Université de Montréal

Estimation des corrélations phylogénétiques entre paramètres d'évolution moléculaire et traits d'histoire de vie

par
Raphaël Poujol

Département de biochimie
Faculté de médecine

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

2012,

© Raphaël Poujol,

.

Université de Montréal
Faculté des études supérieures

Ce mémoire intitulé:

Estimation des corrélations phylogénétiques entre paramètres d'évolution moléculaire et traits d'histoire de vie

présenté par:

Raphaël Poujol

a été évalué par un jury composé des personnes suivantes:

Sylvie Hamel,	président-rapporteur
Nicolas Lartillot,	directeur de recherche
Mathieu Blanchette,	membre du jury

Mémoire accepté le: . . . . . .

RÉSUMÉ

Depuis quelques années, l'évolution moléculaire cherche à caractériser les variations de l'intensité de la sélection grâce au rapport entre taux de substitution synonyme et taux de substitution non-synonyme (dN/dS). Cette mesure, dN/dS, a permis d'étudier l'histoire de la variation de l'intensité de la sélection au cours du temps ou de détecter des épisodes de la sélection positive. Nous verrons qu'elle est influencée par les variations de taille et de structure populationnelle au cours du temps. Les méthodes comparatives, quant a elle, permettent de mesurer les corrélations entre caractères quantitatifs le long d'une phylogénie. Elles sont également utilisées pour tester des hypothèses sur l'évolution corrélée des traits d'histoire de vie, mais pour être employées pour étudier les corrélations entre traits d'histoire de vie, masse, taux de substitution où dN/dS.

Nous proposons ici une approche combinant une méthode comparative basée sur le principe des contrastes indépendants et un modèle d'évolution moléculaire, dans un cadre probabiliste bayésien. Nous reconstruisons le long d'une phylogénie, les valeurs ancestrales des traits et de dN/dS. Nous estimons conjointement les covariances entre traits ainsi qu'entre traits et paramètres du modèle d'évolution moléculaire. L'idée est ainsi de pouvoir tester des hypothèses portant sur les liens entre un gène et, par exemple, la longévité. Nous avons ensuite proposé de prendre simultanément en compte un grand nombre de gènes et de mesurer les covariances entre chacun d'eux et chaque trait quantitatif du modèle. Dans ce but, un modèle hiérarchique a été implémenté dans le cadre de *coevol* [58], il permet de mesurer précisément les paramètres communs a tous les gènes. Un travail de parallélisation des calculs donne la liberté d'augmenter la taille du modèle jusqu'à l'échelle du génome.

Nous étudions ici les placentaires, pour lesquels beaucoup de génomes complets et de mesures phénotypiques sont disponibles. À la lumière des théories sur les traits d'histoire de vie, notre méthode devrait permettre de caractériser l'implication de groupes de gènes dans les processus biologiques liés aux phénotypes étudiés.

Mots clés: inférence bayésienne, phylogénie, modélisation.

ABSTRACT

In recent years, molecular evolution sought to characterize the variation and intensity of selection through the ratio between non-synonymous and synonymous substitution rates (dN/dS). The dN/dS measure was either used to study the history of the variation of the intensity of selection over time or to detect episodes of positive selection. The correlation between selection and variation of the effective population size interferes with these measurements. This is addressed by the comparative method that can model these correlations between quantitative traits along a phylogeny. It is also used to test the hypotheses of correlated evolution of life history traits, like the body mass, or the substitution rate. We propose an approach combining the comparative method based on the principle of independent contrasts and a model of molecular evolution in a Bayesian probabilistic framework. By integrating along a phylogeny both ancestral reconstructions of life history traits and molecular parameters of the model we are able to estimate the covariance among these. Using the previously developed software coevol, a hierarchical model was built. This model allows the simultaneous analysis of multiple genes within a single model. Parallel calculations allow increasing the size of the model to the genome scale. We studied those placental mammals, for which complete genomes and phenotypic data were available. Based on the main theories of life history traits, we expect our method to be able to characterize the association of groups of genes to the studied phenotypes.

Keywords ageing, Bayesian, phylogenetic, modelling.

TABLE DES MATIÈRES

LISTE DES TABLEAUX

LISTE DES FIGURES

LISTE DES ANNEXES

LISTE DES SIGLES

ADN	acide désoxyribonucléique
ARN	acide ribonucléique
ATP	Adénosine triphosphate
DAG	Graphe Acyclique Dirigé
MLSP	Maximum Life Span Potential
Mas	Masse corporelle
Lon	Longévité
Gen	Temps de génération
MCMC	Markov Chain Monte Carlo

NOTATION

dN	taux de substitution non synonyme
dS	taux de substitutions synonymes
ω	ratio de dN et de dS
N_e	Taille de population efficace

à moi même

REMERCIEMENTS

Je tiens à remercier sans aucune mesure mon superviseur, Nicolas Lartillot qui a su admirablement me pousser au travail avec simplicité. Il a aussi su faire grandir ma curiosité, et m'a fait aimer ce domaine étonnant qu'est l'évolution moléculaire. Il est parvenu à m'amener au bout de ce diplôme, et je ne l'oublierais pas.

Je remercie aussi tous les gens du laboratoire Robert Cedergren qui ont accompagné ces années, avec qui les discussions ont été toujours enrichissantes, plus particulièrement Hervé Philippe et tous ses collaborateurs, et sans oublier ma collègue, Sahar à qui je souhaite le meilleur pour la suite.

Merci aussi à notre incroyable secrétaire Elaine Meunier, et au fond de recherche sans qui rien n'aurait été possible.

Enfin je tiens à citer ces gens qui ont comptés dans ma vie à Montréal, d'abord ceux que j'oublie, puis Bruno, Stéphanie, Caroline, Cassandre, Hanno, Roxane, Mathieu, Marie, la famille Jeanson, Marco, Elsa, William et Florence.

AVANT-PROPOS

Ce travail, effectué dans le laboratoire de Nicolas Lartillot, fut le développement d'une méthode basée sur plusieurs champs de la biologie. Le domaine d'étude est principalement l'évolution moléculaire, c'est-à-dire l'étude des mécanismes génétiques sous-jacents à la diversification du vivant sur une échelle évolutive large. Le but est ici de faire se rencontrer les génétiques populationnelles et évolutives, examinant le même phénomène, mais à des échelles temporelles différentes. Ce champ d'études est combiné à celui de l'analyse comparative, qui cherche à comprendre les causes des variations des caractères phénotypiques entre espèces.

Plus précisément, la méthode développée ici permet d'estimer les corrélations entre différents processus variant le long des lignages. Il y a deux types de processus : d'une part les caractères continus, ou phénotypes, et, d'autre part, les paramètres de l'histoire substitutionnelle des séquences. La méthode intègre donc un modèle à codons, reconstituant différentes histoires de substitutions de codons le long des lignages. Ce modèle comprend aussi des paramètres dont les variations le long des lignages sont des processus continus, analogues aux processus représentant des phénotypes. Il est donc possible d'estimer les corrélations entre ces différents processus afin de comprendre le lien entre l'évolution de certains gènes et les variations de certains phénotypes. La méthode se base sur une topologie fixe d'un arbre phylogénétique, un alignement de séquences codantes et une série de caractères dont les valeurs sont connues pour les espèces contemporaines. Ma maîtrise est d'une part la participation à l'élaboration de cette méthode, d'autre part des modifications au modèle et sa parallélisation dans un cadre de modèle probabiliste.

Les caractères phénotypiques étudiés sont ceux dont la variabilité entre espèces éveille la curiosité des scientifiques. Dans un premier temps, nous nous sommes intéressés aux mécanismes impliqués dans le vieillissement des mammifères, en étudiant l'évolution de la longévité sur leur phylogénie. L'hypothèse est ici de caractériser les gènes utiles au maintien de l'organisme, en comparant les variations de la sélection purificatrice sur les gènes avec les variations de la longévité. Finalement, les résultats présentés ici concernent aussi la masse et le temps de génération et se concentrent sur la sous-classe des placentaires.

Du fait de la complexité du sujet, ce mémoire est composé de fragments divers, mais tous nécessaires à la compréhension du thème de ce travail. Nous introduirons dans un premier temps les méthodes comparatives, puis discuterons rapidement des théories existantes

sur l'évolution des traits d'histoire de vie et de la masse corporelle. Une explication des théories neutralistes de l'évolution permettra de comprendre l'intérêt de l'utilisation de dN/dS comme une mesure de sélection purificatrice. Nous expliquerons ensuite les modèles probabilistes et le cadre Bayésien par chaîne de Markov Monte-Carlo (M.C.M.C.) dans lesquels nous travaillons. Enfin, nous définirons formellement les différents modèles développés. Les résultats permettront de discuter de la pertinence des différents modèles.

CHAPITRE 1

INTRODUCTION

1.1 Méthodes comparatives

Nos connaissances sur l'évolution se sont forgées grâce à la comparaison des espèces entre elles. Par exemple, la diversité des tailles de becs chez les pinsons des Galápagos fut un des éléments permettant à Charles Darwin d'établir sa théorie de l'évolution. L'analyse des différences et des similitudes entre les espèces a permis l'étude des relations de parentés entre les êtres vivants, d'abord à partir de caractères morphologiques, puis des données moléculaires.

Les méthodes comparatives s'intéressent, indépendamment de la reconstruction de l'arbre du vivant, à la manière dont les caractères phénotypiques évoluent. Les méthodes comparatives permettent d'étudier des tendances globales dans un groupe d'espèces, par exemple si un trait d'histoire de vie à une valeur d'équilibre ou si certains changements ont eu lieu de manière graduelle ou abrupte. On peut aussi s'intéresser aux estimations des valeurs ancestrales de certains traits. D'autre part, les relations phylogénétiques entre les caractères permettent de formuler des hypothèses sur les relations biologiques entre ces caractères. On cherche alors à déterminer les corrélations ou les covariations entre traits d'histoire de vie, le long des lignages des phylogénies.

Les méthodes comparatives [38] formalisent le problème statistique global soulevé par les questions énoncées précédemment. Elles s'intéressent à l'analyse de tous les types de caractères, les caractères discrets comme le nombre de vertèbres, ou les caractères quantitatifs, dits continus, comme la taille du tibia. Étant donné deux caractères quantitatifs, on cherche à connaître leur corrélation, à la fois en direction et en significativité.

En biologie animale, de nombreuses corrélations prennent la forme d'une allométrie. Une allométrie est une relation entre un trait et la masse corporelle prenant la forme d'une loi de puissance. Par exemple, si M représente la masse, il existe une allométrie pour un trait d'histoire de vie X s'il existe un α tel que :

$$X = nM^{\alpha}, \tag{1.1}$$

avec n réel quelconque, que l'on notera dans ce mémomire :

$$X \sim M^{\alpha}. \tag{1.2}$$

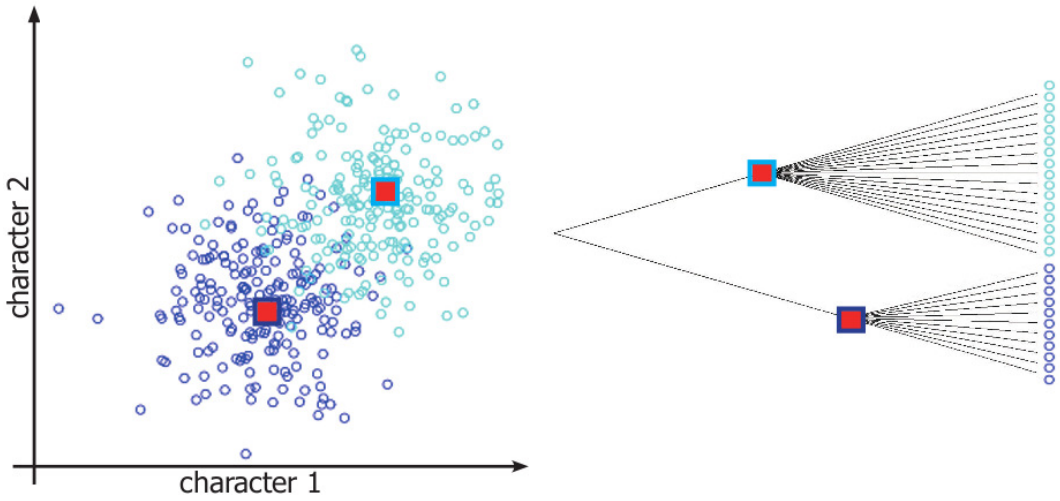
On appelle isométrie le cas où $\alpha = 1$. Constater des allométries permet de mettre en place et de tester des hypothèses sur les processus évolutifs. Une allométrie phylogénétique positive ($\alpha > 0$) se traduirait par des variations allant dans le même sens le long des lignages d'une phylogénie entre un trait d'histoire de vie et la masse corporelle, c'est-à-dire que, les variations de M et de X ont le plus souvent le même signe, pas nécessairement la même amplitude. Inversement, il existe des allométries négatives ($\alpha < 0$) comme entre le taux métabolique et la masse chez les vertébrés [10]. Le taux métabolique est ici la quantité d'énergie dépensée par unité de temps et par cellule.

Plus généralement, dans ce travail, on s'intéressera à des lois d'échelle $X \sim Y^{\alpha}$ entre traits quantitatifs continus. Par abus de langage, nous parlerons ici aussi d'allométrie entre Y et X, même si Y n'est pas forcement la masse corporelle. Nous utiliserons parfois la transformée logarithmique pour des questions pratiques : $\log X \sim \alpha \log Y$.

Les premières méthodes comparant les données physiologiques entre espèces ne se préoccupaient pas des relations phylogénétiques entre celles-ci. Elles utilisaient les caractères disponibles pour tirer des conclusions, mais ont toujours tacitement considéré les observations comme statistiquement indépendantes [10]. En fait, les ensembles d'individus ont des relations de dépendance du fait de leur histoire évolutive. L'intégration de cette histoire dans les méthodes comparatives s'est révélée essentielle [30] [31].

En 1985, Joseph Felsenstein publie "Phylogenies and the comparative method". Dans cet article, il pointe l'importance de prendre en compte la non-indépendance statistique des espèces contemporaines ou l'*inertie phylogénétique*. Une part de la ressemblance entre deux espèces reflète la conséquence de l'histoire commune de ces deux espèces. Il donne, dans cet article, l'exemple de la figure 1.1 où une corrélation apparemment significative est le seul fait de la structure hiérarchique de la phylogénie. À l'époque, les méthodes conventionnelles d'analyse statistique surestimaient la significativité des résultats. La remarque de Felsenstein fut prise comme une attaque sérieuse et son message fut taxé de "nihilisme". Finalement, il a ouvert un nouveau champ d'études qui touche maintenant jusqu'à l'évolution moléculaire.

Figure 1.1 – Cet exemple montre le cas particulier de deux hypothétiques traits d'histoire de vie évoluant de manière indépendante et continue le long des branches d'une phylogénie très particulière et montrant une corrélation induite seulement par l'inertie phylogénétique (adaptation libre de Felsenstein 25)



Dans ce même article, Joseph Felsenstein propose la première méthode prenant en compte l'inertie phylogénétique appelée la méthode des *contrastes indépendants* [25]. Cette idée essentielle a servi de base à la plupart des méthodes comparatives ultérieures. La méthode des contrastes indépendants s'applique sur des traits quantitatifs, c'est-à-dire qui varient de façon continue. Intuitivement, le principe est de reconstruire les valeurs des caractères étudiés le long de l'arbre, puis, de considérer les variations le long de chaque branche indépendamment. Si les valeurs des caractères des espèces modernes sont corrélées, les variations des caractères sur chacune des branches peuvent, elles, être considérées comme des quantités indépendantes.

En réalité, la méthode ne s'applique pas sur une reconstruction spécifique, mais intègre sur tous les scénarios possibles, pondérés par leur probabilité sous un modèle d'évolution de trait. En pratique, on fait l'hypothèse du mouvement brownien. Un mouvement brownien est un processus diffusif qui représente la limite de toute marche aléatoire non biaisée dont tous les pas élémentaires sont de variance finie. C'est un cadre général, et plutôt neutre, dans lequel on peut décrire des variables continues comme les traits d'histoire de vie. Il est utilisé en phylogénie pour sa simplicité mathématique. Plus précisément, pour chaque intervalle de temps, le processus subit un incrément indépendant dont la variance dépend de la taille de l'intervalle de temps et d'un paramètre propre au processus, la variance par unité de temps. À chaque instant, la valeur du processus augmente ou diminue avec

la même probabilité, le seul paramètre à estimer pour ce processus est l'amplitude des mouvements, la variance du processus. Ce modèle est détaillé dans la partie 4.3.
En intégrant sur toutes les valeurs possibles, on peut extraire de chaque branche de l'arbre une variation, un contraste pour chaque trait d'histoire de vie. Ces contrastes sont a priori statistiquement indépendants et sont tirés d'une distribution normale de moyenne nulle. Il suffit alors d'utiliser les méthodes de statistique classique pour calculer une régression ou une covariance entre plusieurs caractères.
La modélisation des traits d'histoire de vie comme des mouvements browniens n'est pas toujours optimale. Mais on peut affirmer que la méthode de contrastes indépendants a recadré la méthode comparative en la basant sur des modèles probabilistes explicites. La modélisation prend en compte les erreurs de mesures des valeurs contemporaines des caractères [24]. Elle peut se baser sur d'autres modèles que le modèle brownien décrit ci-dessus, comme le modèle brownien directionnel [70] qui permet de modéliser des tendances générales ou un processus d'Ornstein-Uhlenbeck qui modélise une attraction vers une valeur d'équilibre [24].
La méthode comparative ainsi élaborée implique une connaissance de la topologie et de la longueur des branches de l'arbre. Cette difficulté est majeure et limitante. Une première solution est déjà proposée par Felsenstein, la méthode des branches sœurs [25]. Le principe est d'utiliser seulement les paires d'espèces voisines afin d'obtenir des différences indépendantes de la topologie. L'avantage de la méthode des branches sœurs est qu'elle ne nécessite pas une connaissance précise des relations profondes d'une phylogénie. Par contre elle n'évite pas la nécessité de connaître la longueur des branches entre les paires afin d'évaluer la force de la dépendance entre ces espèces.
En 2002, une méthode bayésienne ayant fortement influencé ce travail a été publiée [43]. La méthode bayésienne permet d'échantillonner simultanément de leur probabilité conjointe a posteriori les covariances, les longueurs des branches, la topologie de l'arbre et les valeurs des traits d'histoire de vie sur l'arbre. La simultanéité des estimations permet d'améliorer les estimations des valeurs ancestrales, en intégrant sur tous les autres paramètres. Le travail de Huelsenbeck et Rannala estime les covariances sous la forme d'une matrice de variance covariance. Cette matrice est symétrique, elle contient la variance de chaque processus et la covariance de chaque paire de processus. Ils appliquent donc la méthode des contrastes indépendants et utilisent une méthode phylogénétique afin de parcourir plusieurs topologies. L'utilisation conjointe de modèle à ADN et de la méthode des

contrastes indépendants permet de mieux intégrer l'ensemble de l'incertitude.

Jusqu'à récemment, les méthodes comparatives étaient uniquement utilisées pour étudier des traits mesurés chez des espèces contemporaines. Mais, en principe, ces méthodes devraient permettre aussi d'étudier les relations entre paramètres d'évolution moléculaire et traits d'histoire de vie. Par exemple, une étude récente s'intéresse à la corrélation probable entre l'amplitude des ailes des chauves-souris et leur variabilité génétique intraspécifique [92]. L'hypothèse émise est que l'amplitude des ailes, en permettant à l'espèce d'étendre son territoire, augmenterait le territoire occupé, l'isolation géographique de sous-groupes et donc le polymorphisme. Notre travail s'intéressera aussi aux corrélations entre paramètres d'évolution moléculaire et traits d'histoire de vie.

1.2 Traits d'histoire de vie

1.2.1 Masse

L'étude de la masse des organismes a été centrale dans le développement des études d'allométrie. La masse est un caractère fondamental et très utilisé pour caractériser grossièrement le phénotype d'une espèce. C'est une caractéristique très facile à mesurer. De fait, il existe une grande quantité d'espèces pour lesquelles une bonne estimation de la masse est disponible. La taille, ou la masse corporelle sont des caractéristiques très documentées, mais aussi très pertinentes. En effet, c'est en fonction de la taille que la plupart des fonctions biologiques varient. De plus c'est une donnée qui varie beaucoup entre espèces : on trouve un éventail de masses corporelles allant du gramme au million de grammes chez les mammifères, les extrêmes étant la musaraigne, de l'ordre de 2 à 10 grammes, et la baleine bleue pesant plusieurs dizaines de tonnes. L'étendue de cette variation est très intéressante pour nous dans une perspective de compréhension des mécanismes écologiques, physiologiques et moléculaires chez les mammifères. Nous donnons ici quelques exemples d'hypothèses ou de questions sur l'évolution de la masse qui intéressent les biologistes.

Une première question est celle de la contrainte métabolique. Selon la loi de Kleiber, connue depuis 1930, le taux métabolique total par individu a un coefficient d'allométrie de 0.75. Le taux métabolique en énergie dépensée par jour et par gramme, qu'on appelle ici métabolisme masse-spécifique, a donc un coefficient d'allométrie d'environ -0.25 chez les mammifères [38]. Il y a donc une allométrie négative entre le métabolisme par unité de masse et la masse. Les plus petits ont un métabolisme masse-spécifique plus élevé, qui corrèle avec une respiration et une circulation sanguine plus grande. Souvent, une part importante de la masse corporelle des grosses espèces est dédiée aux réserves caloriques, ce qui implique moins de maintenance. La longévité étant aussi négativement corrélée au métabolisme masse-spécifique, on peut imaginer un effet d'usure des mitochondries [2]. À force de produire de l'ATP, celles-ci accumulent peut-être des dommages, ce qui aurait pour effet de réduire l'espérance de vie.

Une deuxième question est celle du lien entre masse et traits d'histoire de vie. Les traits d'histoire de vie sont des descripteurs du développement de l'individu : le temps de gestation, le temps de sevrage, la fécondité, la longévité, le temps de génération. Chez les

mammifères, la vie d'un organisme peut être assez bien résumée par deux valeurs. Tous les mammifères atteignent une phase de stabilité entre la maturité sexuelle et le début du vieillissement. La longévité et le temps de génération permettent donc de caractériser la façon dont l'organisme atteint sa taille adulte, puis maintient ses fonctions biologiques. On s'attend d'ailleurs à une allométrie positive entre la longévité et le temps de génération. Elle a été empiriquement vérifiée [89]. La masse corrèle positivement avec ces deux caractères.

Un troisième point important est l'allométrie négative de la masse avec la taille de la population. La règle de Damuth estime le coefficient de l'allométrie entre ces deux quantités autour de -0.75 [101]. La loi de Kleiber posant un coefficient de 0.75 pour le taux métabolique, on peut donc en déduire que le produit entre la taille de population et le métabolisme par individu est théoriquement constant. La cause principale, très intuitive, serait la contrainte imposée par la productivité totale du milieu de vie. Dans un environnement donné, si les ressources disponibles sont fixes, alors la taille de la population est limitée par ces ressources. La consommation de ressources par individu est reliée à son métabolisme total. Donc, plus la taille des individus est grande, plus la population est limitée par l'environnement. Cette idée est à mettre en lien avec la loi des carrés et des cubes énoncée d'abord par Galilée en 1638. La loi indique que quand un objet croît, sa surface augmente quadratiquement alors que son volume augmente cubiquement. Sur un territoire limité, la production de nourriture est proportionnelle à la surface, alors que l'absorption d'un individu augmente beaucoup plus que la surface occupée.

Cette hypothèse de corrélation négative entre masse et taille de population est très utile pour expliquer certaines observations. En particulier, on sait que la *mégafaune* est plus vulnérable aux activités humaines [11] et on pense que les espèces les plus grosses ont de plus grandes probabilités d'extinction, en partie à cause d'un plus grand aléa démographique, de la moindre fécondité, mais aussi à cause d'une sélection moins efficace. Cet amoindrissement de l'efficacité de la sélection, conséquence de leurs plus petites tailles de populations, sera discuté plus loin dans le chapitre 2 consacré à l'évolution moléculaire.

La classification des espèces sur une échelle allant des plus gros vers les plus petits est pertinente à plusieurs égards. Un lien entre cette classification et la stratégie d'histoire de vie des espèces à été proposé. Nous expliquons donc ici le paradigme r/K que nous utiliserons dans ce travail.

1.2.2 Stratégie r/K

Le paradigme r/K est un des paradigmes qui à été proposé pour expliquer les relations entre traits d'histoire de vie et la masse. Il permet de décrire la stratégie évolutive de l'espèce par rapport à son environnement. Le paradigme r/K fut proposé en 1967 pour décrire quantitativement la stratégie de différentes populations insulaires [66]. Ces stratégies de sélection qui déterminent la manière dont sont infléchies les probabilités de reproduction des individus sont aussi appelées sélection-r et sélection-K. Nous présentons ici les deux extrêmes d'un continuum de stratégies évolutives. Grossièrement, la sélection-K se place dans un environnement sans prédateurs, où un individu est avantagé par une vie plus longue et une fécondité réduite. La sélection-r se fait dans un environnement avec une prédation élevée, l'individu est alors avantagé par une maturité sexuelle rapide avec des litières de grande taille. En général, les gros organismes suivent une sélection-K et les petits une sélection-r.

En réalité, ce n'est pas la seule prédation qui modifie la stratégie évolutive, c'est plutôt la manière dont l'espérance de vie des individus est limitée par l'environnement. Par exemple, la non-stabilité des ressources, les maladies, ou les comportements sociaux permettant de résister à la prédation, modifient la vulnérabilité des individus à des facteurs de décès externes. La sélection naturelle joue, pour chaque individu, sur la maximisation de la part de la prochaine génération constituée de ses propres descendants. Dans les deux cas extrêmes, les fonctions biologiques en jeu seront très différentes : dans un environnement plus favorable, il est plus rentable d'investir dans quelques descendants qui auront une bonne probabilité de se reproduire et d'espacer les enfantements au cours d'une vie plus longue. Le groupe des carnivores, par exemple, rassemble ces caractéristiques. Dans un environnement plus incertain, il est intéressant de maximiser la taille des portées, quitte à investir moins dans chaque petit, puisque la plupart n'atteindront de toute manière pas la maturité sexuelle. Les lagomorphes sont de bons représentants de cette stratégie.

1.2.3 Quelques théories sur la biologie du vieillissement

Une attention particulière est portée dans ce travail à la longévité. C'est cette caractéristique qui nous a poussés en premier lieu à étudier les interactions entre traits d'histoire de vie et paramètres d'évolution moléculaire. Le but premier était de déterminer l'implication

des différents gènes de l'organisme dans la lutte contre le vieillissement. Il s'est ensuite avéré évident que les corrélations entre traits d'histoire de vie pouvaient être suffisamment fortes pour biaiser les résultats, et devaient donc être prises en compte. Par ailleurs, le cadre méthodologique développé ici peut finalement s'appliquer à d'autres problèmes de corrélation, tant que celles-ci concernent des caractères quantitatifs. Nous allons ici donner un aperçu des théories existantes sur les mécanismes évolutifs qui expliquent la diversité des durées de vie chez les mammifères.

On peut schématiquement distinguer deux courants de pensée qui divergent sur ce sujet. Les partisans de la *mort programmée* s'opposent a ceux de l'*usure naturelle*. Pour ces derniers, le principe du vieillissement semble tout d'abord être celui de l'accumulation de dommages irréversibles menant à la mort. Au contraire, la théorie de la mort programmée défend l'hypothèse d'un mécanisme intrinsèque au développement des individus qui favoriserait, au moment venu, leur obsolescence. L'argument sous-jacent est encore sensible aujourd'hui. Il s'agit d'une vision de l'évolution comme un mécanisme œuvrant dans le but de la persistance de l'espèce. Pourtant, dès les premières théories de la génétique, la sélection naturelle est vue comme le résultat de la seule compétition entre individus. Dans ce contexte, on voit mal comment un mécanisme programmant la mort des individus a pu être mis en avant par la sélection naturelle.

Plusieurs découvertes ont donné du crédit à l'idée de l'usure naturelle. La découverte des radicaux libres en est un bon exemple. Les radicaux libres sont des espèces moléculaires, dans un état transitoire, avec au moins un électron qui n'est pas couplé. Ils sont, entre autres, formés pendant le cycle de Krebs et sont très destructeurs pour l'intérieur d'une cellule. Ils réagissent facilement avec les protéines, l'ADN, et les acides gras. Une autre remarque soutenant l'hypothèse de l'usure naturelle est l'allométrie existant entre le métabolisme et la masse, prédisant une usure plus rapide des mitochondries par les radicaux libres qu'elles produisent en fabricant de l'énergie. Il existe aussi d'autres indices allant dans le sens de cette hypothèse, comme la découverte d'une augmentation de l'insolubilité et de l'agrégation des protéines avec l'âge [48].

Mais le fait qu'un organisme soit capable de se générer à partir d'une seule cellule provenant d'une lignée germinale immortelle semble contradictoire avec la dégradation inexorable des cellules de ce même organisme au cours du temps. Par exemple, les antioxydants sont des défenses efficaces, pouvant être produits par une cellule, contre les radicaux libres. Le raccourcissement systématique des télomères à chaque mitose semble être une

manière d'assurer la mort programmée des cellules et donc peut être aussi de l'individu [37]. L'hypothèse de la mort programmée avance ainsi que la longévité des individus est régulée d'un point de vue évolutif.

Mais revenons un moment sur les prémices de ces théories. "Organic Bodies are perishable. While life maintains the appearance of immortality in the constant succession of similar individuals, the individuals themselves pass away". Cette phrase de Johannes Muller, un physiologiste allemand du début du XIX^{e} siècle montre les premières interrogations suscitées par le concept d'évolution appliqué au phénomène du vieillissement. Elle a servi d'introduction à August Weissmann en 1891 pour la conférence qui fut le point de départ de beaucoup d'interrogations sur l'évolution de la durée de vie [97]. Ce biologiste allemand est reconnu comme un précurseur des théories sur l'évolution de la longévité. Il a mis en place plusieurs concepts utiles à la compréhension de ces théories, en particulier celui de *soma*.

Chez un organisme pluricellulaire, le soma est l'ensemble des cellules dont la descendance est vouée à disparaître avec l'individu. Elles peuvent se reproduire seulement grâce à des mitoses. La lignée somatique est à mettre en opposition avec le *germen*, ou lignée germinale. La lignée germinale est constituée de la généalogie des cellules germinales, elle suit un cycle composé de quelques mitoses, une méiose puis une fécondation. Cet angle de vue permet d'apprécier différemment ce qu'est le temps de génération, en fait il est le rythme auquel la lignée germinale recommence un nouveau cycle.

August Weissmann est un fervent détracteur de l'hypothèse de transmission des caractères acquis, il avance une théorie d'un mécanisme acteur du vieillissement, intrinsèque au soma de chaque individu. Ce sont des idées à la base de la théorie de la mort programmée. Il parle déjà de corrélations entre masse, métabolisme et longévité. Il avance un élément important : du point de vue de la lignée germinale, la longévité d'un individu importe peu pour la survie de sa lignée. C'est ce qui s'imposera dans la théorie du soma jetable proposée par Thomas Kirkwood [53] un siècle plus tard. Cette théorie met en exergue les arbitrages entre soma et germen au cours de l'évolution.

La théorie du soma jetable se base sur le paradigme r/K, et le fait que des arbitrages ont lieu en permanence entre l'importance sélective donnée aux différents traits d'histoire de vie. C'est en fonction de ce que l'environnement offre comme chances de survie, c'est-à-dire l'espérance de vie extrinsèque, que l'espérance de vie biologique de chaque individu peut devenir un atout. À une échelle évolutive, les variations de la longévité reflètent la

manière dont la lignée germinale investit des ressources dans le soma. Rappelons que du point de vue de la lignée germinale, ce qui importe, c'est de maximiser le taux net de reproduction. Les caractères des individus permettant de lutter contre le vieillissement ne sont donc sous pression de sélection que dans la mesure où un allongement de la vie augmente le nombre total de descendants, c'est-à-dire la valeur reproductive de l'individu.
Les fluctuations de l'environnement induisent donc des variations du type de sélection sur l'échelle r/K. La stratégie r donne peu d'importance à la longévité relativement à d'autres fonctions biologiques. La stratégie K permet au contraire à la longévité d'augmenter la compétitivité de reproduction de l'individu.
Cette base à permis de proposer deux théories concernant l'effet de ces modes de sélection sur l'ADN : la pléiotropie antagoniste et l'accumulation des mutations.
La théorie de la pléiotropie antagoniste de Georges C. William [102] suggère que la plupart des protéines sont *pléiotropes* dans un organisme, ce qui veut dire qu'elles jouent un rôle dans plusieurs voies biochimiques. Cette théorie s'applique donc sur des gènes ou des fonctions cellulaires pléiotropes qui jouent un rôle dans au moins deux processus biologiques opposés. La pléiotropie antagoniste postule qu'il y aura sélection positive sur la fonction la plus utile pour augmenter le taux de reproduction individuel. Par exemple, la sélection naturelle de type K va favoriser les fonctions liées à la fécondité, au détriment de la lutte contre le vieillissement. Cette hypothèse suggère que les gènes en question sont constamment soumis à une sélection positive dont la direction varie en fonction de l'environnement. En fait, cette théorie est à placer dans un contexte sélectionniste, c'est-à-dire l'idée que la totalité du génome est optimisée à tout moment par la sélection naturelle. Cette théorie permet d'extrapoler en supposant que l'arbitrage fait par la sélection entre les différentes fonctions biologiques favoriserait les protéines et les voies chimiques dont les fonctions ne peuvent pas à être simultanément sous sélection positive dans une espèce.
La deuxième proposition, qu'on appelle ici théorie de l'accumulation des mutations fut mise de l'avant par Peter Medawar en 2006, dans un article explicitement intitulé "Aging is no longer an unsolved problem in biology" [39]. Cette hypothèse correspond aux théories de l'évolution moléculaire utilisées dans ce travail. C'est, là encore, une manière d'interpréter l'arbitrage que fait la sélection entre les différentes fonctions biologiques. La théorie de l'accumulation des mutations de Peter Medawar suggère que les gènes permettant de lutter contre le vieillissement de l'organisme sont soumis à une pression de sélection qui varie en fonction de la stratégie r/K de l'espèce en question. Si une fonc-

tion biologique particulière ne permet plus d'augmenter significativement la fécondité des individus, les mutations nucléotidiques qui affectent cette fonction ne seront plus sous une forte sélection purificatrice. Seule la durabilité de l'organisme permise par l'environnement définit ainsi la durée de vie. Les gènes ralentissant le vieillissement vont donc accumuler des mutations dans le cas de la sélection *r*. Dans le cas de sélection *K*, d'autres gènes accumuleront sûrement des mutations. Le relâchement de la pression de sélection peut parfois aller jusqu'à la transformation d'un gène en pseudogène.

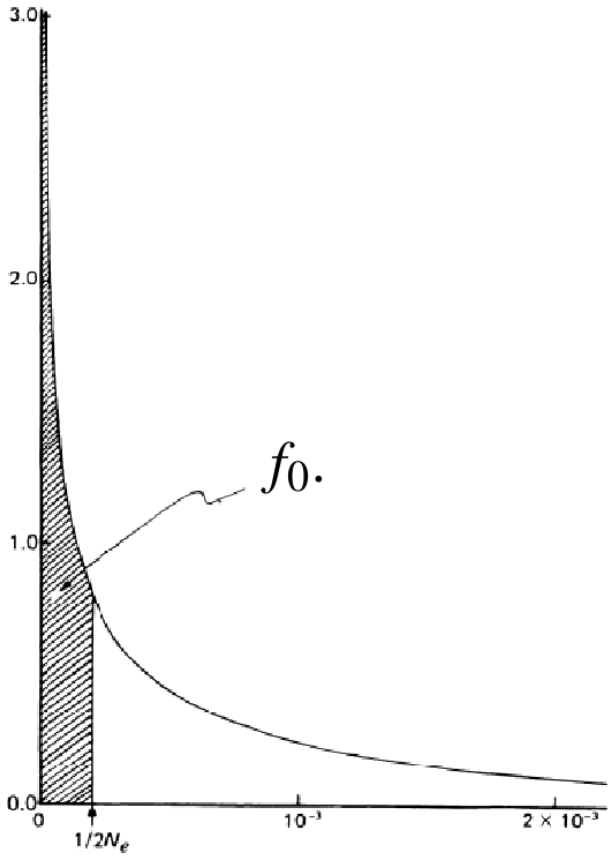
La théorie du soma jetable décrit donc l'évolution de la longévité maximale du soma d'une espèce comme suivant les variations de la longévité permise par l'environnement de l'espèce. L'histoire évolutive d'un gène impliqué dans les mécanismes de maintien du soma est liée à l'histoire de la longévité de ce lignage. L'évolution de la pression de sélection sur un gène devrait donc nous permettre de reconnaître les gènes impliqués dans le maintien du soma. A la lumière de ces théories, notre approche a d'abord été d'utiliser les méthodes de biologie comparative pour tester des hypothèses sur la longévité. Essentiellement nous tentons de mesurer les corrélations entre la force de la sélection par gène et l'histoire de la longévité. Mais avant de parler de ces interactions, nous allons présenter quelques principes fondamentaux de l'évolution moléculaire.

CHAPITRE 2

ÉVOLUTION MOLÉCULAIRE

2.1 La Théorie Neutraliste

Figure 2.1 – Distribution de fréquence des coefficients de sélection parmi les mutants parmi différents sites de l'ADN codant. La partie hachurée représentant la fraction des mutations *quasi neutres*. La distribution a une moyenne de 0.001 et un paramètre de forme de $\beta = 0.5$ [51]



En 1984 Motoo Kimura publie *The Neutral theory of molecular evolution* [52] et pose ainsi les principes de la théorie neutre de l'évolution moléculaire. Cette théorie est basée sur le concept de mutations *quasi neutres*, c'est-à-dire des mutations dont le désavantage sélectif est suffisamment faible pour être négligeables devant les effets stochastiques. Plus précisément la dynamique du comportement de ces mutations se rapproche de celle des mutations neutres dans un modèle populationnel de Wright Fisher. Leur fréquence et leur

destin ultime sont entièrement déterminés par les effets stochastiques de la dérive génétique. La théorie de Motoo Kimura est que ces mutations quasi neutres représentent une grande part des différences moléculaires observées entre espèces. Elles auraient donc un rôle prépondérant dans l'évolution des génomes.

Cette hypothèse s'oppose donc fortement à l'idée *sélectionniste* d'un génome modelé uniquement par l'adaptation. Le sélectionnisme est un courant qui découle directement du Darwinisme et de la découverte de l'ADN. Schématiquement, le sélectionnisme cherche à expliquer toute la variabilité intra et inter spécifique du génome par l'adaptation. Ce courant de pensée postule que les mutations délétères sont systématiquement éliminées par la sélection naturelle, tandis que les mutations qui présentent un avantage reproductif pour l'individu représentent la plus grande partie de la variation observée entre les espèces. D'un autre côté, la théorie neutraliste suggère qu'une grande partie des variations observées n'est pas la conséquence de l'adaptation. Les probabilités qu'une mutation avantageuse disparaisse ou qu'une mutation délétère augmente en fréquence jusqu'à se fixer dans la population ne sont pas négligeables. La théorie neutraliste cherche en particulier à quantifier avec finesse la part de mutations délétères qui ont une probabilité proche d'une mutation neutre de se fixer dans la population.

C'est la génétique des populations qui a introduit l'idée d'un coefficient de sélection s. Le coefficient de sélection est la valeur de l'avantage reproductif associé à une mutation donnée. Si w est la valeur sélective d'un individu, c'est-à-dire le ratio du nombre de ses descendants par le nombre de descendants moyens, alors s le coefficient de sélection se définit :

$$s = \frac{w_1 - w_0}{z_0}. \tag{2.1}$$

Avec w_1 la valeur sélective moyenne d'un individu avec la mutation, w_0 la valeur sélective moyenne d'un individu sans la mutation. La fréquence d'une mutation dans une population varie selon son coefficient de sélection, mais aussi selon quantité d'autres facteurs. La probabilité de fixation d'une mutation dans une population en fonction de s et de N_e, la taille efficace de la population vaut, pour une population diploïde :

$$p_{fixation} = \frac{1 - \exp^{-2s}}{1 - \exp^{-4N_e}}. \tag{2.2}$$

Nous utilisons ici une mesure communément appelée N_e. À l'échelle évolutive, nous définissons N_e comme la taille de population efficace de long terme. N_e prends en compte

le type de reproduction, le sexe-ratio, les fluctuations dans la taille réelle de population, la répartition géographique des individus. Généralement plus petite que la véritable taille de population N, la taille de population efficace représente la taille de population équivalente dans un modèle de Wright-Fisher dans lequel les générations ne se chevauchent pas, la taille de chaque génération est fixe et la reproduction est panmictique [49]. Tous les aspects de la sélection naturelle peuvent influencer la taille de population efficace : dérive génétique, sélection positive, déséquilibre de liaison... Chez les mammifères, plusieurs études ont trouvé une allométrie négative pour N_e. Autrement dit, la taille efficace augmente quand la masse diminue. Les rongeurs auraient une grande taille de population efficace, contrairement aux primates, par exemple [65]. Ceci implique une plus grande accumulation de mutations délétères chez les primates que chez les rongeurs.

La probabilité de fixation d'une mutation est donc dépendante, non seulement de la valeur de son coefficient de sélection, mais aussi de la taille efficace de la population. Pour simplifier, une valeur positive de s rend plus probable une fixation de la mutation correspondante. À l'inverse s très négatif fait tendre cette probabilité vers 0. Pour une mutation proche de la neutralité, avec $s \sim 0$, la probabilité de sa fixation tend vers :

$$\lim_{s \to 0} p_{fixation} = \frac{1}{2N_e}, \tag{2.3}$$

donc plus la taille de population efficace est petite, plus la probabilité de fixation d'une mutation quasi neutre est grande. Intuitivement, une certaine part de mutations légèrement délétères pourraient donc se fixer dans une population avec une probabilité d'autant plus grande que la population est petite. Formellement, une mutation est dite quasi neutre quand :

$$|2N_e s| < 1. \tag{2.4}$$

La conséquence qui nous intéresse particulièrement est du côté des mutations délétères :

$$-\frac{1}{2N_e} < s < 0. \tag{2.5}$$

Les mutations proches de la neutralité ont alors une chance approximativement équivalente de se fixer dans la population que des mutations dont le coefficient de sélection serait nul. Nous nous intéressons à la proportion relative des mutations quasi neutres, notée f_0. La répartition de ces coefficients de sélection, imaginée par M. Kimura, est représentée sur

la figure 2.1. [51] Selon cette distribution, la part des mutations quasi neutre est grande, même avec des grandes populations.

On définit ici le taux de substitution ρ et le taux de mutation μ. Le taux de mutation est défini comme le nombre moyen de mutations qui apparaissent à une génération et par individu. Par exemple, l'ordre de grandeur du taux de mutation a été estimé à $10^{-9} < \mu < 10^{-7}$ mutations par paire de base chez l'humain [65, p.97]. À l'échelle de la population on a donc $2N_e.\mu$ mutations par paire de base qui apparaissent à chaque génération.

Le taux de substitution, ρ, est le nombre moyen de mutations qui se fixent dans une population par unité de temps, c'est-à-dire le produit du nombre de mutations par la probabilité de fixation d'une mutation :

$$\rho = 2N_e \mu p_{fix}. \tag{2.6}$$

Pour les mutations fortement délétères, on a $p_{fix} = 0$ et pour les mutations quasi neutres : $p_{fix} = \frac{1}{2N_e}$ comme donné par l'équation 2.3. Donc le taux de substitution des mutations quasi neutres peut s'écrire :

$$\rho = \mu f_0. \tag{2.7}$$

La théorie quasi neutre postule une grande valeur de f_0, tandis que les sélectionnistes l'imaginent négligeable. Depuis la distribution des mutations délétères de la figure 2.1 a été beaucoup corroborée [22], et des analyses empiriques ont validé l'idée d'une distribution dite leptokurtique, c'est-à-dire qui passe rapidement de valeurs très élevées à des valeurs très basses au fur et à mesure que s s'éloigne de zéro. [50]

La théorie neutre de l'évolution était peu soutenue, à l'époque, et l'hypothèse d'horloge moléculaire permit d'avancer de nouveaux arguments.

2.2 Horloge moléculaire

En 1962, Zuckrekandl et Pauling en comparant l'hémoglobine entre plusieurs espèces de vertébrés, observent une corrélation entre le nombre de différences de deux séquences du gène et l'âge de la divergence des espèces. Cette constatation permit de proposer l'hypothèse de l'horloge moléculaire [104]. Elle permit de faire de la datation et fut un argument pour l'utilisation des matrices de distance en inférence phylogénétique.

Selon cette hypothèse, le nombre de différences entre les séquences est en moyenne proportionnel au temps qui les sépare de leur ancêtre commun. Cette hypothèse était sur-

prenante dans une perspective sélectionniste. Le sélectionnisme impliquerait en effet la variation erratique du taux de substitution, en fonction des besoins adaptatifs de la population. Or, on sait que l'évolution morphologique et donc l'adaptation ne se font pas du tout au même taux dans les différents organismes. Un bon exemple estle cœlacanthe, dont l'évolution morphologique est très faible depuis plusieurs dizaines de millions d'années.
À première vue, la perspective neutraliste semble mieux expliquer cette observation. En effet, l'hémoglobine ayant la même fonction chez tous les vertébrés, on s'attend à un f_0 constant. L'argument de Motoo Kimura était donc que dans ce contexte, un taux de mutation à peu près constant entre espèces a pour conséquence un taux de substitution $\rho = \mu f_0$ constant. L'hypothèse d'une horloge moléculaire fut donc un moteur important du développement des idées neutralistes de Motoo Kimura, corroborant le fait que la majorité des changements le long des lignages sont dus à des mutations effectivement neutres.
Depuis il est clair que l'horloge moléculaire ne s'applique qu'approximativement [63]. On parle à présent d'horloge moléculaire relâchée, car les variations du taux de substitution sont graduelles, et permettent de dater les événements de spéciation, mais avec une grande plus incertitude. Les variations détectées de ρ ont été attribuées à divers facteurs. Il y a une allométrie entre le taux de substitution, le métabolisme et le temps de génération [79] [76]. Mais surtout, μ varie beaucoup entre les espèces.
Il y a deux types de facteurs causant des modifications du taux de mutation de l'ADN, les erreurs lors de la copie et les dommages causés par l'environnement. Les dommages de l'ADN, sont causés par les rayons UV ou par les radicaux libres par exemple. Certaines observations montrent que les plantes dans un environnement où il y a plus d'énergie ont un taux de substitution plus grand. L'autre source de mutations est le taux d'erreur durant la réplication de l'ADN. Par exemple, chez les vertébrés, les espèces avec des temps de génération courts ont un taux de substitution plus élevé. [6] Cette constatation est la fait du nombre de divisions cellulaires entre deux cellules germinales, par unité de temps.
Cependant, la théorie neutraliste reste encore valable, mais, plutôt que d'essayer de prouver la constance de f_0 ou ρ, des outils plus sophistiqués ont dus être mis en place afin d'établir des relations entre les variations simultanées de μ, f_0 et ρ.

2.3 Théorie quasi neutre et taille efficace de population

La taille efficace de population, N_e, montre également de grandes variations inter espèces, et dans le cas qui nous intéresse, chez les placentaires. C'est un point important, car la fraction de mutation quasi neutre f_0 dépend de N_e, comme on peut le voir sur la figure 2.1. Historiquement, une théorie pour sauver l'explication neutraliste de l'horloge moléculaire fut proposée par Ohta et Kimura. Elle se base sur une corrélation entre différents caractères qui laisseraient approximativement ρ constant. L'opposition r/K, permet de faire l'hypothèse que l'augmentation de la masse dans le cas de la stratégie K se fait souvent conjointement avec deux choses. D'abord une augmentation du temps de génération et donc une diminution du taux de mutation, ensuite une diminution de N_e impliquant une augmentation de f_0. L'allométrie négative entre μ et f_0 laisserait constant le taux de substitution ρ.

Toutefois, cette description schématique ne prend pas en compte les coefficients d'allométrie entre les différents caractères quantitatifs impliqués. En réalité, ni μ, ni N_e ne sont constants, leurs variations ne se compensent pas, et de toute manière ρ n'est pas constant, donc l'horloge moléculaire n'a pas à être préservée .

Le pourcentage de mutations proches de la neutralité alimente encore le débat opposant sélectionnistes et neutralistes, et des interprétations différentes découlent des variations de ρ. Les variations de μ rendent assez difficile des prédictions quantitatives du comportement du taux de substitution dans les protéines. Par contre, dans les séquences codantes, il est très intéressant de décomposer le taux de substitution en deux composantes. La première, dS, concerne les mutations silencieuses, c'est-à-dire conservatives au niveau de l'acide aminé. dS peut être supposé approximativement équivalent au taux de substitution neutre et donc égale à μ. La deuxième, dN, le taux de mutations non-synonymes, concerne les mutations qui modifient la protéine codée par le gène. A priori, en grande majorité les mutations non synonymes dans une séquence codante ont des coefficients de sélection négatifs quasi neutres ou fortement délétères. La valeur de dN est donc liée à la force de la sélection purificatrice qui agit sur le gène en question. Le rapport entre dN et dS permet a priori d'annuler les effets taille de population par exemple. On peut alors poser $dN = \mu.f_0$, ce qui permet d'écrire :

$$\omega = \frac{dN}{dS} = f_0 \tag{2.8}$$

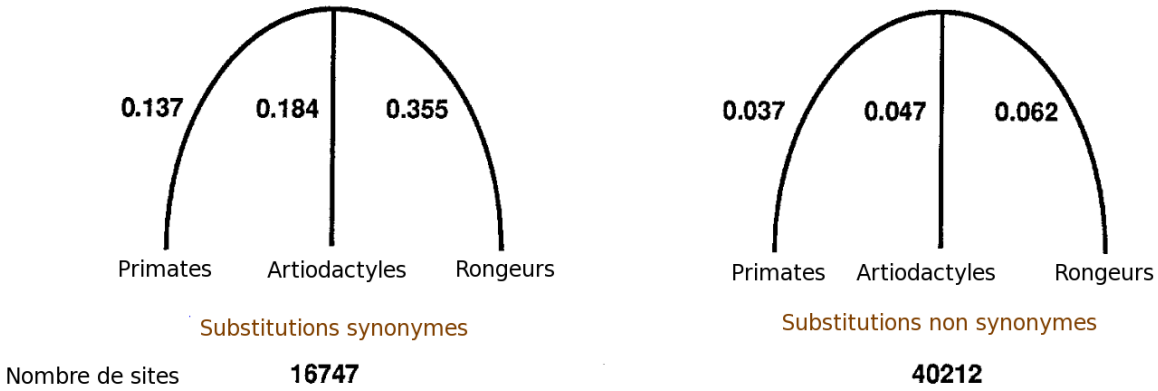
Cette estimation de f_0 est bien le contraste entre dN et dS. On l'appellera couramment ω dans ce mémoire. Son avantage principal est qu'il pourrait éliminer les effets purement dus aux variations du taux de mutations. Pour cette raison, ω, peut permettre d'évaluer la force de la pression de sélection et les variations de cette pression entre lignage ou entre positions [47].

La pression de sélection qu'il y a sur un gène varie selon le coefficient de sélection moyen des mutations dans ce gène. On attend que le ratio ω soit plus grand que l'unité quand la séquence est soumise à une sélection adaptative, c'est à dire sous une sélection positive. Inversement, un dN inférieur à dS c'est-à-dire un ω plus petit que l'unité implique une conservation forte de la séquence. On parle alors de sélection purificatrice, c'est-à-dire que les mécanismes de génétique des populations ont comme rôle essentiel d'évacuer les substitutions non synonymes. En vérité, on attend très peu de sélection positive visible sous l'hypothèse quasi neutre, car la plupart des substitutions sont délétères.[55]

L'intérêt porté à cette mesure de la pression de sélection que représente dN/dS a permis de mettre en place plusieurs nouveaux modèles pour étudier l'évolution moléculaire. Les modèles de ce type sont appelés modèles à codons et seront décrits dans la section 4.2. Succinctement, ils utilisent des matrices de séquences nucléotides alignés selon leurs traductions en acide aminé. pour reconstruire l'histoire des substitutions de codons. Dans un premier temps, ils furent utilisés pour l'inférence de la topologie de l'arbre du vivant, puis en évolution moléculaire. Les modèles à codons [36, 72] intègrent les deux niveaux de compréhension des séquences d'ADN codant et, utilisent la redondance du code génétique pour inférer dN et dS. L'histoire des substitutions de chaque codon sur un lignage est modélisée comme un processus de Markov et les modèles permettent d'estimer le rapport ω.

La théorie quasi neutre de l'évolution a utilisé le rapport ω pour vérifier la consistance de l'hypothèse neutraliste. En 1994, Tomoko Ohta, collègue de Motoo Kimura, l'utilise ainsi qu'illustré sur la figure 2.2. C'est déjà une utilisation des méthodes comparatives proche des contrastes indépendants pour étudier des paramètres d'évolution moléculaire. On voit clairement une allométrie positive entre dN et dS. Le rapport de ces deux taux, ω corrèle négativement et significativement avec les estimations de taille de population de ces groupes. Comme on s'y attendait, on trouve un ω plus faible, correspondant à une forte pression de sélection purificatrice sur le génome dans les populations de grande taille

Figure 2.2 – Cette figure adaptée de Ohta 80, montre sur une phylogénie non racinée les taux de substitutions de 49 gènes orthologues



efficace comme les rongeurs. Inversement les primates qui ont de petites tailles de population, ont un ω plus élevé. Il y a donc bien une augmentation de la part des mutations délétères qui peuvent être considérées comme neutres quand la taille de la population efficace diminue.

Les fluctuations de la taille de population efficace, et ses corrélations avec ρ ou ω, sont donc des caractères quantitatifs à étudier. Nous avons vu aussi dans la section 1.2 qu'il existe des corrélations entre les différents traits d'histoire de vie et N_e. Il peut donc être intéressant d'introduire en plus de caractères phénotypiques, certains paramètres d'évolution moléculaire comme, dS ou dN/dS, dans le cadre des méthodes comparatives.

2.4 Notre travail : coevol

C'est principalement suite à l'abandon de la théorie de l'horloge moléculaire que les méthodes comparatives ont été utilisées pour étudier les paramètres d'évolution moléculaire. Le but était d'abord de comprendre quels sont les facteurs qui affectent le taux de substitution nucléotidique. Les hypothèses étaient nombreuses pour expliquer les variations entre différents groupes taxonomiques. Plusieurs études se sont intéressées aux corrélations entre traits quantitatifs et paramètres d'évolution moléculaire, les méthodes utilisées ont évolué avec le temps, nous en décrivons quelques-unes ici.

En 1992, une étude de Martin et al. [68] a compilé diverses mesures du taux de substitution de séquences mitochondriales de vertébrés à partir de différentes études. Le taux

de substitution de chaque branche terminale à ensuite été comparé aux traits de vie de l'espèce correspondante. Ce papier montre une allométrie négative entre le taux de substitution ρ et la masse. Mais la forte association entre les différents traits d'histoire de vie ne permet pas d'en conclure par une simple relation de causalité.

En 1996, une étude propose de tester l'effet du temps de générations sur le taux de substitution ρ [64]. Une corrélation positive signifierait que le nombre de générations par unité de temps augmente le taux de mutation et donc ρ. Une corrélation négative pourrait signifier, à travers une augmentation du nombre de mitoses séparant deux cellules germinales, que le taux de mutation μ est plus élevé chez les gros. Mais les résultats de l'étude soutiennent un lien entre le nombre de générations par unité de temps et le nombre de substitutions par unité de temps.

En 2004 un article propose de caractériser la sélection naturelle grâce à une estimation des variations de dN et de dS [88]. Ils utilisent ici un modèle à codon permettant des variations de ω sur toutes les branches de l'arbre. Ils s'intéressent par la même occasion aux différences entre le taux de substitution pour les transversions et pour les transitions. Leur travail est dans un cadre Bayésien et permet d'estimer simultanément les temps de divergence, les variations des taux de substitutions synonymes et non synonymes comme suivant une distribution normale bivariée. Ils comparent les corrélations entre ω et le temps de génération, pour le gène COX1 chez les primates. Une importante constatation est que le gène COX1 ne montre pas les mêmes variations de pression de sélection que l'ensemble des 49 gènes étudié par Ohta, (figure 2.2) L'histoire des la pression de sélection est donc spécifique à chaque gène.

Méthodologiquement, les problèmes que pose l'étude conjointe des traits quantitatifs et des paramètres d'évolution moléculaire sont nombreux. Les données permettant de vérifier ces allométries sont obtenues par des méthodes différentes présentant certains inconvénients. Les estimations des données morphologiques sont connues pour la plupart des feuilles de l'arbre et sont estimées ponctuellement, pour chaque spéciation. D'un autre côté, les paramètres d'évolution moléculaire sont calculés sur des intervalles de temps, leurs valeurs instantanées restant théoriques. Les biais concernant les paramètres d'évolution moléculaires sont multiples : erreur de séquençage, mauvaise détection de l'orthologie, qualité de l'alignement, violation de modèle. Il faut ajouter que, de manière générale, les processus estimés sur une phylogénie sont estimés de moins en moins précisément en remontant dans le temps.

Mon laboratoire d'accueil s'est donc consacré au développement d'un outil permettant de reformuler les questions citées plus haut dans une méthode probabiliste plus intégrée. Le développement de *coevol* s'est fait dans un contexte bayésien, permettant à la fois de définir précisément les distributions a priori des paramètres inconnus et d'estimer simultanément tous ces paramètres. Le modèle multivarié construit dans *coevol* permet d'inclure à la fois la reconstruction de traits d'histoire de vie et des substitutions des séquences d'ADN. Ce modèle intègre la méthode des contrastes indépendants et opère des régressions multiples grâce à des matrices de covariance. Cet article est publié et est disponible dans l'appendice I. Les aspects plus spécifiques de la structure du modèle sous-jacent sont rappelés dans le chapitre 3, avec toutes les variantes spécifiques développées dans le cadre de cette maîtrise. Nous verrons qu'il est même envisageable d'estimer les variations de N_e à travers sa relation avec f_0 avec une estimation globale de la pression purificatrice sur le génome.

CHAPITRE 3

MÉTHODES

3.1 Modèles probabilistes

Ce chapitre est le coeur de ce travail. Il est consacré aux méthodes qui ont permis de mettre en place, tant conceptuellement que concrètement, les différents modèles.
Dans un premier temps, nous décrirons les modèles probabilistes, les modèles graphiques, le MCMC, pour finir par l'algorithme de . Nous décrirons ensuite son implémentation dans un cadre orienté objet. Enfin, nous détaillerons la structure des trois modèles sur lesquels la méthode est appliquée.

3.1.1 Principes généraux

Le cadre des modèles probabilistes permet de formaliser les hypothèses scientifiques. En général, les observations et les résultats d'expériences permettent d'estimer la validité d'une hypothèse ou de la réfuter. Les données connues du problème, appelées D, sont alors l'ensemble des résultats expérimentaux permettant d'estimer les paramètres inconnus θ. La structure du modèle M doit définir, de manière formelle, les relations entre D et θ. Dans un cadre probabiliste, ceci est fait en définissant $p(D|\theta,M)$. Cette probabilité des données sachant le modèle et les valeurs des paramètres est appelée vraisemblance. La structure du modèle M est souvent omise pour simplifier les notations, la vraisemblance se note plutôt : $p(D|\theta)$, la dépendance à M restant sous-entendue. Les méthodes couramment utilisées cherchent à estimer θ, tel qu'il maximise la valeur de la vraisemblance.

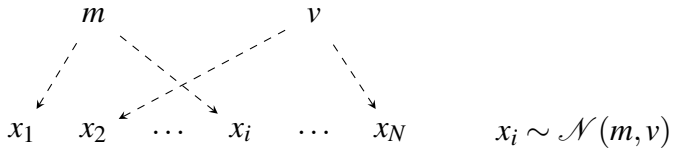


Figure 3.1 – Exemple de modèle probabiliste

La figure 3.1 permet d'illustrer les principes énoncés plus haut par un modèle simple où x_i pour $i = 1...N$ est une série d'observations. Le modèle pose l'hypothèse d'une distribution

normale de moyenne m et de variance v de laquelle sont indépendamment tirées toutes ces observations. Ici le modèle à donc 2 paramètres inconnus : $\theta = \{m, v\}$ et les données du modèle sont $D = \{x_1 ... x_N\}$. La vraisemblance du modèle s'écrit donc :

$$p(D|\theta) = \prod_i^N p(x_i|m, v). \tag{3.1}$$

Selon la méthode du maximum de vraisemblance, m et v sont estimés en maximisant $p(D|\theta)$, aboutissants à :

$$m' = \frac{\sum_i x_i}{N} \qquad v' = \frac{\sum_i (x_i - m')^2}{N}. \tag{3.2}$$

3.1.2 Le théorème de Bayes

La formule de Bayes date du XVIII^{e}ème siècle [3]. Elle établit de façon simple la relation entre les probabilités conditionnelles de deux évènements. Dans le contexte des modèles probabilistes, elle permet de calculer $p(\theta|D)$: la probabilité a posteriori de θ, les paramètres inconnus d'un modèle, sachant les données D d'un problème, avec

$$p(\theta|D) = \frac{p(D|\theta) \cdot p(\theta)}{p(D)} \tag{3.3}$$

où, $p(D|\theta)$ est la vraisemblance, $p(\theta)$ la probabilité a priori des paramètres et $p(D)$ la probabilité marginale des données, aussi appelée vraisemblance marginale. Cette dernière est particulièrement difficile à calculer. Fort heureusement, pour un jeu de données fixées D, la vraisemblance marginale $p(D)$ est constante, on peut alors poser la relation de proportionnalité :

$$p(\theta|D) \propto p(D|\theta) \cdot p(\theta), \tag{3.4}$$

qui permet d'évaluer les probabilités a posteriori relatives de plusieurs valeurs alternatives de θ

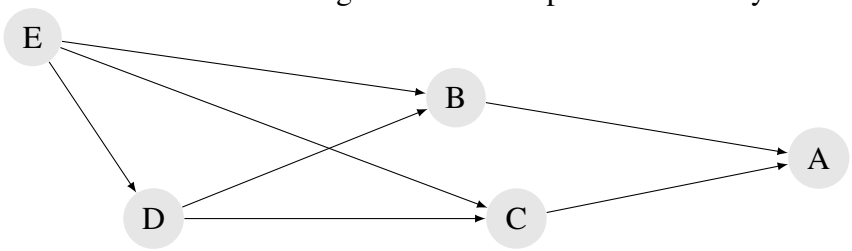
3.1.3 Modèles graphiques

Nous travaillerons ici dans un cadre de réseaux bayésiens, un type précis de modèle graphique. Les modèles graphiques permettent d'unifier la théorie des modèles probabilistes

et la théorie des graphes. Leur but est de tirer avantage des particularités du graphe pour représenter les dépendances conditionnelles entre les variables du modèle. En pratique, quand on cherche à optimiser les calculs, s'appuyer sur une telle représentation permet d'optimiser les calculs et de simplifier le stockage en mémoire des probabilités conditionnelles.

Un modèle probabiliste peut être représenté comme un réseau bayésien, qui peut se décrire par un graphe dirigé acyclique et orienté (Directed Acyclic Graph : DAG). Les noeuds du graphe correspondent aux variables aléatoires du modèle. Les arêtes représentent la dépendance entre variables aléatoires, elles sont donc orientées. Les données sont vues comme des variables aléatoires dont la valeur est fixée. En général, les hyperparamètres sont à la racine du modèle et les données sont les feuilles du modèle, comme on peut le voir dans le graphe de la figure 3.1. Le graphe est acyclique : un cycle signifierait une dépendance entre une variable et elle-même.

Figure 3.2 – Exemple de réseau bayésien



La figure 3.2, est un autre exemple plus abstrait de réseau bayésien où sont donc décrites les relations de dépendances entre variables. Par exemple, la probabilité de A est indépendante de E sachant les valeurs de B et C. De façon générale, c'est-à-dire indépendamment du graphe de la figure 3.2, les probabilités jointes des variables pourraient s'écrire :

$$p(A,B,C,D,E) = p(E) \cdot p(D|E) \cdot p(C|D,E) \cdot p(B|C,D,E) \cdot p(A|B,C,D,E) \tag{3.5}$$

Mais dans le cas présent, les relations d'indépendance représentées par le graphe de la figure 3.2 permettent de simplifier, ici pour A et B :

$$p(A,B,C,D,E) = p(E) \cdot p(D|E) \cdot p(C|D,E) \cdot p(B|D,E) \cdot p(A|B,C) \tag{3.6}$$

De manière générale, la distribution de probabilité conditionnelle d'une variable ne dépend directement que de la valeur de ses ancêtres immédiats, ses parents. L'implication

concrète de cette propriété est qu'elle permet de les probabilités jointes de deux réalisations du modèle qui ne diffèrent que par la valeur d'une variable (par exemple ici B), il faut d'abord définir l'ensemble des noeuds qui dépendent directement de cette variable, les *fils* de cette variable, ainsi que la variable elle-même (ici A et B). Il suffit alors faire le rapport des produits des probabilités de ces variables avant et après le changement de valeur de la variable d'intérêt, ici B. Ce principe nous sera très utile pour implémenter des stratégies computationnelles efficaces dans un cadre de chaînes de Markov Monte-Carlo.

3.2 Monte-Carlo

Notre méthode est du type MCMC (*Markov Chain Monte Carlo*) et s'appuie sur le paradigme bayésien pour approximativement échantillonner de la probabilité a posteriori du modèle. Un algorithme de type Monte-Carlo utilise un générateur de nombres pseudoaléatoires. Un MCMC utilise une méthode de Monte-Carlo pour construire une chaîne de Markov $\theta_n, n \in \mathbb{N}$ dans l'espace du paramètre θ. Dans notre cas, le MCMC, échantillonne les paramètres inconnus selon leur probabilité. La distribution stationnaire de cette chaîne de Markov, correspond à la distribution de probabilité voulue.
Plus précisément, notre MCMC est basé sur une combinaison d'algorithmes de type Métropolis-Hastings et d'échantillonnage de Gibbs.

3.2.1 L'algorithme de Métropolis

La méthode de Métropolis-Hastings est un algorithme de type Monte-Carlo. Il utilise les principes bayésiens pour estimer la distribution maximale a posteriori de l'ensemble du modèle.
Notre méthode utilise l'état précédent du modèle pour construire aléatoirement un nouvel état, comme décrit dans la figure 3.3. Chaque nouvel état est obtenu en deux étapes : supposons tout d'abord un état courant θ_n à l'étape n. Un nouvel état du modèle est proposé de façon aléatoire, à partir de l'état θ_n. On appelle cet état proposé θ^*. La probabilité de proposer l'état θ^* sachant l'état courant θ_n est notée $q(\theta_n \to \theta^*)$. C'est la *densité* de la proposition des mouvements. Par la suite ce nouvel état sera validé selon des règles définies plus loin. S'il est validé, nous définissons θ^* comme l'état suivant $\theta_{n+1} = \theta^*$. Dans le cas où le nouvel état est refusé, l'algorithme conserve l'état précédent : $\theta_{n+1} = \theta_n$. La dépendance de chaque état à l'état précédent, fait de l'ensemble final une chaîne de Markov.
La deuxième étape est l'acceptation de θ^* en fonction de sa probabilité a posteriori. Elle doit d'abord calculer le rapport entre la probabilité du modèle avec le nouvel état et avec l'état précédent. Ce rapport sera donc la probabilité de la présence du nouvel état dans l'échantillon final. Le principe est donc de comparer, grâce au théorème de Bayes, les probabilités a posteriori du modèle dans l'état initial et dans l'état après le mouvement.

On calcule ici ce rapport de ces probabilités r :

$$r = \frac{p(\theta^*|D)}{p(\theta_n|D)} = \frac{p(D|\theta^*) \cdot p(\theta^*)}{p(D|\theta_n) \cdot p(\theta_n)} \tag{3.7}$$

Si r est plus grand que 1, et donc que le nouvel état est plus probable que l'ancien, il est accepté. Sinon il est accepté avec la probabilité r. En pratique, pour déterminer si on accepte le nouvel état, on tire u, un nombre pseudoaléatoire d'une distribution uniforme entre 0 et 1. si $u < r$, le mouvement est accepté. Par exemple, θ^* est cinq fois moins probable que θ_n, alors on acceptera le mouvement avec la probabilité $r = 0.2$.
Dans sa version primitive, le principe de l'algorithme repose sur la symétrie de la densité de propagation :

$$q(\theta_n \to \theta^*) = q(\theta^* \to \theta_n) \tag{3.8}$$

Autrement dit : la probabilité de passer d'un état vers un autre doit équivaloir à la probabilité de revenir de ce nouvel état vers l'état courant. W. Keith Hastings a apporté à cet algorithme une amélioration en permettant la non-symétrie d'un mouvement

$$q(\theta_n \to \theta^*) \neq q(\theta^* \to \theta_n). \tag{3.9}$$

Il suffit alors de corriger le rapport utilisé :

$$r = \frac{p(\theta^*|D) \cdot q(\theta_n \to \theta^*)}{p(\theta_n|D) \cdot q(\theta^* \to \theta_n)} \tag{3.10}$$

Ainsi, l'algorithme de Métropolis-Hastings définit une chaîne de Markov sur l'espace des paramètres du modèle. Cette chaîne de Markov est réversible et a pour distribution d'équilibre $p(\theta|D)$. Ceci peut être vérifié en appliquant le principe du bilan détaillé [74]
L'algorithme de Métropolis-Hastings ne garantit qu'une chose : que la distribution stationnaire de la chaîne de Markov soit égale à $p(\theta|D)$. En pratique, la chaîne est initiée à partir d'une configuration quelconque, θ_0, et il faut laisser la chaîne se rapprocher suffisamment de son état d'équilibre. On doit donc éliminer de l'échantillon final les premières valeurs de paramètres visités par la chaîne $\theta_n, n = 0...B$ jusqu'à un certain seuil B. il faut donc déterminer B tel qu'à partir de θ_B, tous les états du modèle pourront être considérés comme approximativement tirés de leur probabilité a posteriori. Le temps de la *convergence*, c'est-à-dire le nombre de *pas* de Métropolis pris par l'algorithme pour atteindre la distribution de sa probabilité a posteriori, ne peut être déterminé à l'avance. Il faut en gé-

néral observer l'évolution de variables clefs et trouver un seuil raisonnable à partir duquel la série temporelle des valeurs prises par cette variable semble stationnaire.

Une fois la convergence atteinte, la collection de valeurs prises par un paramètre représente un échantillon tiré approximativement de la distribution de probabilité a posteriori marginale sur ce paramètre. On peut alors résumer cette distribution marginale en calculant, à partir de l'échantillon, la moyenne, la variance ou encore l'intervalle de crédibilité. Ces valeurs prennent implicitement en compte l'incertitude sur l'ensemble des autres paramètres du modèle. Les mouvements de Métropolis doivent permettre de potentiellement parcourir l'ensemble des valeurs que peuvent prendre les variables du modèle. L'amplitude des mouvements de Métropolis appliqués doit être suffisamment élevée pour permettre une convergence et éviter de rester bloqué autour d'un maximum local, mais assez petite pour que les mouvements soient souvent acceptés avec une probabilité raisonnable. En pratique, une petite amplitude raisonnable doit permettre aux mouvements de Métropolis d'être acceptés entre 30% et 70% des fois.

Les mouvements s'appliquent tour à tour sur des sous-ensembles du vecteur de paramètres. L'ordonnancement de ces mouvements peut être périodique, ou être aléatoire. Par exemple, le modèle présenté à la figure 3.1 contient seulement deux paramètres inconnus. Les mouvements pourraient alternativement concerner la variance et la moyenne à estimer.

Figure 3.3 – Algorithme de Métropolis-Hastings

Require: $\theta_n, n \in \mathbb{N}$

1: **loop**
2: define θ^* from a kernel $\theta_n \to \theta^*$
3: compute :

$$r = \frac{p(\theta^*|D) \cdot q(\theta_n \to \theta^*)}{p(\theta_n|D) \cdot q(\theta^* \to \theta_n)}$$

4: Draw u a random number from a uniforn distribution between 0 and 1
5: **if** $u < r$ **then**
6: $\theta_{n+1} = \theta^*$
7: **else**
8: $\theta_{n+1} = \theta_n$
9: **end if**
10: **end loop**

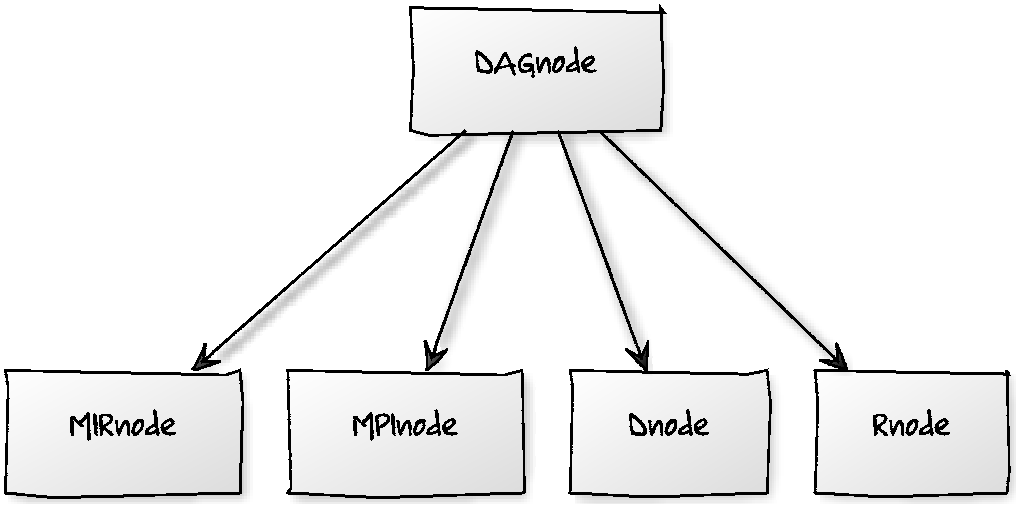
3.3 Implémentation

Coevol [58] à été implémenté dans le cadre d'un modèle graphique bayésien. Le tout à été fait en programmation-objet *C++*. Notre implémentation se base sur la structure du modèle graphique telle que définie précédemment. Le modèle est créé depuis la racine vers les feuilles en assurant des liens de parenté bien déterminés. Les méthodes propres à la structure de graphe sont implémentées dans la classe `DAGnode`. C'est la *brique* de base du modèle graphique, chaque noeud du graphe est un `DAGnode`, associé à un unique objet de type `BaseType`. La classe `BaseType` définissant tous les types de valeurs que le modèle peut contenir : entier, nombre réel, vecteur, matrice, valeur bornée etc ... Le modèle sera donc construit à partir d'instances de classes spécialisées dérivant de `DAGnode`. Ce modèle est ensuite encapsulé dans une série d'objets permettant la construction d'objets plus complexes, l'initialisation de la chaîne de Markov, le cycle des mouvements de Métropolis et la sauvegarde régulière des valeurs es paramètres visités.

3.3.1 DAGnode

La classe abstraite `DAGnode` permet de définir les caractéristiques d'un noeud du graphe. Ses attributs sont principalement deux listes : celle des parents et celle des enfants du noeud. Ces listes définissent la relation d'ordre entre les noeuds et de ce fait, les arrêtes dirigées du graphe.
Les parcours de graphe se font récursivement pour la plupart des opérations, étant donné que les relations de dépendances entre noeuds du graphe définissent un pré ordre qui doit être respecté lors du parcours du graphe. Une difficulté est que le graphe est un ensemble partiellement ordonné avec plusieurs racines, c'est un *poset*. Le parcours récursif est donc bidirectionnel, il est basé sur un principe de récursion, avec des optimisations qui permettent de limiter la taille de la pile d'appels.
La classe `DAGnode` définit aussi quelques méthodes qui permettent de définir le sous-graphe des noeuds dépendant d'un noeud donné dans le cas d'un mouvement de Métropolis. De cette classe quatre autres sont dérivées, comme sur la figure 3.4. Il y a donc quatre différents types de noeuds dans le graphe `Rnode`, `Dnode`, `MPInode` et `MIRnode`. Deux sont spécifiques à la parallélisation, `MPInode` et `MIRnode`, une autre permet d'économiser certains calculs, la classe `Dnode`, elle définit les noeuds *déterministes*. La classe

Figure 3.4 – Diagramme des classes dérivant de DAGnode



`Rnode` définit quant à elle, les noeuds probabilistes.

3.3.1.1 Rnode

La propriété des modèles graphiques qui nous intéresse particulièrement ici est de minimiser le temps passé dans le calcul du rapport de Métropolis-Hastings pour deux configurations du modèle : θ_n et θ^*. Dans le cas classique, les deux configurations ne diffèrent que par la valeur d'une variable, un noeud, les taches a accomplir pour dérouler l'algorithme de Métropolis Hastings peuvent être décrites comme suit :
Il faut tout d'abord faire la liste des noeuds dont la probabilité a posteriori est affectée par le mouvement. Ces noeuds sont : le noeud d'intérêt et l'ensemble de ses noeuds enfants. Pour chacun d'entre eux, il faut calculer le rapport entre la probabilité a posteriori avant le changement de valeur de la variable d'intérêt. On évite ainsi le calcul pour toutes les variables qui ne dépendent pas directement de la variable à laquelle un mouvement de Métropolis a été appliqué. Lors d'une étape de l'algorithme, une variable du modèle va suivre les instructions suivantes.

- La variable applique un mouvement de Métropolis sur la valeur qu'elle stocke.
- Elle recalcule sa probabilité en fonction de ses parents.
- Elle demande à toutes les variables dépendant d'elle même de calculer le rapport de sa probabilité a posteriori avant et après.

- Elle collecte tous ces rapports, les combines avec le rapport de Hastings.
- Elle applique la règle Métropolis Hastings et accepte, ou non, le mouvement.

Par exemple, sur la figure 3.2, un mouvement de Métropolis sur le noeud D va ensuite utiliser le calcul du ratio de Métropolis des noeuds D, B et C. Il s'agit donc de calculer la probabilité $p(D|E)$, $p(B|D,E)$ et $p(C|D,E)$ avant et après le changement de valeur du noeud D.

3.3.1.2 Dnode

Le besoin de pouvoir créer des noeuds déterministes, instances de la classe `Dnode`, vient de l'attention portée à l'd'optimisation des calculs. Parfois, plusieurs noeuds probabilistes de type `Rnode`, ont besoin d'effectuer certains calculs lourds combinant les valeurs de plusieurs noeuds en amont. Par exemple, si le déterminant d'une matrice est en jeu, l'objet `Dnode` contiendra seulement la valeur de ce déterminant. Le noeud du graphe de type `Dnode` est dédié à mettre a jour le résultat du calcul et le rendre disponible aux noeuds en aval. Il est donc inséré entre les noeuds en amont et les noeuds fils. Il ne modifie pas la relation d'ordre entre les noeuds du graphe, mais complexifie un peu la propagation des informations dans le modèle.

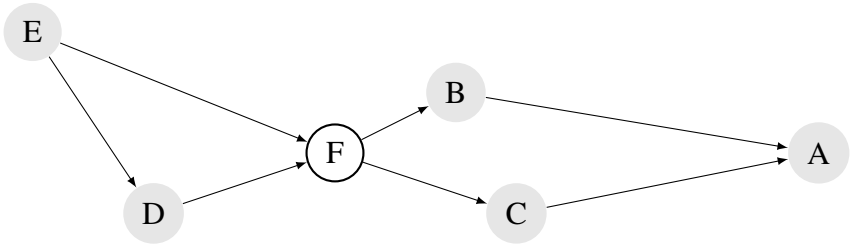


Figure 3.5 – Exemple de noeud déterministe

La figure 3.5 est un modèle graphique représentant le même modèle graphique qu'à l'exemple de la figure 3.2. Un noeud déterministe est placé entre deux groupes de noeuds fortement connectés. Les variables B et D dépendent de E et D via une fonction $f(E,D)$, ce qui permet de simplifier l'expression de la probabilité du modèle

$$p(A,B,C,D,E) = p(E) \cdot p(D|E) \cdot p(C|F) \cdot p(B|F) \cdot p(A|B,C) \tag{3.11}$$

Ici, le calcul fait par le noeud déterministe F est disponible pour B et C. Le noeud déter-

ministe F dont la valeur, son état, est simplement fonction de ses parents ($F = f(D,E)$) ne se voit associé aucune part de probabilité totale du modèle. Par exemple, un mouvement de Métropolis sur le noeud D va propager un signal au noeud F, qui va d'abord mettre à jour son état avant de signaler le mouvement à B et C. Les mises à jour effectuées, B et C envoient leurs contributions a D qui les agrègent au rapport de Métropolis final.

3.3.1.3 MPInode & MIRnode

Le but de cette maîtrise est de développer des modèles hiérarchiques. Ces modèles doivent permettre d'étudier un grand nombre de gènes et de garder la puissance apportée par toutes ces données pour estimer avec précision les paramètres qui ne sont pas spécifiques aux gènes. La structure hiérarchique du modèle peut être décrite comme une partition d'un graphe en $N+1$ sous-graphes. La partie supérieure est constituée d'hyperparamètres correspondant à la reconstruction des paramètres propres aux lignages et communs à tous les gènes. Ce sont, par exemple, les caractères phénotypiques, les paramètres d'évolution moléculaire communs à tout le génome. Les N sous-graphes isomorphes dans leurs structures sont connectés de manière équivalente aux hyperparamètres et correspondent, dans nos modèles, à un gène. La grande quantité de données nous a obligés à répartir le modèle sur plusieurs processeurs. Pour cela, nous avons utilisé *Open MPI* (*Message Passing Interface*. La parallélisation a été mise en oeuvre en s'appuyant sur le paradigme maître esclave correspondant à la structure hiérarchique du modèle.
Chaque noeud du sous-graphe du processeur maître qui a un noeud enfant sur les noeuds maîtres doit pouvoir propager les mises a jour de ses valeurs et les composantes normalement transmises entre les noeuds du graphe. Pour ce faire, chacun des noeuds de la partie supérieure qui a des descendants sur les esclaves est reproduit à l'identique par une instance de la classe `MIRnode` sur chaque noeud esclave. Chaque noeud esclave possède alors un ensemble de noeuds "miroirs" imitant le comportement des noeuds du maître. Du coté maître tous les noeuds qui possèdent des doubles d'eux même sur chaque esclave sont connectés à l'unique instance de la classe `MPInode`, comme montré voir sur la figure 3.6.
L'objet unique de type `MPInode` reçoit des ensembles de noeuds qui demandent à propager des calculs, les collectes et les envoie par paquet tous les esclaves.

Par exemple, sur la figure 3.6, un mouvement du noeud D, provoquera l'envoi de la nou-

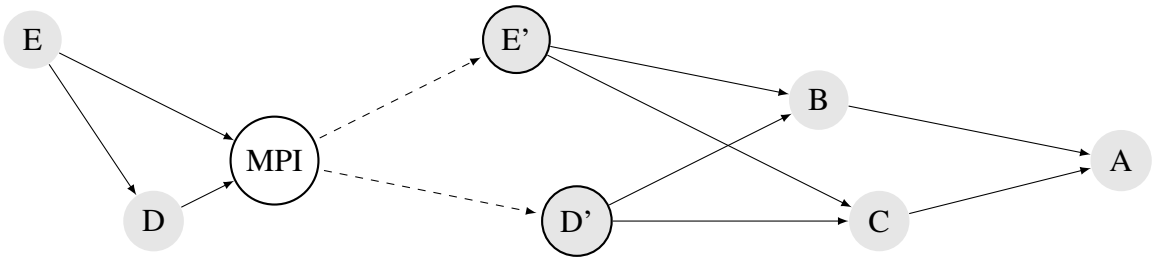


Figure 3.6 – Exemple de graphe parallélisé

velle valeur de D vers D' par le noeud MPI. Sur chaque esclave, D' se comportera alors comme un noeud probabiliste et propagera la demande des contributions respectives de B et C avant de renvoyer les contributions agrégées au processus maître.

3.3.2 La structure du modèle

Le graphe créé avec des instances de `DAGnode` doit être un graphe connexe. La classe `probModel` a comme rôle de construire, initialiser et contenir le graphe. Certaines formes prédéfinies, des gabarits, permettent de créer facilement des structures complexes formées d'instances `DAGnode`, par exemple un tableau de variables indépendantes et identiquement distribuées, ou alors, des structures d'arbre, que nous utilisons beaucoup dans un contexte évolutionniste. Certains gabarits permettent donc d'associer une variable par branche ou par noeud d'un arbre donné comme entrée du modèle.
L'objet de type `probModel` devra donc contenir la suite d'instructions nécessaires à la construction de tous les noeuds et à la définition de leurs dépendances. Dans le cas d'une parallélisation, il faudra définir précisément les liens entre les noeuds et leurs copies. L'instance de `probModel` contient la suite d'instructions correspondant à un ordonnanceur, il contient une suite de mouvement de Métropolis-Hastings. Chaque mouvement du M.C.M.C.inscrit dans cet ordonnanceur y est défini par le noeud probabiliste concerné du graphe, l'amplitude du mouvement et le poids donné au mouvement. L'ensemble des mouvements de Métropolis ainsi définis correspond a un *cycle* de notre MCMC. Les différents poids donnés aux mouvements sont proportionnels au nombre de mouvements que chaque cycle va appliquer sur chaque noeud.
L'instance de `probModel` est elle même contenue dans un un objet de type `Chain` permettant la gestion des fichiers. Le modèle sauvegarde sa configuration détaillée régulière-

ment, l'arrêt et la reprise de la chaîne sont donc possibles. Dans un deuxième temps, les paramètres d'intérêts sont extraits des fichiers et la structure du modèle est utilisée pour extraire distribution marginale, approximant la probabilité a posteriori, des valeurs d'intérêt. Il est alors possible d'obtenir les caractéristiques de la distribution, typiquement : une moyenne, une variance et un écart de confiance à 95%.

La parallélisation a été faite en utilisant *Open MPI* (*Message Passing Interface*). L'avantage de la parallélisation est de répartir les calculs et la mémoire nécessaire a ceux-ci sur plusieurs processeurs. Le désavantage est que la communication entre processeurs est très lente par rapport aux temps d'accès mémoire. Il faut agréger l'information avant de la faire passer processeurs, ce qui est permis par les objets de type `MIRnode` et `MPInode`. Il faut aussi optimiser la répartition des gènes entre processeurs, l'algorithme utilisé est glouton, il place un par un, tous les alignements du plus grand au plus petit, dans les processeurs les moins pleins.

CHAPITRE 4

DESCRIPTION DES MODÈLES PHYLOGÉNÉTIQUES

Nous sommes maintenant en mesure de décrire en détail les modèles que nous avons implémentés. Nous allons donc, dans cette section, détailler les trois différents modèles en introduisant au fur et à mesure les différents concepts qui y sont associés.
Comme mentionné auparavant, l'idée de fond est de généraliser la méthode comparative pour tester les corrélations entre paramètres d'évolution moléculaire et traits d'histoire de vie. Mon travail porte plus précisément sur les corrélations dN/dS et traits. Or ces corrélations impliquent des effets spécifiques aux gènes et des effets globaux, dus à la dynamique populationnelle. L'idée a donc été de construire des modèles hiérarchiques modélisant les effets globaux afin de mieux caractériser les effets locaux.
Dans un premier temps, après avoir introduit les modèles à codons et les processus browniens, je redécrirai ici le modèle simple gène de *coevol*, mais uniquement avec dN/dS. Ensuite je proposerai deux versions de modèle hiérarchiques qui cherchent à discerner les effets globaux.

4.1 Notations

Le modèle est construit à partir d'une série de :

- P noms d'espèces et d'un arbre binaire enraciné. Les espèces sont les étiquettes des feuilles de l'arbre, qui contient donc $2P-1$ noeuds et $2P-2$ branches. Par convention, le noeud directement ancestral au noeud j est noté $Pa(j)$ et la branche reliant les noeuds j et $Pa(j)$ sera indicée j. Par convention également, la racine de l'arbre est indexée 0 et les feuilles sont indexées de 1 à P. Le temps de divergence associé à chaque noeud j est noté T^j avec $0 \leq j < 2P-1$. L'intervalle de temps séparant les évènements de spéciation correspondant aux noeuds délimitant la branche j est noté ΔT^j.

- Une liste de L caractères phénotypiques continus tels que la masse ou la longévité. Nous nous intéresserons ici à des caractères ayant entre eux des relations allomé-

triques log linéaires. La valeur strictement positive C_k^j exprime la valeur du phénotype k avec $0 \leq k < L$, pour toute feuille j, telle que $1 \leq j < P$. C est une matrice de taille $L \times (2P-1)$ définie pour tous les noeuds de l'arbre. Les noeuds internes sont de valeurs inconnues, les valeurs aux feuilles sont, en principe, déterminées par les données empiriques, on peut donc définir la matrice C' des données aux feuilles. C'est une matrice $L \times P$, qui doit contenir un minimum de données manquantes.

- Une série d'alignements de séquences codantes, indicés $i = 1...N$. Pour le premier modèle, on aura un seul alignement ($N = 1$) puis les deux autres modèles hiérarchiques utiliseront une série d'alignement ($N > 1$). Chaque alignement à un nombre d'espèces compris entre 2 et P et un nombre de positions respectif.

4.2 Modèle à Codon

En 1994, Muse and Gaut proposent de modéliser les substitutions de codons comme un processus de Markov en temps continu. Pour cela nous utilisons d'abord une matrice de substitution nucléotidique 4×4. Elle est supposée être un modèle général réversible en temps (GTR) (Simon Tavaré 1986) :

$$Q = \begin{pmatrix} * & \rho_{AC}\pi_C & \rho_{AG}\pi_G & \rho_{AT}\pi_T \\ \rho_{CA}\pi_A & * & \rho_{CG}\pi_G & \rho_{CT}\pi_T \\ \rho_{GA}\pi_A & \rho_{GC}\pi_C & * & \rho_{GT}\pi_T \\ \rho_{TA}\pi_A & \rho_{TC}\pi_C & \rho_{TG}\pi_G & * \end{pmatrix} \tag{4.1}$$

où π_n la probabilité à l'équilibre n, et $\rho_{n_1 n_2}$ le taux relatif de substitution du nucléotide n_1 vers n_2. En imposant par ailleurs que $\sum_{n_2} \rho_{n_1 n_2} = 1$ et $\sum_n \pi_n = 1$. ρ est donc un vecteur à 5 degrés de liberté et π un vecteur a 3 degrés de liberté. Une distribution de Dirichlet uniforme leur sert, à tous deux, de probabilité a priori (*prior*). Les éléments diagonaux sont calculés de manière à ce que la somme des entrées de chaque ligne soit égale à 0 :

$$Q_{n_1 n_1} = -\sum_{n_2 \neq n_1} \rho_{n_1 n_2} \pi_{n_2} \tag{4.2}$$

La matrice Q est normalisée, à partir de cette matrice de substitution nucléotidique, nous pouvons définir la matrice 61×61 définissant les taux de substitution entre chaque paire

de codons c_1 et c_2, différant en une seule position nucléotidique.

$$R_{c_1c_2} = \begin{cases} Q_{n_1n_2} & \text{si } c_1 \text{ et } c_2 \text{ sont synonymes} \\ \omega \cdot Q_{n_1n_2} & \text{si } c_1 \text{ et } c_2 \text{ ne sont pas synonymes} \end{cases} \tag{4.3}$$

Seules les substitutions entre deux codons qui diffèrent par exactement un nucléotide sont considérées. Si c_1 et c_2 n'ont pas deux positions identiques, $R_{c_1c_2} = 0$. La valeur de ω est donc le rapport entre le taux de substitution synonyme et le taux de substitution non synonyme. En l'absence de sélection positive, selon les principes de la théorie quasi neutre, ω estime ainsi la proportion de mutations non synonymes effectivement neutres, comme expliqué au chapitre 2. Pour compléter la description du processus de substitution, il reste à définir le taux absolu de substitution synonyme, λ. Les deux paramètres λ et ω sont modélisés comme des processus browniens, décrits dans le paragraphe suivant.

4.3 Processus browniens

Dans nos modèles, un certain nombre de traits quantitatifs varient de façon continue au cours du temps, sur chaque lignage de l'arbre. Par exemple : des traits d'histoire de vie comme la masse, ou certains paramètres du processus de substitutions comme le taux de substitution synonyme λ, ou le rapport ω. Afin que nos processus, à valeur positive, puissent varier dans l'espace des réels, nous leur imposons une transformation logarithmique.

Pour introduire le concept de processus brownien, nous définissons dans un premier temps le processus représentant les variations du taux de substitution λ,

$$X(t) = \ln \lambda(t). \tag{4.4}$$

On suppose que $X(t)$ est un processus brownien, dont la variance par unité de temps est appelée μ. Ce processus est caractérisé par des incréments normalement distribués de variance proportionnelle à μ. Plus précisément, la variation de $X(t)$ sur une période de temps finie est telle que

$$X(T) \sim \mathcal{N}(X(0), \mu T), \tag{4.5}$$

ou, de façon équivalente :

$$X(T) - X(0) \sim \mathcal{N}(0, \mu T). \tag{4.6}$$

En particulier si l'on considère une branche j de l'arbre, alors on note X^j, la valeur de $X(t)$ à l'extrémité la plus récente de la branche. De manière équivalente, on note $X^{Pa(j)}$ la valeur de $X(t)$ à l'extrémité la plus ancienne de la branche. Par conséquent :

$$X^j - X^{Pa(j)} \sim \mathcal{N}(0, \mu \Delta T^j), \tag{4.7}$$

ou, de façon plus compacte :

$$\Delta X^j \sim \mathcal{N}(0, \mu \Delta T^j). \tag{4.8}$$

Le modèle instancie les valeurs du processus seulement aux noeuds de l'arbre. Idéalement, pour calculer les probabilités de substitution le long de la branche il faudrait connaître l'ensemble du *pont brownien*, c'est-à-dire les valeurs prises par $\lambda(t)$ à chaque instant le long des branches. Les taux de substitution le long d'une branche donnée s'expriment alors :

$$e^{(\int \lambda(t)dt) \cdot R}. \tag{4.9}$$

En pratique, on fait une approximation en calculant une moyenne de $\lambda(t)$ par branche. Nous avons fait le choix de la moyenne arithmétique pour calculer $\overline{\lambda}^j$:

$$\overline{\lambda}^j = \frac{1}{2}\left(e^{X^j} + e^{X^{Pa(j)}}\right) \tag{4.10}$$

Cette approximation est assez grossière, mais des simulations suggèrent qu'elle n'affecte pas trop les résultats obtenus en pratique avec ces modèles [58]. Les probabilités de transition entre codons sont maintenant données par la matrice exponentielle :

$$e^{\overline{\lambda}^j \cdot \Delta T^j \cdot R} \tag{4.11}$$

Le taux de substitution synonyme λ, est le seul processus univarié dans tous les modèles considérés dans ce travail. Nous avons utilisé, comme prior pour μ, variance de $X(t)$ par unité de temps, une distribution exponentielle de moyenne 1. La valeur à la racine est tirée d'une distribution uniforme.

4.4 Matrice de covariance

Les *processus browniens multivariés* généralisent les processus browniens introduits dans la section précédente. Ils peuvent être paramétrés grâce à une matrice de covariance, ou alternativement, vus comme des processus browniens distincts reliés entre eux à via des régressions linéaires. Les deux représentations sont utilisées dans nos modèles. Pour illustrer la représentation par matrice de covariance, considérons l'exemple des traits d'histoire de vie. Nous avons $k = 1...L$ traits quantitatifs, observés chez les P espèces contemporaines. Ces valeurs correspondent à la matrice C' définie précédemment (section 4.1). On peut alors imaginer que ces traits sont le résultat d'un processus brownien log-normal multivarié qui "courent" le long des branches de l'arbre phylogénétique.

$$Y(t) = Y_k(t) \text{ avec } k = 1...L \tag{4.12}$$

Ce processus est contraint par les données : il doit correspondre aux feuilles de l'arbre aux valeurs observées, spécifiées par la matrice C'. Définissons donc la transformation logarithmique de ces valeurs pour tout noeud j de l'arbre ($j \in [0, 2P-2]$) :

$$Y_k^j = \ln C_k^j \tag{4.13}$$

les variations du processus sur une période de temps T sont donc un vecteur de dimension k distribué suivant une normale multivariée :

$$Y(T) - Y(0) \sim \mathcal{N}(0, \Delta T \Sigma), \tag{4.14}$$

ou encore :

$$\Delta Y^j \sim \mathcal{N}(0, \Delta T^j \Sigma), \tag{4.15}$$

où Σ est une matrice de covariance de dimension $L \times L$. Pour désigner individuellement les entrées de la matrice Σ, nous utilisons la notation suivante :

$$\Sigma = \begin{pmatrix} \langle Y_1, Y_1 \rangle & \langle Y_1, Y_2 \rangle & \dots & \langle Y_1, Y_L \rangle \\ \langle Y_2, Y_1 \rangle & \langle Y_2, Y_2 \rangle & \dots & \langle Y_2, Y_L \rangle \\ \dots & \dots & \dots & \dots \\ \langle Y_L, Y_1 \rangle & \langle Y_L, Y_2 \rangle & \dots & \langle Y_L, Y_L \rangle \end{pmatrix} \tag{4.16}$$

Ainsi, $\langle Y_k, Y_k \rangle$ est la variance par unité de temps du processus $Y_k(T)$ et $\langle Y_k, Y_l \rangle$ avec $k \neq l$ la covariance entre $Y_k(T)$ et $Y_l(T)$. A strictement parler, c'est la covariance entre les variations de Y_k et Y_l.
Σ est a priori distribuée suivant une distribution de Wishart Inverse, de paramètre Σ_0 [91], à L degrés de liberté, Σ_0 définie comme la matrice identité de même dimension. La valeur à la racine Y_k^0 de chacun des traits d'histoire de vie a une prior uniforme entre les limites -100 et $+100$. L'utilisateur a aussi la possibilité de donner la moyenne et la variance d'une distribution normale, qui se substituera à la prior uniforme pour chacun des paramètres Y_k^0.

4.5 Modèle simple-gène

Le modèle simple-gène est construit avec un seul alignement, $N = 1$, et un nombre L de traits d'histoire de vie. Les temps de divergence du chronogramme sont échantillonnés d'une distribution de probabilité a priori uniforme. Le rapport entre le taux de substitution synonyme et le taux de substitution non synonyme est noté Y_0 en appliquant la transformation habituelle.

$$Y_0(t) = \ln \omega(t) \tag{4.17}$$

Comme ci-dessus $Y_k, k = 1...L$ représente les logarithmes naturels des traits d'histoire de vie. Y_0 est maintenant considéré comme une des composantes du processus multivarié Y. La matrice de covariance, Σ, contient maintenant les covariances respectives entre ω et les traits quantitatifs : $\langle Y_0, Y_k \rangle$ pour $k = 1...L$.
En résumé, les principales composantes du modèle sont donc :

- ρ le taux relatif de substitution nucléotidique
- π la probabilité à l'équilibre nucléotidique
- λ le taux de substitution synonyme
- T les temps de divergence de la phylogénie
- Y_k le processus multivarié
- Σ la matrice de covariance

Ce modèle est quasiment similaire au modèle présenté dans notre papier de 2010 *coevol* [58] en annexe I. La différence majeure est que λ le taux de substitution synonyme est un processus univarié, séparé de Y. À l'inverse, dans *coevol*, λ, ω et les traits d'histoire de vie sont rassemblés dans un seul processus multivarié de dimension $L+2$.

4.6 Modèle hiérarchique simple

Après avoir développé le premier modèle, nous avons multiplié les tests, indépendamment sur plusieurs alignements, et avons constaté des différences substantielles dans la reconstruction des traits d'histoire de vie en fonction du gène utilisé. En effet la matrice de covariance et les processus qui en dépendent s'influencent mutuellement et selon l'alignement l'estimation de la valeur ancestrale, par exemple, n'est pas la même. Or bien que chaque alignement corresponde à une histoire spécifique du gène, il n'existe, bien entendu, qu'une seule histoire des traits de vie de l'espèce

L'idée est donc ici de rassembler dans un seul modèle plusieurs alignements. C'est un *modèle hiérarchique* que nous construisons ici. Chacune des histoires substitutives propres à chaque gène dépend à la fois de paramètres spécifiques et de paramètres communs à tous les gènes. Le modèle hiérarchique simple utilise donc une série d'alignements A_i avec $0 \leq i < N$. Les données sont au même niveau et partagent les hyperparamètres, qui représentent le "tronc commun" du modèle. Cette structure, utilisée pour la parallélisation du modèle, nous a poussés à dissocier d'une part, les paramètres *gènes spécifiques* du modèle propre à chaque alignement et tous indicés par *i*, d'autre part : des paramètres *globaux* considérés comme homogènes à l'échelle du génome entier.

À chaque gène correspond ici une part du modèle contenant un processus $\omega_i(t)$. Chacun des ω_i corrèle avec les traits d'histoire de vie, cependant, ils ne sont pas inclus dans la matrice Σ. Les relations entre ω_i et les autres processus de la matrice sont définis comme une série de régressions linéaires. Ainsi le long d'une branche j la variation :

$$\Delta \ln \omega_i^j = \sum_{k=1}^{L} \alpha_{ik} \cdot \Delta Y_k^j + \varepsilon_i^j \tag{4.18}$$

- α_{ik} est l'estimation du coefficient de régression linéaire entre ω_i et le trait de vie k.
- pour rappel, $\Delta \ln \omega_i^j$ la variation logarithme du trait de vie k sur la branche j.

- $\varepsilon \sim \mathcal{N}(0, \sigma_i^2 \Delta T^j)$ est l'erreur résiduelle de la régression linéaire.

Le coefficient α_{ik} n'est pas équivalent à $\langle \ln \omega_i, \ln C_k \rangle$, la covariance entre la pression de sélection ω_i et le trait de vie C_k. Celle-ci peut néanmoins être retrouvée de la façon suivante :

$$\langle \ln \omega_i, \ln C_k \rangle = \sum_{l=0}^{L} \alpha_{il} \langle \ln C_l, \ln C_k \rangle \tag{4.19}$$

Les deux valeurs expriment des choses légèrement différente. La covariance accumule les effets conjugués des différentes corrélations. Elle ne prend pas en compte l'incertitude liée, par exemple, aux covariances fortes entre traits de vie. Les corrélations sont en général moins significatives, mais prenne en compte les effets d'interactions entre processus. Pour simplifier nous n'utilisons dans nos résultats que des corrélations.
$\ln \omega_i^0$, les valeurs à la racine, sont tirées d'une distribution gamma telle que :

$$\ln \omega_i^0 \sim Gamma(\beta_0, \beta_1), \tag{4.20}$$

où β_0 est le paramètre de forme et β_1 le paramètre d'échelle estimé une seule fois pour tous les gènes dans le modèle.

4.7 Modèle $\Omega-$hiérarchique

Le modèle précédent estime indépendamment les variations des $\omega_i(t)$, pour chaque gène. Comme nous l'avons dit dans la section 2.3, ω_i peut se décomposer en deux parties : d'une part les variations spécifiques à chaque gène, qui reflètent potentiellement les variations de la sélection sur ce gène, et d'autre part les variations qui sont causées par des facteurs environnementaux et par les fluctuations populationnelles. Cet argument nous a poussés à ajouter au modèle un processus $\ln \Omega$ qui représente les variations partagées par tous les gènes.
Ces variations globales sont représentées par un processus $\ln \Omega$. On pose donc :

$$\begin{cases} Y_0(t) = \ln \Omega(t) \\ Y_k(t) = \ln C_k(t) \end{cases} \tag{4.21}$$

Y_k est maintenant défini pour $k = 0...L$ et Ω est maintenant considéré comme une composante du processus multivarié Y. Σ, la matrice de covariance, en plus des covariances entre

traits, contient, les covariances entre Ω et les traits quantitatifs : $\langle Y_0, Y_k \rangle$ pour $k = 1...L$.
Enfin, les variations de ω_i sont reformulées comme la combinaison de deux effets, un effet global et un effet spécifique au gène :

$$\Delta \ln \omega_i^j = \Delta \ln \Omega^j + \sum_{k=1}^{L} \alpha_{ik} \cdot \Delta Y_k^j + \varepsilon_i^j, \tag{4.22}$$

où $\varepsilon \sim \mathcal{N}(0, \sigma_i^2 \Delta T^j)$ est l'erreur résiduelle de la régression linéaire. La valeur absolue de Ω n'étant d'aucune utilité, nous avons arbitrairement fixé sa valeur à la racine : $\Omega^0 = 1$. Le modèle est ainsi identifiable.
Ce nouveau paramètre du modèle est intéressant puisqu'il représente explicitement les effets de variation du dN/dS partagé par tous les gènes. Mais sa signification reste difficile à discuter. A priori, sur un ensemble de séquences représentatives du génome, on s'attend a ce qu'il reflète les variations de N_e. Nous discuterons plus loin de la pertinence de ce nouveau paramètre Ω

CHAPITRE 5

AUTRES MÉTHODES

5.1 Nettoyage des données

Nous avons utilisé la base de données Orthomam [84] dans sa version 6 (septembre 2010). Elle contient 12177 alignements de gènes orthologues. Pour chacun d'entre eux, un ensemble de séquences nucléotidiques codantes a été conceptuellement traduit en acide aminé puis rétrotraduit en codons. Cette méthode aboutit à des alignements qui respectent les cadres de lectures des gènes, et nous permet de travailler au niveau des substitutions de codons. Ces alignements ont été constitués à partir des 36 mammifères dont le génome complet a été publié sur le site du projet *Ensembl* [41]. Nous avons choisi de nous intéresser seulement aux placentaires, réduisant le nombre d'espèces dans l'arbre à 33, et nous avons mis en place une méthode heuristique pour nettoyer le jeu de donnée. Les deux marsupiaux et le monotrème ont été retirés après le nettoyage du jeu de données. Ce nettoyage a pour but d'annoter les *mauvaises* positions, conséquences d'erreurs de séquençages, erreurs d'annotation, mauvaise détection d'orthologie, décalage lors de l'alignement. Les mauvaises positions seront retirées du jeu de donnée. Une difficulté est que, pour conserver une couverture représentative du génome mammifère, il faut minimiser le nombre de positions retirées. Notre procédure a pour but de retirer de chaque alignement les espèces et les morceaux d'alignement qui semblent anormalement différents de ce qui est attendu.

- Tout d'abord, nous avons utilisé *HMMER* [26] qui construit des profils de chaîne de Markov cachés [85]. Chaque alignement sert de modèle pour construire un profil. Chaque séquence de l'alignement est comparée au profil correspondant, des scores pour chaque position nucléotidiques sont donnés par *HMMER*. Les positions connues sont remplacées par des positions inconnues quand une série de scores est trop éloignée du score maximal. La procédure parcourt la séquence et détecte les suites de positions nucléotidiques dont le score moyen est plus bas qu'un certain seuil.

- Quand l'opération précédente retire plus du quart des positions informatives d'une séquence, elle est éliminée de l'alignement.

- Nous utilisons alors *Gblock* [12] qui retire les positions d'un alignement qui contiennent trop de positions non informative.
- Nous ne gardons que les alignements dont la taille est entre 160 et 2000 codons.

Les reconstructions de ω sur une série d'alignements, après et avant filtrage, ont montré une nette amélioration de la distribution des variations du mouvement brownien correspondant, notamment, pour ce qui a trait à la variance moyenne des mouvements browniens telle que déterminé par un test de normalité (le test de D'Agostino). Le *pipeline* permettant à la procédure entière d'être exécuté sans intervention extérieure a été fait en *bash*. Cette méthode utilisant les Chaînes de Markov a été, par la suite, modifiée et utilisés sur un jeu de données de séquences d'acides aminés portant sur la résolution de l'arbre des mollusques dans une étude en cours de publication avec Hervé Philippe et Béatrice Roure.

5.2 Test de permutation

La méthode présentée permet d'analyser simultanément une grande quantité de gènes. Il est donc intéressant de chercher à savoir si certains groupes fonctionnels de gènes ont des résultats significativement différents des autres. Pour constituer les groupes de gènes, nous avons choisi d'utiliser Gene Ontology [1] associant a chaque gène une liste de termes. Chaque terme est, par exemple, le domaine cellulaire où la protéine correspondante à été observée, la fonction biologique dans laquelle cette protéine est impliquée. Nos résultats associent à chaque gène g des valeurs V_i^g avec $i \in \mathcal{N}$, indiquant par exemple la probabilité a posteriori de la covariance entre masse et pression de sélection. Le problème est donc de déterminer si un terme t, annotant un groupe de gènes de taille N^t, a une moyenne $\bar{V}_i^t$ significativement plus petite pour une valeur i donnée. Le test de permutation est très intuitif. L'idée est de tirer aléatoirement N^t gènes et de comparer la moyenne de ces gènes à $\bar{V}^t$ pour la valeur en question. La probabilité que cette moyenne soit supérieure est d'autant plus grande que $\bar{V}_t$ est petit. Il suffit donc de répéter cette opération un maximum de fois et de conserver la fréquence à laquelle $\bar{V}_t$ est plus petite. Cette méthode permet d'obtenir une *p-value* pour chaque terme, pour chaque valeur d'intérêt. Elle a été codée en *PERL*

CHAPITRE 6

RÉSULTATS ET DISCUSSION

Ce chapitre est consacré aux résultats que nous discuterons au fur et à mesure de leur présentation. Le premier jeu de donnée concerne 17 gènes orthologues chez 77 placentaires, associés à trois traits quantitatifs : la masse, la longévité et la maturité sexuelle femelle. Nous avons calculé les variations de ω à l'aide du modèle simple-gène, successivement sur les 17 gènes, ainsi que sur leur concaténation. Nous utilisons ensuite les deux modèles hiérarchiques pour montrer leurs différences sur le même jeu de données. À travers ces résultats nous décrivons les avantages et les inconvénients de chacun des trois modèles introduits précédemment.

Dans un deuxième temps, nous détaillerons les résultats d'une étude plus ambitieuse où nous avons utilisé un très grand nombre de gènes alignés à partir de génomes complets pour 33 mammifères placentaires et avons regardé leurs corrélations avec les trois mêmes traits de vie que précédemment. Pour ceci nous nous sommes focalisés sur le modèle $\Omega-$hiérarchique pour des raisons expliquées plus loin. Nous avons donc obtenu des résultats par gènes et par trait de vie que nous ne pouvons montrer ici. Afin d'extraire l'information, nous avons regroupé les gènes selon leur annotation telle que spécifiée par la base de données *Gene Ontology*. À l'aide d'un test multiple, nous pouvons alors faire ressortir les fonctions biologiques les plus significatives dans nos résultats.

6.1 Comparaison des modèles

Nos modèles comprennent toujours une composante qui estime la covariance entre traits quantitatifs, s'inspirant de la méthode des contrastes indépendants et intégrant donc sur l'ensemble de la phylogénie. Ils reconstruisent pour cela l'histoire des variations de ces traits le long de l'arbre. Plus précisément, les variations du logarithme de la valeur de chaque trait sont décrites par un processus brownien courant le long de l'arbre. Les valeurs aux feuilles sont connues et le processus est donc contraint à ces valeurs. Sur la figure 6.1 nous montrons l'arbre utilisé dans la première étude, ainsi qu'une représentation de la distribution marginale de la longévité aux nœuds de cet arbre, au cours du temps

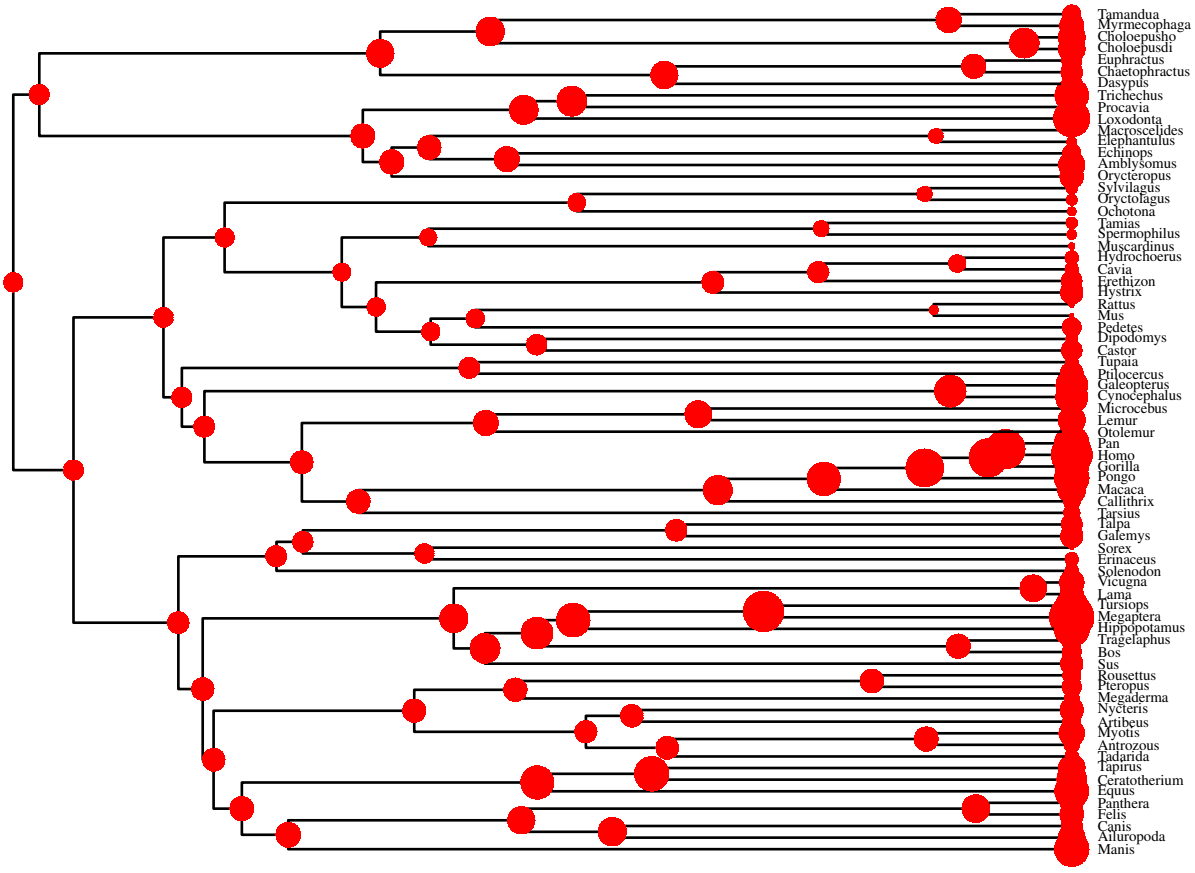


Figure 6.1 – Une reconstruction de la longévité

selon le modèle simple gène. L'aire de chaque cercle est proportionnelle à la valeur de la longévité de l'ancêtre correspondant. Les bornes inférieures et supérieures des intervalles de crédibilité à 95%, pour chaque valeur estimée, sont représentées à chaque cladogénèse par les différents niveaux de rouge, comme *manis* ou *galeopterus*. Certaines valeurs aux feuilles ne sont pas connues, ces données manquantes sont donc estimées par le modèle.

Lorsque plusieurs traits sont analysés simultanément, ils forment alors un processus brownien log-normal multivarié. La variance de chaque processus et les covariances entre composantes définissent une matrice de covariance Σ. L'estimation simultanée de la matrice et du processus permet au modèle d'estimer les corrélations entre traits tout en corrigeant pour l'inertie phylogénétique et l'incertitude sur les valeurs ancestrales des traits. Réciproquement, la reconstruction des valeurs ancestrales prend en compte les corrélations entre traits.

En plus des traits d'histoire de vie, le processus multivarié intègre aussi des paramètres d'évolution moléculaire. La modélisation de l'évolution des séquences est donc la deuxième

composante du modèle. Plus précisément, la méthode reconstruit l'histoire des substitutions de codons le long des branches de la phylogénie à partir d'un ou plusieurs alignements de séquences nucléotidiques codantes. En fait, elle intègre sur toutes les histoires substitutives probables selon un processus de substitution Markovien défini au niveau des codons et paramétré par $\omega = dN/dS$. Le modèle utilisé estime donc le rapport entre substitution synonyme et substitution non synonyme, ω. Ce rapport peut être interprété comme une mesure de la force de la sélection appliquée à un instant donné sur la séquence en question. C'est un paramètre d'évolution moléculaire qui varie au cours du temps sur chaque branche de l'arbre, théoriquement, en fonction de deux facteurs : l'importance fonctionnelle du gène, mais aussi la dynamique populationnelle telle que résumée par le concept abstrait de taille *efficace* de population, notée N_e. Si N_e diminue la sélection est moins efficace par rapport à la dérive génétique et donc le $\frac{dN}{dS}$ monte comme cela à été discuté précédemment au paragraphe 2.1

L'importance fonctionnelle de chaque gène pris individuellement est l'objet principal de notre étude. Pour cela, nous regarderons les covariances de ω et de chacun des traits quantitatifs ou traits d'histoire de vie. Cette covariance nous permet d'essayer de qualifier et de quantifier le lien entre les fonctions moléculaires des gènes et les phénotypes décrits par le trait. Par exemple, une allométrie significativement négative entre le ω_i d'un gène i et la longévité signifie que la masse augmente quand ω diminue, et donc, que la sélection purificatrice est plus forte sur ce gène quand la masse augmente. Inversement, le gène accumule plus de mutations délétères quand la longévité diminue. Une interprétation possible de cette allométrie est que cette l'implication de la protéine codée par le gène i dans les fonctions ou les traits covariant positivement avec la longévité. Plus précisément, on pourra supposer que le gène en question est potentiellement utile au maintien des fonctions cellulaires ou physiologiques de l'organisme dans les phases plus avancées de sa vie. En d'autre termes, nous pourrons ainsi définir un ensemble de gènes potentiellement impliqués dans la lutte contre le vieillissement. Ajoutons qu'une covariance significativement positive pourra aussi permettre de formuler des hypothèses.

En pratique, les effets de changement de taille efficace et de variation de l'importance fonctionnelle sont superposés dans la variation reconstruite de $\omega_i(t)$. Tout l'enjeu est donc de séparer leurs contributions respectives.

6.1.1 Description des données

Nous avons utilisé ici trois traits d'histoire de vie à partir de la base de données en ligne *AnAge* [17]

- **Gen :** Le premier trait de vie utilisé est la maturité sexuelle femelle. C'est une borne inférieure du temps de génération moyen dans une espèce donnée. En pratique, la maturité sexuelle est largement utilisée comme approximation pour le temps de génération.
- **Mas :** Nous avons ensuite utilisé la masse corporelle d'un individu adulte.
- **Lon :** Finalement le record de longévité enregistré, qui quantifie la longévité intrinsèque de l'espèce. Cette donnée est souvent critiquée, car elle est fortement biaisée par les conditions environnementales de l'espèce. La sélection utilitaire des espèces par l'homme favorise en général le vieillissement des individus tout en augmentant le nombre d'individus dont on connaît la durée de vie. Malgré tout, cette valeur est considérée comme un bon estimateur de la longévité, dans le cas des mammifères. En effet, à l'intérieur d'une espèce, le vieillissement commence à un âge assez précis et le record de longévité est en général du même ordre que cet âge.

De manière générale, malgré leur imprécision, les données disponibles permettent une estimation qualitativement suffisante pour mesurer les variations des traits d'histoire de vie le long de l'arbre. Ceci est d'autant plus vrai avec des traits pour lesquels des variations extrêmes existent chez les espèces contemporaines, comme dans notre cas, chez les placentaires. Les 20 valeurs inconnues, ainsi que la longévité humaine, considérée comme surestimée à cause de la taille de l'échantillon, seront imputées par le modèle.
Nous avons mené 4 différents types d'analyse sur un jeu de données constitué de 17 alignements de séquences d'ADN nucléaire codant pour autant de protéines. Les 17 protéines : ADORA3, ADRA2B, ADRB2, APOB, ATP7A, BDNF, BRCA1, CNR1, GHR, PNOC, RAG1, RAG2, RBP3 (ex-IRBP), S1PR1 (ex-EDG1), TYR, VWF, and ZFX proviennent de l'article [59]. Les alignements couvrent 73 euthériens, ou mammifères placentaires. La topologie de l'arbre de la figure 6.1 suppose les afrothériens et les xénarthres monophylétiques, c'est l'hypothèse des clades Atlantogenata. L'élimination des monotrèmes et des marsupiaux pour ne travailler que sur les placentaires permet de travailler sur un arbre plus homogène en spéciations. L'arbre sur lequel porte cette étude comporte

en fait peu de longues branches, ce qui favorise une meilleure information des composantes du processus multivarié à chaque nœud de l'arbre et élimine de l'incertitude. Les temps de divergences, bien qu'ils puissent en principe être coestimés dans le cadre de notre modèle, ont été ici estimés à partir d'autres séquences et sont fixes dans toutes nos analyses. Nous avons utilisé quelques calibrations fossiles dans l'estimation des longueurs de branche. La raison de ce choix est pratique : le temps de convergence des analyses par MCMC s'allonge quand les temps de divergences sont estimés simultanément avec les autres paramètres.

6.1.2 Résultats pour la concaténation

Nous avons d'abord lancé deux chaînes sur la concaténation des 17 gènes, avec le modèle multivarié, dit simple-gène, de *coevol* [58]. Dans cette concaténation contenant 5039 codons, chacune des 73 espèces contient environ 16.5% d'indels ou de positions non informées. À supposer que cette concaténation est représentative de l'ensemble des séquences codant pour des protéines, le but de cette analyse est alors d'approximer les tendances globales des variations de ω, dans le génome nucléaire, au cours de l'évolution des placentaires.

Mais revenons tout d'abord sur la figure 6.1. La marginalisation des estimations a posteriori de la longévité nous montre plusieurs choses. Tout d'abord, on trouve une longévité à la racine de valeur moyenne 19.77 ans, avec un intervalle de crédibilité à 95% allant de 13.35 ans à 27.64 ans. La moyenne des âges connus aux feuilles est d'environs 22 ans, la différence avec la moyenne de la distribution a posteriori n'est pas significative, mais cette tendance correspondrait à un ancêtre plus jeune que la moyenne des observations. Cet arbre montre une augmentation de la longévité. Ces résultats pourraient s'expliquer par le biais d'une extinction préférentielle des gros mammifères, observée surtout dans les tropiques [28].

	Gen	Mas	Lon	ω
Gen	2.07	3.39	0.51	0.30
Mas	>0.99	13.90	1.75	0.11
Lon	>0.99	>0.99	0.56	0.04
ω	0.99	0.67	0.78	0.23

Tableau 6.I – Matrice de covariance sur la concaténation des alignements

Le tableau 6.I montre les variances par unité de temps de chacune des différentes composantes du processus multivarié, ainsi que les covariances et la probabilité a posteriori d'une covariance positive pour chaque paire de composantes telles qu'estimées par le modèle. La probabilité a posteriori d'une covariance positive est dans la partie sur fond gris dans les tables. Nous avons marginalisé sur l'ensemble des réalisations de Σ pour deux chaînes indépendantes.

Ces estimations de covariance ainsi obtenues entre le temps de génération, la masse, la longévité et ω permettent de confirmer plusieurs hypothèses. En premier lieu, les covariances entre traits d'histoire de vie sont significativement positives. C'est un constat qui a déjà été relevé de nombreuses fois [10] et qui correspond à une explication classique en terme d'histoire de vie. On observe en effet, parmi les placentaires connus aujourd'hui, une corrélation entre la masse et la longévité des individus. Nous suggérons ici que c'est à travers l'influence de l'environnement que les traits de vie évoluent, et notamment selon les stratégies de survie r/K comme expliqué à la section 1.2.2. L'explication que nous proposons pour la corrélation positive entre longévité et temps de génération justifie la corrélation fortement positive entre ces deux composantes du processus multivarié.

La masse corrèle très bien avec ces deux composantes, on suppose que pour permettre aux fonctions biologiques de la lignée somatique de durer, la masse doit augmenter. D'autres traits de vie comme la fécondité, le métabolisme, le rythme cardiaque, la consommation calorique corrèlent positivement ou négativement chez les mammifères [10]. Intuitivement on peut penser qu'une masse importante augmente la résistance aux prédateurs.

Nous nous tournons maintenant vers la composante ω. Elle est hypothétiquement liée à N_e dans ce cas puisque la concaténation étudiée comprend des gènes dont la fonction est très différente de sorte qu'elle peut être considérée comme représentative du protéome nucléaire entier. La covariance est significativement positive entre ω et le temps de génération. Ce résultat montre que les branches où le temps de génération augmente sont celles où la probabilité relative de fixation des mutations non synonymes sur les 17 gènes augmente aussi. Si les 17 gènes sont représentatifs des génomes placentaires et que l'on considère que la grande majorité des mutations non synonymes sont délétères, alors l'augmentation du temps de génération est concomitante à la diminution de l'intensité de la sélection purificatrice. Inversement, la diminution de la maturité sexuelle correspond à une plus grande conservation des gènes. Les covariances entre les autres traits de vie et ω sont positives. Nous interprétons ici cette covariance comme étant le résultat d'une corrélation

négative entre masse et taille efficace de population.

6.1.3 Résultats du modèle simple gène

L'analyse précédente sur la concaténation ne nous a pas permis de détecter d'effet spécifique à chaque gène. Une première approche pour détecter de tels effets consiste tout simplement à utiliser le même modèle indépendamment sur chacun des gènes. Lors de cette expérience, nous avons donc lancé deux chaînes pour chacun des 17 gènes, soit 34 chaînes. Le modèle a donc convergé de manière indépendante sur chacun des gènes.

identifiant	GEN	MAS	LON	identifiant	GEN	MAS	LON
ADORA3	0.81	0.83	0.80	ADRA2B	0.61	0.52	0.84
ADRB2	0.83	0.90	>0.99	APOB	0.99	0.98	0.98
ATP7A	0.90	0.97	0.96	BDNF	0.59	0.68	0.47
BRCA1	0.52	0.48	0.65	CNR1	0.14	0.12	0.13
GHR	0.17	0.16	0.16	PNOC	0.48	0.49	0.48
RAG1	0.94	0.99	>0.99	RAG2	0.92	0.94	0.95
RBP3	0.23	0.14	0.02	S1PR1	0.61	0.63	0.98
TYR	0.85	0.91	0.98	VWF	0.59	0.80	0.58
ZFX	0.44	0.34	0.40	concaténation	0.99	0.67	0.78

Tableau 6.II – Part de la distribution a posteriori de la covariance qui est positive entre ω et traits d'histoire de vie

La table 6.II montre les probabilités a posteriori d'une covariance positive, pour chaque gène, pour chaque trait de vie. Ils permettent de formuler des hypothèses sur l'importance fonctionnelle de chacun des gènes.

Ici le gène RBP3 (*Retinol-binding protein 3*) corrèle très négativement avec la longévité. Il serait donc très conservé chez les espèces qui vivent longtemps, et sous une pression de sélection faible chez les espèces dont la longévité est petite. Le gène RBP3 est impliqué dans le transport des rétinoïdes et joue ainsi un rôle dominant dans la vision. La relation entre l'acuité de la vision et la longévité ne semble pas immédiate, mais permet d'envisager certaines hypothèses.

Le gène GHR, *Growth Hormone Receptor* n'est pas significativement corrélé aux trois traits de vie, mais la tendance est qu'il est maintenu sous une forte pression de sélection chez les gros. On s'attend donc à trouver une implication de ce gène dans des processus biologiques liés à la stratégie K. Cette hormone est aussi utile pour la croissance rapide

des plus petits, ce qui explique peut être la non-significativité de sa corrélation.

	Gen	Mas	Lon	ω
Gen	2.08	3.62	0.58	0.09
Mas	>0.99	14.15	1.81	0.23
Lon	>0.99	>0.99	0.57	0.07
ω	0.62	0.63	0.66	0.33

Tableau 6.III – Matrice de covariance moyenne obtenue grâce au modèle simple gène sur chacun des 17 alignements

La table 6.1.3 montre la moyenne des marginalisations de Σ pour toutes les 34 chaînes. Les résultats de la table 6.1.3 sont consistants avec les résultats de la concaténation de la table 6.I. Cela suggère que la part des variations de ω qui est propre à chaque histoire de substitution de gène s'est ajoutée aux variations globales par opposition à la concaténation. On constate néanmoins une perte de significativité des covariances moyennes entre traits de vie et ω, notamment avec le temps de génération. On remarque aussi que la variance moyenne des processus ω est plus haute que dans le cas de la concaténation.
Du fait que le modèle estime pour chaque gène l'ensemble des paramètres, les paramètres qui ne sont pas spécifiques à un gène sont estimés en fonction d'une part du génome peu significative. Par exemple, si on compare la distribution de l'âge à la racine estimé pour le gène ADRB2 et pour le gène RBP3, deux gènes dont les covariances avec la longévité sont très différentes, alors, le gène ADRB2 on trouve une distribution de longévité, en années, dont 95% se trouve dans l'intervalle $[12.49, 29.19]$, tandis que pour le gène RBP3, l'intervalle est $[7.95, 19.72]$. Évidemment l'histoire de la longévité des mammifères est unique et l'imprécision observée est nuisible à nos résultats. Mais plus fondamentalement, on observe que des corrélations sont positives sur l'ensemble des gènes. Or on sait que c'est à travers l'effet de la taille efficace, N_e.
L'idée est donc de soustraire les effets des variations du N_e afin de déterminer la part des variations spécifique à chacun des gènes. Cette constatation nous a suggéré de développer plutôt des modèles hiérarchiques.

6.1.4 Modèle hiérarchique simple

Dans ce premier modèle hiérarchique, tous les alignements sont conditionnés par un même jeu d'hyperparamètres qui sont les variables globales. L'ensemble des variables

globales, partagées par tous les gènes, est donc : les temps de divergence entre espèces, la reconstruction du taux de mutation synonyme μ, l'histoire des traits de vie, la matrice de covariance Σ ainsi que la distribution a priori des valeurs de ω_i à la racine. Par contre ici les variations de ω_i le long des branches sont reconstruites indépendamment pour chaque gène. Chaque $\omega_i(t)$ corrèle avec les traits d'histoire de vie comme indiqué dans l'équation 4.18.

	Gen	Mas	Lon	ω
Gen	2.28	3.98	0.56	0.01
Mas	>0.99	14.47	1.80	-0.05
Lon	>0.99	>0.99	0.54	0.05
ω	0.53	0.46	0.70	0.61

Tableau 6.IV – Matrice de covariance obtenue à partir du modèle hiérarchique simple

Ce modèle hiérarchique, dit simple, a permis de faire trois simulations MCMC sur l'ensemble de 17 alignements. Les moyennes, pour les trois simulations, sont données dans la table 6.IV. Les variances et covariances des 3 traits d'histoire de vie sont les marginalisations de la matrice Σ. Le coefficient α propre à chaque gène et à chaque trait de vie peut donc être utilisé pour retrouver la covariance équivalente comme expliqué dans l'équation 4.19. Chaque simulation à permis d'obtenir les 17×3 covariances entre ω et les traits quantitatifs, leur moyenne est donc reportée comme covariance dans la table 6.IV. La variance est ici la moyenne des variances de chacun des processus, elle est plus élevée, car les parts de variations propres a chacune des covariances ne sont pas retirées. Ce modèle a l'avantage d'estimer une seule fois le taux de substitution synonyme λ ainsi que le processus multivarié contenant les traits d'histoire de vie.

6.1.5 Modèle $\Omega-$hiérarchique

Dans le modèle hiérarchique précédent, chaque gène partageait une même tendance à avoir des corrélations positives entre leur dN/dS et les traits d'histoire de vie. Cette tendance est captée par les α_i.

Toutefois en réalité cette tendance est due à une variable cachée $N_e(t)$ qui subit des variations le long de l'arbre et qui est supposée corrélée négativement avec les traits d'histoire de vie. Ceci suggère que, plus fondamentalement, c'est l'histoire de $N_e(t)$ à travers l'arbre qui doit être estimée et partagée. Cette idée est à l'origine de ce modèle hiérarchique, mais

N_e est appelé ici Ω, c'est la part de la pression de sélection qui est commune à tous les gènes.

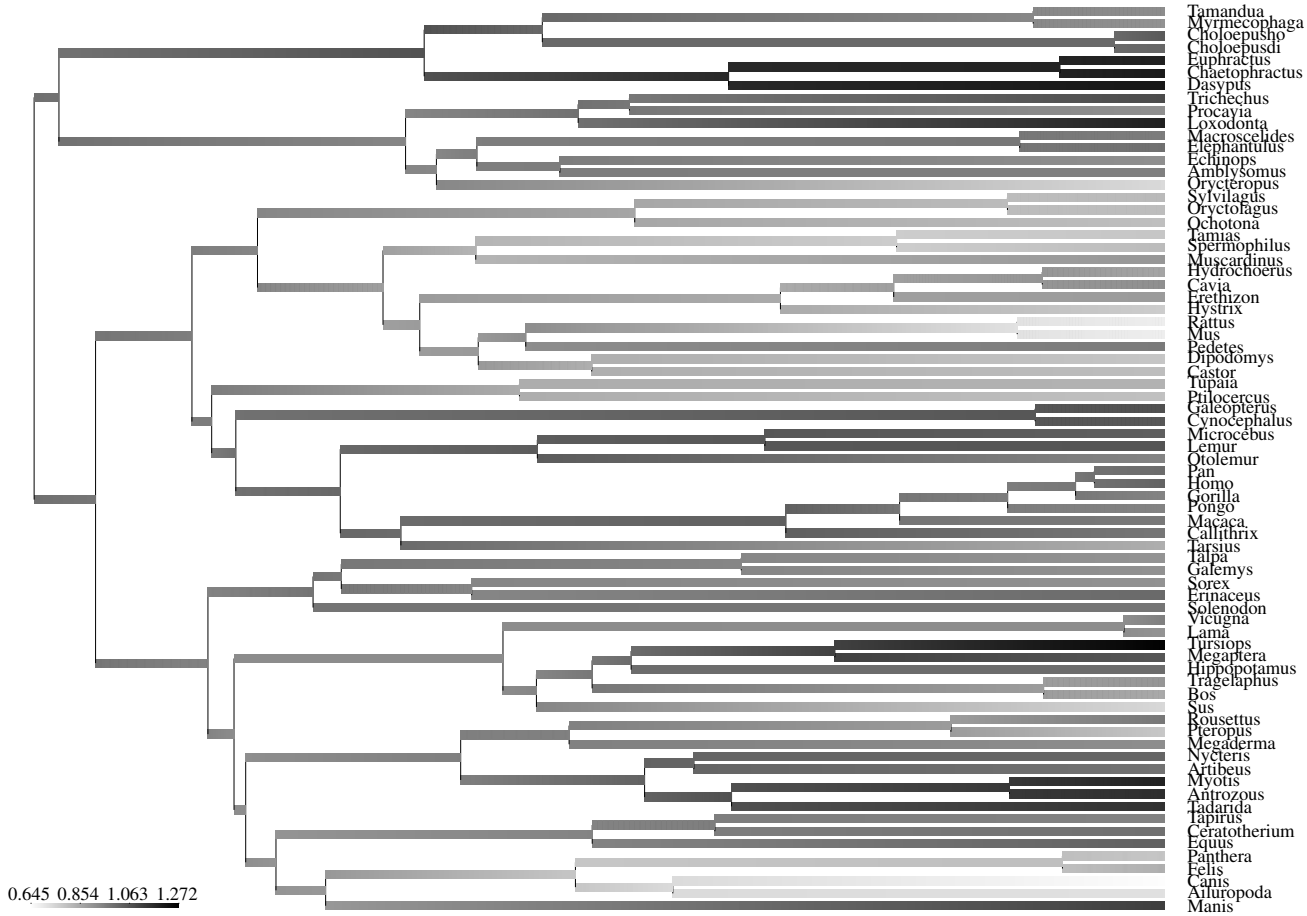
	Gen	Mas	Lon	Ω
Gen	2.23	3.98	0.58	0.01
Mas	>0.99	14.66	1.85	-0.08
Lon	>0.99	>0.99	0.55	0.02
Ω	0.54	0.36	0.68	0.09

Tableau 6.V – Matrice de covariance obtenue grâce au modèle Ω−hiérarchique sur les alignements

Dans ce modèle, les coefficients α_i montrent cette fois la part de variation spécifique à chaque gène qui corrèle à chacun des traits de vie.
La table 6.V montre les résultats obtenus grâce au modèle Ω−hiérarchique. C'est donc une moyenne sur deux simulations de la matrice Σ. Ces résultats sont à comparer avec la table 6.IV. On retrouve des résultats presque identiques concernant les covariances entre traits de vie. La composante Ω est maintenant la part du rapport entre le taux de substitution synonyme et le taux de substitution non synonyme qui est commune à tous les gènes. Ω montre les mêmes covariances que les moyennes des processus ω du modèle hiérarchique simple.

La partie globale des variations de dN/dS, Ω, est supposée être lié aux variations de taille efficace de population. La figure 6.2 montre ces variations sur l'ensemble de la phylogénie des placentaires. Sur cette figure les branches où Ω décroît, devraient correspondre à des augmentations de taille efficace de population. Inversement, les branches où Ω croit, sont supposé indiquer une taille efficace de population en diminution.
Sur une telle échelle évolutive, les violations de modèles, les erreurs de séquençage ou d'alignement pourraient influencer les résultats. Nous constatons malgré tout, des résultats à peu près concordants avec les connaissances des groupes représentés sur cet arbre. Les rongeurs ont de petites valeurs de Ω conformément à ce que l'on pourrait attendre, étant donné leurs grandes tailles de population. Les ordres des cétacés et la famille des *Dasypodidae* ont effectivement des tailles de population plus réduites. On observe une forte augmentation de Ω chez les éléphants de savane *Loxodonta Africana*, correspondant, là aussi, à une petite taille de population. D'un autre coté on constate de très fortes valeurs de Ω pour les dauphins (*tursiops*) qui ne s'expliquent pas si facilement en terme de taille

Figure 6.2 – Variation du Ω pour le modèle Ω−hiérarchique



de population. En effet, les dauphins n'ont pas spécialement de taille de population plus petite que les baleines, leur espèce soeur, or les Ω respectifs On constate aussi des valeurs très fortes de Ω sur les branches courtes, ce qui pourrait être dû à un biais du modèle ou à des erreurs d'alignements.

Les valeurs des covariances entre gènes et traits d'histoire de vie sont montrées dans le tableau 6.VI. Ce tableau donne pour les trois derniers modèles les valeurs des covariances entre le ω du gène k et le trait d'histoire de vie i, noté α_i^k, moyennés sur, au moins, deux simulations. Pour chacun des α_i^k on y donne la moyenne de la covariance et la probabilité a posteriori que cette covariance soit positive. Par exemple, la probabilité a posteriori que $\alpha_{RAG2}^{\text{LONG}}$ soit positif est élevée au-dessus du seuil de significativité pour le modèle simple-gène, le modèle hiérarchique simple, mais pas pour le modèle Ω−hiérarchique.

Une première remarque générale est que, pour le modèle simple-gène, beaucoup de gènes semblent une histoire de ω corrélée avec un trait de vie. Toutefois cette tendance disparaît

quand on prend en compte les effets globaux. On suppose que ce sont les effets de la variation de N_e qui apparaissaient indépendamment pour chaque alignement dans le modèle $\Omega-$hiérarchique. La distribution des covariances pour la longévité par exemple se trouve dans un intervalle plus petit dans les deux modèles hiérarchiques et plus centrée sur zéro dans le modèle $\Omega-$hiérarchique. En résumé les covariations de ω et des traits propres à chaque gène sont effectivement dominées par des effets de variation de N_e mais d'un autre coté, les modèles hiérarchiques semblent capables de découpler cet effet en effets résiduels spécifiques à chaque gène. Il s'agit donc la d'un succès de la méthode développée ici.

Deux gènes, RAG2 et ADRB2, conserve une corrélation positive,ou marginalement significative, avec la longévité, à travers les trois modèles. Cela signifie que ω, notre mesure de pression de sélection purificatrice, augmente dans les lignages où la longévité augmente et diminue quand conjointement à la longévité, pour ces deux gènes. On peut en déduire que le rôle de ces gènes n'est pas lié à la lutte contre le vieillissement et qu'ils s'avèrent peu importants chez les espèces a longue durée de vie.

L'allométrie la plus négative chez ces 17 gènes est celle qui existe entre le ω du gène RBP3 et la longévité.. On peut supposer que c'est RBP3 qui est le plus fortement impliqué dans la résistance de l'organisme au vieillissement, c'est une sous unité de polymérase à ARN.

Les probabilités a posteriori entre les deux modèles hiérarchiques varient peu, on suppose que l'effet global est assez peu significatif pour la plupart des gènes. Le modèle $\Omega-$hiérarchique est plus pertinent, mais force est de constater qu'il ne reste pas d'allométries très fortes. Les covariances négatives, qui nous intéressent particulièrement, ne sont pas en nombre suffisant pour en tirer des conclusions satisfaisantes. C'est pourquoi nous avons poussé la méthode en utilisant des alignements de gènes orthologues issus de génomes complets, comme nous allons le voir par la suite.

Tableau 6.VI – Probabilité a posteriori d'être positive des allométries entre gènes et traits d'histoire de vie

	Maturité femelle						Masse						Longevité					
	coevol		H		Ω-H		coevol		H		Ω-H		coevol		H		Ω-H	
	$\tilde{\alpha}$	pp	$\tilde{\alpha}$	pp	$\tilde{\alpha}$	pp	$\tilde{\alpha}$	pp	$\tilde{\alpha}$	pp	$\tilde{\alpha}$	pp	$\tilde{\alpha}$	pp	$\tilde{\alpha}$	pp	$\tilde{\alpha}$	pp
ORA3	0.28	0.81	0.17	0.77	0.18	0.78	0.54	0.83	0.28	0.73	0.28	0.70	0.09	0.80	0.11	0.92	0.09	0.83
RA2B	0.06	0.62	0.04	0.60	0.07	0.65	0.03	0.53	0.00	0.51	0.02	0.52	0.08	0.84	0.06	0.79	0.04	0.70
RB2	0.28	0.83	0.20	0.83	0.19	0.83	0.71	0.91	0.48	0.84	0.43	0.81	0.34	>0.99	0.19	0.98	0.15	0.94
OB	0.28	>0.99	-0.01	0.49	0.02	0.56	0.56	0.99	-0.05	0.44	-0.02	0.47	0.09	0.99	0.03	0.70	0.01	0.60
7A	0.18	0.91	-0.14	0.19	-0.13	0.21	0.58	0.98	-0.41	0.09	-0.41	0.10	0.11	0.97	-0.02	0.36	-0.05	0.23
NF	0.07	0.59	0.13	0.72	0.10	0.68	0.32	0.68	0.27	0.71	0.21	0.65	-0.01	0.48	0.12	0.89	0.09	0.81
CA1	0.00	0.53	0.11	0.74	0.09	0.67	-0.01	0.49	0.03	0.52	-0.02	0.46	0.02	0.65	0.01	0.54	-0.01	0.46
R1	-0.11	0.15	-0.08	0.36	-0.08	0.36	-0.36	0.12	-0.37	0.23	-0.40	0.22	-0.06	0.14	-0.01	0.47	-0.04	0.34
R	-0.15	0.18	-0.02	0.45	-0.02	0.44	-0.30	0.16	-0.21	0.29	-0.25	0.26	-0.06	0.17	0.02	0.65	-0.01	0.45
OC	0.00	0.48	-0.09	0.36	-0.06	0.38	0.00	0.49	-0.31	0.22	-0.31	0.22	0.00	0.48	0.00	0.54	-0.02	0.37
G1	0.55	0.95	0.07	0.62	0.07	0.64	1.67	0.99	0.06	0.54	0.03	0.52	0.38	>0.99	0.06	0.77	0.03	0.65
G2	0.18	0.93	0.18	0.85	0.17	0.83	0.49	0.94	0.39	0.86	0.34	0.82	0.10	0.95	0.13	0.97	0.10	0.92
23	-0.22	0.24	-0.27	0.15	-0.23	0.19	-0.72	0.15	-0.62	0.10	-0.59	0.12	-0.22	0.02	-0.03	0.38	-0.06	0.29
R1	0.07	0.61	-0.06	0.42	-0.04	0.46	0.18	0.63	-0.14	0.40	-0.15	0.39	0.22	0.98	0.06	0.75	0.03	0.65
R	0.16	0.85	0.04	0.60	0.05	0.63	0.50	0.92	0.13	0.64	0.11	0.61	0.19	0.98	0.10	0.91	0.08	0.83
F	0.04	0.59	-0.08	0.36	-0.06	0.39	0.32	0.80	-0.20	0.30	-0.19	0.30	0.02	0.58	0.03	0.66	0.00	0.52
K	-0.02	0.44	-0.02	0.48	-0.02	0.48	-0.41	0.34	-0.18	0.38	-0.20	0.37	-0.03	0.40	0.03	0.62	0.00	0.52

6.2 Une analyse plus large

Nous utilisons ici, le modèle $\Omega-$hiérarchique pour une analyse du génome complet des placentaires. Les résultats présentés ici poussent la méthode un peu au-delà de ces limites doivent être interprété avec précaution. Les ressources nécessaires aux résultats ont été conséquentes, plusieurs années-calcul offertes par le réseau québécois de calcul de haute performance (RQCHP). Afin d'accélérer la convergence, nous avons estimé séparément puis fixé plusieurs paramètres du modèle.
Les résultats obtenus restent peu significatifs, mais permettent d'effectuer des tests de permutation sur des ensembles de gènes.

6.2.1 Description des données

Nous avons utilisé ici *Orthomam* [84], dans sa version de juillet 2010. Elle contient des alignements de gènes orthologues pour les 36 espèces de placentaire séquencées entièrement à la date de sa dernière mise a jour. Il y a donc 12777 alignements de séquences codantes, alignés par codons. Ces alignements ont été nettoyés selon la méthode décrite dans la section 5.1. Nous avons finalement choisi de garder 5135 alignements contenant au moins 30 espèces et 200 codons. La topologie de l'arbre des placentaires utilisée est représentée sur la figure 6.3.
Les temps de divergence et les valeurs ancestrales de la masse et des traits d'histoire de vie ont été fixés à des valeurs obtenues préalablement.

6.2.2 Résultats

Les deux chaînes dont les résultats sont présentés ici ont effectué environ 5000 cycles effectuant chacun plusieurs dizaines de millions de mouvements de Métropolis-Hastings sur l'ensemble des paramètres du modèle. La parallélisation à été faites sur 1020 cœurs pendant environ 250 heures. Nous avons "brulé" les 1000 premiers cycles, c'est-à-dire qu'ils ont été retirés et que les moyennes marginales ont été obtenues sur les cyles restants.
Tout d'abord,la matrice de covariance obtenue est montrée dans la table 6.VII. Les trois caractères quantitatifs utilisés étant fixés dans le modèle, on retrouve, sans étonnement,

des valeurs de covariance significativement positive entre eux. Les covariances entre Ω et les traits de vie sont peu significatives, mais toutefois étonnantes. En effet cet estimateur de N_e devrait corréler positivement avec la masse de manière significative selon les hypothèses exposées préalablement. On remarque même une allométrie plus ou moins négative entre le temps de génération et la Ω, c'est-à-dire que le modèle suggère des variations du temps de génération allant dans le même sens que les variations de taille efficace de population. Ceci semble contradictoire avec le paradigme r/K.

	Gen	Mas	Lon	Ω
Gen	2.41	3.04	0.50	-0.15
Mas	>0.99	23.03	1.80	-0.02
Lon	0.99	>0.99	1.21	0.09
Ω	0.27	0.48	0.71	1.61

Tableau 6.VII – Matrice de covariance pour OrthoMam

Une explication à ces covariances surprenantes bien que non significatives peut-être à chercher dans les variations de la proportion de *GC* dans le génome. On sait que la somme des taux de guanine et de cytosine varie plus ou moins systématiquement et en fonction de la distance avec le centromère. Il se peut que ce biais naturel s'ajoute aux différents facteurs pour modifier le taux de substitution synonyme.

Ensuite, nous présentons sur la figure 6.3 la reconstruction de Ω sur l'arbre des placentaires. Cette reconstruction se veut d'autant plus précise qu'elle est faite sur un assez grand nombre de gènes pour être représentative du génome. Toutefois elle met en évidence un point faible du modèle. La valeur de Ω utilisée pour chaque branche est la moyenne des valeurs aux deux extrémités de la branche. Les valeurs ponctuelles sont donc chacune utilisées dans le calcul des trois branches les entourant et ne représentent pas forcément la valeur instantanée aux noeuds. Ceci indique des effets de discrétisation alternant entre valeurs hautes et basses le long d'un lignage qui soulèvent quelques doutes sur la qualité de la reconstruction.

Cependant on peut lire certaines choses intéressantes. Par exemple, les valeurs dans le groupe des hominidés, représentés ici par le gorille, l'homme et le chimpanzé, nous suggèrent des diminutions indépendantes de la taille efficace de population. On y remarque aussi une diminution moins importante chez l'homme. On peut faire l'hypothèse d'un

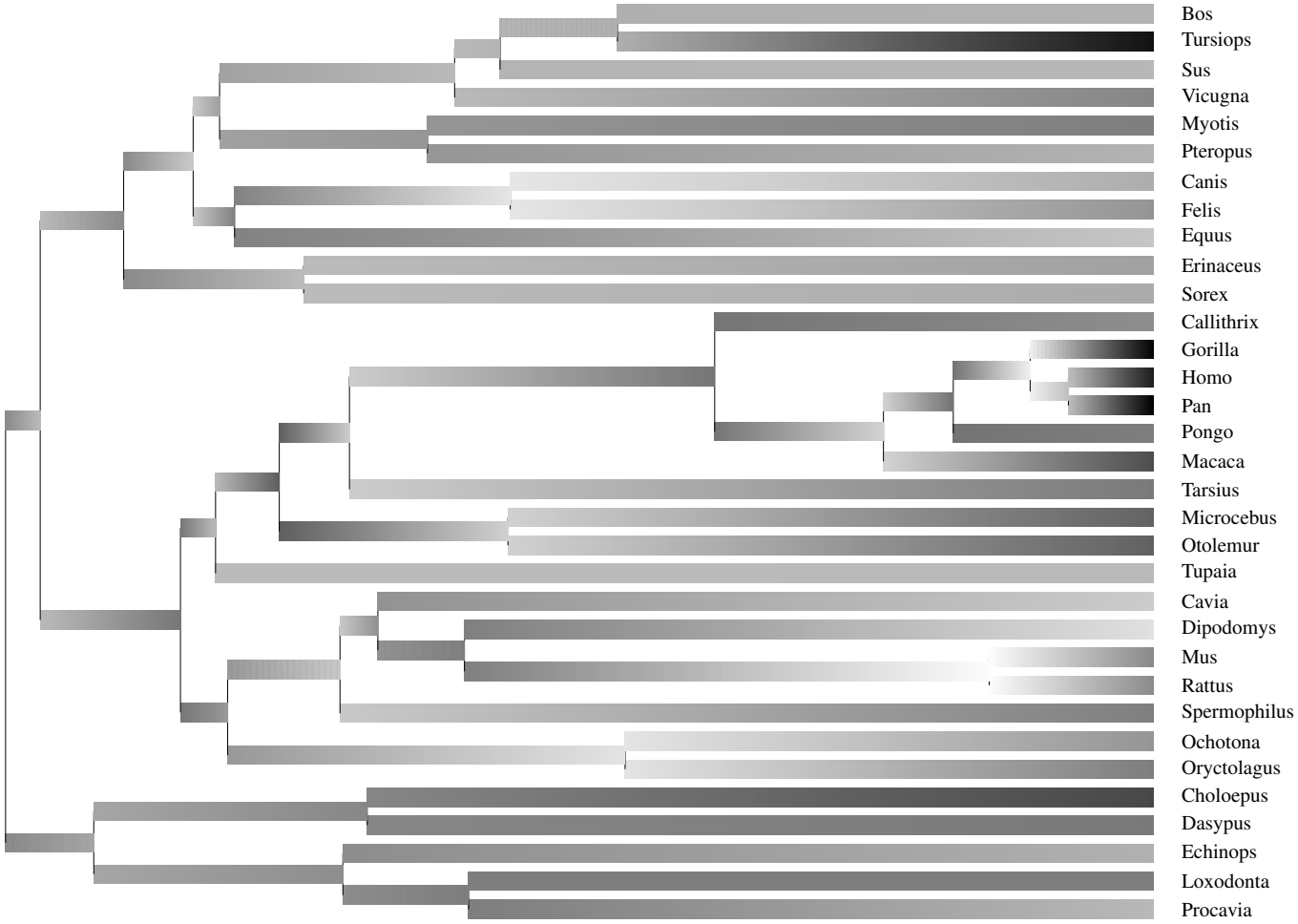


Figure 6.3 – Variations de Ω pour OrthoMam

grand nombre de tris de lignage incomplet, c'est-à-dire de gènes pour lesquels le polyallélisme à été conservé lors de deux spéciations successives. Un tri de lignage incomplet a parfois pour effet de reconstruire une histoire du gène différente de l'histoire des espèces. On pense que plusieurs gènes orthologues dans la famille des hominidés ont été polymorphes pendant les deux spéciations qu'on peut voir sur l'arbre de la figure 6.3.

Chez les rongeurs, la taille efficace de population serait bien plus grande, particulièrement chez le rat et la souris chez qui on remarque une forte diminution de Ω. Comme sur la figure 6.2, on voit une forte augmentation de Ω chez le dauphin (*tursiops)*, on peut imaginer qu'ici, des évènements récurrents de goulots d'étranglement ont fait diminuer la taille de population efficace.

Les valeurs des covariances associées à chaque gène ne sont pas significativement signées et ne permettent pas de faire ressortir certains gènes. Nous avons alors utilisé *GO slim*,

une version réduite au maximum de *Gene Ontology* contenant seulement 105 concepts associés à des gènes. Nous avons choisis les annotations de *GO slim* sur le génome humain pour créer 105 catégories et nous avons utilisé le test de permutation décrit dans la section 5.2 pour déterminer quels concepts ont des valeurs significativement hautes ou basses.

identifant Gene Ontology	nombre de gènes	test de permutation $\alpha < 0$	$\bar{\alpha}$	définition du concept : localisation processus biologique ou composant cellulaire
GO :0005829	651	0.9927	0.9707	cytosol
GO :0005886*	553	0.9999	0.9999	plasma membrane
GO :0009058*	465	0.9929	0.9801	biosynthetic process
GO :0006810*	457	0.9999	0.9999	transport
GO :0044281*	408	0.9999	0.9999	small molecule metabolic process
GO :0009056*	306	0.9999	0.9999	catabolic process
GO :0006629*	209	0.9999	0.9999	lipid metabolic process
GO :0055085*	202	0.9999	0.9999	transmembrane transport
GO :0007267*	144	0.9999	0.9966	cell-cell signaling
GO :0016192*	137	0.9999	0.9944	vesicle-mediated transport
GO :0005739*	132	0.9999	0.9995	mitochondrion
GO :0042592*	128	0.9999	0.9999	homeostatic process
GO :0005856	99	0.9960	0.9522	cytoskeleton
GO :0006520*	94	0.9999	0.9999	cellular amino acid metabolic process
GO :0005615*	94	0.9995	0.9931	extracellular space
GO :0005576*	87	0.9999	0.9999	extracellular region
GO :0061024*	74	0.9999	0.9959	membrane organization
GO :0006091*	74	0.9999	0.9999	generation of precursor metabolites and energy
GO :0034655*	72	0.9991	0.9877	nucleobase, nucleotide and nucleic acid catabolism
GO :0005768*	64	0.9999	0.9989	endosome
GO :0050877*	62	0.9998	0.9891	neurological system process
GO :0016023*	62	0.9999	0.9927	cytoplasmic membrane-bounded vesicle
GO :0051186*	57	0.9999	0.9999	cofactor metabolic process
GO :0006399	33	0.9997	0.9668	tRNA metabolic process
GO :0006790*	29	0.9999	0.9992	sulfur compound metabolic process
GO :0005764	27	0.9999	0.9804	lysosome
GO :0005635*	26	0.9999	0.9955	nuclear envelope
GO :0006412*	21	0.9999	0.9997	translation
GO :0007005	16	0.9999	0.9952	mitochondrion organization
GO :0051604	14	0.9999	0.9687	protein maturation
GO :0019748	14	0.9999	0.9548	secondary metabolic process
GO :0007034	14	0.9999	0.9943	vacuolar transport
GO :0005777*	14	0.9999	0.9999	peroxisome
GO :0000228	14	0.9999	0.9774	nuclear chromosome
GO :0005773	6	0.9999	0.9843	vacuole
GO :0005578	6	0.9999	0.9652	proteinaceous extracellular matrix
GO :0021700	5	0.9999	0.9828	developmental maturation

Tableau 6.VIII – Liste des concepts GO pour lesquels l'allométrie entre longévité et Ω est significativement moins grande qu'attendu sous l'hypothèse nulle

Dans le tableau 6.IX nous montrons les concepts dont la moyenne des covariances calculées à parti de α^{LONG} sont significativement plus basses qu'attendue sous l'hypothèse nulle. Nous avons choisi de ne garder que les concepts dont, à la fois, la valeur moyenne

de la covariance et de la probabilité a posteriori est plus basse qu'attendue avec une probabilité supérieure à 0.95. D'après les hypothèses sous-jacentes à notre modèle, nous obtenons une liste de concepts recouvrant l'ensemble des fonctions biologiques, processus cellulaires ou composants de la cellule dont l'optimalité est favorable au vieillissement des individus. Les concepts marqués d'un astérisque ($*$) sont présents dans le tableau équivalent concernant la masse corporelle ou le temps de génération. Ces concepts concernent la membrane cellulaire, l'espace intercellulaire, ainsi que la production d'énergie. Ils sont supposés correspondre aux cibles de la stratégie de sélection K favorisant à la fois le vieillissement, des temps de génération plus longs et une augmentation de la masse corporelle. On peut penser que c'est surtout la sélection pour une plus grande masse corporelle qui est visible dans ces résultats.

Regardons maintenant les concepts qui apparaissent comme spécifiquement liés à l'histoire de la longévité chez les placentaires. Ces concepts annotent des groupes de gènes qui sont soumis à une sélection purificatrice plus intense chez les espèces plus longévives et moins intenses chez les espèces qui vivent moins longtemps. Selon notre modèle et la théorie du soma jetable exposée dans la section 1.2.3, nous avons la liste des cibles préférentielles de la sélection pour allonger la durée de vie chez les placentaires. Il n'est pas étonnant de trouver les processus métaboliques secondaires qui regroupent les réactions chimiques qui ne sont pas nécessaires à la croissance et jouent souvent un rôle dans le système immunitaire. Nous trouvons aussi les gènes impliqués dans l'activité des lysosomes or, chez l'humain, on sait que, les maladies lysosomales sont très majoritairement dégénératives, ce qui laisse à penser que le maintien des fonctions de cet organite est important pour le maintien du soma. Nous trouvons aussi l'ARN de transfert ou le cytosquelette.

Ce tableau montre bien que les interprétations de nos résultats sont nombreuses, mais qu'il serait aventureux de sur estimer leur pertinence. En outre, il est probablement biaisé de chercher à justifier nos résultats en regardant les fonctions connues des gènes. D'autre part, les biais du modèle, les erreurs dans la détection de gènes orthologues, dans le séquençage ou l'alignement, les épiphénomènes évolutifs sont nombreux. Nos résultats sont donc intéressants, mais doivent être interprétés avec précaution. Ils posent cependant les bases d'une approche méthodologiquement bien construite.

identifant Gene Ontology	nombre de gènes	test de permutation $\alpha < 0$	$\bar{\alpha}$	définition du concept : localisation (l) processus biologique (bp) ou composant cellulaire (cc)
			longévité	
GO :0005634	680	0.9964	0.9994	nucleus
GO :0051276	148	0.9702	0.9943	chromosome organization
			temps de génération	
GO :0005634	680	0.9987	0.9998	nucleus
GO :0005654	279	0.9863	0.9973	nucleoplasm
GO :0051276	148	0.9659	0.9999	chromosome organization
GO :0007049	124	0.982	0.9998	cell cycle
			masse	
GO :0005634	680	0.9974	0.9999	nucleus
GO :0005654	279	0.9712	0.9885	nucleoplasm
GO :0051276	148	0.9792	0.9981	chromosome organization
GO :0007049	124	0.9727	0.9977	cell cycle

Tableau 6.IX – Liste des concepts GO pour lesquels l'allométrie avec Ω est significativement plus élevée

CHAPITRE 7

CONCLUSION

Ce travail a donc été le développement d'une méthode d'évolution moléculaire permettant la reconstruction de l'histoire substitutive et l'analyse comparative de caractères continus et sur une phylogénie. La méthode a pour caractéristique d'être bayésienne et d'avoir été implémentée en parallèle, ce qui permet de travailler à des échelles plus grandes. Le développement du logiciel *coevol* puis, à partir de ces sources, de la version parallélisée *paracoevol* a été mon activité principale lors de cette maîtrise.

Les applications de *coevol* ont permis de soulever des questions comme le biais en guanine et cytosine. La parallélisation effectuée, quant à elle, devrait permettre de tester de manière plus poussée les effets de la sélection au niveau de chaque gène. Les résultats s'apparentent à de l'annotation de génome, mais l'annotation fonctionnelle d'un gène au niveau d'un taxon reste une notion à définir plus précisément.

L'utilisation de grandes quantités de données pose un problème qualitatif. Il n'est pas possible de vérifier à la main la qualité des alignements. Malgré l'efficacité de notre méthode de nettoyage, le biais dans les résultats persiste, et est difficile à mesurer.

La possibilité d'utiliser de grandes quantités de données moléculaires dans des analyses bayésiennes ouvre de nombreuses perspectives pour répondre à des questions sur l'origine de la vie. Nombre de mesures de covariance phylogénétique sont envisageable, joignant des variables environnementales, des phénotypes ou des paramètres de l'histoire moléculaire du vivant. On peut penser à la composition nucléotidique de l'ADN, la composition en acide aminé du protéome, à la température de l'atmosphère, des océans ou au taux d'oxygène, et à d'autres phénotypes comme le métabolisme ou la taille du système nerveux, ou à des caractéristiques sociales comme le grégarisme. L'histoire évolutive de toutes ces composantes révèle leurs liens biologiques et peuvent nous apprendre beaucoup de choses sur le vivant.

BIBLIOGRAPHIE

[1] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin et G. Sherlock. Gene ontology : tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.*, 25(1):25–29, May 2000.

[2] Robert S. Balaban, Shino Nemoto et Toren Finkel. Mitochondria, oxidants, and aging. *Cell*, 120(4):483 – 495, 2005. ISSN 0092-8674.

[3] T. Bayes, R. Price et J. Canton. *An Essay Towards Solving a Problem in the Doctrine of Chances*. C. Davis, Printer to the Royal Society of London, 1763. URL `http://books.google.ca/books?id=Xi2wpwAACAAJ`.

[4] S. Blanquart et N. Lartillot. A Bayesian compound stochastic process for modeling nonstationary and nonhomogeneous sequence evolution. *Mol. Biol. Evol.*, 23: 2058–2071, 2006.

[5] B. Boussau, S. Blanquart, N. Lartillot et M. Gouy. Parallel adaptations to high temperatures in the archaean eon. *Nature*, 456:942–945, 2008.

[6] L. Bromham. Why do species vary in their rate of molecular evolution ? *Biol. Lett.*, 5(3):401–404, Jun 2009.

[7] L. Bromham, A. Rambaut et P. H. Harvey. Determinants of rate variation in mammalian DNA sequence evolution. *J. Mol. Evol.*, 43:610–621, 1996.

[8] W. M. Brown, M. Jr. George et A. C. Wilson. Rapid evolution of animal mitochondrial DNA. *Proc. Natl. Acad. Sci. U.S.A.*, 76:1967–1971, 1979.

[9] M. A. Butler et A. A. King. Phylogenetic comparative analysis : a modeling approach for adaptive evolution. *Am. Nat.*, 164:683–695, 2004.

[10] William Calder. *Size, function, and life history*. Dover Publications, Mineola, N.Y, 1996. ISBN 0486691918.

[11] Marcel Cardillo, Georgina M. Mace, Kate E. Jones, Jon Bielby, Olaf R. P. Bininda-Emonds, Wes Sechrest, C. David L. Orme et

Andy Purvis. Multiple causes of high extinction risk in large mammal species. *Science*, 309(5738):1239–1241, 2005. URL `http://www.sciencemag.org/content/309/5738/1239.abstract`.

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

[13] L. Chao et D. E. Carr. The molecular clock and the relationship between population size and generation time. *Evolution*, 47:688–690, 1993.

[14] N. Cooper et A. Purvis. Body size evolution in mammals : complexity in tempo and mode. *Am. Nat.*, 175:727–738, 2010.

[15] A.P. Dawid et S. L. Lauritzen. Hyper Markov laws in the statistical analysis of decomposable graphical models. *Ann. Statist.*, 21:1272–1317, 1993.

[16] A. P. de Koning, W. Gu et D. Pollock. Rapid likelihood analysis on large phylogenies using partial sampling of substitution histories. *Mol. Biol. Evol.*, 27:249–265, 2010.

[17] J. P. de Magalhaes et J. Costa. A database of vertebrate longevity records and their relation to other life-history traits. *J. Evol. Biol.*, 22(8):1770–1774, Aug 2009.

[18] A.P. Dempster. Covariance selection. *Biometrics*, 28:157–175, 1972.

[19] R. Diaz-Uriarte et T. Jr. Garland. Effects of branch length errors on the performance of phylogenetically independent contrasts. *Syst. Biol.*, 47:654–672, 1998.

[20] P.A.M. Dirac. *The principles of quantum mechanics*. Oxford University Press, 1982.

[21] A. Dobra, C. Hans, B. Jones, J.R. Nevins, G. Yao et M. West. Sparse graphical models for exploring gene expression data. *J. Multiv. Analysis*, 90:196–212, 2004.

[22] A. Eyre-Walker et P. D. Keightley. The distribution of fitness effects of new mutations. *Nat. Rev. Genet.*, 8(8):610–618, Aug 2007.

[23] J. Felsenstein. Evolutionary trees from dna sequences : a maximum likelihood approach. *J. Mol. Evol.*, 17(6):368–376, 1981.

[24] J. Felsenstein. Comparative methods with sampling error and within-species variation : contrasts revisited and revised. *Am. Nat.*, 171(6):713–725, Jun 2008.

[25] Joseph Felsenstein. Phylogenies and the comparative method. *The American Naturalist*, 125(1):pp. 1–15, 1985. ISSN 00030147. URL http://www.jstor.org/stable/2461605.

[26] R. D. Finn, J. Clements et S. R. Eddy. HMMER web server : interactive sequence similarity searching. *Nucleic Acids Res.*, 39(Web Server issue):29–37, Jul 2011.

[27] E. Fontanillas, J. J. Welch, J. A. Thomas et L. Bromham. The influence of body size and net diversification rate on molecular evolution during the radiation of animal phyla. *BMC Evol. Biol.*, 7:95, 2007.

[28] S. A. Fritz, O. R. Bininda-Emonds et A. Purvis. Geographical variation in predictors of mammalian extinction risk : big is bad, but only in the tropics. *Ecol. Lett.*, 12(6):538–549, Jun 2009.

[29] N. Galtier, N. Tourasse et M. Gouy. A nonhyperthermophilic common ancestor to extant life forms. *Science*, 283:220–221, 1999.

[30] Jr. Garland, Theodore, Paul H. Harvey et Anthony R. Ives. Procedures for the analysis of comparative data using phylogenetically independent contrasts. *Systematic Biology*, 41(1):pp. 18–32, 1992. ISSN 10635157. URL http://www.jstor.org/stable/2992503.

[31] T. Garland, A. F. Bennett et E. L. Rezende. Phylogenetic approaches in comparative physiology. *J. Exp. Biol.*, 208(Pt 16):3015–3035, Aug 2005.

[32] T. J. Garland, A. F. Bennett et E. L. Rezende. Phylogenetic approaches in comparative physiology. *J. Exp. Biol.*, 208:3015–3035, 2005.

[33] A. Gelman, J. B. Carlin, H. S. Stern et D. B. Rubin. *Bayesian data analysis*. Chapman and Hall/CRC, 2004.

[34] J. H. Gillespie. *The causes of molecular evolution*. Oxford University Press, 1991.

[35] J. F. Gillooly, A. P. Allen, G. B. West et J. H. Brown. The rate of DNA evolution : effects of body size and temperature on the molecular clock. *Proc. Natl. Acad. Sci. U.S.A.*, 102:140–145, 2005.

[36] N. Goldman et Z. Yang. A codon-based model of nucleotide substitution for protein-coding DNA sequences. *Mol. Biol. Evol.*, 11:725–736, 1994.

[37] N. M. Gomes, O. A. Ryder, M. L. Houck, S. J. Charter, W. Walker, N. R. Forsyth, S. N. Austad, C. Venditti, M. Pagel, J. W. Shay et W. E. Wright. Comparative biology of mammalian telomeres : hypotheses on ancestral states and the roles of telomeres in longevity determination. *Aging Cell*, 10(5):761–768, Oct 2011.

[38] P.H. Harvey et M.D. Pagel. *The comparative method in evolutionary biology*. Oxford series in ecology and evolution. Oxford University Press, 1998. ISBN 9780198546405. URL `http://books.google.ca/books?id=RkVRAAAAMAAJ`.

[39] R. Holliday. Aging is no longer an unsolved problem in biology. *Ann. N. Y. Acad. Sci.*, 1067:1–9, May 2006.

[40] E. A. Housworth, E. P. Martins et M. Lynch. The phylogenetic mixed model. *Am. Nat.*, 163(1):84–96, Jan 2004.

[41] T. J. P. Hubbard, B. L. Aken, S. Ayling, B. Ballester, K. Beal, E. Bragin, S. Brent, Y. Chen, P. Clapham, L. Clarke, G. Coates, S. Fairley, S. Fitzgerald, J. Fernandez-Banet, L. Gordon, S. Graf, S. Haider, M. Hammond, R. Holland, K. Howe, A. Jenkinson, N. Johnson, A. Kahari, D. Keefe, S. Keenan, R. Kinsella, F. Kokocinski, E. Kulesha, D. Lawson, I. Longden, K. Megy, P. Meidl, B. Overduin, A. Parker, B. Pritchard, D. Rios, M. Schuster, G. Slater, D. Smedley, W. Spooner, G. Spudich, S. Trevanion, A. Vilella, J. Vogel, S. White, S. Wilder, A. Zadissa, E. Birney, F. Cunningham, V. Curwen, R. Durbin, X. M. Fernandez-Suarez, J. Herrero, A. Kasprzyk, G. Proctor, J. Smith, S. Searle et P. Flicek. Ensembl 2009. *Nucleic Acids Research*, 37(suppl 1):D690–D697, 2009.

[42] J. Huelsenbeck et B. Rannala. Frequentist properties of Bayesian posterior probabilities of phylogenetic trees under simple and complex substitution models. *Syst. Biol.*, 53:904–913, 2004.

[43] J. P. Huelsenbeck et B. Rannala. Detecting correlation between characters in a comparative analysis with uncertain phylogeny. *Evolution*, 57(6):1237–1247, Jun 2003.

[44] J. P. Huelsenbeck, B. Rannala et J. P. Masly. Accommodating phylogenetic uncertainty in evolutionary studies. *Science*, 288(5475):2349–2350, 2000.

[45] J. G. Inoue, P. C. J. Donogue et Z. Yang. The impact of the representation of fossil calibration on bayesin estimation of species divergence times. *Syst. Biol.*, 59: 74–89, 2010.

[46] H. Jeffreys. *Theory of probability*. Oxford University Press, 1961.

[47] R. W. Jobson, B. Nabholz et N. Galtier. An evolutionary genome scan for longevity-related natural selection in mammals. *Mol. Biol. Evol.*, 27(4):840–847, Apr 2010.

[48] R. Jones. Protein aggregation increases with age. *PLoS Biol.*, 8(8):e1000449, 2010.

[49] G. Jordan. Analysis of alignment error and sitewise constraint in mammalian comparative genomics. 2011.

[50] P. D. Keightley et A. Eyre-Walker. What can we learn about the distribution of fitness effects of new mutations from DNA sequence data ? *Philos. Trans. R. Soc. Lond., B, Biol. Sci.*, 365(1544):1187–1193, Apr 2010.

[51] M. Kimura. Model of effectively neutral mutations in which selective constraint is incorporated. *Proc. Natl. Acad. Sci. U.S.A.*, 76(7):3440–3444, Jul 1979.

[52] Motoo Kimura. *The Neutral theory of molecular evolution*. Cambridge University Press, Cambridge Cambridgeshire New York, 1984. ISBN 0521317932.

[53] T. B. Kirkwood et R. Holliday. The evolution of ageing and longevity. *Proc. R. Soc. Lond., B, Biol. Sci.*, 205(1161):531–546, Sep 1979.

[54] H. Kishino, J. L. Thorne et W. Bruno. Performance of a divergence time estimation method under a probabilistic model of rate evolution. *Mol. Biol. Evol.*, 18:352–361, 2001.

[55] S. Kryazhimskiy et J. B. Plotkin. The population genetics of dN/dS. *PLoS Genet.*, 4(12):e1000304, Dec 2008.

[56] R. Lanfear, J. A. Thomas, J. J. Welch, T. Brey et L. Bromham. Metabolic rate does not calibrate the molecular clock. *Proc. Natl. Acad. Sci. U.S.A.*, 104:15388–15393, 2007.

[57] N. Lartillot. Conjugate Gibbs sampling for Bayesian phylogenetic models. *J. Comput. Biol.*, 13:1701–1722, 2006.

[58] N. Lartillot et R. Poujol. A phylogenetic model for investigating correlated evolution of substitution rates and continuous phenotypic characters. *Mol. Biol. Evol.*, 28(1):729–744, Jan 2011.

[59] Nicolas Lartillot et Frédéric Delsuc. Joint reconstruction of divergence times and life-history evolution in placental mammals using a phylogenetic covariance model. *Evolution*, page no, 2012. URL http://dx.doi.org/10.1111/j.1558-5646.2011.01558.x.

[60] T. Lepage, D. Bryant, H. Philippe et N. Lartillot. A general comparison of relaxed molecular clock models. *Mol. Biol. Evol.*, 24:2669–2680, 2007.

[61] G. Letac et H. Massam. Wishart distributions for decomposable graphs. *Ann. Statist.*, 35:1278–1323, 2007.

[62] W.-H. Li, D. L. Ellsworth, J. Krushkal, B. H.-J. Chang et D. Hewett-Emmett. Rates of nucleotide substitution in primates and rodents and the generation-time effect hypothesis. *Mol. Phyl. Evol.*, 5:182–187, 1996.

[63] W. H. Li et M. Tanimura. The molecular clock runs more slowly in man than in apes and monkeys. *Nature*, 326(6108):93–96, 1987.

[64] Wen-Hsiung Li, Darrell L. Ellsworth, Julia Krushkal, Benny H.-J. Chang et David Hewett-Emmett. Rates of nucleotide substitution in primates and rodents and the generationâĂŞtime effect hypothesis. *Molecular Phylogenetics and Evolution*, 5 (1):182 – 187, 1996. ISSN 1055-7903.

[65] M. Lynch. *The origins of genome architecture*. Sinauer Associates, 2007. ISBN 9780878934843. URL http://books.google.ca/books?id=7NAPAQAAMAAJ.

[66] R.H. MacArthur et E.O. Wilson. *The Theory of Island Biogeography*. Princeton Landmarks in Biology. Princeton University Press, 2001. ISBN 9780691088365. URL http://books.google.ca/books?id=a10cdkywhVgC.

[67] K. V. Mardia, J. T. Kent et J. M. Bibby. *Multivariate analysis*. Academic Press, 1979.

[68] A P Martins et S R Palumbi. Body size, metabolic rate, generation time, and the molecular clock. *Proceedings of the National Academy of Sciences*, 90(9):4087–4091, 1993. URL http://www.pnas.org/content/90/9/4087.abstract.

[69] E. P. Martins et T. F. Hansen. The statistical analysis of interspecific data : a review and evaluation of phylogenetic comparative methods. Dans E. Martins, éditeur, *Phylogenies and the comparative method in animal behavior*, pages 22–75. Oxford University Press, Oxford, 1996.

[70] Emilia P. Martins et Thomas F. Hansen. Phylogenies and the comparative method : A general approach to incorporating phylogenetic information into the analysis of interspecific data. *The American Naturalist*, 149(4):pp. 646–667, 1997. ISSN 00030147. URL http://www.jstor.org/stable/2463542.

[71] L. Mateiu et B. Rannala. Inferring complex dna substitution processes on phylogenies using uniformization and data augmentation. *Syst. Biol.*, 55:259–269, 2006.

[72] S. V. Muse et B. S. Gaut. A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitutions, with applications to chloroplast genome. *Mol. Biol. Evol.*, 11:715–724, 1994.

[73] B. Nabholz, S. Glémin et N. Galtier. Strong variations of mitochondrial mutation rate across mammals–the longevity hypothesis. *Mol. Biol. Evol.*, 25:120–130, 2008.

[74] Radford M. Neal. Circularly-coupled markov chain sampling. Rapport technique, Dept. of Statistics and Dept. of Computer Science, University of Toronto, 2002.

[75] R. Nielsen. Mapping mutations on phylogenies. *Syst. Biol.*, 51:729–739, 2002.

[76] S. I. Nikolaev, J. I. Montoya-Burgos, K. Popadin, L. Parand, E. H. Margulies et S. E. Antonarakis. Life-history traits drive the evolutionary rates of mammalian coding and noncoding genomic elements. *Proc. Natl. Acad. Sci. U.S.A.*, 104(51): 20443–20448, Dec 2007.

[77] T. Ohta. Slightly deleterious mutant substitutions in evolution. *Nature*, 252:315–354, 1973.

[78] T. Ohta. Mutational pressure as the main cause of molecular evolution and polymorphisms. *Nature*, 252:351–354, 1974.

[79] T. Ohta. An examination of the generation-time effect on molecular evolution. *Proc. Natl. Acad. Sci. U.S.A.*, 90(22):10676–10680, Nov 1993.

[80] T. Ohta. Synonymous and nonsynonymous substitutions in mammalian genes and the nearly neutral theory. *J. Mol. Evol.*, 40(1):56–63, Jan 1995.

[81] M. Pagel. Inferring the historical patterns of biological evolution. *Nature*, 401: 877–884, 1999.

[82] K. Popadin, L. V. Polishchuk, L. Mamirova, D. Knorre et K. Gunbin. Accumulation of sightly deleterious mutations in mitochondrial protein-coding genes of large versus small mammals. *Proc. Natl. Acad. Sci. U.S.A.*, 104:13390–13395, 2007.

[83] B. Rannala et Z. Yang. Inferring speciation times under an episodic molecular clock. *Syst. Biol.*, 56:453–466, 2007.

[84] V. Ranwez, F. Delsuc, S. Ranwez, K. Belkhir, M. K. Tilak et E. J. Douzery. OrthoMaM : a database of orthologous genomic markers for placental mammal phylogenetics. *BMC Evol. Biol.*, 7:241, 2007.

[85] Richard Durbin, Sean Eddy, Anders Krogh, and Graeme Mitchison. *Biological sequence analysis : probabalistic models of proteins and nucleic acids*. Cambridge University Press, Cambridge, UK New York, 1998. ISBN 0521620414.

[86] N. Rodrigue, H. Philippe et N. Lartillot. Uniformization for sampling realizations of Markov processes : applications to Bayesian implementations of codon substitution models. *Bioinformatics*, 24:56–62, 2008b.

[87] A. Roverato. Hyper inverse Wishart distribution for non-decomposable graphs and its application to Bayesian inference for Gaussian graphical models. *Scand. J. Statist.*, 29:391–411, 2002.

[88] T. K. Seo, H. Kishino et J. L. Thorne. Estimating absolute rates of synonymous and nonsynonymous nucleotide substitution in order to characterize natural selection and date species divergences. *Mol. Biol. Evol.*, 21:1201–1213, 2004.

[89] J. R. Speakman. Body size, energy metabolism and lifespan. *J. Exp. Biol.*, 208(Pt 9):1717–1730, May 2005.

[90] J. R. Speakman. Correlations between physiology and lifespan–two widely ignored problems with comparative studies. *Aging Cell*, 4:167–175, 2005.

[91] Sawyer Stanley. Wishart distributions and inverse-wishart sampling. website : http ://wustl.edu/, 2007.

[92] Peter J. Taylor, Steven M. Goodman, M. Corrie Schoeman, Fanja H. Ratrimomanarivo et Jennifer M. Lamb. Wing loading correlates negatively with genetic structuring of eight afro-malagasy bat species (molossidae). *Acta Chiropterologica*, 14(1):53–62, Jun 2012. ISSN 1508-1109. URL http://dx.doi.org/10.3161/150811012X654268.

[93] J. A. Thomas, J. J. Welch, M. Woolfit et L. Bromham. There is no universal molecular clock for invertebrates, but rate variation does not scale with body size. *Proc. Natl. Acad. Sci. U.S.A.*, 103:7366–7371, 2006.

[94] J. L. Thorne et H. Kishino. Divergence time and evolutionary rate estimation with multilocus data. *Syst. Biol.*, 51:689–702, 2002.

[95] J. L. Thorne, H. Kishino et I. S. Painter. Estimating the rate of evolution of the rate of molecular evolution. *Mol. Biol. Evol.*, 15:1647–1657, 12 1998.

[96] D. M. Weinreich. The rates of molecular evolution in rodent and primate mitochondrial DNA. *J. Mol. Evol.*, 52:40–50, 2001.

[97] A. Weismann, E.B. Poulton, S. Schönland et A.E. Shipley. *Essays upon heredity and kindred biological problems*. Numbéro vol. 1 dans Essays Upon Heredity and Kindred Biological Problems. Clarendon press, 1891. URL http://books.google.ca/books?id=Hc45AAAAMAAJ.

[98] J. J. Welch, O. R. P. Bininda-Edmonds et L. Bromham. Correlates of substitution rate variation in mammalian protein-coding sequences. *BMC Evol. Biol.*, 8:53, 2008.

[99] J. J. Welch, A. Eyre-Walker et D. Waxman. Divergence and poymorphism under the nearly neutral theory of molecular evolution. *J. Mol. Evol.*, 67:418–426, 2008.

[100] J.J. Welch et D. Waxman. Calculating independent contrasts for the comparative study of substitution rates. *J. Theor. Biol.*, 251:667–678, 2008.

[101] E. P. White, S. K. Ernest, A. J. Kerkhoff et B. J. Enquist. Relationships between body size and abundance in ecology. *Trends Ecol. Evol. (Amst.)*, 22(6):323–330, Jun 2007.

[102] G. C. Williams. Pleiotropy, natural selection, and the evolution of senescence. *Sci. Aging Knowl. Environ.*, 2001(1):cp13, 2001.

[103] Z. Yang et B. Rannala. Bayesian estimation of species divergence times under a molecular clock using multiple fossil calibrations with soft bounds. *Mol. Biol. Evol.*, 23(1):212–226, 2006.

[104] E. Zuckerkandl et L. Pauling. Molecules as documents of evolutionary history. *J. Theor. Biol.*, 8(2):357–366, Mar 1965.

1

Annexe I

A phylogenetic model for investigating correlated evolution of substitution rates and continuous phenotypic characters

Abstract

The comparative approach is routinely used to test for possible correlations between phenotypic or life-history traits. To correct for phylogenetic inertia, the method of independent contrasts assumes that continuous characters evolve along the phylogeny according to a multivariate Brownian process. Brownian diffusion processes have also been used to describe time variations of the parameters of the substitution process, such as the rate of substitution or the ratio of synonymous to non-synonymous substitutions.

Here, we develop a probabilistic framework for testing the coupling between continuous characters and parameters of the molecular substitution process. Rates of substitution and continuous characters are jointly modelled as a multivariate Brownian diffusion process of unknown covariance matrix. The covariance matrix, the divergence times and the phylogenetic variations of substitution rates and continuous characters are all jointy estimated in a Bayesian Monte Carlo framework, imposing on the covariance matrix a prior conjugate to the Brownian process so as to achieve a greater computational efficiency. The coupling between rates and phenotypes is assessed by measuring the posterior probability of positive or negative covariances, while divergence dates and phenotypic variations are marginally reconstructed in the context of the joint analysis.

As an illustration, we apply the model to a set of 410 mammalian cytochrome b sequences. We observe a negative correlation between the rate of substitution and mass and longevity, which was previously observed. We also find a positive correlation beween $\omega = dN/dS$ and mass and longevity, which we interpret as an indirect effect of variations of effective population size, thus in partial agreement with the nearly-neutral theory. The method can easily be extended to any parameter of the substitution process, and to any continuous phenotypic or environmental character.

Introduction

Phylogenetic comparative methods are an essential tool in ecological and evolutionary analyses. One of their essential aims is to empirically investigate the evolutionary variations of phenotypic characters and life-history traits, and to test hypotheses about the underlying evolutionary mechanisms. A large diversity of empirical questions can be addressed, for instance, concerning the existence of specific trends in the direction of evolution, or the shape and the intensity of the correlations among phenotypic characters and life-history traits [38]. When measuring correlations between characters, the problem is often reduced to performing linear regressions, possibly after applying some transformation to the variables under investigation (e.g. a logarithmic transformation). An implicit assumption behind this approach is that the variables, once transformed, are linearly correlated. This assumption may either be justified on empirical grounds, as in the case of allometric relations between life-history traits and body mass [10], equivalent to linear relations upon logarithmic transformation, or may be retrospectively assessed using goodness-of-fit tests.

Because species are phylogenetically related, the data obtained in a given set of taxa cannot be considered as independent, and therefore simple regression analysis does not apply. To address this problem, general methods accounting for phylogenetic dependencies have been proposed, the most popular being the method of independent contrasts [25]. The independent contrast method is statistical in essence, relying on a probabilistic model assuming that the continuous characters under study evolve along the lineages of the phylogenetic tree according to a multivariate Brownian diffusion process. The framework is generally used to test the null hypothesis that the characters are not correlated, or to estimate the covariance matrix specifying the correlation structure among the phenotypic characters, corrected for phylogenetic inertia. The idea has been reformulated in a generalized linear model framework [70], revised to account for intra-specific variation [24, 40], extended to processes other than the Brownian diffusion process [9], and extensively applied in ecology and evolution [32, 38, 69, 81].

Comparative analyses need not be restricted to phenotypic and ecological traits, but can also be applied to parameters of the substitution process, such as the substitution rate, the ratio of non-synonymous to synonymous substitutions $\omega = dN/dS$, or the nucleotide or

amino-acid composition of sequences. The correlations between substitution parameters and phenotypic or life-history traits uncovered in this manner may provide fundamental empirical clues about the mechanisms of molecular evolution. For example, the fact that the mitochondrial dN/dS negatively correlates with body size (and thus indirectly with population size) in mammals was interpreted as an indirect confirmation of the nearly neutral theory [82]. Many other studies have investigated the influence of metabolism, body size, generation time, or longevity, on rate variation in nuclear or mitochondrial DNA in mammals, animals, or plants [7, 27, 35, 56, 62, 68, 79, 93, 98].

Correlations between molecular, phenotypic and ecological characters can also be extrapolated back in time, thus allowing inference of the history of phenotypic evolution based on ancestral sequences reconstructed using phylogenetic methods. In this direction, correlations between temperature and the GC content of ribosomal RNA stems, or the amino-acid composition of the proteome, were used to infer the history of variations of the optimal growth temperature along the tree of eubacteria and archaea [5, 29].

All these are but a few examples suggesting that further development of the comparative method, so as to jointly analyze morphological traits and substitution parameters in one single unified statistical framework, would provide essential empirical leverage for a more global understanding of evolution encompassing the molecular, phenotypic and ecological dimensions. However, several methodological points need to be addressed, so as to optimize power and avoid potential pitfalls.

A first issue is how to deal with uncertainty. Proceeding sequentially, first estimating the substitution parameters, then assessing their correlation with phenotypic characters, and finally extrapolating the correlation onto ancestral nodes, raises a potentially important problem of error propagation. The estimated correlation coefficients should ideally integrate the uncertainty about the substitution parameters and the divergence times. Conversely the reconstructed phenotypic histories should account for the uncertainty about the estimated correlation coefficients and the substitution parameters. All these problems can in principle be naturally dealt with in a joint estimation framework. In this direction, a Bayesian method for estimating covariance between phenotypic characters while accounting for the uncertainty concerning the topology and the branch lengths of the tree has been developed previously [43]. This method could be extended to allow for more general covariance analyses including phenotypic characters as well as substitution parameters.

Of note, joint estimation would also have the advantage of allowing for cross-talk between components of the model, something which would otherwise not be possible in a sequential method, and which would allow the reconstruction of phenotypic and molecular histories, or the estimation of divergence times [100], to borrow strength from each other via the inferred covariations.

Second, what should normally be compared to the phenotypic characters of a given set of species are the instant values of the rate of substitution in those species at present time [100]. In practice, what is often measured is the average substitution rate on the terminal branches of the phylogenetic tree. In principle, using sufficiently closely related taxa would allow as high a time resolution as desired. However, for fixed sequence length, this would be at the expense of the accuracy of the estimates of the substitution rates. A more satisfactory approach to the estimation of instant rates of substitution is to model rates as continuously evolving parameters. This has been first proposed in the context of molecular dating methods, using Brownian diffusion processes [54, 60, 83, 94, 95, 103]. Most often, the rate has implicitly been considered as a nuisance variable, modelled not so much for its own sake, but rather for integrating out the effects of its variations on divergence date estimates. Nevertheless, relaxed clock models naturally provide, as a by-product, an estimate of the instant value of the rates, in particular at the leaf nodes of the tree, and those rates can in principle be used as the variables to be regressed against the continuous phenotypic traits of interest.

Third, unlike phenotypic characters, rates, and more generally molecular evolutionary parameters, can be inferred also at internal nodes of the phylogenetic tree, using non-stationary models of substitution. For an optimal statistical power, the correlation between substitution parameters should therefore be assessed in a more general framework, which would not exclusively rely on the values estimated at the leaves. In this direction, [88] extended the idea of the Brownian relaxed clock to model the continuous variations of the rates of synonymous and of non-synonymous substitution. In their analysis, they modelled these rates as two independent log-normal Brownian diffusion processes, although they suggested that modelling them as a general bi-variate process would be a straightforward generalization of the approach.

In the present article, we develop a model combining the ideas of [25], [43] and [88]. In this model, all the molecular evolutionary parameters of interest (of total number K),

and all the phenotypic characters, environmental variables or life-history traits (of total number L), are modelled as one single multivariate diffusion process of dimension $K+L$. The covariance matrix of this multivariate process and the divergence times of the underlying phylogeny are considered as unknown. The history of the multivariate process, the covariance matrix and the divergence times are jointly estimated in a Bayesian framework, using Monte-Carlo estimation methods. We illustrate the method by applying it to the analysis of the correlations between the rate of substitution, $\omega = dN/dS$, together with several phenotypic characters and life-history traits (generation time, mass and longevity), in a multiple alignment of cytochrome b sequences in mammals.

Methods

Definitions and Notations

The method relies on a combination of aligned coding sequences and phenotypic characters for a set of P taxa. The alignment $\mathbf{D}$ is made of P coding sequences of length $3N$ nucleotides (N codon positions), and the phenotypic data are summarized in a $L \times P$ matrix $\mathbf{C}$ such that C_i^j is the value taken by phenotypic character $i = 1..L$, in taxon $j = 1..P$. In the following, taxa, branches and nodes will always be referred to using upper indices, and phenotypic characters using lower indices.

The taxa are related through a rooted bifurcating phylogenetic tree. The topology is assumed known and will never explicitly appear in the equations. Variables associated to the nodes of the tree are upper-indexed by $j = 0..2P-2$, with the convention that root has index 0, leaf nodes have index $j = 1..P$, and internal nodes have index $j = P+1..2P-2$. If $j > 0$, we refer to the index of the node immediately ancestral to node j as j_{up}. Similarly, upper indices $j = 1..2P-2$ will be used to refer to branches, with the convention that branch j is the branch immediately ancestral to node j. The divergence times are noted $\mathbf{T} = (T^j)_{j=0..2P-2}$. They are relative to the age of the root, i.e. $T^0 = 1$ and $T^j = 0$ for the leaf nodes ($j = 1..P$). For $j > 0$, $\Delta T^j = T^{j_{up}} - T^j$ is the time interval represented by the branch leading to node j.

Codon substitution model

We consider a codon substitution process, such as originally proposed by [72]. First, a general time-reversible Markov process is defined at the nucleotide level, by a 4×4 instant rate matrix R specifying a rate of transition between any pair (n_1, n_2) of nucleotides :

$$R_{n_1 n_2} = \frac{1}{Z} \rho_{n_1 n_2} \pi_{n_2},$$

where Z is the normalization constant, ρ is the set of relative exchangeability parameters constrained to sum to 1 (5 degrees of freedom), and π is the set of nucleotide equilibrium frequencies (3 degrees of freedom). The rate of substitution between any pair of codons (b_1, b_2) differing only at one position and with respective nucleotides n_1 and n_2 at that position, is then defined to be equal to :

$$\begin{aligned} Q_{b_1 b_2} &= \lambda_S R_{n_1 n_2}, \qquad \text{if } b_1 \text{ and } b_2 \text{ are synonymous,} \\ Q_{b_1 b_2} &= \lambda_N R_{n_1 n_2}, \qquad \text{if } b_1 \text{ and } b_2 \text{ are non-synonymous.} \end{aligned}$$

The rate of substitution between any two codons differing at more than one position is assumed to be equal to 0 [72]. The parameters λ_S and λ_N are the rates of synonymous and non-synonymous substitution.

Alternatively, one can define $\omega = \lambda_N / \lambda_S$, and express the rates of substitution in terms of λ_S and ω :

$$\begin{aligned} Q_{b_1 b_2} &= \lambda_S R_{n_1 n_2}, \qquad \text{if } b_1 \text{ and } b_2 \text{ are synonymous,} \\ Q_{b_1 b_2} &= \lambda_S \omega R_{n_1 n_2}, \qquad \text{if } b_1 \text{ and } b_2 \text{ are non-synonymous.} \end{aligned}$$

Both formulations will be explored in the following, as they are not strictly equivalent in the context of the present work. They will be called the (λ_S, λ_N) and the (λ_S, ω) parameterizations.

Multivariate process.

The L phenotypic characters are assumed to evolve continuously along the lineages of the phylogenetic tree. In addition, some of the parameters of the substitution process (he-

reafter called substitution parameters) also vary continuously along the lineages. In the specific case developed in this article, there are two such parameters : either λ_S and ω, or λ_S and λ_N. In more general settings, one could consider the variations of any set of K independent substitution parameters. The substitution process (which is here homogeneous across sites) will then be described by an instant rate matrix $\mathbf{Q}$, itself depending on the K substitution parameters : $\mathbf{Q} = \mathbf{Q}(y_1, \ldots, y_K)$. We wish to model the variations in time of y_k, for $k = 1..K$, and to correlate these variations with those of the L continuous phenotypic characters.

Accordingly, we define a multivariate diffusion process $X(t)$, of dimension $M = K + L$, running along the branches of the tree. The mth. component of $X(t)$ is noted $X_m(t)$, for $m = 1..M$. By convention, the first K components of the process describe the variations of the K substitution parameters, and the last L map to the phenotypic characters. Note that we might have to impose a transformation on the phenotypic characters and the substitution parameters. In the following, the continuous characters that we will consider are life-history traits such as body mass, longevity and generation-time, for which a logarithmic transformation is justified based on known allometric relations [10]. In the case of the synonymous and non-synonymous substitution rates, we follow [88], and impose a logarithmic transformation also in their case. Thus :

$$\begin{aligned} X_1(t) &= \ln \lambda_S(t), \\ X_2(t) &= \ln \omega(t), \\ X_{l+2}(t) &= \ln C_l(t), l = 1..L, \end{aligned}$$

in the (λ_S, ω) parameterization, or

$$\begin{aligned} X_1(t) &= \ln \lambda_S(t), \\ X_2(t) &= \ln \lambda_N(t), \\ X_{l+2}(t) &= \ln C_l(t), l = 1..L, \end{aligned}$$

in the (λ_S, λ_N) formulation.

The stochastic process $X(t)$ is assumed to be multivariate Brownian. The use of a Brownian motion entails several important assumptions. First, it is a Markovian process. Se-

cond, it does not display any trend in the direction of its variations. And third, the rate of change per unit of time is constant, and is completely determined by an $M \times M$ symmetric definite covariance matrix Σ. Between time t and time $t + dt$, the process undergoes a random increment drawn from a normal distribution of mean 0 and variance Σdt. The total variation of the process over a finite time t can be seen as an infinite sum of such normally distributed random increments, and is thus also normally distributed, of mean 0, and variance Σt [25] :

$$X(t) - X(0) \sim N(0, \Sigma t). \tag{I.1}$$

The Brownian motion does not have a stationary distribution.

As a way of explicitly referring to the actual meaning of each entry of the covariance matrix Σ, in the following, we will use a bra-ket notation [20]. For instance, assuming that we have only $K = 1$ phenotypic character, and that we are working under the (λ_S, ω) parameterization, the entries of the matrix will be noted :

$$\Sigma = \begin{pmatrix} \langle \lambda_S, \lambda_S \rangle & \langle \lambda_S, \omega \rangle & \langle \lambda_S, C_1 \rangle \\ - & \langle \omega, \omega \rangle & \langle \omega, C_1 \rangle \\ - & - & \langle C_1, C_1 \rangle \end{pmatrix}.$$

Strictly speaking, these entries correspond to the covariances between the logarithm of the variations of each pair of variables. However, for short, we will more simply refer to this entry as the covariance between λ_S and ω.

The phylogenetic multivariate process

Since we cannot instantiate the process at all times along the phylogeny, we only consider the values of $X(t)$ at the nodes of the tree. We note X^j the instant value of the process at node j, and $\mathbf{X} = (X^j)_{j=1..2P-2}$ the set of values at all nodes except the root. The Brownian process is Markovian, and we can therefore express the joint probability of $\mathbf{X}$, given the initial value X^0 at the root, as :

$$p(\mathbf{X} \mid X^0, \mathbf{T}, \Sigma) = \prod_{j=1}^{2P-2} p(X^j \mid X^{j_{up}}, \Delta T^j, \Sigma), \tag{I.2}$$

where the finite-time transition probabilities are given by equation I.1.

Conditioning the last L dimensions of the multivariate process $X(t)$ on the values observed at leaf nodes (i.e. in extant species) is straightforward, as it is done by setting : $X^j_{l+K} = \ln C^j_l$ for $l = 1..L$ and $j = 1..P$. To introduce a coupling of the first K components of the multivariate process with the substitution process, one would like to set : $y_k(t) = e^{X_k(t)}$ at all times. Doing this would define a non-stationary substitution process, of instant rate matrix $\mathbf{Q}(t) = \mathbf{Q}(y_1(t), \ldots, y_K(t))$. However, we cannot instantiate the values of $X(t)$ at all times, but only at the nodes of the tree, and integrating the likelihood over all possible realizations of $X(t)$ conditional on the values at the nodes seems in most cases completely intractable. Instead, we make the approximation consisting in assuming a constant value for y^j_k along a given branch j, equal to some average of the values of the process at both ends. Doing this for all k defines a constant rate matrix on branch j : $\mathbf{Q}^j = \mathbf{Q}(y^j_1, \ldots, y^j_K)$, and thus, the substitution model reduces to a set of branch-specific substitution matrices $\mathbf{Q}^j$.

There are several possible ways the averages over branches can be computed. First, a very simple approximation, already used in most implementations of relaxed clock models [60, 95], consists of taking the *arithmetic* average :

$$y^j_k = \frac{1}{2}\left(e^{X_k^{j_{up}}} + e^{X_k^j}\right).$$

An alternative method can be proposed. In the case of the Brownian process, the most likely path (or geodesic) going from $X^{j_{up}}$ at time $T^{j_{up}}$ to X^j at time T^j, is the straight line, and therefore, it would make sense to take the mean value of $e^{X(t)}$ along this geodesic, which is equal to :

$$y^j_k = \frac{e^{X_k^{j_{up}}} - e^{X_k^j}}{X_k^{j_{up}} - X_k^j}.$$

We will refer to this latter averaging method as the *geodesic* average. Its main advantage, compared to the arithmetic average, is to more properly account for the convexity of the exponential function. Both approximations are admittedly crude, but since they are quite different in their formulation, using both of them in turn, and comparing the results, will allow some check of the robustness of the method to these finite-time approximations.

Priors

We set a uniform prior on relative divergence times. Analyses were performed with and without calibrations. Without calibrations, divergence dates are simply mesured relative to the root, as in [60]. With calibrations, we proceed as in [54], i.e. we use a gamma density for the age of the root, and conditional on this age, we impose a uniform density on relative ages, truncated so as to be compatible with the intervals specified by the calibrations.

A uniform Dirichlet distribution was imposed on the nucleotide frequencies π and the nucleotide exchangeabilities ρ, a truncated uniform prior defined on $[-100,100]$ for the root state X^0, and an inverse Wishart prior distribution on the covariance matrix Σ, parameterized by $\Sigma_0 = \kappa I_M$, where I_M is the identity matrix of dimension M, and with $q = M+1$ degrees of freedom. As for κ, we tried 2 different values, $\kappa = 1$ and $\kappa = 10$, and we checked that the results were not sensitive to this choice.

The inverse Wishart distribution can be defined as follows [67]. If one samples q independent and identically distributed multivariate normal random variables of dimension M, $Z_i \sim N(0,\Sigma_0^{-1})$ for $i = 1..q$, and computes the scatter matrix

$$M = \sum_{i=1}^{q} Z_i Z_i',$$

then, $\Sigma = M^{-1}$ is, by definition, distributed according to an inverse Wishart of mean Σ_0 and with q degrees of freedom : $\Sigma \sim W^{-1}(\Sigma_0, q)$. The probability density is :

$$p(\Sigma \mid \Sigma_0, q) \propto \mid \Sigma_0 \mid^{\frac{q}{2}} \mid \Sigma \mid^{-\frac{q+M+1}{2}} e^{-\frac{1}{2} tr(\Sigma_0 \Sigma^{-1})},$$

where we have dropped numerical constants. The choice of the inverse Wishart is motivated by the fact that it is conjugate to the multivariate normal distribution, a property which is key to the efficiency of the estimation strategy.

An alternative version of the model is obtained by enforcing all non-diagonal entries of the covariance matrix to be equal to 0. This can be seen as an alternative prior on Σ, with support restricted to the set of diagonal matrices. To make this *diagonal* model as close as possible to the fully *covariant* model introduced thus far, the prior on the entries of the diagonal matrices can be chosen to be the same as the marginal priors of those same

entries in the inverse Wishart. Technically, these are inverse gamma distributions of shape parameter $\alpha = 2$ and scale parameter $\beta = \kappa/2$ [67].

MCMC sampling

Samples from the posterior distribution are obtained by Markov chain Monte Carlo (MCMC). For divergence times, internal nodes are taken one by one (in an order defined by a recursive traversal of the tree). For node j, a simple additive move is applied to T^j within the constraints defined by immediately upstream and downstream nodes. The two vectors ρ and π are updated using a simple move constrained so as to keep the sum of all components equal to 1 [57]. In the simplest version of this mechanism, we randomly choose a pair of 2 entries of the vector to be resampled (say, the two entries π_a and π_b of π), and set $x = \pi(a) + \pi(b)$, and $y = \pi(a)$. We then propose $y' = y + \varepsilon(U - 0.5)$, If y' falls outside of the interval $[0,x]$, we reflect it back. Finally, we set $\pi'(a) = y'$ and $\pi'(b) = x - y'$, thus preserving the total sum of the two stationary probabilities. The Hastings ratio is 1. A generalized version of this sum-constrained mechanism randomly draws d non-overlapping pairs of entries of the vector to be resampled, and simultaneously proposes a compensated move independently on each pair.

Concerning the covariance matrix Σ and the multivariate process X, we implemented two update schemes. The first is a simple alternation between Metropolis-Hastings updates of X conditional on Σ, and conversely, of Σ conditional on X (and all other parameters). Specifically, for the multivariate process, internal nodes of the tree are visited one by one. For node j, one among three types of moves are proposed, each with probabilty 1/3. According to move number 1, one entry of X^j is chosen uniformly at random, and, if this entry is not clamped, an additive move is performed on it (otherwise, nothing happens). According to move number 2, all entries of X^j (if not clamped) are moved by a same random amount $\varepsilon(U - 0.5)$. According to move number 3, all entries that are not clamped are simultaneously moved, each by a different random amount $\varepsilon(U_m - 0.5)$, for $m = 1..M$. Concerning the covariance matrix Σ, an entry l,m, such that $l \leq m$ is chosen uniformly at random, an additive move is proposed : $\Sigma'_{lm} = \Sigma_{lm} + \varepsilon(U - 0.5)$, and the symmetry of the matrix is restored by setting Σ'_{ml} equal to Σ'_{lm}. The move is immediately refused if the resulting matrix is not positive definite. Otherwise, the Metropolis decision rule is applied. Note that this is equivalent to setting a posterior density of 0 on all symmetric but

not definite positive matrices.

However, this simple alternate update scheme is not efficient, due to potentially strong correlations between X and Σ in their joint distribution. A much more efficient approach relies on the conjugate relation between the Inverse Wishart and the multinomial distributions to analytically integrate away the covariance matrix Σ.

To see this, we first make a change of variables, so as to define the branchwise independent contrasts $\mathbf{Y} = (Y^j)_{j=1..2P-2}$:

$$Y^j = \frac{X^j - X^{j_{up}}}{\sqrt{\Delta T^j}}.$$

These contrasts are i.i.d. from a multivariate normal distribution :

$$Y^j \sim N(0, \Sigma).$$

With this change of variable, any density defined on $\mathbf{Y}$ corresponds to a density on $\mathbf{X}$ according to :

$$p(\mathbf{X}) = p(\mathbf{Y}) \mid J \mid,$$

where $\mid J \mid$ is the jacobian of the transformation :

$$\mid J \mid = \prod_{j=1}^{2P-2} (\Delta T^j)^{-\frac{M}{2}}.$$

Next, we rely on the fact that the inverse Wishart distribution is conjugate to the multivariate normal distribution. We can write, up to a normalization constant, the prior on Σ :

$$p(\Sigma \mid \Sigma_0) \propto \mid \Sigma \mid^{-\frac{q+M+1}{2}} e^{-\frac{1}{2} tr(\Sigma_0 \Sigma^{-1})}, \tag{I.3}$$

and the sampling probability of $\mathbf{Y}$:

$$p(\mathbf{Y} \mid \Sigma) \propto \prod_{j=1}^{2P-2} \frac{1}{\sqrt{\mid \Sigma \mid}} e^{-\frac{1}{2} Y'^j \Sigma^{-1} Y^j} \tag{I.4}$$

$$\propto \mid \Sigma \mid^{-(P-1)} e^{-\frac{1}{2} tr(A \Sigma^{-1})}, \tag{I.5}$$

where we define the sample covariance matrix

$$A = \sum_{j=1}^{2P-2} Y^j Y'^j. \tag{I.6}$$

By Bayes theorem, the posterior on Σ, conditional on a particular realization of X (and thus Y) is proportional to the product of equations I.3 and I.5 :

$$p(\Sigma \mid \mathbf{Y}, \Sigma_0) \propto \mid \Sigma \mid^{-\frac{q+M+2P-1}{2}} e^{-\frac{1}{2}tr((\Sigma_0+A)\Sigma^{-1})}, \tag{I.7}$$

from which we see that the posterior is also an inverse Wishart, of parameter $\Sigma_0 + A$ and with $q + 2P - 2$ degrees of freedom.

By identification, the marginal probability density on $\mathbf{Y}$ can be obtained :

$$p(\mathbf{Y} \mid \Sigma_0, q) = \int p(\mathbf{Y} \mid \Sigma) p(\Sigma \mid \Sigma_0, q) d\Sigma \tag{I.8}$$

$$\propto \frac{\mid \Sigma_0 \mid^{\frac{q}{2}}}{\mid \Sigma_0 + A \mid^{\frac{q+2P-2}{2}}}, \tag{I.9}$$

where we have dropped unimportant numerical constants. Finally, the marginal probability of X is obtained by multiplying equations I.9 and I.3 :

$$p(\mathbf{X} \mid \Sigma_0, q) = p(\mathbf{Y} \mid \Sigma_0, q) \mid J \mid . \tag{I.10}$$

The two important results coming out from this mathematical derivation are contained in equation I.7 and I.10. Equation I.10 tells us that we can compute the probability of a particular configuration of the stochastic process $\mathbf{X}$ integrated over all possible values of Σ. This means that we can devise a MCMC sampler working on the reduced parameter space $(\mathbf{X}, \mathbf{T}, \pi, \rho)$, not including Σ. The equilibrium distribution of this sampler is the marginal posterior distribution :

$$p(\mathbf{X}, \mathbf{T}, \pi, \rho \mid D, C, \Sigma_0, q) \propto p(D, C \mid \mathbf{X}, \rho, \pi) p(\mathbf{X} \mid \Sigma_0, q) p(\mathbf{T}) p(\pi) p(\rho), \tag{I.11}$$

where $p(\mathbf{X} \mid \Sigma_0, q)$ is given by equation I.10. This posterior distribution is conditional on the multiple alignment D, the matrix of continuous characters C, but also on the constants

of the model Σ_0 and q.

In a second step, for each value of $\mathbf{X}$ obtained from this reduced MCMC (burn-in excluded), a scatter matrix A can be computed (equation I.6), and a value of Σ can then be sampled from the distribution given by equation I.7. Since, at equilibrium, the values of $\mathbf{X}$ sampled from the reduced MCMC are from the posterior distribution :

$$\mathbf{X} \sim p(\mathbf{X} \mid D, C, \Sigma_0, q), \tag{I.12}$$

and since Σ is conditionally independent of the data (D and C) given X :

$$p(\Sigma, \mathbf{X} \mid D, C, \Sigma_0, q) \quad = \quad p(\Sigma \mid \mathbf{X}, \Sigma_0, q)\, p(\mathbf{X} \mid D, C, \Sigma_0, q), \tag{I.13}$$

sampling Σ from equation I.7 results in Σ and $\mathbf{X}$ being sampled from their joint posterior distribution. In practice, our sampler works on the reduced parameter vector, but resamples Σ on the fly, each time before saving a new parameter configuration.

Data augmentation.

In addition to the conjugate sampling method just described, the overall MCMC framework relies on data augmentation [16, 57, 71]. At any time, a complete substitution history (or mapping) is specified, for all sites and over the whole tree, and all the update mechanisms described above are performed conditional on the current mapping. Periodically, the mapping is refreshed, i.e. it is resampled conditional on the current parameter values, using a combination of two algorithms described elsewhere [75, 86]. This MCMC strategy allows for drastic simplifications of the computations [57, 71]. First, it does not use the pruning algorithm [23], except for resampling the substitution mappings. This is a substantial advantage in the present case, where the state-space of the substitution process has size 61. Second, conditional on the current substitution mapping, the probability of the substitution mapping depends on fairly compact sufficient statistics (total waiting time in each possible codon, number of transitions between each pair of codons, and in the case of the root, total number of occurrences of each codon), that have here to be computed separately for each branch, but can be summed over all sites of the alignment [16, 57, 71]. The cost of all update mechanisms except the data augmentation step itself are therefore vir-

tually independent of the number of sites, which is a great advantage for long sequences. Using fast-access associative-array representations of the 61 vectors and 61×61 arrays of sufficient statistics, which are potentially sparse for smaller alignments, makes the implementation efficient across the whole range of sequence length.

The frequency at which the substitution mapping is refreshed, now the limiting step, is tuned so that the MCMC sampler spends between one tenth and one half of the total computing time refreshing mappings, while the rest of the time is distributed over all other update operators. In practice, all parameters are each resampled several hundred times between each update of the substitution mapping, and one point is saved before each such update. The burn-in is determined visually, and the chain is run for approximately 1000 points. Each analysis was run at least twice independently.

The implementation was checked using three different methods [4] : (1) the program was run using alternative sampling methods (using conjugate or regular sampling for the covariance matrix, using sufficient statistics or directly recomputing the probability of the substitution histories), and we checked that the equilibrium distributions obtained under the different methods were indistinguishable ; (2) the MCMC was run with no data, to visually check whether the model was indeed sampling from the prior ; and (3) 100 replicates were simulated from the prior, using the (λ_S, ω) parameterization and assuming one continuous character, and reanalyzed under the model, so as to obtain for each replicate a sample of 1000 points approximately from the posterior. For each replicate, and for a series of 8 summary statistics, the true value of the statistic was ranked against the sample from the posterior. The 100 ranks thus obtained (expressed in percentiles) should follow a uniform distribution [42], which we visually checked, and quantitatively assessed by the Kolmogorov uniformity test. The summary statistics were the the total length of the tree, the mean value of omega along the tree, and the 6 independent entries of the 3×3 covariance matrix.

Post treatment

Once a sample approximately from the posterior is obtained, marginal estimates of any parameter of the model are readily computed. Concerning the reconstructed chronology and phenotypic histories, we estimate, for each node of the tree, a posterior mean and a

95% credibility interval for its date and for each phenotypic character. The same argument applies to substitution parameters.

Concerning the covariance matrix Σ, for each entry, we simply report the posterior mean. For non-diagonal entries $k \neq l$, the correlation coefficient is defined as :

$$r_{kl} = \frac{\Sigma_{kl}}{\sqrt{\Sigma_{kk}\Sigma_{ll}}}.$$

The reported correlation coefficients are obtained by applying this formula separately for each point sampled approximately from the posterior distribution, and then taking the average. The posterior probability (pp) of a positive correlation ($r_{kl} > 0$) is also estimated based on the observed frequency at which the r_{kl} parameter was found to be positive.

Finally, the slope of the linear regression between two components k and l can be estimated. Here, we use the major axis method, which estimates the slope of the major axis of the bivariate ellipsoid formed by the joint distribution of the two variables of interest [38] :

$$\frac{\partial X_l}{\partial X_k} = \frac{\Sigma_{ll}^2 - \Sigma_{kk}^2 + \sqrt{(\Sigma_{ll}^2 - \Sigma_{kk}^2)^2 + 2\Sigma_{kl}^2}}{2\Sigma_{kl}}.$$

This slope is computed for each point sampled from the posterior, thus giving a distribution from which an average and a 95% credibility interval are then immediately obtained.

For multiple regressions, one is interested in knowing the covariance between k and l, for constant m, which is given by :

$$\Sigma_{kl;m} = \Sigma_{kl} - \frac{\Sigma_{km}\Sigma_{lm}}{\Sigma_{mm}}.$$

Again, the formula is applied for each point from the posterior, so as to obtain a distribution from which to compute a mean and a posterior probabilty for assessing significance.

The software program runs under the Linux or MacOS operating system. It is freely available from our website http ://www.phylobayes.org.

Data set

We analyzed an alignment of cytochrome b sequences of 410 therian species, 29 marsupials and 381 placentals (1146 nucleotide positions, or equivalently, $N = 382$ codon positions), obtained from [73]. Three life-history traits were investigated : age of female at maturity, taken as a proxy for generation time, adult weight, as a proxy for body mass, and maximum recorded lifespan, as a proxy for longevity. The values of these three characters were obtained from the AnAge database [17]. The 410 sequences correspond to all marsupials and placentals of the initial data set of [73] for which the three life-history traits are documented in the AnAge database. We also extracted from this large alignment a reduced data set restricted to carnivores (67 taxa). We used the fossil calibrations reported in [73], except for the 5 involving taxa absent from our data set (i.e. *Bradypus*/*Dasypus*, *Apodemus*, *Gerbillus*, *Bolomys*/*Acodon* and *Neofiber*/*Ondatra*). The prior on the age of the root was defined as an exponential of mean 150 Million years.

Simulations

A first series of simulations was conducted to assess whether the method is able to recover reasonable estimates of the covariance matrix in practical situations. To ensure realism, the simulations were based on a tree, a set of divergence times and a mutation rate matrix estimated on the carnivore data set. In addition, and as an attempt to address the problem of the discretization error induced by the finite-time averages computed along each branch, the simulations were performed using a more sophisticated version of the model, in which each branch is subdivided in 50 small segments of equal length. The simulation proceeds step by step, successively along each segment, which results in smaller discretization errors, and is therefore closer to the ideal situation in which the codon substitution process is supposed to change continuously along the branches.

For 50 replicates, random covariance matrices were sampled from the prior (using $\kappa = 1$), and were used to simulate a complete history of the multivariate process along the tree, and from this realization of the process, a codon-alignment of 342 coding positions (1146 aligned nucleotides) and a data-matrix for one continuous character, using the (λ_S, ω) model with arithmetic averages as the simulation model. To avoid numerical problems and

unrealistic substitution rates, the replicates for which one of the branches had a length or a value of ω greater than 5 were discarded. The simulated data sets were then analyzed under the (λ_S, ω) model. To test the effects of the approximations due to the finite-time averages taken over branches, we analyzed the data using either the arithmetic or the geodesic averaging schemes. For each replicate, and each covariance parameter of the 3×3 covariance matrix, the 95% credibility interval obtained under each model was compared to the true value.

In theory, when the replicates have been simulated and analyzed under the same model, and with the same prior, the Bayesian credibility intervals have a simple frequentist interpretation, namely, that 5% of the true values are expected to fall outside the 95% credibility intervals [42]. In the present case, since the simulation and estimation model are different, and since we condition the simulations, but not the analysis, on a fixed, predefined, chronogram and a fixed mutation matrix, we do not expect this property to strictly hold. On the other hand, as long as the parameters that are fixed across simulations are not too atypical *a priori*, the frequentist property is expected to hold approximately. Another point of interest is how far from the true value the mean estimated covariances will be, which can be checked visually.

In both cases, using either the arithmetic or the exponential averaging method, a strong correlation between the true and the estimated covariances is observed (Figure 1). The choice between linear or geodesic averaging seems to have a rather small influence on the estimation (compare Figure 1B,D,F with A,C,E), although the geodesic method appears to be slightly more accurate. Specifically, the 95% credibility intervals encompass the true value except in 10% of the cases (16 out of 150 estimated covariances) when using the arithmetic average, and 6% of the cases (9 out of 150) when using the geodesic average.

A second series of simulations was aimed at assessing the rate of false positives of the method. When estimating a covariance matrix on a true data set, one naturally wants to assess how confident to be about the fact that the covariance between 2 parameters of interest is indeed positive (or negative). In a Bayesian framework, the posterior probabilty (pp) that the covariance between the two parameters of interest is positive is supposed to measure exactly this confidence. Note that, by symmetry, the prior probability of a positive covariance is 0.5, and therefore, the model does not *a priori* favor any particular direction.

In principle, the posterior probability is not to be interpreted in frequentist terms, i.e. $1-pp$ is not supposed to be an equivalent of the p-value of a frequentist test in which the null hypothesis would be that the covariance is in fact equal to zero. Nevertheless, it is natural to expect that the method does not produce false positives too often, i.e. does not often give a high posterior probability for a positive or a negative covariance, when applied to data that have in fact been simulated under a null covariance model.

To assess this on a more empirical ground, we first estimated the parameters of the diagonal model (i.e. with all covariances set to 0) on the carnivore data set, and with the 3 continuous life-history traits (generation time, mass and longevity). We then resimulated data under the posterior predictive distribution, i.e. we simulated 100 replicates of the data set, each replicate consisting of a codon alignment of 342 coding positions (1146 aligned nucleotides) and a set of continuous phenotypic characters, always under the assumption of no correlation between the $M = 5$ components of the process. Next, we applied the fully covariant model on each replicate, and measured the posterior probability of a positive covariance between each $M(M-1)/2 = 10$ pairs of entries of the multivariate process. In this way, we can assess the frequency at which posterior probabilities are more extreme than a given threshold. Since we do not have any prior expectation about the sign of the covariance, for a given hreshold α, we measure the frequency at which either $pp > 1-\alpha/2$, or $pp < \alpha/2$.

The results are presented in Table 1, for several values of α. Whether the data are simulated and tested under the same model, or whether different approximation schemes are used for simulation and analysis, the test, as seen in a frequentist perspective, seems slightly conservative (i.e. the rate of false positives at the α level appears to be less than α). The specific approximation scheme does not seem to have a strong impact on the behavior of the test. A point of great practical importance is that, for a very low threshold ($\alpha = 0.0001$), no false positives were seen among the 100 replicates, thus for all 1000 covariances tested. This means that, if anything, the method does not seem to result in apparently strongly significant, albeit in fact spurious, correlations. Altogether, although more extensive simulations and more definitive theoretical results would probably be needed to add further weight to this conclusion, the present empirical analysis suggests that we can be confident in the posterior probabilities associated to the observed correlations.

Results

To illustrate the method, we applied it to two alignments of cytochrome b sequences of 67 carnivores, and 410 therian mammals [73]. The phenotypic or life-history characters were generation time, mass and longevity, and the substitution parameters were the rates of synonymous substitution λ_S and the ratio of non-synonymous over synonymous substitution ω.

Covariance analysis

The estimated covariance matrix is reported in Table 2, together with the correlation coefficients and the posterior probability for each non-diagonal entry to be positive.

In therians, mass, generation time, and longevity are strongly and positively correlated with each other ($pp > 0.99$). The rate of synonymous substitution λ_S is negatively correlated with mass ($pp < 0.01$), and with longevity ($pp = 0.01$). No correlation is observed with generation time ($pp = 0.30$). Similarly, ω is positively correlated with mass ($pp > 0.99$), with longevity ($pp = 0.99$), but again not with generation time ($pp = 0.35$).

In carnivores ω is also correlated with mass ($pp > 0.99$), marginally with longevity ($pp = 0.94$) and, unlike in therians, marginally also with generation time ($pp = 0.93$). On the other hand, in carnivores, λ_S does not seem to correlate with any of the three life-history traits (Table 2). Using either the geodesic or the arithmetic averaging procedure, or using $\kappa = 1$ or $\kappa = 10$ for the inverse Wishart prior, did not seem to have any influence on the inference (not shown).

Using fossil calibrations, in the case of therians, led to a global enhancement of the estimated covariance matrix (Table 3). In particular, the variance per unit of time of λ_S is larger by nearly 50%, which cleary indicates that the variations of the mutation rate in mitochondrial DNA are underestimated when divergence dates are not properly calibrated, as previously suggested [73]. Interestingly, the calibrated analysis also yields a significantly negative correlation between λ_S and ω, which was not observed in the analysis without calibrations. All other estimates are very similar, whether or not calibrations are used (Table 3).

An analysis was also conducted under the (λ_S, λ_N) parameterization (Table 4). The results

are concordant with those obtained under the (λ_S, ω) parameterization, i.e. λ_S does not correlate with life-history traits, and λ_N correlates with mass, and marginally with longevity and generation time in carnivores. In therians, a negative correlation between λ_S and mass and longevity is recovered. As for λ_N, it shows a marginal positive correlation with mass and longevity. Of interest, λ_S and λ_N are found to be positively correlated in therians ($pp = 0.99$), and marginally in carnivores ($pp = 0.92$).

Some of the methods of standard linear regression and analysis of variance have a direct equivalent in the present case. In particular, the slope of the pairwise relation between two variables can be estimated (see methods). For instance, in the case of Therians, the slope of the logarithmic variations of generation-time versus mass is estimated at 0.20, with a 95% credibility interval (95% CI) at $[0.16, 0.25]$. In the case of longevity as a function of mass we obtain 0.14 (95% CI $[0.11, 0.17]$). The estimated slopes were very similar, with or without calibrations, under $\kappa = 1$ or 10, and using the arithmetic or the geodesic averaging method. They are smaller than the coefficients of 0.25 and 0.20 often reported for these allometric scaling relations [10]. On the other hand, a direct linear regression on the life-history traits of the 410 therian taxa yields a slope of 0.22 for generation-time versus mass, and of 0.17 for longevity versus mass, which suggests that the discrepancy may come from the particular taxonomic level or sampling presently considered, and not from the model.

Another quantity of interest is the square of the correlation coefficient between 2 components r^2_{kl}, which expresses how much of the total variation of l is explained by k, and vice-versa. Here, the squared correlation coefficient is $(-0.18)^2 = 0.032$ between λ_S and mass, and $(-0.27)^2 = 0.073$ between ω and mass (Table 2). In other words, the variations of body weight explain 3% of the variations of the rate of substitution, and 7% of the variations of ω. Mass and longevity are the most strongly correlated characters, explaining 25% of the variation. In therians, the correlation coefficient between λ_S and λ_N (Table 4) is 0.35, i.e. the rate of synonymous substitution explains $(0.35)?2 = 12\%$ of the variations in the rate of non-synonymous substitutions.

Multiple regression analysis can also be used (see methods), for instance, to attempt to discriminate between mass and longevity as the primary factor correlating with λ_S (Table 4). Under constant mass, no residual correlation is observed between longevity and λ_S ($pp = 0.16$). In contrast, under constant longevity, the covariance between λ_S and mass is

still significantly negative ($pp = 0.02$). Thus, according to the present analysis, mass, and not longevity, seems to be the main explanatory variable for substitution rate variations in therians.

Phylogenetic reconstruction of the phenotypes

The divergence times and the variations in body mass along the phylogeny were reconstructed under the fully covariant model for carnivores (Figure 2) and for therians (not shown). Overall, the 95% credibility intervals are large. Specifically, the common ancestor of carnivores is inferred to have an adult body mass between 1.8 and 24.7 kg. From there, the evolutionary trends are very different, depending on the suborders considered : mass progressively increases in the lineage leading to Ursidae and Pinnipedia, being estimated at 46 (95% credibility interval $[11, 8210]$) kg in their common ancestor, while decreasing to less than 1kg in Herpestidae (mongooses).

To measure the impact of potential interactions between molecular and phenotypic reconstructions through the covariance matrix, an inference was also conducted under the diagonal model, obtained by constraining the non-diagonal entries of the matrix to be equal to zero. To quantify the differences between the two reconstructions, we computed, for each node, a deviation index as follows :

$$z^j = \frac{2(E_{cov}[\ln C^j] - E_{diag}[\ln C^j])}{\sqrt{V_{cov}[\ln C^j]} + \sqrt{V_{diag}[\ln C^j]}}, \tag{I.14}$$

where $E[.]_{cov}$ and $E[.]_{diag}$ are the sample means, $V[.]_{cov}$ and $V[.]_{diag}$ the sample variances, under the covariant and the diagonal models, and C^j is the value reconstructed at node j for the character of interest. This deviation is loosely analogous to a z-score, although it is not meant as a measure of significance, but only as a heuristic measure of the difference between the estimates obtained under the two models.

The differences between the reconstructions inferred under the covariant and the diagonal models are small (Figure 3). In therians the first 10 highest deviation indices are all positive, between 0.9 and 1, and all of them fall in the group of Cricetidae. Thus, there is a signal in the multiple sequence alignment indicating that early Cricetidae may have been larger than what is inferred just based on the phenotypic data and the phylogenetic

tree. For instance, the ancestor of Cricetidae is inferred to have a mass of 151 (95% CI $[58, 376]$) grams under the covariant model, instead of 97 (95% CI $[32, 279]$) grams under the diagonal model. An opposite trend is observed in carnivores, with a maximum deviation index of -0.5 for the ancestor of Ursidae, suggesting that, in this case, covariance between substitution rates and body mass results in a downward correction of body mass for the ancestor of Ursidae. Likewise, the most recent common ancestor of Ursidae and Pinnipedia, which comes third (z = -0.4), is inferred with a mass of 46 (95% CI $[11, 8210]$) kg under the covariant model, instead of the 73 (95% CI $[18, 317]$) kg found under the diagonal model. In all cases, however, the differences between the covariant and the diagonal model are small compared to the credibility intervals, and may just as well be a stochastic fluctuation, or a consequence of inaccurate divergence time reconstruction.

Discussion

Comparative analyses of molecular and phenotypic characters are a key aspect of molecular evolutionary studies. In this direction, what we propose here is the first fully integrated method dealing with the nuisances caused by phylogenetic dependences and by the various sources of uncertainty about the phenotypic and molecular history.

Probably the most immediate advantage of the method developed here is its practical simplicity. Essentially, the entire procedure reduces to a one-step analysis in which all the available evidence is given as an input, and estimates of all potentially interesting aspects of the problem (covariances, divergence times, phenotypic histories) are obtained as the output. The posterior probabilities offer a simple and natural method for evaluating the significance of the observed correlations.

Joint estimation of all parameters in a Bayesian framework has another important advantage. When estimating a parameter of interest, in particular the covariance matrix, the uncertainty about all other parameters (e.g. on reconstructed rates of substitution, or on divergence times), is automatically accounted for [33, 43, 44]. This is particularly important in analyses of single genes such as cytochrome b, for which the sampling error associated with the estimation of substitution rates is potentially large. Integrating over divergence times may also be important, as errors on branch lengths have been shown to result in inflated type I errors [19].

On the other hand, accounting for the uncertainty about the nuisance parameters by integrating over the prior raises the issue of prior sensitivity. In the present case, one can point out at least two components of the model for which prior sensitivity may be an issue. First, divergence times are potentially more sensitive to the prior [45] than previously suggested [60], indicating that the present framework should be extended to accomodate alternative priors on divergence dates, in particular the birth-death prior [103], and that the robustness of the analysis to the choice of the prior should be thoroughly assessed.

Another point concerning prior sensitivity is the choice of κ, the prior mean variance parameter for each component of the multivariate process. Defining a sensible value for κ is particularly difficult, since we have absolutely no relevant prior information about the scale of the rate of change of the substitution parameters, nor of the phenotypic characters. In the present case, we just checked that the analysis was robust with respect to the choice of $\kappa = 1$ or 10, which is probably good enough given that the posterior mean values obtained for the variance parameters are within this range (Table 2, diagonal coefficients). However, this approach is not totally satisfactory, conceptually speaking. An alternative would be to work in a hierarchical Bayes framework, and use an uninformative prior, such as Jeffreys' prior [46], possibly accomodating different values κ_m for each component $m = 1..M$ of the process. Another approach, more in the spirit of empirical Bayes, would consist in optimizing the marginal likelihood of the overall model according to the hyperparameter κ (or the hyperparameters κ_m).

More fundamentally, the choice of an inverse Wishart distribution as our prior on the covariance matrix was motivated exclusively by computational arguments. The inverse Wishart is conjugate to the multivariate normal distribution, thus allowing us to integrate away the covariance matrix from the MCMC sampler (see methods). In practice, the improvement brought by the conjugate sampling method seems to be essentially dependent on the dimension M of the multivariate process, which is expected, given that the number of independent parameters represented by the covariance matrix increases as M^2. Thus, the improvement is minor for $M = 3$ (i.e. two substitution parameters combined with one continuous character), significant for $M = 5$, with a burn-in three times as long under the non-conjugate sampling method than under the conjugate one, and essential for larger values of M : for $M > 10$, we were unable to obtain convergence using the non-conjugate method. This is also true if we increase the dimension of the process by having more

substitution parameters allowed to vary along the lineages (not shown).

On the other hand, alternative priors could be imagined. In particular, conditional independence between certain pairs of variables could be modelled more directly, and perhaps more adequately, by allowing the corresponding non-diagonal entries of Σ^{-1} to be equal to zero with positive prior probability. This can be seen as a reformulation, in a comparative context, of covariance selection models [18, 21]. Reversible-jump Monte Carlo methods might have to be developed in order to sample from such models. Alternatively, under certain conditions, it might be possible to develop covariance-selection priors while preserving conjugacy [15, 61, 87]. In both cases, the model would offer a direct estimate of the marginal posterior probability of conditional independence between each pair of variables, and may also have a better fit, thanks to its smaller effective number of parameters.

Apart from the question of the prior, our method makes several assumptions and approximations, all of which may deserve further discussion. First, we approximate the continuously changing substitution process by a piecewise constant process, using average substitution matrices, one for each branch. We have proposed two alternative approximations for these average matrices, and checked by simulations that these approximations did not result in large estimation errors. On the other hand, our checks do not offer any guarantee that the approximation will be acceptable in all circumstances. A possibility would be to use less extreme discretization schemes, for instance by sampling the values of the multivariate process at intermediate points along the branches, and not only at their extremities. An acceptable compromise between granularity and computational cost may then be found empirically, on a case-by-case basis.

Second, we have assumed that logarithmic transformations would be adequate to reduce the problem to one of estimating linear correlations between variables. In the present case, a logarithmic transformation is probably the most obvious choice to make in the case of life-history traits, for which known allometric relations are equivalent to log-linear correlations between variables [10]. Concerning substitution rates, the case is less obvious, although the fact that rates can display variations on several orders of magnitude [73] strongly argues in favor of a change of variable akin to a logarithmic transformation. In principle, alternative transformations of the variables could be proposed, and compared by computing Bayes factors, or alternatively, could be averaged over using a hierarchi-

cal model. An even more advanced approach would consist in developing non-parametric methods able to estimate the transformation directly from the data.

Finally, the assumptions implied by the use of a Brownian diffusion process, namely a constant rate of change, and an absence of trend in the direction of the changes, could be relaxed by implementing alternative stochastic processes, such as the Ornstein-Uhlenbeck process [9], or burst models, allowing for a varying rate of phenotypic evolution in different regions of the tree [14].

Insights into molecular evolutionary mechanisms

We have introduced two alternative parameterizations of the covariance model, in terms of either λ_S and ω, or λ_S and λ_N. Ideally, since the process is multivariate normal, and since $\ln \omega = \ln \lambda_N - \ln \lambda_S$ is a log-linear relation preserving the normality of the process, the two representations should be equivalent, the two covariance matrices being related by the following change of variables :

$$\begin{aligned}
\langle C, \omega \rangle &= \langle C, \lambda_N \rangle - \langle C, \lambda_S \rangle, \\
\langle \lambda_S, \omega \rangle &= \langle \lambda_S, \lambda_N \rangle - \langle \lambda_S, \lambda_S \rangle, \\
\langle \omega, \omega \rangle &= \langle \lambda_S, \lambda_S \rangle + \langle \lambda_N, \lambda_N \rangle - 2\langle \lambda_S, \lambda_N \rangle,
\end{aligned}$$

where C is any phenotypic character. On the other hand, since the prior on Σ is not invariant by this change of variable, for finite data, the result of the estimation will depend on the chosen representation. Which representation is more convenient depends on the question being addressed. The prior is centered on the diagonal model, and thus is neutral with respect to positive or negative covariance among the substitution parameters and the phenotypic characters. Therefore, the choice should mainly depend on which variables we consider as *a priori* independent.

The justification of the (λ_S, ω) parameterization is mechanistic. Assuming that selection on synonymous substitutions is negligible, variations of λ_S will mostly be due to variations of the mutation rate λ, and will be independent of population size. On the other hand, if the mutation rate is not too high, so that interferences between non-neutral polymorphisms are negligible, ω will be independent of the mutation rate λ, and will be equal to the

fraction of effectively neutral non-synonymous mutations. This fraction is expected to be a decreasing function of effective population size, as slightly deleterious mutations that would otherwise be nearly neutral in species with a small effective size may find themselves purified away in species with a larger effective size [51, 52, 77, 78, 99]. Using the (λ_S, ω) parameterization therefore amounts to assuming that such a nearly-neutral model applies to the data at hand.

On the other hand, if we are more suspicious about the validity of the nearly-neutral model, in particular if we suspect that non-synonymous substitutions may not be limited by the mutation rate, but instead by ecological adaptive opportunities [34], then considering λ_S and λ_N as the two *a priori* independent variables may turn out to be more adequate. Using the (λ_S, λ_N) parameterization can also be seen as a way of testing whether λ_S and λ_N are indeed positively correlated, as would be expected under the nearly-neutral model, or more generally if the mutation rate is limiting. For those reasons, we think it is important to propose a software program in which the two alternative parameterizations are available.

The positive correlation observed between λ_S and λ_N in our analyses (Table 4) is consistent with the nearly-neutral model. The correlation is significant, albeit perhaps a bit weak, with the variations of λ_S explaining only 12% of those of λ_N in therians. This may simply be due to a lack of power, owing to the small size of the alignment (small number of positions). Alternatively, it could be the consequence of adaptive phenomena partially decoupling λ_N from λ_S. More extensive analyses, in particular using longer sequences, would be needed here.

Also in favor of the nearly-neutral model, we see a positive correlation between ω and mass and longevity, both at the level of one single order (carnivores) and at the more global scale of therians (Table 2). This correlation can be interpreted as an indirect effect of variations of effective population size, itself negatively correlated with mass and longevity, in agreement with previous observations [82, 96]. What may appear more intriguing is that population size is known to also be correlated with generation time in mammals [13]. Yet, in the present case, we do not observe any correlation between ω and generation time in therians, and we see it only marginally in carnivores (Table 2).

We also observe a negative correlation between λ_S and mass and longevity, which is consistent with a previous analysis [73]. Two alternative explanations can be proposed

for this correlation. First, it could be an indirect effect of metabolism, larger animals having a lower metabolism [35, 68]. However, several analyses have already questioned the metabolic rate hypothesis, suggesting that the correlation with metabolic rate was at best indirect, being mediated by a body-size effect [7, 56]. Alternatively, the mutation rate could be under adaptive regulation, linked to the necessity of restricting mitochondrial somatic mutations in large and long-living mammals [73, 90]. The question could be further investigated under the present framework, and using multiple regression, to discriminate between mass and metabolic rate.

Finally, we do not observe any correlation between λ_S and generation time, neither in carnivores nor in therians, and whether or not fossil calibrations are used. A strong generation-time effect has often been reported previously, but mostly for nuclear sequences [62, 63]. In contrast, the generation-time effect was found to be weaker in mitochondrial sequences [73]. Nevertheless, the fact that we could not observe any correlation between either λ_S or ω and generation-time, despite the fact that such correlations would be plausible, should probably be further investigated.

Reconstructing phenotypic evolution

Reconstructing phenotypic evolution based on a joint analysis of phenotypic and molecular data is one of the most exciting perspectives opened by the present method. Joint estimation implies that potentially relevant information is shared across the different components of the parameter vector. Via the covariance matrix, a potentially interesting cross-talk may therefore occur between substitution rates and divergence times [100], or between rates and the reconstructed phenotypic history.

In the present case, however, we have not seen much influence of the covariance structure of the model on divergence times, nor on phenotypic reconstructions. This could be due to several factors, although a likely explanation in the present case is indicated by the squared correlation coefficients estimated for this data set (Table 2 and 3). Since λ_S explains only around 3%, and ω about 7%, of the variations of mass and longevity, we should not expect λ_S and ω to have a strong influence on the phenotypic reconstruction. Such a weak coupling between rates and life-history traits could be an intrinsic property of the evolution of mammalian mitochondrial sequences, i.e. rates may have other hidden

determinants that happen to dominate their overall fluctuations. Alternatively, it could be a consequence of the large uncertainty associated with the estimation of substitution rates, itself a consequence of the small number of aligned positions used in the present study. In the latter case, a comparative analysis conducted with the entire proteome of mammalian mitochondrial genomes should lead to higher correlations, and may therefore help reveal significant interaction between rates and phenotypes.

Perspectives

Our observations concerning the pattern of molecular and phenotypic evolution in mammals are very preliminary. Their aim was merely to introduce the method, and it is clear that, in order to draw more definitive conclusions about the several points of biological interest raised above, much more ambitious analyses need to be conducted.

First, not only are the variations of λ_S and of ω for cytochrome b subject to large stochastic errors, due to the small number of aligned positions, but they may also express specific adaptations of cytochrome b in particular lineages. Since variations of life-history traits are expected to have genome-wide consequences on the pattern of molecular evolution, one possible approach would consist in integrating the signal over several genes, so as to average out gene-specific idiosyncrasies, and recover only the global trends. The power of comparative analyses also crucially depends on a dense taxonomic sampling, and therefore the two criteria, many genes and many taxa, should ideally be met simultaneously.

Second, mitochondrial sequences are particularly saturated [8], owing to a very high mutation rate in mammals, and more generally in metazoans. Saturation has probably eroded a significant proportion of the molecular evolutionary signal in the deeper part of the mammalian tree. Using the less saturated nuclear sequences, and more generally, analyzing several genetic units subject to different evolutionary regimes, should significantly increase the power of the comparative approach.

Finally, the method could be extended to many other potentially interesting substitution parameters. For instance, investigating the correlation between the transition-transversion ratio, the equilibrium GC content, and phenotypic characters may help discriminate between alternative hypotheses about the determinants of the mutation pressure, or the population genetic mechanisms underlying the substitution process. More generally, the me-

thod developed here could in principle be used for investigating a wide diversity of potential correlations between phenotypes and sequences, thereby providing many stimulating empirical observations helping us to better understand the mechanisms of molecular evolution, and to reconstruct the evolution of phenotypic and life-history traits.

Acknowledgments

We wish to thank Benoît Nabholz, Sylvain Glémin and Nicolas Galtier for providing the data, Nicolas Rodrigue, Fredrik Ronquist and two anonymous reviewers for their useful comments on the manuscript. We also thank the Réseau Québecois de Calcul de Haute Performance for computational resources. This work was funded by the Natural Science and Engineering Research Council of Canada.

Tables

Table 1. Rate of false positives[a].

	α				
averaging method	0.100	0.050	0.010	0.001	0.0001
arithmetic	0.050	0.022	0.002	0.001	0.000
geodesic	0.049	0.021	0.000	0.000	0.000

[a] frequency, over 100 simulations under the diagonal model, at which the posterior probability of a positive covariance is less than $\alpha/2$ or greater than $1-\alpha/2$ (see text for details.)

Table 2 : Covariance analysis for carnivores (left), and for therians (right) under the (λ_S, ω) parameterization[a].

	CARNIVORES					THERIANS				
COVARIANCE	λ_S	ω	mat.[b]	mass	long.[c]	λ_S	ω	mat.[b]	mass	long.[c]
λ_S	0.93	-0.25	-0.01	0.08	-0.06	0.59	-0.15	-0.03	**-0.30**	**-0.07**
ω	-	1.09	0.28	**0.90**	0.13	-	1.02	-0.03	**0.58**	**0.13**
maturity	-	-	0.98	**0.95**	**0.18**	-	-	0.81	**0.77**	**0.19**
mass	-	-	-	4.31	**0.38**	-	-	-	4.54	**0.61**
longevity	-	-	-	-	0.31	-	-	-	-	0.34
CORRELATION	λ_S	ω	mat.[b]	mass	long.[c]	λ_S	ω	mat.[b]	mass	long.[c]
λ_S	-	-0.24	-0.01	0.04	-0.11	-	-0.19	-0.04	**-0.18**	**-0.16**
ω	-	-	0.24	**0.41**	0.23	-	-	-0.03	**0.27**	**0.22**
maturity	-	-	-	**0.46**	**0.33**	-	-	-	**0.40**	**0.37**
mass	-	-	-	-	**0.33**	-	-	-	-	**0.49**
POSTERIOR PROB.[d]	λ_S	ω	mat.[b]	mass	long.[c]	λ_S	ω	mat.[b]	mass	long.[c]
λ_S	-	0.11	0.47	0.60	0.21	-	0.02	0.30	**< 0.01**	**0.01**
ω	-	-	0.93	**0.99**	0.94	-	-	0.35	**> 0.99**	**0.99**
maturity	-	-	-	**> 0.99**	**> 0.99**	-	-	-	**> 0.99**	**> 0.99**
mass	-	-	-	-	**> 0.99**	-	-	-	-	**> 0.99**

[a]covariances estimated using the geodesic averaging procedure, and $\kappa = 10$. Entries in bold correspond to a posterior probability of a positive covariance smaller than 0.025 or greater than 0.975. [b]maturity. [c]longevity. [d]posterior probability of a positive covariance.

Table 3 : Covariance analysis for therians, under the (λ_S, ω) parameterization and using fossil calibrations[a].

	THERIANS				
COVARIANCE	λ_S	ω	mat.[b]	mass	long.[c]
λ_S	0.77	**-0.21**	-0.04	**-0.40**	**-0.09**
ω	-	1.07	-0.04	**0.66**	**0.16**
maturity	-	-	0.99	**0.90**	**0.22**
mass	-	-	-	5.23	**0.69**
longevity	-	-	-	-	0.39
CORRELATION	λ_S	ω	mat.[b]	mass	long.[c]
λ_S	-	**-0.24**	-0.05	**-0.20**	**-0.16**
ω	-	-	-0.04	**0.28**	**0.25**
maturity	-	-	-	**0.40**	**0.36**
mass	-	-	-	-	**0.48**
POSTERIOR PROB.[d]	λ_S	ω	mat.[b]	mass	long.[c]
λ_S	-	**0.01**	0.27	**< 0.01**	**0.01**
ω	-	-	0.33	**> 0.99**	**0.99**
maturity	-	-	-	**> 0.99**	**> 0.99**
mass	-	-	-	-	**> 0.99**

[a]covariances estimated using the geodesic averaging procedure, and $\kappa = 10$. Entries in bold correspond to a posterior probability of a positive covariance smaller than 0.025 or greater than 0.975. [b]maturity. [c]longevity. [d]posterior probability of a positive covariance.

Table 4 : Covariance analysis for carnivores and therians under the (λ_S, λ_N) parameterization[a].

	CARNIVORES					THERIANS				
COVARIANCE	λ_S	λ_N	mat.[b]	mass	long.[c]	λ_S	λ_N	mat.[b]	mass	long.[c]
λ_S	1.04	0.29	-0.03	0.07	-0.07	0.62	**0.30**	-0.02	**-0.32**	**-0.08**
λ_N	-	1.13	0.26	**0.91**	0.08	-	1.18	-0.05	0.28	0.06
maturity	-	-	0.98	**0.94**	**0.18**	-	-	0.82	**0.78**	**0.20**
mass	-	-	-	4.31	**0.38**	-	-	-	4.56	**0.61**
longevity	-	-	-	-	0.31	-	-	-	-	0.34
CORRELATION	λ_S	λ_N	mat.[b]	mass	long.[c]	λ_S	λ_N	mat.[b]	mass	long.[c]
λ_S	-	0.27	-0.03	0.03	-0.13	-	0.35	-0.03	**-0.19**	**-0.17**
λ_N	-	-	0.25	**0.41**	0.13	-	-	-0.05	0.12	0.09
maturity	-	-	-	**0.46**	**0.33**	-	-	-	**0.40**	**0.37**
mass	-	-	-	-	**0.33**	-	-	-	-	**0.49**
POSTERIOR PROB.[d]	λ_S	λ_N	mat.[b]	mass	long.[c]	λ_S	λ_N	mat.[b]	mass	long.[c]
λ_S	-	0.92	0.44	0.58	0.17	-	**0.99**	0.34	**< 0.01**	**< 0.01**
λ_N	-	-	0.93	**0.99**	0.81	-	-	0.29	0.95	0.88
maturity	-	-	-	**> 0.99**	**0.99**	-	-	-	**> 0.99**	**> 0.99**
mass	-	-	-	-	**> 0.99**	-	-	-	-	**> 0.99**

[a]covariances estimated using the geodesic averaging procedure, and $\kappa = 10$. Entries in bold correspond to a posterior probability of a positive covariance smaller than 0.025 or greater than 0.975. [b]maturity. [c]longevity. [d]posterior probability of a positive covariance.

Table 5 : Multiple regression analysis in therians[a].

	CONSTANT MASS					CONSTANT LONGEVITY				
COVARIANCE	λ_S	ω	mat.[b]	mass	long.[c]	λ_S	ω	mat.[b]	mass	long.[c]
λ_S	0.74	-0.17	0.03	-	-0.04	0.75	**-0.18**	0.01	**-0.24**	-
ω	-	0.98	-0.15	-	0.07	-	0.99	-0.13	**0.37**	-
maturity	-	-	0.83	-	**0.10**	-	-	0.86	**0.51**	-
mass	-	-	-	-	-	-	-	-	4.00	-
longevity	-	-	-	-	0.30	-	-	-	-	-
POSTERIOR PROB.[d]	λ_S	ω	mat.[b]	mass	long.[c]	λ_S	ω	mat.[b]	mass	long.[c]
λ_S	-	0.04	0.67	-	0.16	-	**0.02**	0.56	**0.02**	-
ω	-	-	0.04	-	0.94	-	-	0.07	**0.99**	-
maturity	-	-	-	-	**> 0.99**	-	-	-	**> 0.99**	-

[a]covariances estimated using the geodesic averaging procedure, fossil calibrations, and $\kappa = 10$. Entries in bold correspond to a posterior probability of a positive covariance smaller than 0.025 or greater than 0.975. [b]maturity. [c]longevity. [d]posterior probability of a positive covariance.

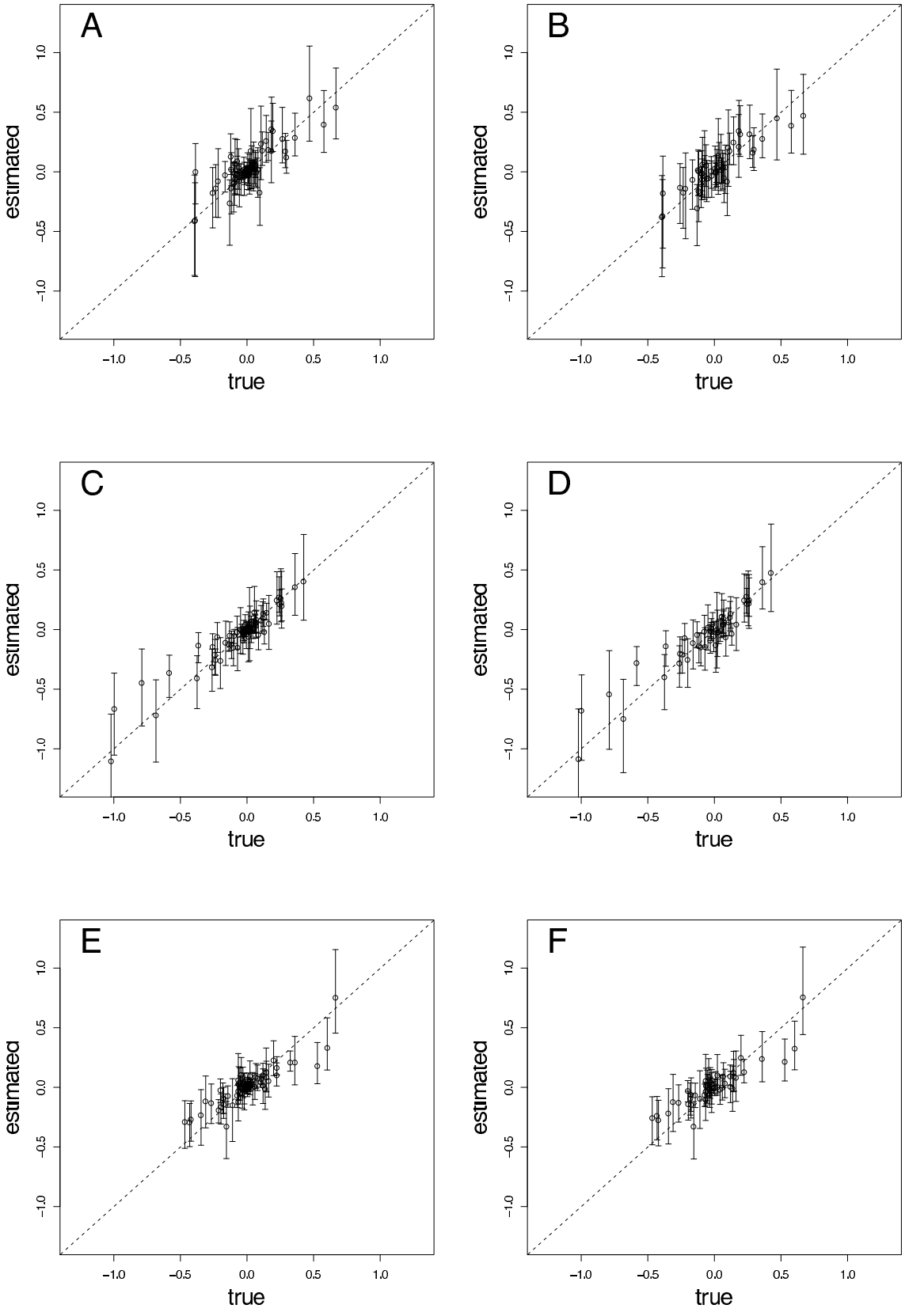
Figure legends

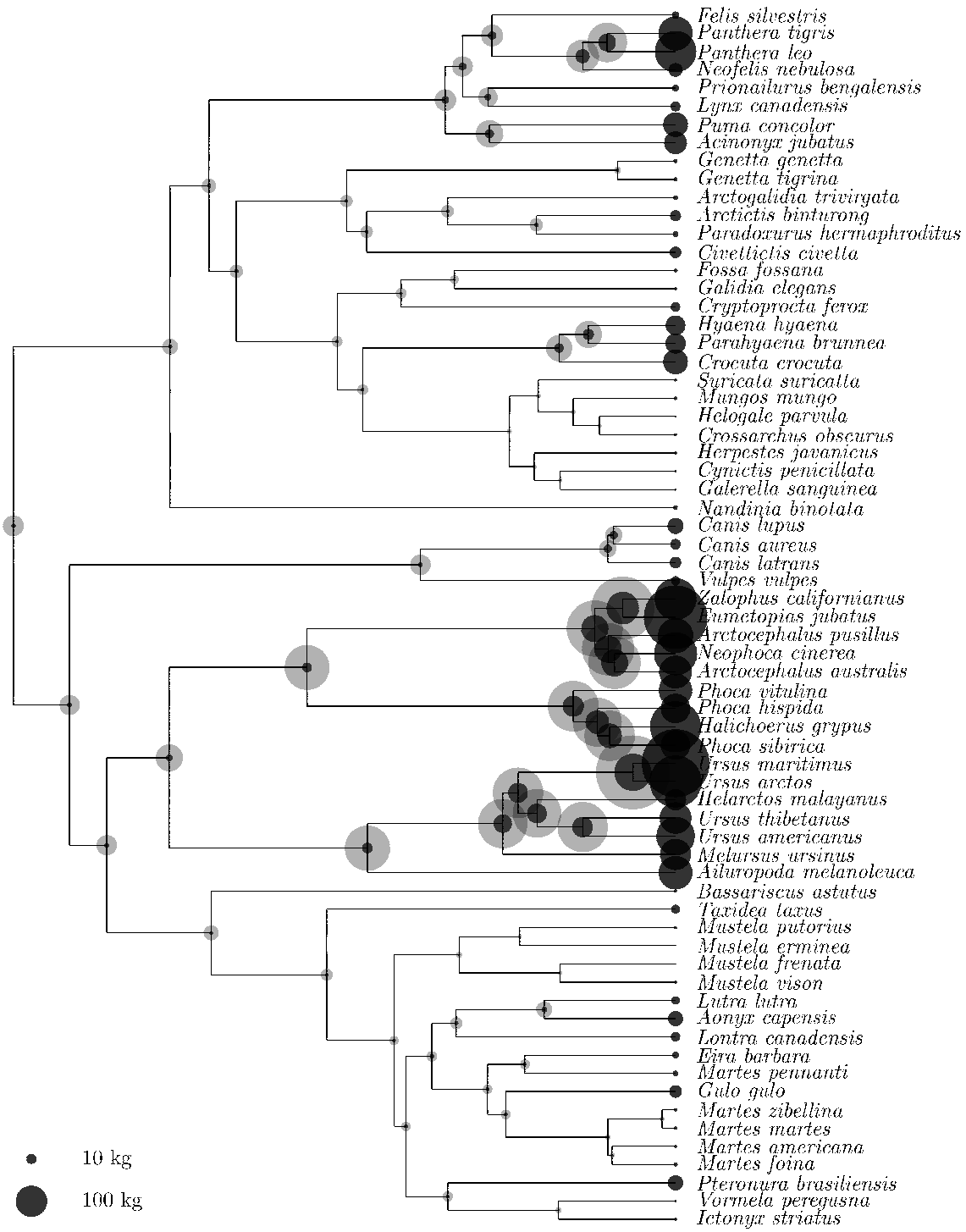
Figure 1 : Comparison between true value (x-axis), posterior mean and 95 % credibility interval (y-axis) for the 3 covariance parameters of the model (A,B : $\langle \lambda_S, \lambda_N \rangle$, C,D : $\langle \lambda_S, C_1 \rangle$, E,F : $\langle \lambda_N, C_1 \rangle$). A,C,E : arithmetic averages, B,D,F : geodesic averages. See text for details

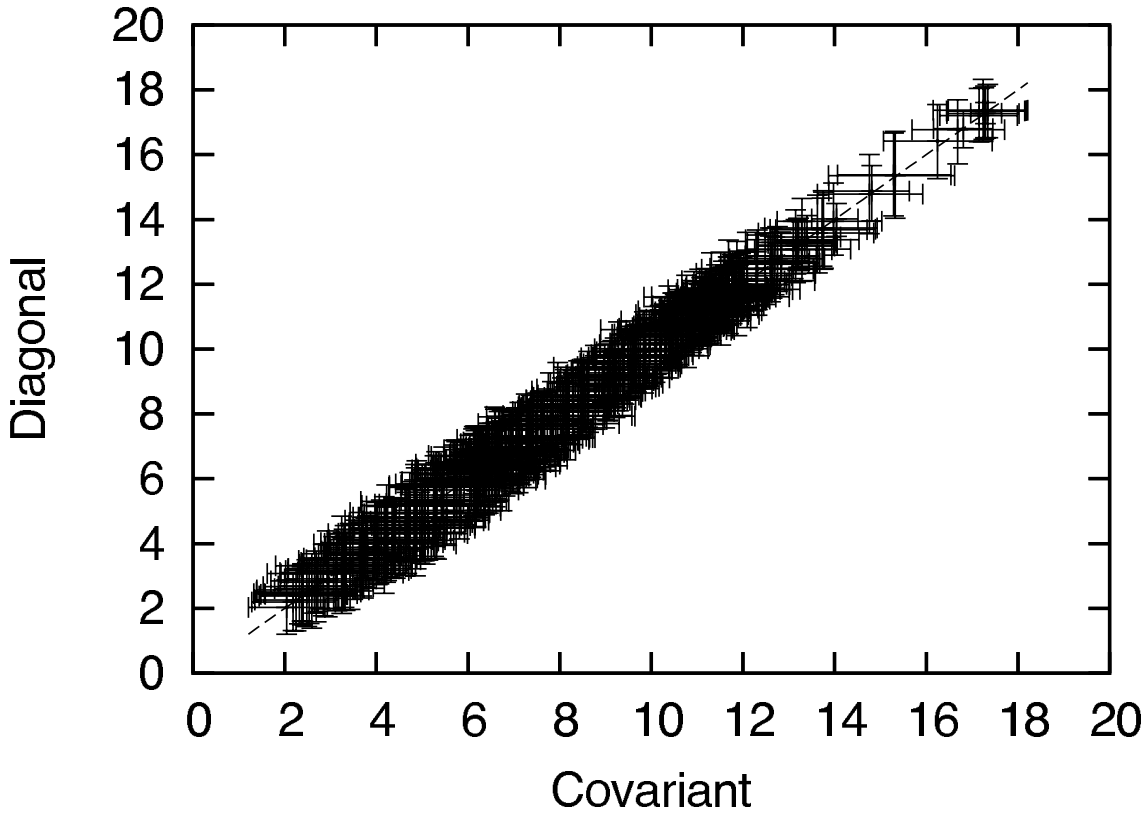
Figure 2 : Reconstruction of the evolution of body mass in carnivores. Disk area is proportional to body mass. Boundaries of the 95% credibility interval at each node are represented by the dark and light shaded disks.

Figure 3 : Comparison between inferred ancestral masses under the covariant (x-axis) and the diagonal (y-axis) model for the therian data set. Error bars correspond to the marginal 95% credibility intervals at each node.

Figures




Felis silvestris
Panthera tigris
Panthera leo
Neofelis nebulosa
Prionailurus bengalensis
Lynx canadensis
Puma concolor
Acinonyx jubatus
Genetta genetta
Genetta tigrina
Arctogalidia trivirgata
Arctictis binturong
Paradoxurus hermaphroditus
Civettictis civetta
Fossa fossana
Galidia elegans
Cryptoprocta ferox
Hyaena hyaena
Parahyaena brunnea
Crocuta crocuta
Suricata suricatta
Mungos mungo
Helogale parvula
Crossarchus obscurus
Herpestes javanicus
Cynictis penicillata
Galerella sanguinea
Nandinia binotata
Canis lupus
Canis aureus
Canis latrans
Vulpes vulpes
Zalophus californianus
Eumetopias jubatus
Arctocephalus pusillus
Neophoca cinerea
Arctocephalus australis
Phoca vitulina
Phoca hispida
Halichoerus grypus
Phoca sibirica
Ursus maritimus
Ursus arctos
Helarctos malayanus
Ursus thibetanus
Ursus americanus
Melursus ursinus
Ailuropoda melanoleuca
Bassariscus astutus
Taxidea taxus
Mustela putorius
Mustela erminea
Mustela frenata
Mustela vison
Lutra lutra
Aonyx capensis
Lontra canadensis
Eira barbara
Martes pennanti
Gulo gulo
Martes zibellina
Martes martes
Martes americana
Martes foina
Pteronura brasiliensis
Vormela peregusna
Ictonyx striatus
10 kg
100 kg


Diagonal
Covariant
0
2
4
6
8
10
12
14
16
18
20