5 et 7 M.A.

On peut conclure de ses résultats, même si l'incertitude est très grande :

- qu'il vaut mieux utiliser un grand nombre d'espèces pour avoir une valeur plus proche de la date de divergence réelle entre les espèces (voir l'estimation de la datation du nœud *Mus musculus/Rattus novergicus*)
- la mauvaise estimation des longueurs de branche par des modèles d'évolution de séquences mal « adaptés » peut également influencer l'âge de divergence
- l'utilisation de dates de calibrations trop récentes a tendance à sous-estimer l'âge réel du nœud, en particulier en présence de branches longues.

Jeu de données-Nb d'espèces	Âge du nœud				
	Laurasia	Placentalia	Euarchontoglires	*Mus / Rattus*	*Homo /Pan*
A-8	77,3 ± 5,6	117,1 ± 13,5	80,5 ± 8,8	24,2 ± 5,5	
A-196	89 ± 6	110 ± 9	91 ± 7	18 ± 4	
B-8	84.9 ± 8.5	137.0 ± 17.7	94.8 ± 12.0	30.3 ± 8.2	
B-196	88.2 ± 7.6	109.5 ± 10.4	89.8 ± 8.2	17.8 ± 3.4	
C-8	55.2 ± 10.7	93.7 ± 17.6	63.6 ± 10.7	18.3 ± 3.2	
C-196	72.0 ± 13.3	89.4 ± 16.4	73.1 ± 13.3	14.5 ± 3.4	
D-8	40.8 ± 9.5	71.6 ± 15.7	47.0 ± 9.7		4.2 ± 1.0
D-196	55.5 ± 13.3	69.3 ± 16.4	56.5 ± 13.4		4.3 ± 1.3

Tableau 12 : Âges des nœuds estimés pour les différents jeux de données (A, B, C et D) sur un alignement de 8 et 196 espèces. Chaque valeur représente l'âge du nœud correspondant ainsi que l'écart-type fourni par le programme Multidivtime.

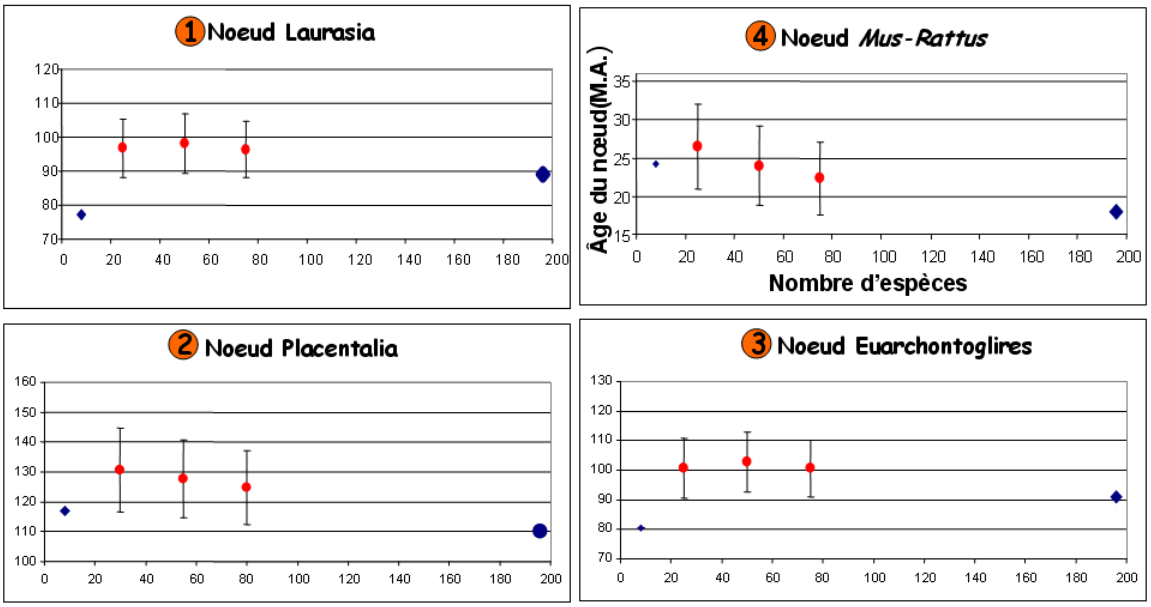

Figure 41 : Datation moléculaire sur le jeu de données A. L'intervalle de calibration a été fixé pour *Equus caballus* et *Rhinoceros unicornis* (54 à 58 M.A.). On retrouve sur la figure l'âge des nœuds estimés (axe-Y) avec l'écart-type (barre d'erreur) en fonction de l'augmentation du nombre d'espèces (axe-X). 1) Âge du nœud à la base des Laurasiatheria; 2) Âge du nœud à la base des placentaires; 3) Âge du nœud à la base des Euarchontoglires; 4) Âge du nœud à la base de *Mus musculus* et *Rattus norvegicus*.

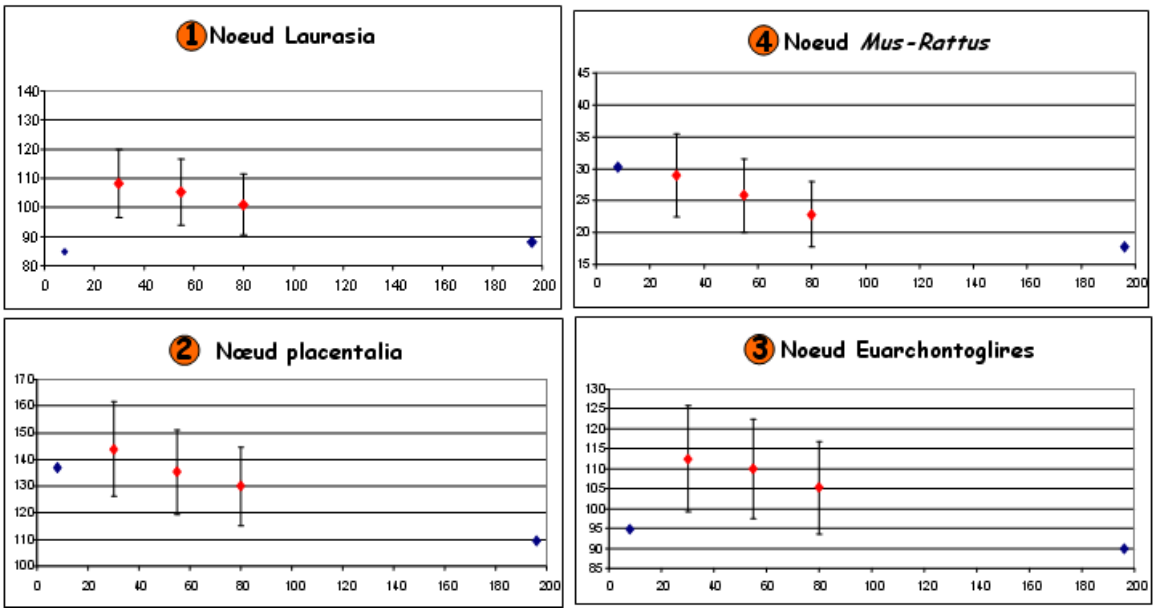

Figure 42 : Datation moléculaire sur le jeu de données B. L'intervalle de calibration a été fixé pour *Felis catus* et *Zalophus californianus* (50 à 63 M.A.). On retrouve sur la figure l'age des nœuds estimés (axe-Y) avec l'écart-type (barre d'erreur) en fonction de l'augmentation du nombre d'espèces (axe-X). 1) Âge du nœud à la base des Laurasiatheria; 2) Âge du nœud à la base des placentaires; 3) Âge du nœud à la base des Euarchontoglires; 4) Âge du nœud à la base de *Mus musculus* et *Rattus norvegicus*.

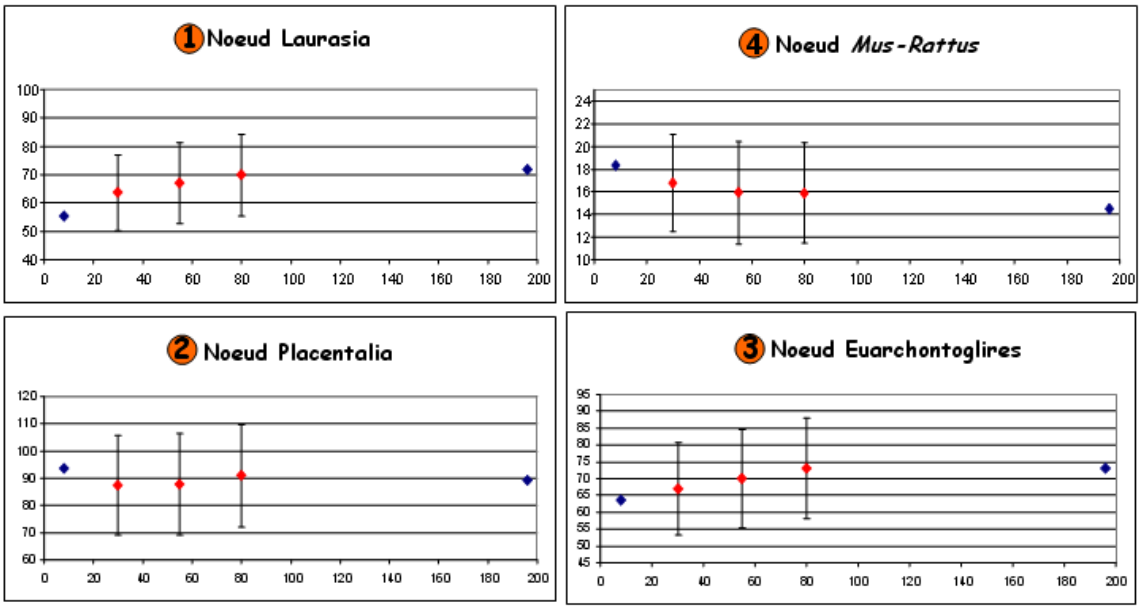

Figure 43 : Datation moleculaire sur le jeu de données C. L'intervalle de calibration a été effectué pour *Homo sapiens* et *Pan paniscus* (5 à 7 M.A). On retrouve sur la figure l'âge des nœuds estimés (axe-Y) avec l'écart-type (barre d'erreur) en fonction de l'augmentation du nombre d'espèces (axe-X). 1) Âge du nœud à la base des Laurasia; 2) Âge du nœud à la base des placentaires; 3) Âge du nœud à la base des Euarchontoglires; 4) Âge du nœud à la base de *Mus musculus* et *Rattus norvegicus*.

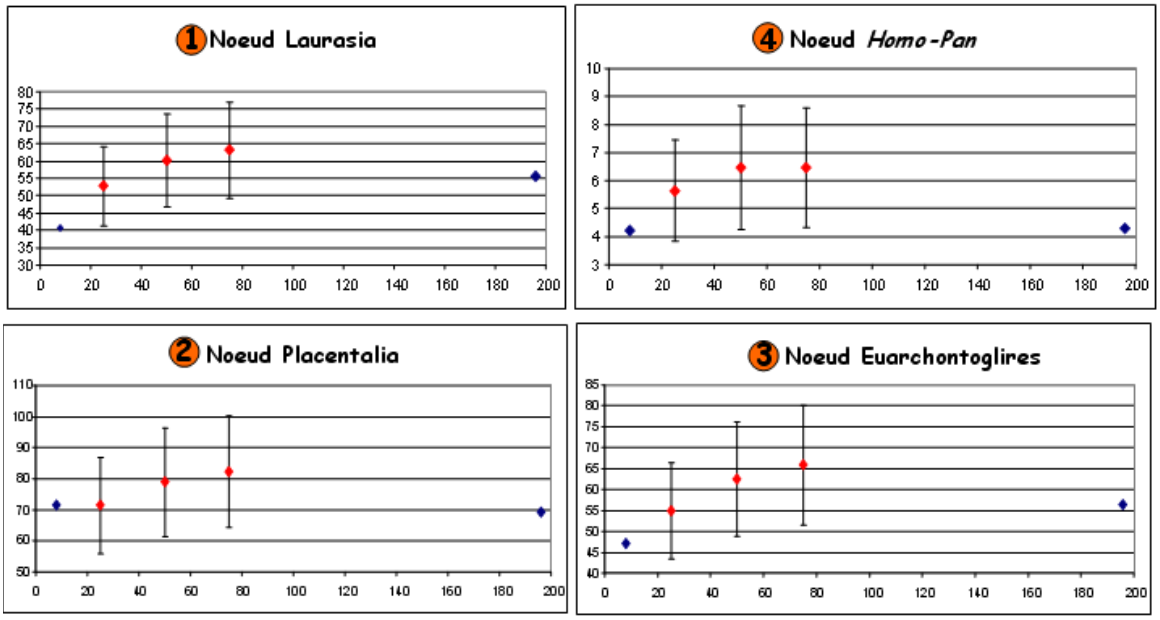

Figure 44 : Datation moleculaire sur le jeu de données D. L'intervalle de calibration a été effectué pour *Mus musculus* et *Rattus norvegicus* (7 à 16 M.A). On retrouve sur la figure l'age des nœuds estimés (axe-Y) avec l'écart-type (barre d'erreur) en fonction de l'augmentation du nombre d'espèces (axe-X). 1) Âge du nœud à la base des Laurasia; 2) Âge du nœud à la base des placentaires; 3) Âge du nœud à la base des Euarchontoglires; 4) Âge du nœud à la base de *Homo sapiens* et *Pan paniscus*.

IV. CONCLUSION & PERSPECTIVES

Le problème de l'estimation des longueurs de branche est rarement discuté lorsqu'il s'agit d'évaluer la fiabilité des datations moléculaires. Cependant, il est connu que les substitutions multiples peuvent avoir, quand elles sont trop fréquentes, des effets dévastateurs sur l'inférence phylogénétique (par exemple, les Microsporidies ont été positionnées à la base des Eucaryotes alors qu'elles appartiennent aux Champignons (Cavalier-Smith 1983)). Il est donc attendu qu'une mauvaise prise en compte des substitutions multiples puisse avoir un impact important sur les longueurs de branche et donc sur les dates moléculaires.

Pour bien mesurer l'ampleur de la sous-estimation des longueurs de branche, il fallait disposer d'un marqueur génétique bien saturé. Notre choix du génome mitochondrial des Mammifères s'est révélé pertinent, car plusieurs résultats (section 3.2, 3.5 et 3.6) ont montré la présence de nombreuses substitutions multiples. Au niveau des acides aminés, leur sous-estimation a une ampleur importante ($\sum LB_5^5$ = 0,73 substitutions par site pour le maximum de parcimonie mais $\sum LB_5^{196}$ = 2,66 pour CATGTR+Γ en bayésien), soit un facteur d'environ 3,5), qui justifie notre étude. En fait, le niveau de saturation au niveau de la troisième position du codon était tellement important que l'on a rencontré des problèmes numériques ne nous permettant pas de poursuivre l'analyse. Il serait donc important d'analyser d'autres jeux de données pour vérifier si les mêmes tendances sont observées, quel que soit le niveau de saturation du marqueur génétique utilisé.

On s'attendait à ce qu'il soit plus facile d'analyser l'impact de la sous-estimation des substitutions multiples sur les longueurs de branche que sur les dates, car les dates sont déduites des longueurs de branche en utilisant une méthode complexe (par exemple basée sur l'autocorrélation des taux). Nous avons effectivement observé des tendances presque toujours claires dans le cas des longueurs de branche (sauf CAT+Γ avec peu d'espèces et pour la troisième position du codon (NT3) pour les analyses

nucléotidiques), alors que, pour les datations, le simple fait de changer la calibration pouvait changer l'impact de l'échantillonnage taxonomique.

Pour les longueurs de branche, on a toujours une augmentation des valeurs de $\sum LB_5$ avec l'augmentation du nombre d'espèces et l'amélioration du modèle. Les résultats des analyses en maximum de vraisemblance montrent que l'augmentation du nombre d'espèces a un impact beaucoup plus grand sur l'estimation d'un nombre supplémentaire de substitutions par site le long des branches par rapport à l'amélioration du modèle, comme par exemple l'ajout d'une distribution gamma. On a effectivement un écart beaucoup plus grand entre l'estimation du nombre de substitutions à peu d'espèces par rapport à un nombre élevé d'espèces peu importe le modèle d'évolution choisi comparé à l'écart observé pour cette estimation entre les différents modèles. L'analyse individuelle de chaque branche a montré que l'augmentation de la valeur du sous-arbre est relié au fait que chacune des branches de l'arbre est brisé par l'ajout d'espèces, ce qui permet alors de détecter des substitutions supplémentaires. Nous avons pu démontrer également que l'augmentation de la valeur de chacune des branches de l'arbre de référence est corrélée aux nombres d'espèces ajoutées dans le groupe taxonomique en question. Il est donc important lorsqu'on veut détecter le nombre de substitutions le long d'une branche pour une espèce en particulier, d'ajouter de façon attentionnée des espèces qui vont briser cette branche de manière régulière.

Néanmoins, il faut noter que le modèle CAT+Γ fournit avec 5 espèces des longueurs de branche beaucoup plus grandes que presque tous les autres modèles, même avec 196 espèces (sauf CATGTR+Γ). Il est très probable qu'il s'agisse d'une sur-estimation par le modèle CAT. Il faut néanmoins noter que la variance dans l'estimation de $\sum LB_5^5$ est très grande (0,33), montrant que le modèle CAT indique de lui-même qu'il n'est pas vraiment apte à estimer les longueurs de branche avec autant de saturation et si peu d'information. Il faut en effet être vigilant vis-à-vis des substitutions inférées par les modèles, car, comme ceux-ci peuvent être mal spécifiés, ils peuvent sur-estimer, et non seulement sous-estimer certains paramètres comme les longueurs de branche. C'est donc le grand avantage d'utiliser le maximum de parcimonie, qui ne peut pas sur-

estimer le nombre total de changements. L'augmentation observée de $\sum LB_5$ avec l'augmentation du nombre d'espèces dans ce cas est donc indiscutable. L'analyse d'alignements avec plus d'espèces (250 environ sont maintenant disponibles) est donc une piste à suivre puisque nous n'avions toujours pas atteint un plateau avec 196 espèces, ce qui suggère que de nouvelles substitutions peuvent être détectées avec plus d'espèces.

Il est aussi important de regarder les relations, et si possible les synergies, entre l'augmentation du nombre d'espèces et l'amélioration du modèle sur l'estimation des longueurs de branche. Comme on peut le voir dans le Tableau 13, ce n'est pas parce qu'une méthode détecte beaucoup de changements avec 5 espèces qu'elle ne continue pas à en découvrir plus à 196. En particulier, le modèle CAT-GTR+Γ a la plus grande longueur à 5 (1.50), mais aussi un des plus grands ratios $\sum LB_5^{196}$ / $\sum LB_5^5$ (1.77). Seuls les modèles de type CAT-GTR sans distribution Γ ont des ratios plus élevés. Cela suggère fortement qu'il existe une bonne synergie entre amélioration du modèle et augmentation du nombre d'espèces dans ce cas. À l'inverse, la distribution Γ est très efficace avec 5 espèces, mais beaucoup moins à 196 (comparer les lignes avec et sans Γ dans le Tableau 13). Cela est probablement dû au fait qu'avec 196 espèces le fait qu'un site évolue rapidement apparaisse très facilement à la simple vue de l'alignement.

	$\sum LB_5^5$	$\sum LB_5^{196} / \sum LB_5^5$
Poisson	0.79	1.62
Mtrev	0.85	1.54
Wag	0.84	1.58
GTR	0.84	1.58
CAT	1.38	1.42
CAT-mtrev	1.07	1.86
CAT-wag	0.94	2.05
CAT-GTR	1.13	2.20
Poisson+Γ	1.00	1.39
Mtrev+Γ	1.11	1.36
Wag+Γ	1.09	1.39
GTR+Γ	1.20	1.52
CAT+Γ	2.35	0.86
CAT-mtrev+Γ	1.39	1.43
CAT-GTR+Γ	1.50	1.77

Tableau 13 : Valeurs des $\sum LB_5^5$ ainsi que le rapport $\sum LB_5^{196} / \sum LB_5^5$ pour différents modèles avec le programme Phylobayes.

L'impact de la saturation sur les datations est plus complexe à interpréter. Théoriquement, les dates doivent être plus fiables si les longueurs de branche sont bien estimées. Le cas de la divergence *Mus musculus*/*Rattus norvegicus* semble confirmer cette idée. Néanmoins, la complexité des méthodes de datation peut faire que des erreurs sur les longueurs de branche peuvent se compenser ou à l'inverse qu'une erreur, plus petite mais mal positionnée dans l'arbre, a un effet déstabilisateur. La très grande variance observée en faisant varier l'échantillonnage taxonomique illustre clairement ces difficultés. Nous avons donc voulu regarder si le modèle avait aussi une grande influence sur les dates. Nous avons donc analysé sept modèles différents avec 196 espèces, la calibration *Equus caballus/Rhinoceros unicornis*, l'autocorrélation log-normale et le logiciel Phylobayes. Le Tableau 14 reporte la valeur moyenne de recouvrement de l'intervalle de confiance des dates inférées pour tous les nœuds internes. Pour chaque modèle, deux chaînes indépendantes ont été analysées et ont toujours un meilleur recouvrement que quand on compare deux modèles, avec des valeurs entre 0,88 et 0,92. La précision est suffisante pour voir que le modèle d'évolution des séquences semble avoir une influence beaucoup moins grande sur les

datations que l'échantillonnage taxonomique. Néanmoins, il faudrait calculer cela, même si ce n'est faisable que sur 4 ou 5 nœuds, et bien vérifier que le recouvrement de l'intervalle de confiance est plus petit pour l'échantillonnage taxonomique. Le contraste avec l'estimation des longueurs de branche est saisissant, car dans ce cas-là les deux approches avaient un impact globalement similaire.

L'inter-relation entre estimation des longueurs de branche et estimation des dates demeurent une question ouverte qu'il faut étudier. Une piste pourrait être de faire des sous échantillons d'espèces en inférant les branches à partir des données ou à partir des longueurs de branche obtenues avec 196 espèces, puis de faire les datations. Cela n'implique de n'utiliser que les longueurs de branche, mais c'est possible avec le logiciel r8s de Sanderson (Sanderson 2003). Un dernier point intéressant du Tableau 14 est que les modèles utilisant une matrice de Poisson (CAT, CAT+Γ et Poisson+Γ) ont un bien meilleur recouvrement entre eux qu'avec ceux ayant une matrice de type GTR. Nous n'avons pas d'explications pour le moment, mais cela pourrait être justifié par l'utilisation de matrices de type GTR, même si elles requièrent beaucoup plus de temps de calcul.

CAT	0,87	0,83	0,85	0,73	0,75	0,72	0,72	0,7	0,67	0,73	0,72	0,89	0,84	0,83	0,81
CAT		0,81	0,83	0,72	0,75	0,73	0,73	0,69	0,66	0,73	0,72	0,88	0,88	0,78	0,78
CAT+Γ			0,88	0,78	0,81	0,78	0,78	0,73	0,7	0,78	0,76	0,85	0,81	0,89	0,88
CAT+Γ				0,78	0,81	0,75	0,75	0,71	0,68	0,75	0,74	0,86	0,83	0,88	0,87
CATGTR+Γ					0,88	0,72	0,73	0,68	0,66	0,7	0,69	0,75	0,74	0,78	0,81
CATGTR+Γ						0,75	0,75	0,69	0,67	0,72	0,71	0,78	0,77	0,8	0,82
GTR+Γ							0,92	0,8	0,78	0,88	0,86	0,75	0,72	0,73	0,74
GTR+Γ								0,81	0,78	0,87	0,86	0,75	0,71	0,74	0,75
MtREV									0,91	0,84	0,85	0,71	0,67	0,71	0,71
MtREV										0,81	0,81	0,68	0,64	0,69	0,69
MtREV+Γ											0,92	0,76	0,71	0,74	0,74
MtREV+Γ												0,75	0,7	0,73	0,73
Poisson													0,88	0,83	0,83
Poisson														0,8	0,8
Poisson+Γ															0,88

Tableau 14 : Moyenne de recouvrement de l'intervalle de confiance des dates inférées pour tous les nœuds internes avec différents modèles d'évolution de séquences.

Comment continuer à améliorer l'estimation des longueurs de branche, à part en augmentant le nombre de taxons (mais cela n'est souvent pas possible à cause des extinctions) ? Plusieurs pistes d'amélioration de modèles sont déjà disponibles et nous les avons évalués avec 5 et 196 espèces (Figure 45). Il apparaît que l'introduction d'une horloge non-corrélée supposant une distribution gamma des longueurs de branche (Ugam) (Drummond et al. 2006) ou un modèle log-normal (ln) (Thorne et al. 1998), n'a que peu d'effets. Rappelons que cela revient à imposer des contraintes sur la distance de la racine aux feuilles. Mais il ne s'agit pas de contraintes très fortes, le temps et le taux pouvant absorber cet aspect, ce qui explique les résultats. De manière intéressante, modéliser la variation de taux entre sites avec un processus de Dirichlet (DP) (Huelsenbeck et al. 2006) au lieu d'une loi Γ a l'air assez efficace, puisqu'on voit une amélioration de l'estimation des longueurs de branche autant à 5 qu'à 196 espèces. Cela ouvre de bonnes perspectives, car le modèle DP réduit le temps de calcul et l'espace mémoire utilisé de manière importante (par un facteur 4 si la distribution Γ est discrétisées en 4 catégories). Finalement, utiliser le modèle covarion (Tuffley et al. 1998), ou un modèle partititionné (modèle pour lequel on considère 12 longueurs par branche, soit une pour chacun des 12 gènes mitochondriaux) (Yang 1996) permettent de prendre en compte l'hétérotachie[10]. Leur utilisation permet de détecter beaucoup plus de substitutions multiples en particulier le modèle covarion+Γ. Cela n'est pas surprenant. En effet, l'hétérogénéité des taux au cours du temps fait qu'un site peut évoluer très vite dans une partie de l'arbre (d'où de nombreuses substitutions multiples) et être constant ailleurs. Le taux moyen de ce site, utilisé dans les modèles homotaches (par opposition à hétérotaches), ne permettra pas de bien inférer ces substitutions.

Outre le fait de prendre en compte l'hétérotachie, il serait également intéressant pour améliorer la détection de substitutions multiples de considérer d'autres caractéristiques des protéines dans les modèles d'évolution de séquences. Un exemple, serait d'utiliser des modèles qui prennent en compte l'interdépendance des positions des acides aminés dans la protéine (Rodrigue et al. 2006), ce qui permettrait probablement d'inférer des longueurs de branche plus proches de la réalité. Il devrait

[10] Variations de vitesses d'évolution pour un même site au sein d'une phylogénie (sa détection permet d'inférer des variations de pressions de sélection sur ce site au cours du temps et entre lignées).

également être possible grâce aux implémentations de Phylobayes de tester des modèles d'horloge différents, comme le modèle CIR (Lepage et al. 2007).

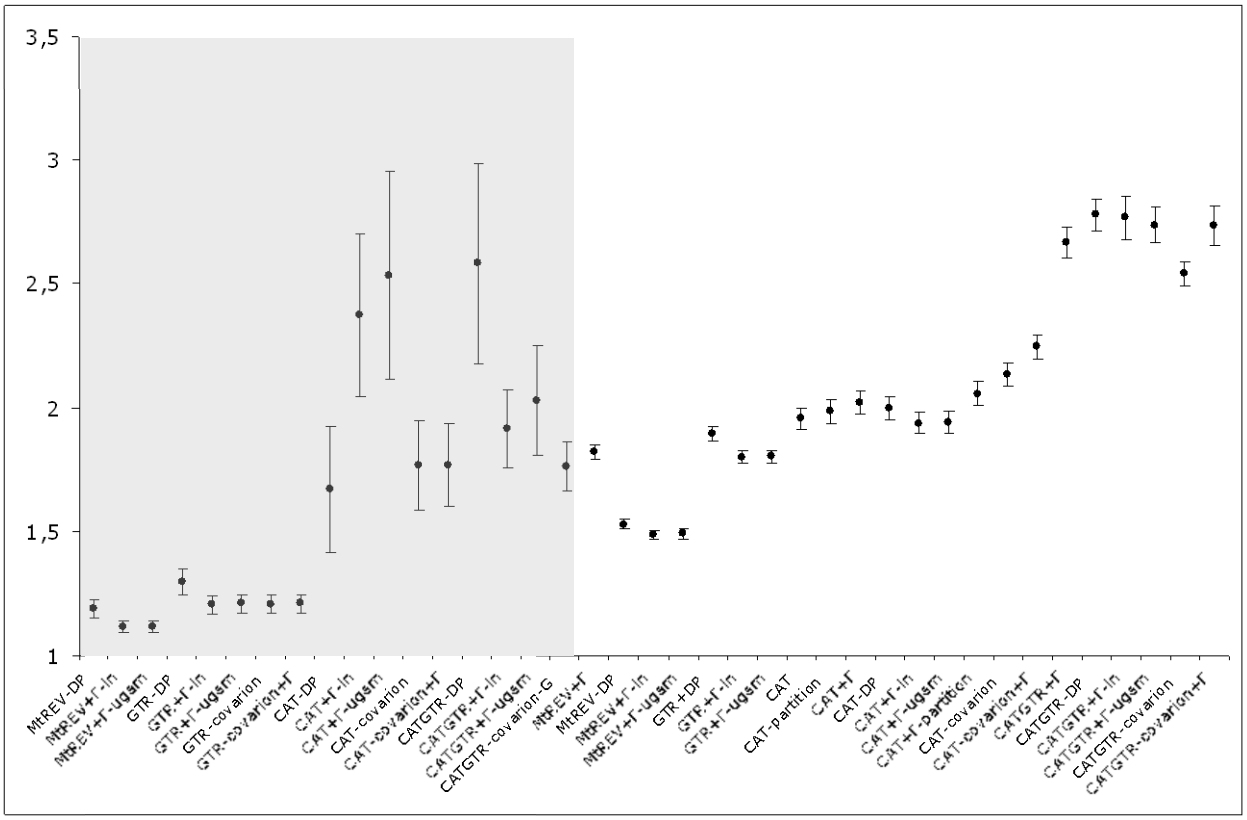

Figure 45 : Valeur de $\sum LB_5^5$ et $\sum LB_5^{196}$ pour différents modèles d'évolution de séquences avec l'écart-type (barre d'erreur). L'axe-X représente les modèles avec Phylobayes tandis qu'on retrouve sur l'axe-Y en gris les valeurs du $\sum LB_5^5$ et en blanc les valeurs pour $\sum LB_5^{196}$.

On voit donc que la marge de progression des modèles est importante. Il faut néanmoins se méfier de la sur-paramétrisation. Par exemple, avec des petits gènes et beaucoup d'espèces, un modèle partitionné aura très peu d'information pour inférer toutes les longueurs de branche de chaque gène. Néanmoins, l'approche bayésienne devrait être assez robuste, même si la convergence des MCMC peut devenir délicate. Cela amène au problème le plus important, celui du temps de calcul. En effet, le meilleur modèle que nous avons utilisé, CAT-GTR-covarion+Γ, est déjà extrêmement demandant, avec plus de 2 mois de calcul pour 196 espèces. Il faut noter qu'il s'agit

d'une analyse phylogénétique facile, puisque la topologie est fixée (cela explique que CAT-GTR-covarion+Γ ne soit pas utilisée pour les inférences phylogénétiques). Le modèle CAT-GTR+covarion+Γ+In demanderait probablement un an de calcul pour fournir des résultats précis. La piste de l'amélioration des modèles est donc difficile.

Notre conseil pour les biologistes intéressés dans les datations moléculaires est donc de surtout chercher à améliorer l'échantillonnage taxonomique (plus de taxons, mais aussi élimination des taxons trop rapides, car plus sujet à homoplasie) et d'ajuster le choix du modèle en fonction du temps de calcul disponible.

BIBLIOGRAPHIE

Adachi, J. and M. Hasegawa (1996). "Model of amino acid substitution in proteins encoded by mitochondrial DNA." J. Mol. Evol. **42**(4): 459-68.

Aguinaldo, A. M., J. M. Turbeville, et al. (1997). "Evidence for a clade of nematodes, arthropods and other moulting animals." Nature **387**(6632): 489-93.

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. Proceedings 2nd International Symposium on Information Theory. Petrov and Csaki. Budapest, Akademia Kiado**:** 267-281.

Alberts, B., A. Johnson, et al. (2002). "Molecular Biology of the Cell." Fourth Edition.

Alexandros, S. (2005). An Efficient Program for Phylogenetic Inference Using Simulated Annealing. Proceedings of the 19th IEEE International Parallel and Distributed Processing Symposium (IPDPS'05) - Workshop 7 - Volume 08, IEEE Computer Society.

Alroy, J. (1999). "The fossil record of North American mammals: evidence for a Paleocene evolutionary radiation." Syst. Biol. **48**: 107-118.

Altschul, S. F., W. Gish, et al. (1990). "Basic local alignment search tool." J Mol Biol **215**(3): 403-10.

Archibald, J. (1999). "Molecular dates and the mammalian radiation." Trends Ecol. Evol. **14**(7): 278.

Archibald, J., L. Bromham, et al. (1999). "Growing up with dinosaurs: molecular dates and the mammalian radiation." Trends Ecol. Evol. **14**: 113-118.

Aris-Brosou, S. and Z. Yang (2002). "Effects of Models of Rate Evolution on Estimation of Divergence Dates with Special Reference to the Metazoan 18S Ribosomal RNA Phylogeny." Syst Biol **51**(5): 703-14.

Aris-Brosou, S. and Z. Yang (2003). "Bayesian Models of Episodic Evolution Support a Late Precambrian Explosive Diversification of the Metazoa." Mol Biol Evol **20**(12): 1947-1954.

Aris-Brosou, S. p. (2007). "Dating Phylogenies with Hybrid Local Molecular Clocks." PLoS ONE **2**(9): e879.

Ayala, F. J. (1999). "Molecular clock mirages." Bioessays **21**(1): 71-5.

Ayala, F. J., A. Rzhetsky, et al. (1998). "Origin of the metazoan phyla: molecular clocks confirm paleontological estimates." Proc. Natl. Acad. Sci. USA **95**(2): 606-11.

Bapteste, E., H. Brinkmann, et al. (2002). "The analysis of 100 genes supports the grouping of three highly divergent amoebae: *Dictyostelium*, *Entamoeba*, and *Mastigamoeba*." Proc. Natl. Acad. Sci. USA **99**(3): 1414-9.

Bayes, T. (1763). "An Essay towards solving a Problem in the Doctrine of Chances." Philosophical Transactions of the Royal Society of London **53**.

Blair, J. E. and S. B. Hedges (2005). "Molecular clocks do not support the Cambrian explosion." Mol Biol Evol **22**(3): 387-90.

Bowring, S., J. Grotzinger, et al. (1993). "Calibrating rates of Early Cambrian evolution." Science **261**: 1293–1298.

Brinkmann, H., M. Giezen, et al. (2005). "An empirical assessment of long-branch attraction artefacts in deep eukaryotic phylogenomics." Syst Biol **54**(5): 743-57.

Brinkmann, H. and H. Philippe (1999). "Archaea sister group of Bacteria? Indications from tree reconstruction artifacts in ancient phylogenies." Mol. Biol. Evol. **16**(6): 817-25.

Brochier, C. (2002). "Phylogénie des eubactéries et mise en evidence des transferts horizontaux." PHD, Université Paris-Sud.

Bromham, L., M. J. Phillips, et al. (1999). "Growing up with dinosaurs: molecular dates and the mammalian radiation." Trends in Ecology and Evolution **14**(3): 113-118.

Bromham, L., A. Rambaut, et al. (1998). "Testing the Cambrian explosion hypothesis by using a molecular dating technique." Proc Natl Acad Sci U S A **95**(21): 12386-9.

Brunet, M., F. Guy, et al. (2002). "A new hominid from the Upper Miocene of Chad, Central Africa." Nature **418**(6894): 145-151.

Burleigh, J. G. and S. Mathews (2004). "Phylogenetic signal in nucleotide data from seed plants: implications for resolving the seed plant tree of life." Am.J. Bot. **91**(10): 1599-1613.

Casella, G. and E. George (1992). "Explaining the Gibbs Sampler." The American Statistician **46**: 167-174.

Castresana, J. (2000). "Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis." Mol Biol Evol **17**(4): 540-552.

Cavalier-Smith, T. (1983). "A 6-kingdom classification and a unified phylogeny." W. Schwemmler and H. E. A. Schenk, eds. Endocytobiology II. de Gruyter, Berlin: 1027-1034.

Conway, M. (2000). "The Cambrian “explosion” : slow-fuse or megatonnage." Proc Natl Acad Sci USA **97**: 4426-9.

Cox, J. C., J. E. Ingersoll, et al. (1985). "A theory of the term structure of interest-rates." Econometrica **53**(2): 385-407.

Cutler, D. J. (2000). "Estimating divergence times in the presence of an overdispersed molecular clock." Mol. Biol. Evol. **17**(11): 1647-60.

Darlu, P. and P. Tassy (1993). "Reconstruction Phylogénétique. Concepts et méthodes." Masson.

Darwin, C. (1859). The origin of species by means of natural selection. London, Murray.

Dayhoff, M. O., R. M. Schwartz, et al. (1978). A model of evolutionary change in proteins. Atlas of Protein Sequences and Structure. M. O. Dayhoff. Washington DC, National Biomedical Research Foundation**:** 345-352.

DeBry, R. (2005). "The systematic component of phylogenetic error as a function of taxonomic sampling under parsimony." Systematic Biology(54): 432-440.

Delsuc, F., H. Brinkmann, et al. (2005). "Phylogenomics and the reconstruction of the tree of life." Nat Rev Genet **6**: 361 - 375.

Delsuc, F. and E. J. P. Douzery (2004). "Les méthodes probabilistes en phylogénie moléculaire. (2) L’approche Bayésienne." Biosystema : « Avenir et pertinence des méthodes d’analyses en phylogénie moléculaire » **22**: 75-86.

Douzery, E. (2002). "Les datations en phylogénie moléculaire : de l'ADN et des protéines jusqu'aux fossiles et réciproquement." HDR, Montpellier.

Douzery, E. J., F. Delsuc, et al. (2006). "[Molecular dating in the genomic era]." Med Sci (Paris) **22**(4): 374-80.

Douzery, E. J., E. A. Snell, et al. (2004). "The timing of eukaryotic evolution: Does a relaxed molecular clock reconcile proteins and fossils?" Proc Natl Acad Sci U S A **101**: 15386-15391.

Drummond, A. J., S. Y. W. Ho, et al. (2006). "Relaxed phylogenetics and dating with confidence." Plos Biology **4**(5): 699-710.

Easteal, S. (1999). "Molecular evidence for the early divergence of placental mammals." Bioessays **21**(12): 1052-1058.

Erwin, D. H., J. W. Valentine, et al. (1997). "The origin of animal body plans." Am. Sci. **85**: 126-137.

Farris, J. (1970). "Methods for computing Wagner trees." Syst. Zool. **19**: 83-92.

Felsenstein, J. (1978). "Cases in which parsimony or compatibility methods will be positively misleading." Syst. Zool. **27**: 401-410.

Felsenstein, J. (1981). "Evolutionary trees from DNA sequences: a maximum likelihood approach." J. Mol. Evol. **17**(6): 368-76.

Felsenstein, J. (1984). "Distance methods for inferring phylogenies: a justification." Evolution Int J Org Evolution **38**: 16-24.

Felsenstein, J. (2001). "Taking variation of evolutionary rates between sites into account in inferring phylogenies." J. Mol. Evol. **53**(4-5): 447-455.

Felsenstein, J. (2004). Inferring phylogenies. Sunderland, MA, USA, Sinauer Associates, Inc.

Fitch, W. (1976). "Molecular evolutionary clocks." Ayala FJ (ed),Molecular evolution **Sinauer Associates, Sunderland**: pp 160-178.

Fitch, W. M. (1971). "Toward Defining the Course of Evolution: Minimum Change for a Specific Tree Topology." Systematic Zoology **20**(4): 406-416.

Fitch, W. M. and J. J. Beintema (1990). "Correcting parsimonious trees for unseen nucleotide substitutions: the effect of dense branching as exemplified by ribonuclease." Mol. Biol. Evol. **7**(5): 438-43.

Fitch, W. M. and M. Bruschi (1987). "The evolution of prokaryotic ferredoxins--with a general method correcting for unobserved substitutions in less branched lineages." Mol. Biol. Evol. **4**(4): 381-94.

Fitch, W. M. and E. Margoliash (1967). "A method for estimating the number of invariant amino acid coding positions in a gene using cytochrome c as a model case." Biochem. Genet. **1**(65).

Fitch, W. M. and E. Margoliash (1968). "The construction of phylogenetic trees. II. How well do they reflect past history?" Brookhaven Symp Biol **21**(1): 217-42.

Fitch, W. M. a. and E. Markowitz (1970). "An improved method for determining codon variability in a gene and its application to the rate of fixation of mutations in evolution." Biochemical Genetics **4**: 579-593.

Foote, M., J. Hunter, et al. (1999). "Evolutionary and preservational constraints on origins of biologic groups: divergence times of eutherian mammals." Science **283**: 1310-1314.

Geman, S. and D. Geman (1984). "Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images." Trans. Pattn. Anal. **6**: 721-741.

Gillespie, J. (1991). "The causes of molecular evolution." Oxford University Press, New York.

Gingerich, P. D. (1986). "Temporal scaling of molecular evolution in primates and other mammals." Mol Biol Evol **3**(3): 205-221.
Goldman, N. (1993). "Simple diagnostic statistical tests of models for DNA substitution." J Mol Evol **37**(6): 650-61.
Goodman M. (1962). " Evolution of the immunologic species specificity of human serum proteins." Hum. Biol. **34**: 104-150.
Goodman M. (1963). "Serological analysis of the systematics of recent hominoids." Hum. Biol. **35**: 377-436.
Graur, D. and W. Li (2000). "Fundamentals of Molecular Evolution." 2nd Ed **Sinauer Associates, Sunderland, MA.**
Graybeal, A. (1994). "Evaluating the Phylogenetic Utility of Genes: A Search for Genes Informative About Deep Divergences Among Vertebrates." Systematic Biology **43**(2): 174-193.
Gribaldo, S. and H. Philippe (2002). "Ancient phylogenetic relationships." Theoretical Population Biology **61**(4): 391-408.
Gu, X. (1998). "Early metazoan divergence was about 830 million years ago." J Mol Evol **47**(3): 369-71.
Gu, X. and W.-H. Li (1992). "Higher rates of amino acid substitution in rodents than in humans." Molecular Phylogenetics and Evolution **1**(3): 211-214.
Hasegawa, M., H. Kishino, et al. (1985). "Dating of the human-ape splitting by a molecular clock of mitochondrial DNA." J Mol Evol **22**(2): 160-74.
Hassanin, A., G. Lecointre, et al. (1998). "The 'evolutionary signal' of homoplasy in protein-coding gene sequences and its consequences for a priori weighting in phylogeny." Comptes Rendus de lAcademie des Sciences Series III Sciences de la Vie **321**: 611-620.
Hastad, O. and M. Bojrkland (1998). "Nucleotide Substitution Models and Estimation of Phylogeny." Mol Biol Evol **15**(11): 1381-1389.
Hastings., W. (1970). "Monte carlo sampling methods using Markov chains and their applications." Biometrika(57): 97-109.
Heath, T., S. Hedtke, et al. (2008). "Taxon sampling and the accuracy of phylogenetic analyses." Journal of Systematics and Evolution **46**(3): 239-257.
Heckman, D. S., D. M. Geiser, et al. (2001). "Molecular Evidence for the Early Colonization of Land by Fungi and Plants." Science **293**(5532): 1129-1133.
Hedges, S., P. Parker, et al. (1996). "Continental breakup and the ordinal diversification of birds and mammals." Nature **381**: 226-229.
Hedges, S. B., J. E. Blair, et al. (2004). "A molecular timescale of eukaryote evolution and the rise of complex multicellular life." BMC Evol Biol **4**: 2.
Hedtke, S. M., T. M. Townsend, et al. (2006). "Resolution of phylogenetic conflict in large data sets by increased taxon sampling." Syst Biol **55**(3): 522-9.
Hendy, M. D. and D. Penny (1989). "A framework for the quantitative study of evolutionary trees." Syst. Zool. **38**: 297-309.
Hillis, D. (1995). "Approaches for assessing phylogenetic accuracy." Systematic Biology(44): 3-16.
Hillis, D. M. (1996). "Inferring complex phylogenies." Nature **383**(6596): 130-1.
Hillis, D. M., J. P. Huelsenbeck, et al. (1994). "Application and accuracy of molecular phylogenies." Science **264**(5159): 671-7.

Hillis, D. M., J. P. Huelsenbeck, et al. (1994). "Hobgoblin of phylogenetics?" Nature **369**(6479): 363-4.

Hirt, R. P., J. M. Logsdon, Jr., et al. (1999). "Microsporidia are related to fungi: evidence from the largest subunit of RNA polymerase II and other proteins." Proc Natl Acad Sci U S A **96**(2): 580-585.

Ho, S. and G. Larson (2006). "Molecular clocks: when times are a-changin." Trends Genet.(22): 79-83.

Ho, S., M. Phillips, et al. (2005). "Accuracy of rate estimation using relaxed-clock models with a critical focus on the early metazoan radiation." Mol Biol Evol. **22**: 1355-1363.

Huelsenbeck, J. P. (1995). "The robustness of two phylogenetic methods: four-taxon simulations reveal a slight superiority of maximum likelihood over neighbor joining." Mol Biol Evol **12**(5): 843-9.

Huelsenbeck, J. P. and D. M. Hillis (1993). "Success of phylogenetic methods in the four-taxon case." Syst Zool **42**: 247-264.

Huelsenbeck, J. P., S. Jain, et al. (2006). "A Dirichlet process model for detecting positive selection in protein-coding DNA sequences." Proc Natl Acad Sci U S A **103**(16): 6263-8.

Huelsenbeck, J. P., B. Larget, et al. (2000). "A compound poisson process for relaxing the molecular clock." Genetics **154**(4): 1879-1892.

Huelsenbeck, J. P. and F. Ronquist (2001). "MRBAYES: Bayesian inference of phylogenetic trees." Bioinformatics **17**(8): 754-755.

Huelsenbeck, J. P., F. Ronquist, et al. (2001). "Bayesian inference of phylogeny and its impact on evolutionary biology." Science **294**(5550): 2310-2314.

Hurvich, C. and C. Tsai (1995). "Model Selection for Extended Quasi-Likelihood Models in Small Samples." Biometrics **51**(3): 1077-1084.

Jacobs, L. and D. Pilbeam (1980). "Of mice and men: fossil-based divergence dates and molecular "clocks."." J. Hum. Evol.(9): 551-555.

Janecek, L., R. Honeycutt, et al. (1996). "Mitochondrial gene sequences and the molecular systematics of the artiodactyl subfamily Bovinae." Mol Phylogenet Evolution **6**: 107-119.

Jobb, G., A. von Haeseler, et al. (2004). "TREEFINDER: a powerful graphical analysis environment for molecular phylogenetics." BMC Evol. Biol. **4**(1): 18.

Jones, D. T., W. R. Taylor, et al. (1992). "The rapid generation of mutation data matrices from protein sequences." Comput Appl Biosci **8**(3): 275-82.

Jukes, T. H. and C. R. Cantor (1969). Evolution of protein molecules. Mammalian protein metabolism. H. N. Munro. New York, Academic Press**:** 21-132.

Kass, R. and A. Raftery (1995). "Bayes Factors." Journal of the American Statistical Association **90**(430): 773-795.

Kim, J. (1996). "General inconsistency conditions for maximum parsimony: effects of branch lengths and increasing numbers of taxa." Syst. Biol. **45**(3): 363-374.

Kimura, M. (1980). "A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide substitutions." J Mol Evol **16**: 111-120.

Kimura, M. and T. Ohta (1971). " On the rate of molecular evolution." J. Mol. Evol. **1**: 1-17.

Kimura M. (1968). "Evolutionary rate at the molecular level." Nature **217**: 624-626.
Kishino, H., J. L. Thorne, et al. (2001). "Performance of a divergence time estimation method under a probabilistic model of rate evolution." Mol Biol Evol **18**(3): 352-61.
Kolaczkowski, B. and J. W. Thornton (2007). "Effects of Branch Length Uncertainty on Bayesian Posterior Probabilities for Phylogenetic Hypotheses." Mol Biol Evol **24**(9): 2108-2118.
Korber, B., M. Muldoon, et al. (2000). "Timing the ancestor of the HIV-1 pandemic strains." Science **288**(5472): 1789-1796.
Kuhner, M. K. and J. Felsenstein (1994). "A simulation comparison of phylogeny algorithms under equal and unequal evolutionary rates." Mol. Biol. Evol. **11**(3): 459-68.
Kumar, S. (2005). "Molecular clocks: four decades of evolution." Nat Rev Genet **6**(8): 654-662.
Kumar, S. and S. B. Hedges (1998). "A molecular timescale for vertebrate evolution." Nature **392**: 917 - 920.
Lanave, C., G. Preparata, et al. (1984). "A new method for calculating evolutionary substitution rates." J Mol Evol **20**(1): 86-93.
Larget, B. and D. Simon (1999). "Markov chain Monte Carlo algorithms for the Bayesian analysis of phylogenetic trees." Mol Biol Evol. **16**: 750-759.
Lartillot, N., H. Brinkmann, et al. (2007). "Suppression of long-branch attraction artefacts in the animal phylogeny using a site-heterogeneous model." BMC Evol Biol **7 Suppl 1**: S4.
Lartillot, N. and H. Philippe (2004). "A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process." Mol. Biol. Evol. **21**(6): 1095-1109.
Lartillot, N. and H. Philippe (2006). "Computing Bayes factors using thermodynamic integration." Syst Biol **55**(2): 195-207.
Lartillot, N. and H. Philippe (2008). "Improvement of molecular phylogenetic inference and the phylogeny of Bilateria." Philos Trans R Soc Lond B Biol Sci **363**: 1463–1472.
Le, S. Q. and O. Gascuel (2008). "An Improved General Amino Acid Replacement Matrix." Mol Biol Evol **25**(7): 1307-1320.
Lecointre, G., H. Philippe, et al. (1993). "Species sampling has a major impact on phylogenetic inference." Mol Phylogenet Evol **2**(3): 205-24.
Lepage, T., D. Bryant, et al. (2007). "A General Comparison of Relaxed Molecular Clock Models." Mol Biol Evol.
Li, C., Chen JY, et al. (1998). "Precambrian sponges with cellular structures. ." Science **279**: 879-882.
Li, P. and J. Bousquet (1992). "Relative-Rate Test for Nucleotide Substitutions between 2 Lineages." Molecular Biology and Evolution **9**(6): 1185-1189.
Li, S., D. Pearl, et al. (2000). "Phylogenetic tree construction using Markov chain Monte Carlo." J Amer Statist Assoc **95**: 493-508.
Lin, J. and M. Nei (1991). "Relative efficiencies of the maximum-parsimony and distance-matrix methods of phylogeny construction for restriction data." Mol. Biol. Evol. **8**(3): 356-65.

Loughran, N., B. O'Connor, et al. (2008). "The phylogeny of the mammalian heme peroxidases and the evolution of their diverse functions." BMC Evolutionary Biology **8**(1): 101.

Mau, B., M. Newton, et al. (1999). "Bayesian phylogenetic inference via Markov chain Monte Carlo methods." Biometrics **55**: 1-12.

Mayrose, I., N. Friedman, et al. (2005). "A Gamma mixture model better accounts for among site rate heterogeneity." Bioinformatics **21 Suppl 2**: ii151-ii158.

Metropolis, N., A. Rosenbluth, et al. (1953). "Equations of state calculations by fast computing machine." Journal Chem. Phys. **21**: 1087-1091.

Moreira, D. and H. Philippe (2000). "Molecular phylogeny: pitfalls and progress." Int Microbiol **3**(1): 9-16.

Page, R. D. M. and E. C. Holmes (1998). Molecular evolution a phylogenetic approach. Oxford, Blackwell Science.

Paradis, E., J. Claude, et al. (2004). "APE: Analyses of Phylogenetics and Evolution in R language." Bioinformatics **20**(2): 289-290.

Peterson, K. J., J. B. Lyons, et al. (2004). "Estimating metazoan divergence times with a molecular clock." Proc Natl Acad Sci U S A **101**(17): 6536-41.

Philippe, H., N. Lartillot, et al. (2005). "Multigene analyses of bilaterian animals corroborate the monophyly of Ecdysozoa, Lophotrochozoa, and Protostomia." Mol Biol Evol **22**(5): 1246-53.

Philippe, H. and J. Laurent (1998). "How good are deep phylogenetic trees?" Curr Opin Genet Dev **8**(6): 616-623.

Phillips, M. J., F. Delsuc, et al. (2004). "Genome-scale phylogeny and the detection of systematic biases." Mol Biol Evol **21**: 1455 - 1458.

Poe, S. (2003). "Evaluation of the strategy of long-branch subdivision to improve the accuracy of phylogenetic methods." Syst Biol **52**(3): 423-8.

Poe, S. and D. L. Swofford (1999). "Taxon sampling revisited." Nature **398**(6725): 299-300.

Pollock, D. and D. Goldstein (1995). "A comparison of two methods for constructing evolutionary distances from a weighted contribution of transition and transversion." Mol Biol. Evol.(12): 713-717.

Pollock, D. D., D. J. Zwickl, et al. (2002). "Increased taxon sampling is advantageous for phylogenetic inference." Syst Biol **51**(4): 664-71.

Pulquerio, M. and R. Nichols (2007). "Dates from the molecular clock: how wrong can we be?" Trends Ecol. Evol.(22): 180-184.

R Development Core Team (2004). "R: a language and environment for statistical computing Version 2.4.1." R Foundation for Statistical Computing **Vienna, Austria**.

Rambaut, A. and L. Bromham (1998). "Estimating divergence dates from molecular sequences." Mol Biol Evol **15**(4): 442-8.

Rannala, B., J. P. Huelsenbeck, et al. (1998). "Taxon sampling and the accuracy of large phylogenies." Syst. Biol. **47**: 702-710.

Robinson, M., M. Gouy, et al. (1998). "Sensitivity of the relative-rate test to taxonomic sampling." Mol. Biol. Evol. **15**(9): 1091-8.

Rodrigue, N., H. Philippe, et al. (2006). "Assessing site-interdependent phylogenetic models of sequence evolution." Mol Biol Evol **23**(9): 1762-75.

Rodriguez-Ezpeleta, N., H. Brinkmann, et al. (2007). "Detecting and overcoming systematic errors in genome-scale phylogenies." Syst Biol **56**(3): 389-99.

Ruiz-Trillo, I., M. Riutort, et al. (1999). "Acoel flatworms: earliest extant bilaterian Metazoans, not members of Platyhelminthes." Science **283**(5409): 1919-23.

Rutschmann, F. (2006). "Molecular dating of phylogenetic trees: A brief review of current methods that estimate divergence times." Diversity & Distributions **12**(1): 35-48.

Saitou, N. and M. Nei (1987). "The neighbor-joining method: a new method for reconstructing phylogenetic trees." Mol. Biol. Evol. **4**(4): 406-25.

Sanders, K. L. and M. S. Y. Lee (2007). "Evaluating molecular clock calibrations using Bayesian analyses with soft and hard bounds." Biology Letters **3**(3): 275-279.

Sanderson, M. J. (1997). "A nonparametric approach to estimating divergence times in the absence of rate constancy." Mol. Biol. Evol. **14**(12): 1218-1231.

Sanderson, M. J. (2002). "Estimating absolute rates of molecular evolution and divergence times: a penalized likelihood approach." Mol Biol Evol **19**(1): 101-9.

Sanderson, M. J. (2003). "r8s: inferring absolute rates of molecular evolution and divergence times in the absence of a molecular clock." Bioinformatics **19**(2): 301-302.

Sanderson, M. J., J. L. Thorne, et al. (2004). "Molecular evidence on plant divergence times." Am. J. Bot. **91**(10): 1656-1665.

Sarich, V. M. and A. C. Wilson (1967). "Immunological time scale for hominid evolution." Science **158**: 1200-1202.

Sarich, V. M. and A. C. Wilson (1973). "Generation time and genomic evolution in primates." Science **179**(78): 1144-7.

Scherer, S. (1989). "The relative-rate test of the molecular clock hypothesis: a note of caution." Mol Biol Evol **6**(4): 436-41.

Schmidt, H. A., K. Strimmer, et al. (2002). "TREE-PUZZLE: maximum likelihood phylogenetic analysis using quartets and parallel computing." Bioinformatics **18**(3): 502-4.

Shaul, S. and D. Graur (2002). "Playing chicken (Gallus gallus): methodological inconsistencies of molecular divergence date estimates due to secondary calibration points." Gene **300**(1-2): 59-61.

Sibley, C. G. and J. E. Ahlquist (1984). "The phylogeny of the hominoid primates, as indicated by DNA-DNA hybridization." J. molec. Evol. **20**: 2-15.

Smith, A. B. (1994). "Systematics and Fossil Records." London: Blackwell Scientific.

Sneath, P. and R. Sokal (1973). "Numerical Taxonomy: The Principles and Practice of Numerical Classification." San Francisco: W. H. Freeman.

Springer, M. S. (1995). "Molecular clocks and the incompleteness of the fossil record." Journal of Molecular Evolution **41**(5): 531-538.

Springer, M. S., W. J. Murphy, et al. (2003). "Placental mammal diversification and the Cretaceous Tertiary boundary." Proceedings of the National Academy of Sciences of the United States of America **100**(3): 1056-1061.

Steiper, M. E., N. M. Young, et al. (2004). "Genomic data support the hominoid slowdown and an Early Oligocene estimate for the hominoidâ€“cercopithecoid divergence." Proceedings of the National Academy of Sciences of the United States of America **101**(49): 17021-17026.

Swofford, D. and W. Maddison (1987). "Reconstructing ancestral character states under Wagner parsimony." Math. Biosci. **87**: 199–229.

Swofford, D. L. (2000). PAUP*: Phylogenetic Analysis Using Parsimony and other methods, Sinauer, Sunderland, MA.

Swofford, D. L., G. J. Olsen, et al. (1996). Phylogenetic inference. Molecular systematics. D. M. Hillis, C. Moritz and B. K. Mable. Sunderland, Sinauer Associates**:** 407-514.

Swofford, D. L., P. J. Waddell, et al. (2001). "Bias in phylogenetic estimation and its relevance to the choice between parsimony and likelihood methods." Syst Biol **50**(4): 525-39.

Tajima, F. (1993). "Simple Methods for Testing the Molecular Evolutionary Clock Hypothesis." Genetics **135**(2): 599-607.

Takahata, N. (2007). "Molecular Clock: An Anti-neo-Darwinian Legacy." Genetics **176**(1): 1-6.

Takezaki, N., A. Rzhetsky, et al. (1995). "Phylogenetic test of the molecular clock and linearized trees." Mol. Biol. Evol. **12**(5): 823-33.

Tamura, K. and M. Nei (1993). "Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees." Mol Biol Evol **10**(3): 512-526.

Tavare, S. (1986). "Some probabilistic and statistical problems in the analysis of DNA sequences." Some mathematical questions in biology - DNA sequence analysis **17**: 57 - 86.

Thorne, J. L. and H. Kishino (2002). "Divergence time and evolutionary rate estimation with multilocus data." Syst. Biol. **51**(5): 689-702.

Thorne, J. L., H. Kishino, et al. (1998). "Estimating the rate of evolution of the rate of molecular evolution." Mol Biol Evol **15**(12): 1647-57.

Tierney, L. (1994). "Markov chains for exploring posterior distributions." Ann Statist. **22**: 1701-1762.

Tuffley, C. and M. Steel (1998). "Modeling the covarion hypothesis of nucleotide substitution." Math Biosci **147**(1): 63-91.

Uzzell, T. and K. Corbin (1971). "Fitting discrete probability distributions to evolutionary events." Science **172**: 1089-1096.

Van de Peer, Y., J. S. Taylor, et al. (2001). "The Ghost of Selection Past: Rates of Evolution and Functional Divergence of Anciently Duplicated Genes." Journal of Molecular Evolution **53**(4): 436-446.

Wang, D. Y., S. Kumar, et al. (1999). "Divergence time estimates for the early history of animal phyla and the origin of plants, animals and fungi." Proc R Soc Lond B Biol Sci **266**(1415): 163-71.

Welch, J. J. and L. Bromham (2005). "Molecular dating when rates vary." Trends in Ecology & Evolution **20**(6): 320-327.

Welch, J. J., E. Fontanillas, et al. (2005). "Molecular dates for the "cambrian explosion": the influence of prior assumptions." Syst Biol **54**(4): 672-8.

Whelan, S. and N. Goldman (2001). "A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach." Mol Biol Evol **18**(5): 691-9.

Wilson, A. C. and V. M. Sarich (1969). "A molecular time scale for human evolution." Proc. Natl. Acad. Sci. USA **63**(4): 1088-93.

Wray, G. A., J. S. Levinton, et al. (1996). "Molecular evidence for deep precambrian divergences among metazoan phyla." Science **274**: 568-573.

WU, C. and W. LI (1985). "Evidence for higher rates of nucleotide substitution in rodents than in man." Proc. Natl. Acad. Sci. USA **82**: 1741-1745.

Xiao, S., Y. Zhang, et al. (1998). "Three-dimensional preservation of algae and animal embryos in a Neoproterozoic phosphorite." Nature (London) **391**: 553-558.

Yang, Z. (1993). "Maximum-likelihood estimation of phylogeny from DNA sequences when substitution rates differ over sites." Mol Biol Evol **10**(6): 1396-401.

Yang, Z. (1996). "Among-site rate variation and its impact on phylogenetic analyses." Trends Ecol Evol **11**: 367-370.

Yang, Z. (1996). "Maximum-likelihood models for combined analyses of multiple sequence data." J. Mol. Evol. **42**: 587-596.

Yang, Z. (1997). Phylogenetic Analysis by Maximum Likelihood (PAML), Version 1.3, Department of Integrative Biology, University of California at Berkeley.

Yang, Z. (2004). "A heuristic rate smoothing procedure for maximum likelihood estimation of species divergence times." Acta Zool Sinica **50**: 645-656.

Yang, Z., R. Nielsen, et al. (1998). "Models of amino acid substitution and applications to mitochondrial protein evolution." Mol Biol Evol **15**(12): 1600-11.

Yang, Z. and B. Rannala (1997). "Bayesian phylogenetic inference using DNA sequences: a Markov Chain Monte Carlo Method." Mol Biol Evol **14**(7): 717-24.

Yang, Z. and B. Rannala (2005). "Branch-length prior influences Bayesian posterior probability of phylogeny." Systematic Biology **54**(3): 455-470.

Yang, Z. and B. Rannala (2006). "Bayesian Estimation of Species Divergence Times Under a Molecular Clock Using Multiple Fossil Calibrations with Soft Bounds." Mol Biol Evol **23**(1): 212-226.

Yang, Z. and A. Yoder (2003). "Comparison of likelihood and Bayesian methods for estimating divergence times using multiple gene loci and calibration points, with application to a radiation of cute-looking mouse lemur species." Syst Biol(52): 705-716.

Yoder, A. D. and Z. Yang (2000). "Estimation of primate speciation dates using local molecular clocks." Mol Biol Evol **17**(7): 1081-90.

Zeng, L.-W., J. Comeron, et al. (1998). "The molecular clock revisited: the rate of synonymous vs. replacement change in Drosophila." Genetica **102-103**(0): 369-382.

Zhou, Y., N. Rodrigue, et al. (2007). "Evaluation of the models handling heterotachy in phylogenetic inference." BMC Evol Biol **7**(1): 206.

Zuckerkandl, E. and L. Pauling (1965). "Molecules as documents of evolutionary history." J Theor Biol **8**(2): 357-66.

Zwickl, D. J. and D. M. Hillis (2002). "Increased taxon sampling greatly reduces phylogenetic error." Syst Biol **51**(4): 588-98.