Université de Montréal

Modélisation des biais mutationnels et rôle de la sélection sur l'usage des codons

par

Simon Laurin-Lemay

département de biochimie et de médecine moléculaire
Faculté de médecine

Thèse préesentée en vue de l'obtention du grade de Philosophiæ Doctor (Ph.D.)
en bio-informatique, orientation biologie moléculaire et évolution

5 octobre 2019

Université de Montréal

Département de biochimie et de médecine moléculaire, Faculté de médecine

Cette thèse intitulée

Modélisation des biais mutationnels et rôle de la sélection sur l'usage des codons

Présentée par

Simon Laurin-Lemay

A été évaluée par un jury composé des personnes suivantes

Sebastian Pechmann
Président-rapporteur

Hervé Philippe
Directeur de recherche

Nicolas Rodrigue
Codirecteur de recherche

Simon Joly
Membre du jury

Laurent Guéguen
Examinateur externe

Résumé

L'acquisition de données génomiques ne cesse de croître, ainsi que l'appétit pour les interpréter. Mais déterminer les processus qui ont façonné l'évolution des séquences codantes (et leur importance relative) est un défi scientifique passant par le développement de modèles statistiques de l'évolution prenant en compte de plus en plus d'hétérogénéités au niveau des processus mutationnels et de sélection.

Identifier la sélection est une tâche qui nécessite typiquement de détecter un écart entre deux modèles : un modèle nulle ne permettant pas de régime évolutif adaptatif et un modèle alternatif qui lui en permet. Lorsqu'un test entre ces deux modèles rejette le modèle nulle, on considère avoir détecter la présence d'évolution adaptative. La tâche est d'autant plus difficile que le signal est faible et confondu avec diverses hétérogénéités négligées par les modèles.

La détection de la sélection sur l'usage des codons spécifiquement est controversée, particulièrement chez les Vertébrés. Plusieurs raisons peuvent expliquer cette controverse : (1) il y a un biais sociologique à voir la sélection comme moteur principal de l'évolution, à un tel point que les hétérogénéités relatives aux processus de mutation sont historiquement négligées ; (2) selon les principes de la génétique des populations, la petite taille efficace des populations des Vertébrés limite le pouvoir de la sélection sur les mutations synonymes conférant elles-mêmes un avantage minime ; (3) par contre, la sélection sur l'usage des codons pourrait être très localisée le long des séquences codantes, à des sites précis, relevant de contraintes de sélection relatives à des motifs utilisés par la machinerie d'épissage, par exemple.

Les modèles phylogénétiques de type mutation-sélection sont les outils de prédilection pour aborder ces questions, puisqu'ils modélisent explicitement les processus mutationnels ainsi que les contraintes de sélection. Toutes les hétérogénéités négligées par les modèles

mutation-sélection de Yang and Nielsen [2008] peuvent engendrer de faux positifs allant de 20% (préférence site-spécifique en acides aminés) à 100% (hypermutabilité des transitions en contexte CpG) [Laurin-Lemay et al., 2018b]. En particulier, l'hypermutabilité des transitions du contexte CpG peut à elle seule expliquer la sélection détectée par Yang and Nielsen [2008] sur l'usage des codons.

Mais, modéliser des phénomènes qui prennent en compte des interdépendances dans les données (par exemple l'hypermutabilité du contexte CpG) augmente de beaucoup la complexité des fonctions de vraisemblance. D'autre part, aujourd'hui le niveau de sophistication des modèles fait en sorte que des vecteurs de paramètres de haute dimensionnalité sont nécessaires pour modéliser l'hétérogénéité des processus étudiés, dans notre cas de contraintes de sélection sur la protéine.

Le calcul bayésien approché (Approximate Bayesian Computation ou ABC) permet de contourner le calcul de la vraisemblance. Cette approche diffère de l'échantillonnage par Monte Carlo par chaîne de Markov (MCMC) communément utilisé pour faire l'approximation de la distribution *a posteriori*. Nous avons exploré l'idée de combiner ces approches pour une problématique spécifique impliquant des paramètres de haute dimensionnalité et de nouveaux paramètres prenant en compte des dépendances entre sites. Dans certaines conditions, lorsque les paramètres de haute dimensionnalité sont faiblement corrélés aux nouveaux paramètres d'intérêt, il est possible d'inférer ces mêmes paramètres de haute dimensionnalité avec la méthode MCMC, et puis les paramètres d'intérêt au moyen de l'ABC. Cette nouvelle approche se nomme CABC [Laurin-Lemay et al., 2018a], pour calcul bayésien approché conditionnel (Conditional Approximate Bayesian Computation : CABC).

Nous avons pu vérifier l'efficacité de la méthode CABC en étudiant un cas d'école, soit celui de l'hypermutabilité des transitions en contexte CpG chez les Eutheria [Laurin-Lemay et al., 2018a]. Nous trouvons que 100% des 137 gènes testés possèdent une hypermutabilité des transitions significative. Nous avons aussi montré que les modèles incorporant l'hypermutabilité des transitions en contexte CpG prédisent un usage des codons plus proche de celui des gènes étudiés. Ceci suggère qu'une partie importante de l'usage des codons peut être expliquée à elle seule par les processus mutationnels et non pas par la sélection.

Finalement nous explorons plusieurs pistes de recherche suivant nos développements méthodologiques : l'application de la détection de l'hypermutabilité des transitions en contexte CpG à l'échelle des Vertébrés ; l'expansion du modèle pour reconnaître des contextes autres que seul le CpG (e.g., hypermutabilité des transitions et transversions en contexte CpG et TpA) ; ainsi que des perspectives méthodologiques d'amélioration de la performance du CABC.

mots-clés : évolution moléculaire, phylogénie, calcul bayésien approché, modélisation phénoménologique et mechanistique, évolution des séquences codantes, usage des codons, évolution des Vertébrés

Abstract

The acquisition of genomic data continues to grow, as does the appetite to interpret them. But determining the processes that shaped the evolution of coding sequences (and their relative importance) is a scientific challenge that requires the development of statistical models of evolution that increasingly take into account heterogeneities in mutation and selection processes.

Identifying selection is a task that typically requires comparing two models: a null model that does not allow for an adaptive evolutionary regime and an alternative model that allows it. When a test between these two models rejects the null, we consider to have detected the presence of adaptive evolution. The task is all the more difficult as the signal is weak and confounded with various heterogeneities neglected by the models.

The detection of selection on codon usage is controversial, particularly in Vertebrates. There are several reasons for this controversy: (1) there is a sociological bias in seeing selection as the main driver of evolution, to such an extent that heterogeneities relating to mutation processes are historically neglected; (2) according to the principles of population genetics, the small effective size of vertebrate populations limits the power of selection over synonymous mutations conferring a minimal advantage; (3) On the other hand, selection on the use of codons could be very localized along the coding sequences, at specific sites, subject to selective constraints related to DNA patterns used by the splicing machinery, for example.

Phylogenetic mutation-selection models are the preferred tools to address these issues, as they explicitly model mutation processes and selective constraints. All the heterogeneities neglected by the mutation-selection models of Yang and Nielsen [2008] can generate false positives, ranging from 20% (site-specific amino acid preference) to 100% (hypermutability of transitions in CpG context)[Laurin-Lemay et al., 2018b]. In particular, the hypermutability

of transitions in the CpG context alone can explain the selection on codon usage detected by Yang and Nielsen [2008].

However, modelling phenomena that take into account data interdependencies (e.g., hypermutability of the CpG context) greatly increases the complexity of the likelihood function. On the other hand, today's sophisticated models require high-dimensional parameter vectors to model the heterogeneity of the processes studied, in our case selective constraints on the protein.

Approximate Bayesian Computation (ABC) is used to bypass the calculation of the likelihood function. This approach differs from the Markov Chain Monte Carlo (MCMC) sampling commonly used to approximate the posterior distribution. We explored the idea of combining these approaches for a specific problem involving high-dimensional parameters and new parameters taking into account dependencies between sites. Under certain conditions, when the high dimensionality parameters are weakly correlated to the new parameters of interest, it is possible to infer the high dimensionality parameters with the MCMC method, and then the parameters of interest using the ABC. This new approach is called Conditional Approximate Bayesian Computation (CABC) [Laurin-Lemay et al., 2018a].

We were able to verify the effectiveness of the CABC method in a case study, namely the hypermutability of transitions in the CpG context within Eutheria [Laurin-Lemay et al., 2018a]. We find that 100% of the 137 genes tested have significant hypermutability of transitions. We have also shown that models incorporating hypermutability of transitions in CpG contexts predict a codon usage closer to that of the genes studied. This suggests that a significant part of codon usage can be explained by mutational processes alone.

Finally, we explore several avenues of research emanating from our methodological developments: the application of hypermutability detection of transitions in CpG contexts to the Vertebrate scale; the expansion of the model to recognize contexts other than only CpG (e.g., hypermutability of transitions and transversions in CpG and TpA context); and methodological perspectives to improve the performance of the CABC approach.

keywords: molecular evolution, phylogeny, approximate bayesian computation, phenomenological and mechanistic modeling, coding sequence evolution, codon usage, vertebrates evolution

Table des matières

Liste des tableaux

Table des figures

Liste des abréviations

A : adénine

ADN : acide désoxyribonucléique

AIC : Akaike information criterion

ARN : acide ribonucléique

ARNt : ARN de transfert

BIC : bayes information criterion

C : cytosine

G : guanine

GTR : modèle general time reversible

HKY : modèle d'Hasegawa, Kishino et Yano (1985)

LBA : long branch attraction

LRT : likelihood ratio test

NER : nucleotide excision repair

OTU : operational taxonomic unit

POLB : DNA polymerase beta

R : purine

ROS : reactive oxygen species

SNVs : single nucleotide variants

T : thymine

Y : pyrimidine

Remerciements

J'aimerais commencer par ma famille, ma conjointe Géraldine qui m'a soutenu et encouragé toutes ces années. Qui a accepté mes absences prolongées à l'étranger et qui a si bien su s'occuper de nos deux garçons lorsque je travaillais en France ou que j'étais (bien souvent) en période de rush derrière mon ordinateur. Géraldine, ta confiance, ta douceur et ta patience à toute épreuve m'ont permis de travailler d'arrache-pied et de dormir sur mes deux oreilles. Merci. Merci aussi à mes deux petites merveilles : Léonard et Malcolm pour vos câlins, vos fous-rires, vos prouesses et votre force de caractère. Vous me ramenez à l'essentiel, la vie et l'amour. Merci à mes parents Ginette et François, leur conjoint Roger et Claire ainsi que mes beaux-parents Claude et Gérard d'être toujours à l'écoute, toujours prêt à aider. Je vous aime. Tous.

Le doctorat est une aventure en soi avec son lot de réussites, de défaites, de défis et autres challenges. J'aimerais prendre le temps de remercier les personnes qui ont étées présentes et qui m'ont permis de passer à travers cette épreuve. On dit parfois que ça prend un village pour élever un enfant, je crois qu'il est également possible dire que ça prend un village pour accompagner un étudiant au doctorat. Ce que Hervé Philippe a pris soin de faire en m'entourant des meilleurs, je parle de Nicolas Rodrigue, Nicolas Lartillot, Béatrice Roure, Ignacio Bravo ainsi que Henner Brinkmann. Merci à Hervé Philippe d'avoir eu confiance en moi, malgré le fait que je n'étais pas un finissant en bioinformatique. Sans cette confiance je ne serais pas en train de déposer ma thèse aujourd'hui. Ce travail, aurait été impossible sans la très grande implication de Nicolas Rodrigue. En effet, cette formation scientifique n'aurait pas été possible sans la générosité des chercheurs qui m'ont encadré, je parle du travail sans relâche de Hervé Philippe, puis de Nicolas Rodrigue, et Nicolas Lartillot. Un énorme merci à vous.

Un grand merci à Sébastien Lemieux qui, durant mon cheminement, a sû m'aider à mettre en perspective ce qui avait été accompli et ce qui devait être fait. J'aimerais aussi remercier Elaine Meunier pour m'avoir aidé à naviguer à travers la structure administrative et chacune des étapes qu'implique le doctorat. Merci à Franz Lang pour m'avoir accueilli dans son laboratoire. Merci aussi à Linda D'Astous pour le support dans cette grande aventure.

J'aimerais aussi remercier Matt Sarrasin, mon collègue depuis maintenant plusieurs années avec qui j'ai pu partager nombreuses discussions concernant la bioinformatique, mais aussi les moments de tous les jours. Merci également à tous les collègues du Centre Robert Cedergren de l'Université de Montréal.

Avant-propos

La motivation initiale était de s'intéresser, à comprendre pourquoi les virus de vertébrés possèdent un usage des codons différent de leurs hôtes et cela peu importe le type de virus. Que ce soit un virus à ADN, à ARN, simple brin, double brin, sens ou non-sens, petits ou grands, les virus qui infectent les vertébrés possèdent un usage des codons différent de celui de leur hôte, malgré le fait que les virus soient fortement dépendants de la machinerie de traduction de leur hôte pour produire leurs protéines. Cependant, cette observation va à l'encontre de l'hypothèse d'un usage des codons optimisé pour une traduction efficace [Ikemura, 1981, 1985, Akashi, 1995]. Pour illustrer le phénomène, lorsque les biochimistes veulent exprimer des gènes viraux, ils doivent optimiser l'usage des codons à celui de leur hôte (c.-à-d. de remplacer les codons synonymes rares par des codons fréquents de leur hôte) pour s'assurer d'obtenir une quantité suffisante de protéines virales afin de conduire leurs études (revue dans Quax et al. [2015]). Il est tentant de proposer que la différence d'usage des codons chez les virus et leur hôte soit une conséquence évolutive du mode de vie parasitaire des virus de mammifères (communications personnelles Hervé Philippe et Ignacio Bravo).

L'approche choisie pour développer la connaissance du système devait nous emmener à identifier les propriétés biophysiques et biochimiques de l'usage des codons qui expliqueraient la raison d'un usage des codons différents chez les virus de vertébrés et chez leurs hôtes et cela dans une perspective évolutive. L'avantage de la perspective évolutive est qu'elle permet de tirer des règles générales, fondées sur des principes de génétique des populations.

La détection de la sélection sur l'usage des codons est une tâche complexe, car le signal évolutif est potentiellement ténu, proportionnel à l'avantage sélectif qu'il confère, et donc facilement confondu avec le biais mutationnel. Mais la question est d'autant plus intéressante, sachant que l'usage des codons des virus joue un rôle dans l'immunité. Li et al. [2012] ont récemment découvert un nouveau mécanisme antiviral de l'immunité innée des mammifères

impliquant l'usage des codons des virus. Ils ont démontré que l'expression de la protéine *schlafen11* réduisait significativement l'abondance des isoaccepteurs fréquents pour le virus et, par conséquent, réduisait significativement aussi la quantité de protéines virales exprimées dans la cellule.

Chez les vertébrés, la détection de la sélection sur l'usage des codons est controversée, de par leur petite taille efficace des populations (effective population size : N_e), alors que chez les bactéries à croissance rapide, potentiellement avec une Ne très grande, la sélection est plus facilement détectable (e.g., [Sharp et al., 2010]). Yang and Nielsen [2008] ont malgré tout détecté de la sélection sur l'usage des codons en analysant plus de 5000 alignements de séquences codantes comprenant cinq espèces de mammifères placentaires au moyen d'un modèle phylogénétique à codon de type mutation-sélection. Ce type de modèle sert précisément à tester des hypothèses quant aux facteurs évolutifs affectant les séquences codantes (e.g., sélection traductionnelle). La première étape du projet de recherche sert à valider ce résultat surprenant. Puis, dans les étapes suivantes, afin de tester nos propres hypothèses, nous avons développé un cadre de travail probabiliste beaucoup plus flexible, ce qui nous emmena vers l'utilisation du calcul bayésien approché (Approximate Bayesian Computation : ABC).

Chapitre 1

Introduction

1.1. Contexte biologique

1.1.1. Notions générales d'évolution

L'étude de l'usage des codons et de leur évolution revêt un caractère tout autant appliqué que fondamental. Du point de vue fondamental, l'usage des codons intervient dans l'étude d'une étape cruciale du système d'information faisant le pont entre **génotype** et **phénotype** [Sapp, 1983]. Nous entendons par génotype l'information qui est transmissible d'une génération à la suivante (incluant les modifications **épigénétiques**) et par phénotype, l'expression du génotype sous la forme d'un individu, ou d'une dimension de celui-ci (e.g., protéine) pour un environnement donné. La plasticité phénotypique permet à un même génotype d'exprimer différents phénotypes selon l'environnement (e.g., le puceron du poids *Acyrthosiphon pisum* (Harris) maintient un polymorphisme sous la pression de sélection de deux prédateurs : une coccinelle *Coccinella septempunctata* et une guêpe parasitoïde *Aphidius ervi* [Losey et al., 1997]). Nous entendons par modifications épigénétiques (e.g., méthylation de l'ADN, modification des histones) une information supplémentaire à celle contenue dans la séquence des bases de l'ADN. Ces modifications sont aussi héritables, en partie du moins, et peuvent permettre de moduler le phénotype. Contrairement au génome, l'**épigénome**, soit l'ensemble des modifications épigénétiques présentes dans le génome, est appelé à varier à de plus petites échelles évolutives, soit de quelques générations. Par exemple, l'hypermutabilité due à la désamination spontanée des cytosines méthylées (noté ici par m^5C) diminue grandement le caractère héritable de cette modification épigénétique [Bird, 1980].

En dehors de systèmes d'étude très bien circonscrits (e.g., séquences codantes et protéines) il est très difficile d'étudier la relation entre génotype et phénotype, d'une part parce que le nombre de phénotypes est très grand ainsi que la variété d'environnements dans lesquels s'exprime le génotype, mais aussi par la très grande stochasticité de tous les processus biologiques impliqués. La relation qui existe entre génotype et phénotype est donc extrêmement complexe et difficile à modéliser, d'autant plus que l'épigénome joue un rôle important dans la modulation du phénotype, par exemple via la modulation de l'expression des gènes (revue dans [Gibney and Nolan, 2010]). En fait, nous allons nous focaliser sur le phénotype au niveau des protéines.

Les **mutations** sont tout changement modifiant les bases de la séquence nucléotidique des génomes, en contraste avec les **épimutations**, comme la méthylation, qui ne changent pas la base. Les mutations possèdent de multiples origines, **endogènes** (e.g., fidélité limitée de la réplication) ou **exogènes** (e.g., exposition à la fumée de cigarette). Elles sont le résultat d'**interactions abiotiques** (e.g., exposition au radon) et **biotiques** (e.g., exposition à des aflatoxines), revue dans Chatterjee and Walker [2017]. L'environnement est donc un facteur qui peut affecter la quantité et la nature des mutations qu'un système biologique peut incorporer. Par contre, la source de mutation principale reste endogène. Le concept de mutation apparaît avant la découverte de la structure en double hélice d'ADN [Watson and Crick, 1953], vers le début du 20e siècle et a permis de mettre en place le concept de gène, lui aussi avant la découverte de la double hélice d'ADN (revue dans [Auerbach, 1976]).

Le **taux de mutation** [Muller, 1928], μ, est le taux de mutation par base par génération ou par unité de temps selon l'origine des mutations. Dans le cas où les mutations sont dépendantes de la réplication, le taux de mutation sera en nombre de générations, alors que si les mutations sont indépendantes de la réplication, le taux de mutation sera fonction du temps [Gangloff et al., 2017]. Le **taux d'évolution neutre** d'une population diploïde est donné par $2 \times N_e \times \mu$. Cette caractéristique est particulièrement importante pour la datation qui s'appuie sur des modèles **d'horloge moléculaire** (Zuckerkandl, E and Pauling, L 1962 : revue dans [dos Reis et al., 2016]). Les chercheurs qui s'intéressent à dater des évènements de l'histoire évolutive préfèrent travailler avec des mutations dites indépendantes du temps de génération pour ne pas avoir à modéliser la manière dont varie le temps de génération au cours de l'évolution [dos Reis et al., 2016].

Les mutations peuvent affecter des **cellules germinales** ou **somatiques**, avec des conséquences différentes. Il est alors question de **mutations germinales** ou de **mutations somatiques**. Les **mutations somatiques** auront un effet instantané sur le phénotype, s'il y a lieu, mais ne seront pas transmises à la génération suivante. Les **mutations germinales**, elles, n'auront un effet qu'à la génération suivante, sauf dans les rares cas où elles affectent la valeur adaptative des gamètes. Les deux types de mutations ont un impact sur l'évolution des systèmes biologiques.

Le **biais mutationnel** est l'appellation donnée au **profil mutationnel** résultant de la somme de tous les processus mutationnels (e.g., désamination des m^5C), incluant les mécanismes de réparation de l'ADN, comme la recombinaison homologue. Donc, différentes combinaisons de processus mutationnels mèneront vers différents profils mutationnels.

Tout au long de la thèse, nous utilisons le concept de **système biologique** pour faire référence à un système :

— où le phénotype est une interprétation de l'information contenue dans le génotype, plus ou moins fidèle, pour un environnement donné,

— où le génotype des individus d'une population change d'une génération à l'autre par l'acquisition de mutations ou de modifications épigénétiques,

— où la participation des individus de la population à la génération suivante se fait avec un certain biais, plus ou moins prononcé, faisant référence au processus de sélection,

— et où la démographie est un facteur déterminant de l'efficacité de la sélection et de l'introduction/rétention de la diversité génétique.

La **théorie quasi-neutre de l'évolution** [Ohta, 1973] nous semble être le modèle le plus utile pour expliquer l'évolution de la majorité des séquences codantes de protéines, ce à quoi nous nous intéressons dans le travail ici présenté. Cette théorie de l'évolution prédit que la plupart des mutations qui affectent les séquences codantes sont délétères et purgées des populations via la **sélection négative ou purificatrice**, alors que les mutations qui confèrent un avantage sélectif sont rares, et généralement fixées via la **sélection positive ou darwinienne**. Cependant, l'efficacité de la sélection est dépendante de la N_e, soit la **taille efficace des populations** ou le nombre réel d'individus participants à la génération suivante. À N_e élevée, la sélection agit de facon importante sur le processus d'évolution. À

faible N_e, par contre, le rôle de la sélection est grandement affaiblit, et le processus d'évolution se dit en **dérive génétique**.

La probabilité de fixation d'une mutation dépend de la valeur adaptative relative des individus qui forment la population et de la N_e. Dans le cas le plus simple, où il a que deux génotypes ou phénotypes dans la population, à partir de la valeur adaptative relative, w_b, d'une nouvelle mutation non-synonyme, il est possible de calculer un coefficient de sélection, s, sachant la valeur adaptative relative des autres individus de la population, w_a. Ainsi, s est obtenu en calculant le rapport de w_b sur w_a. Si la valeur adaptative relative est plus grande pour la nouvelle mutation non-synonyme, w_b, alors $s > 1$, nous sommes en présence de sélection positive. Inversement, si la valeur adaptative relative $w_b << w_a$, nous sommes en présence de sélection négative avec $s << 1$. Ce modèle peut être complexifié, et permettre de tester plusieurs génotypes différents. Dans le contexte de génétique des populations la probabilité de fixation d'une mutation neutre ($s = 1$) est approximée par $1/2N_e$ si le système biologique est diploïde. De la même manière, la probabilité de fixation d'une mutation non-neutre est approximé par $2s/1-e^{4N_e s}$ [Fisher, 2000, Wright, 1931]. Par la suite, il est possible de calculer le rapport des probabilités de fixation des mutations non-neutres aux mutations neutres donné par le facteur de fixation :

$$\frac{S}{1 - e^{-S}}, \tag{1.1.1}$$

où $S = 4N_e s$.

Ainsi, lorsque la mutation proposée est faiblement délétère, par exemple $S = 10^{-5}$, et que la N_e est grande, $N_e = 10^6$, la sélection négative l'emporte sur la dérive génétique ($4.5 \times 10^{-10} << 10^{-6}$) et empêche la fixation de la mutation. Alors que si la N_e est petite, $N_e = 10^3$, l'effet de la sélection ne sera pas différent de l'effet de la dérive génétique ($1 \times 10^{-3} \sim 1/2N_e$), la mutation faiblement délétère pourra se fixer dans la population.

Les populations sont la structure dans laquelle est organisée la diversité génétique. Des facteurs environnementaux abiotiques (e.g., climat : Myers [2018]) et biotiques (e.g., présence de pathogènes : Myers [2018]) ont un impact sur la démographie des populations (N_e). À certains moments de l'histoire des populations, celles-ci prennent de l'expansion, et accumulent des mutations car les desce

ants ont simplement plus de chances de survivre [Peischl and Gilbert, 2018]. Par exemple, actuellement chez *Homo sapiens*, il y a une augmentation de la diversité génétique due à

l'**expansion démographique** que vit l'espèce. Alors que suite à des goulots d'étranglement, il y a érosion de la diversité génétique (e.g., sortie de l'Afrique par *Homo sapiens* : e.g., [Amos and Hoffman, 2010]). D'autres facteurs peuvent affecter la démographie et la diversité génétique, comme le niveau trophique du système biologique étudié (e.g., Nair et al. [2016]). Par contre, le fardeau génétique tel qu'introduit par Muller [1950] serait en croissance chez *Homo sapiens*, et cela surtout dans les pays industrialisés expliquant la croissance de l'incidence des cancers [You and Henneberg, 2018]. Historiquement, l'emphase a été mise sur l'identification des facteurs environnementaux (e.g., l'utilisation de la cigarette) pour expliquer l'incidence des cancers [You and Henneberg, 2018]. par contre, en réduisant le taux de mortalité et en augmentant le taux de reproduction par la fécondation assistée [Saniotis and Henneberg, 2011], la médecine autoriserait la conservation de mutations délétères au sein de la population [You and Henneberg, 2018]. À travers les lunettes de l'évolution, ce type de conflit n'est pas suprenant : ici il y a un conflit entre la valeur adaptative de l'individu à soigner et celle de la population (le concept de conflit en évolution est revue dans [Queller and Strassmann, 2018]). Quel est donc l'apport de la médecine à la diversité génétique de *Homo sapiens* ?

La diversité génétique observée est le résultat d'un processus d'évolution en deux temps, où des mutations sont tout d'abord générées créant une population de génotypes [Haldane, 1927]. Puis, dans un second temps, certains génotypes se reproduiront avec un plus grand succès ; la réalité est un peu moins propre que le voudrait ce modèle (e.g., les générations se chevauchent). Les contraintes de sélection intrinsèques (contraintes moléculaires) et extrinsèques (contraintes écologiques) sont dictées par le fonctionnement des systèmes biologiques et du réseau d'interactions qu'ils entretiennent (revue dans [Fragata et al., 2019]). La relation entre génotype et phénotype est d'autant plus difficile à élucider lorsqu'elle doit intégrer des niveaux de complexité grandissante qui vont de la molécule (e.g., protéines) à la population/communauté, voir l'écosystème.

Les mutations sont le moyen par lequel le paysage adaptatif peut être exploré. Le paysage adaptatif ([Wright, 1932]) est un modèle qui permet de conceptualiser la relation complexe qui existe entre le génotype (ou phénotype) d'un individu ou d'une population (figure 1.1 : axe des x et des y) et sa valeur adaptative (figure 1.1 : axe des z) à un instant donné de l'évolution. Il existe plusieurs modèles de paysage adaptatif, et lorsque conditionnés aux données, ils peuvent, avec différentes capacités, permettre de prédire la valeur adaptative

d'un génotype/phénotype (revue dans [Fragata et al., 2019]). C'est actuellement un champ de recherche très important. Par exemple, des chercheurs se sont récemment intéressés à caractériser la valeur adaptative des variants adjacents par une simple mutation pour un domaine fonctionnel du gène *BRCA1* chez l'humain. Le paysage adaptatif ainsi dépeint avec CRISPR/Cas9 confirme la pathogénécité observée en clinique de certains génotypes [Findlay et al., 2018].

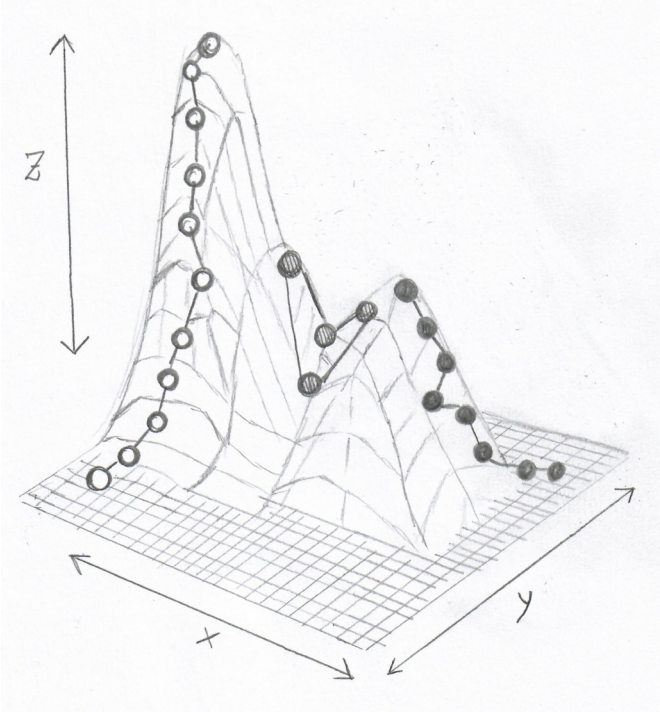

Figure 1.1. Paysage adaptatif vallonné (rugged) où l'axe des x et y correspondent à deux dimensions du phénotypes (traits) et l'axe z à leur valeurs adaptatives (adapté à partir de [Payne and Wagner, 2019]). Crédits : Géraldine Philippin

La capacité d'explorer le paysage adaptatif, ou l'**évolvabilité** au sens de Wagner and Altenberg [1996], est donc une caractéristique fondamentale des systèmes biologiques. L'évolvabilité est donc l'aptitude d'un système biologique à générer des variants qui pourront être sélectionnés (revue dans [Fragata et al., 2019]). En parallèle, la **robustesse** des systèmes biologiques leur permet de faire face à la nécessité de vivre dans des environnements changeants, la robustesse aux mutations étant un cas particulier. L'architecture des génomes est un moyen par lequel les systèmes biologiques peuvent résister aux mutations. Par exemple, la ploïdie et les duplications de gènes permettraient de réduire l'effet des mutations délétères

en assurant la production de protéines fonctionnelles. Certains avancent que la duplication de gènes pourrait plutôt fragiliser le système puisque les gènes dupliqués deviennent mutuellement dépendants, dans la régulation de leur abondance respective par exemple [Diss et al., 2017].

La relation entre génotype et phénotype est d'autant plus complexe qu'un trait phénotypique implique habituellement plusieurs *loci*, il est alors question d'**épistasie** [Bateson and Mendel, 1909.]. De plus, plusieurs génotypes différents peuvent générer le même phénotype par des systèmes de compensations. La capacité d'exploration du paysage adaptatif peut permettre aux systèmes biologiques de survivre dans un environnement où les contraintes écologiques changent (e.g., changement climatique et émergence de maladies infectieuses).

En d'autres mots, les systèmes biologiques possèdent la capacité d'explorer le paysage adaptatif et d'atteindre des valeurs adaptatives de hauts niveaux, sans pour autant nécessairement trouver la meilleure valeur adaptative. L'enchaînement des évènements évolutifs pour atteindre ce sommet correspondant à la meilleure valeur adaptative, peuvent être tout simplement trop improbables. Il est intéressant de noter que les plus hauts niveaux du paysage adaptatif peuvent être atteints sans l'effet de la sélection positive, donc via les processus mutationnels [Stoltzfus and McCandlish, 2017, Sackman et al., 2017]. Au cours du temps, l'environnement abiotique et biotique change, ce qui a pour conséquence de modifier le paysage adaptatif et ainsi la valeur adaptative d'un génotype/phénotype. Il est évidemment nécessaire d'ajouter un axe au paysage adaptatif pour apprécier la relation entre génotype/phénotype et valeur adaptative dans un environnement où les interactions changent (figure : 1.2). Cela fait référence au concept de mer adaptative (seascape : Merrell [1994], Mustonen and Lassig [2009]).

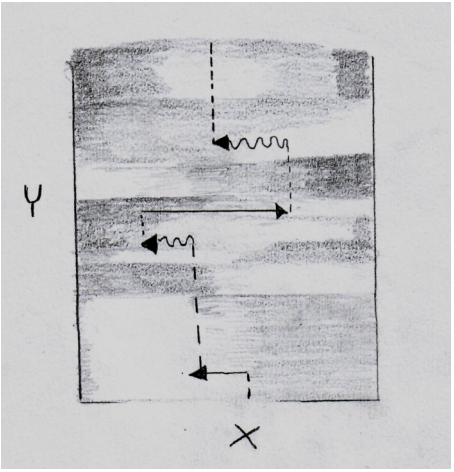

Figure 1.2. Mer adaptative : l'axe des x représente la valeur adaptative d'un génotype/phénotype, alors que l'axe des y permet d'ajouter une variation dans le temps ou dans l'espace (modifié à partir de [Mustonen and Lassig, 2009]). Le gradient gris foncé à clair représente un gain de valeur adaptative. Les flèches droites indiquent un gain de valeur adaptative alors que les flèches en serpent indiquent une perte de valeur adaptative après un déplacement sur l'axe des y. Crédits : Géraldine Philippin

Les systèmes biologiques sont composés de modules fonctionnels. C'est l'assemblage et le maintient de ces modules qui définit les systèmes biologiques. Le concept de **modularité** est donc très important à la compréhension de l'évolution. Les modules sont quant à eux des abstractions identifiées *a posteriori* qui nous permettent d'étudier et comprendre un système biologique. Ce sont ces modules qui sont mutés et dupliqués, voir perdus au cours de l'évolution. Les gènes sont certainement des modules de l'évolution. En fait, est-ce que tous les traits qui forment un phénotype ne sont pas eux-même des modules de l'évolution ? Le processus mutationnel par exemple, c'est-à-dire le module du processus mutationnel, par lequel le paysage adaptatif est exploré, est lui-même un trait phénotypique qui évolue (e.g., [Lynch, 2010b]). L'évolvabilité, la robustesse et la modularité sont donc eux-mêmes des traits qui font la spécificité des systèmes biologiques et sont donc le résultat de l'évolution.

Les trais phénotypiques appartiennent à des **niveaux de complexité** différents (e.g., molécule, chromosome, cellule, tissus, etc). Le nombre de modules qui participent à un assemblage de modules donne une idée de la complexité de cet assemblage. Les propriétés des assemblages ne sont pas la somme des propriétés des modules qui les composent. Des

nouvelles structures émergent de nouvelles fonctions. Le concept de niveau de complexité, ici fait référence au fait qu'il est de plus en plus difficile de prédire le phénotype à partir du génotype plus le phénotype étudié est loin de l'information génétique et que de nombreuses couches de stochasticité sont impliquées.

Nous avons brièvement survolé toute la complexité de la relation génotype/phénotype et de leur évolution, en particulier pour les systèmes biologiques multicellulaires. Dans le cadre de ce travail, nous abordons cette thématique en nous focalisant sur une petite partie des systèmes biologiques, les protéines. Une telle approche volontairement réductionniste nous permet en particulier d'étudier de nombreuses données expérimentales de qualité et de pouvoir utiliser des modèles mécanistiques de l'évolution des protéines qui incorporent les aspects de mutation et de sélection.

1.1.2. Le système stochastique de l'expression des protéines

Plusieurs étapes sont nécessaires à la production des protéines. Chacune de ces étapes nécessite l'intervention de mécanismes moléculaires spécifiques, voir la figure 1.3. La dynamique moléculaire de l'expression des gènes est un processus hautement stochastique, dont l'aspect le plus étudié est le niveau de la transcription des gènes. Ainsi, toutes les étapes jusqu'au produit final introduisent des variations (revue dans [Drummond and Wilke, 2009]) : erreurs de l'ARN polymérase (au point que seul un transcrit incorrect soit à même de produire la protéine prédite à partir de la séquence codante [Duffy et al., 2008]), erreurs dans la maturation de l'ARNm (en particulier au niveau de l'épissage), erreurs lors de l'initiation de la traduction, erreurs lors de l'élongation de la chaîne polypeptidique, erreurs lors de la terminaison de la traduction, sans parler des multiples modifications posttraductionnelles possibles [Drummond and Wilke, 2009]. Une cellule produit donc une immense diversité de protéines fonctionnelles et non-fonctionnelles ; il est estimé qu'environ 30% des protéines sont dégradées juste après leur production chez *Homo sapiens* [Schubert et al., 2000].

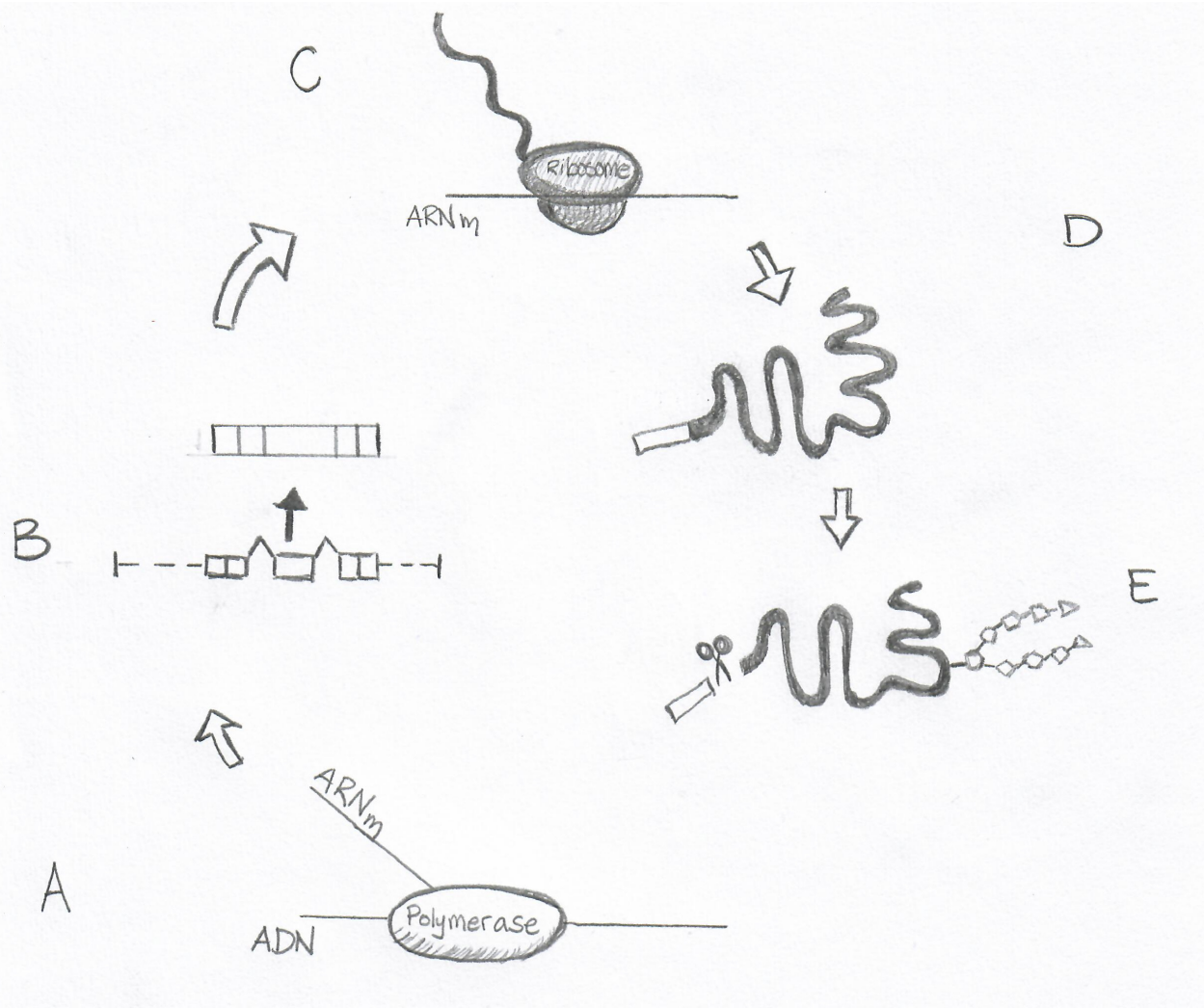

Figure 1.3. Erreur au niveau de l'expression des gènes d'eukaryotes (adapté à partir de Drummond and Wilke [2009]). L'erreur peut être au niveau (A) de la transcription, (B) de l'épissage, (C) de la traduction, (D) du repliement ou encore (E) au niveau des modifications posttraductionnelles. Crédits : Géraldine Philippin

La **transcription** de l'ADN en ARN est conduite par une ARN polymérase II chez les eucaryotes de l'extrémité 5' à l'extrémité 3' (la lecture du brin d'ADN antisens se fait dans la direction 3' à 5' chez les eucaryotes), avec une erreur à toutes les 10^4 bases (revue dans Alberts et al. [2002]). L'initiation nécessite la reconnaissance de la boîte TATA par un facteur de transcription, *TFIID*, qui permettra la mise en place du complexe d'initiation de la transcription. Mais en amont de l'initiation de la transcription, l'ADN empaqueté dans les nucléosomes organisés en chromatine (ADN, histones et autres protéines) doit être rendu

accessible ; des protéines activatrices et médiatrices se chargent de cette tâche (revue dans Alberts et al. [2002]).

Une fois l'ARN transcrit, l'**épissage** est le processus par lequel les **introns** (régions non-codantes intra-géniques), chez les eucaryotes, sont excisés. Les **exons**, les différentes parties codantes des gènes (régions codantes intra-géniques), sont alors joints pour former des $ARNm$ qui serviront de matrice à la production des protéines de la cellule. Parmi les signaux disponibles chez les mammifères pour réguler l'épissage il y a les **activateurs d'épissage exoniques** (exonic splice enhancers) et les **inhibiteurs d'épissage exoniques** (exonic splice silencers). Cette modularité des gènes permet la génération d'une grande variété de protéines appelées **isoformes**. La transcription, l'épissage et la coiffe de l'ARN, se font simultanément. L'ARN est coiffé d'une guanine tri-phosphate à l'extrémité 5' et est polyadenylé à extrémité 3'. La présence de ces modifications assure l'intégrité du messager avant l'initiation de la traduction via des mécanismes de contrôle de qualité (e.g., dégradation des ARN non-coiffés par les exonucleases), dans le cas contraire l'$ARNm$ est appelé à être dégradé (revue dans Alberts et al. [2002]).

Avant de parler de l'étape de l'élongation impliquant le décodage de l'information contenue dans l'$ARNm$ par le complexe ribosomique, il est nécessaire de présenter quelques propriétés fondamentales du code génétique. La traduction s'initie sur le codon de départ (AUG : méthionine) au site A (site de l'**a**minoacylation) par le chargement du premier $ARNt$. Puis les acides aminés chargés sur les $ARNt$ sont transférés du site P (site **p**eptidile) au site A, ce qui fait l'élongation de la chaîne polypeptidique. Puis l'$ARNt$ libre d'acide aminé quitte le site P pour le site E, la voix de sortie (**e**xit) du ribosome. L'élongation continue jusqu'à ce que le signal de terminaison soit rencontré sur l'$ARNm$: soit les codons de terminaison UGA, UAG et UAA (code génétique universel).

Plusieurs mécanismes moléculaires du contrôle de la qualité permettent d'éviter les erreurs d'incorporation pendant l'étape de l'élongation. Parmi les causes possibles d'erreur il y a : (1) la présence fortuite d'un codon de terminaison trop tôt dans le messager (Nonsense-Mediated Decay : revue dans [Drummond and Wilke, 2009]), (2) une élongation trop lente (No-Go-Decay :[Harigaya and Parker, 2010]), ou encore (3) l'absence du codon de terminaison (Non-Stop-Decay : revue dans [Drummond and Wilke, 2009]).

Une fois la traduction initiée, suivant le codon situé au site A, des $ARNt$ tenteront d'occuper ce site. Mais, idéalement, l'affinité des $ARNt$ complémentaires étant suffisamment grande, seul un $ARNt$ complémentaire restera assez longtemps pour permettre la formation de la liaison peptidique, étape suivie par la translocation du codon vers le site P. Il y a donc une part de stochasticité dans le processus d'appariement des $ARNt$, un codon tous les $10^3 - 10^4$ codons est mal traduit lors de la seule étape de l'élongation (revue dans Drummond and Wilke [2009]). Plus la concentration des $ARNt$ complémentaires est faible, plus la probabilité qu'un $ARNt$ plus ou moins proche mais non-complémentaire soit impliqué dans la liaison peptidique augmente. Il est même possible que le ribosome se déplace légèrement par rapport à l'$ARNm$, entraînant soit un changement de cadre de lecture dans la protéine finale, soit un détachement du ribosome, produisant une protéine tronquée, revue dans Drummond and Wilke [2009]. Si la concentration en $ARNt$ complémentaire est grande, un tel codon occupera rapidement le site A, ce qui permettra une traduction fidèle et rapide. À l'inverse, si elle est faible, la fidélité sera basse et le taux d'élongation faible, justifiant l'avantage sélectif des mutations vers l'usage des codons fréquents des protéines fortement exprimées au sens de [Ikemura, 1981, 1985, Akashi, 1995].

Les systèmes biologiques eucaryotes utilisent généralement moins de 50 **isoaccepteurs** différents [Goodenbour and Pan, 2006]. Un isoaccepteur correspond à chacun des anticodons différents présents parmi les gènes d'ARNt pouvant charger le même acide aminé. Alors comment l'élongation peut-elle se faire lorsqu'un codon n'a pas son anticodon spécifique ? Grâce au mécanisme du **wobble** [Crick, 1966], des liaisons non-Watson-Crick entre troisième position du codon et première position de l'anticodon sont possibles, augmentant le nombre de codons reconnus par un anticodon.

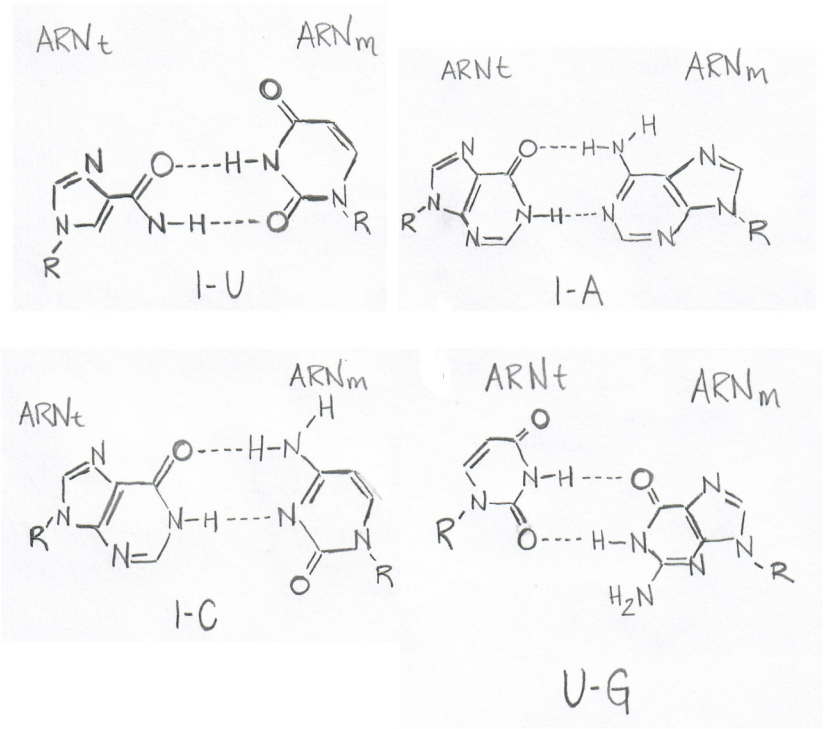

Figure 1.4. Appariement non-Watson-Crick entre deux paires de base appartenant à l'ARNt et l'ARNm. Les appariements possibles sont l'hypoxanthine-uracil (I-U), l'hypoxanthine-adenine (I-A), l'hypoxanthine-cytosine (I-C) et uracil-guanine (U-G) ou guanine-uracil (G-U). Crédits : Géraldine Philippin

Le nombre d'**isodécodeurs** varie lui aussi, c'est-à-dire que pour un même anticodon la séquence de l'*ARNt* (le corps) peut être différente. Pour obtenir le nombre d'isodecodeurs [Goodenbour and Pan, 2006], il faut compter le nombre de gènes différents ayant le même anticodon. Chez *Homo sapiens* par exemple, il y a un seul isoaccepteur pour certains acides

aminés qui possèdent deux codons : soit la phénylanaline, la tyrosine, l'histidine, l'asparagine, l'acide aspartique et la cystéine [Pan, 2018]. Pour les autres acides aminés, ceux-ci possèdent plusieurs isoaccepteurs : deux pour l'acide glutamique, la lysine et la glutamine, trois pour l'isoleucine, la valine, la thréonine, l'alanine, la glycine et la proline, et finalement quatre pour la sérine et cinq pour la leucine et l'arginine [Pan, 2018]. Il est intéressant de noter que le nombre de gènes d'$ARNt$ par isoaccepteur et le nombre d'isodécodeurs chez les mammifères sont très semblables, mais pas chez *Sacaromyces cerevisiae*, *Caenorhabditis elegans* et *Drosophila melanogaster* des espèces à plus grande N_e (Figure 2B dans Goodenbour and Pan [2006]). Est-ce que cette observation peut suggérer la présence d'une pression de sélection négative plus importante sur le corps des $ARNt$ des espèces à grande N_e, permettant de conserver un rapport isodecodeurs / nombre de gènes par isoaccepteur petit ? Est-ce que la sélection sur les $ARNt$ ne serait pas à l'origine de ces différences entre espèces à grande N_e et à petite N_e ? Une particularité liée à la diversité des isodécodeurs relève de la spécificité de leur expression dans différents tissus. Par exemple, l'expression de l'isodécodeur $ARNt^{arginine}(TCT)$ sans intron est importante dans les cellules du système nerveux central [Parisien et al., 2013]. L'expression de l'isodécodeur $ARNt^{arginine}(TCT)$ avec intron dans le système nerveux central de *Mus musculus* affecte la dynamique de la traduction empêchant le repliement adéquat des protéines [Parisien et al., 2013].

Plusieurs modifications posttranscriptionnelles sont possibles au corps des $ARNt$, mais aussi à l'anticodon, et plus particulièrement à la position wobble de l'anticodon [Parisien et al., 2013]. Ces modifications joueraient un rôle fonctionnel dans la réponse au changement d'environnement, la réponse au stress notamment, en modulant plusieurs aspects de l'expression des protéines, comme la dynamique de la traduction, la stabilité de l'$ARNt$, la localisation, l'affinité d'appariement au ribosome, les capacités de décodage [Parisien et al., 2013]. Parmi les modifications les plus importantes, il y a la modification de l'adénosine en position wobble à l'inosine, ce qui étend la variété de codons pouvant être reconnu (de $A : U$ à $I : U$, $I : C$ et $I : A$). Tel le phénomène wobble, les modifications posttrancriptionnelles augmentent la robustesse face aux mutations en permettant la reconnaissance de plusieurs codons, alors que des $ARNt$ spécifiques pour chacun des codons d'un acide aminé ne sont pas disponibles. Beaucoup de modifications sont possibles : par exemple les $ARNt$ de la tyrosine

peuvent arborer toutes les modifications suivantes : N2-methylguanine, dihydrouridine, 3-(3-amino-3-carboxypropyl)-uridine, N2,N2-dimethylguanine, galactosylqueuosine, pseudouridine, N1-methylguanine, N1-methylpseudouridine, N7-methylguanine, 5-methylcytosine, 5-methyluridine et N1-methyladenosine. Mais l'étude des modifications des ARNt est difficile, puisqu'il n'est actuellement pas possible d'isoler les différents isodécodeurs présents dans le cytoplasme, mais seulement d'en faire l'analyse dans son ensemble [Parisien et al., 2013].

L'abondance des isoaccepteurs est une propriété déterminante dans la capacité des systèmes biologiques à produire leurs protéines. Par exemple, les protéines fortement exprimées possèdent un usage des codons correspondant aux isoaccepteurs les plus fréquents (proxy de leur expression) assurant ainsi une traduction efficace (sur le plan de la fidélité et de la rapidité), et cela à travers le vivant (hypothèse de la sélection traductionnelle : [Ikemura, 1981, Bulmer, 1987, Plotkin et al., 2004, Dittmar et al., 2006]). Bulmer [1987] parle alors de coévolution entre le pool d'ARNt et l'usage des codons des protéines fortement exprimées. L'hypothèse de la sélection traductionnelle propose que l'ajustement entre le pool d'ARNt et l'usage des codons soit le résultat de la sélection positive : soit par des duplications/délétions qui auraient affecté le nombre de copies des gènes d'$ARNt$, soit par des mutations qui auraient changé l'usage des codons des protéines fortement exprimées. Comme nous l'avons vu plus haut dans le texte, la capacité d'un système biologique à sélectionner un trait phénotypique dépend de l'importance de l'avantage adaptatif qu'il confère, mais aussi de la N_e.

Le rôle de l'usage des codons ne se limite pas à sa correspondance avec la composition en isoaccepteurs du pool d'$ARNt$. Chez la levure, la cinétique de l'élongation peut être modulée par l'organisation spatiale de l'usage des codons le long de l'$ARNm$ et par la disponibilité en $ARNt$ chargés [Cannarrozzi et al., 2010]. Par exemple, l'usage des codons correspondant à la composition du pool d'$ARNt$ serait favorisé jusqu'à la fin 3' du messager pour augmenter la vitesse de l'élongation et réduire le niveau d'infidélité [Cannarrozzi et al., 2010], alors que l'extrémité 5' de l'$ARNm$ serait enrichie en codons dont les $ARNt$ sont rares pour assurer une initiation lente de la traduction. Cette traduction lente permet d'éviter le détachement des ribosomes, au prix d'une traduction potentiellement moins fidèle [Mitarai and Pedersen, 2013]. À cette étape de l'élongation, la demande locale sur les $ARNt$ chargés étant importante

serait assurée localement par le recyclage des $ARNt$ et donc l'utilisation groupée des mêmes codons [Cannarrozzi et al., 2010].

La production de protéines ne correspondant pas à ce qui est codé dans le génome est généralement délétère : (1) ces protéines incorrectement traduites peuvent avoir perdu la fonction moléculaire attendue, ou pire en avoir une autre ; (2) elles peuvent ne pas se replier correctement, ce qui peut mener à la formation d'agrégats, souvent dommageables pour la cellule [Drummond and Wilke, 2008]. La fidélité de la traduction peut évoluer très rapidement (*E. coli* [Kurland, 1992]) maintenant que les cellules s'accommodent facilement de taux d'erreurs assez différents.

1.1.3. Évolution du code génétique

La structure du code génétique serait parmi les plus robustes aux mutations délétères qui auraient pu être conçues (revue dans Koonin and Novozhilov [2017]). Les mutations synonymes n'impliquent pas de changement d'acides aminés, alors que les mutations non-synonymes impliquent des changements d'acides aminés. La **dégénérescence** du code génétique, c'est-à-dire le fait que plusieurs codons codent pour le même acide aminé, est loin d'être aléatoire. Par exemple, les codons d'un même acide aminé sont généralement adjacents par une seule mutation. Aussi, les acides aminés codés par deux codons, ne diffèrent que par le dernier nucléotide ; mais il s'agit toujours soit d'une purine (R = A ou G) soit d'une pyrimidine (Y = C ou U), donc adjacent par une transition. Les transitions sont les mutations les plus fréquentes [Gojobori et al., 1982, Kumar, 1996, Wakeley, 1996, Petrov and Hartl, 1999, Rosenberg et al., 2003, Lynch, 2010a, Duchene et al., 2015]. Les changements d'acides aminés résultant des transitions seraient plus conservateurs que ceux engendrés par les transversions [Wakeley, 1996, Rosenberg et al., 2003, Keller et al., 2007] mais cette question reste controversée [Stoltzfus and Norris, 2016]. La sélection qui affecte les transversions non-synonymes serait donc plus forte que celle affectant les transitions non-synonymes. La plupart des événements non-synonymes qui impliquent une seule base se font entre acides aminés biochimiquement semblables [Grantham, 1974] tels, que le passage d'une valine (GUN) à une isoleucine (AUH), deux acides aminés hydrophobes.

L'émergence et le maintien du code génétique relève aussi de l'évolution. Par exemple, dans certains groupes taxonomiques la structure du code génétique est instable : chez les

Saccharomycotina, incluant la plupart des levures, plusieurs **réassignations** de codons ont été détectées. C'est-à-dire, que le codon sera reconnu par un $ARNt$ chargé d'un acide aminé différent de celui prédit par le code génétique universel. Le codon CTG est habituellement reconnu par un $ARNt$ chargé d'une leucine (code génétique universel), alors que dans certains clades des levures il est reconnu par un $ARNt$ chargé d'une sérine (deux clades distincts) ou d'une alanine dans un autre clade [Krassowski et al., 2018]. Chacune de ces réassignations serait un évènement indépendant dû à une mutation de l'anticodon des $ARNt^{sérine}$ et $ARNt^{alanine}$ permettant ainsi de reconnaître le codon CUG [Krassowski et al., 2018]. Dans le clade CTG comprenant *Candida albicans*, une mutation (G_{37}) de l'$ARNt^{sérine}$ permet aussi le chargement de leucine dans 3% des cas. Il est intéressant de noter que dans le second clade CTG, où ce dernier codon est décodé par un $ARNt$ chargé d'une sérine, plusieurs espèces possèdent toujours le gène codant pour l'$ARNt^{leucine}(CAG)$; à noter par contre que les résultats de spectrométrie de masse ne montrent que la présence de $ARNt^{sérine}(CAG)$ [Krassowski et al., 2018]. Les auteurs proposent la sélection comme mécanisme évolutif pour expliquer la perte de $ARNt^{leucine}(CAG)$. D'autre part, certaines espèces des clades sans réassignation détectée ne possèdent plus l'$ARNt^{leucine}(CAG)$ ou encore le possèdent, mais avec le plus grand intron jamais trouvé dans un $ARNt$, le rendant probablement non fonctionnel. Est-ce que la dérive génétique ne pourrait pas être à l'origine de la perte de l'$ARNt^{leucine}(CAG)$ dans les clades avec réassignation, un peu à l'image de ce qui arrive dans les génomes extrêmement réduits des endosymbiontes bactériens [McCutcheon and Moran, 2012] ?

1.1.4. Les processus mutationnels

Le concept de mutation a été développé durant les trois premières décennies du 20e siècle (revue dans Auerbach [1976]) sans connaître le medium de l'information génétique : l'acide nucléique comme médium n'est démontré que dans les années 1940 [Avery et al., 1944], et la structure en double hélice dans les années 1950 [Watson and Crick, 1953]. Depuis, la caractérisation des mutations et de leur impact sur le phénotype est un domaine de recherche important de la biologie moléculaire en pleine expansion (e.g., [Zou et al., 2018, Findlay et al., 2018]).

À l'échelle des bases de l'ADN, les altérations possibles sont les remplacements de bases simples ou multiples (e.g., impliquant deux ou plusieurs bases adjacentes), les insertions

et les délétions. Parmi les mécanismes moléculaires endogènes reconnus pour générer des changements de bases de l'ADN chez les mammifères, il y la réplication de l'ADN, la réparation de l'ADN, la désamination spontanée de certaines bases de l'ADN, la dépurination/dépyrimination, la transcription, la recombinaison et l'exposition aux dérivés réactifs de l'oxygène (revue dans Chatterjee and Walker [2017]). Parmi les sources exogènes les plus communes de mutation, il y a les radiations ionisantes (e.g., désintégration du radon 222), les rayons UV, les produits chimiques (e.g., les agents oxydatifs, les agents alkylants tel la fumée de cigarette, les composés aromatiques, les toxines telles les aflatoxines) ainsi que les stress liés à des changements d'environnements (e.g., froid), revue dans Chatterjee and Walker [2017]. À l'échelle des chromosomes, des régions entières peuvent être éliminées (délétions), ou encore inversées (inversions), voire dupliquées (duplications) ainsi que déplacées d'une région génomique à une autre (translocations). Les transferts de gènes sont aussi une source importante de modification du contenu génomique, à considérer comme des mutations.

Les mutations peuvent affecter un brin de l'ADN en particulier (e.g., brin non-transcrit), alors que d'autres mutations peuvent survenir sur les deux brins de l'ADN sans discrimination (e.g., désamination de la cytosine). Il existe aussi une variation spatiale dans les processus mutationnels [Smith et al., 2018], ou encore près des points chauds de recombinaison [Arbeithuber et al., 2015]. Le taux de mutation et la nature même des mutations varient au cours de l'évolution pour des raisons endogènes et exogènes aux systèmes biologiques.

Il serait impossible d'étudier les processus mutationnels sans parler des mécanismes moléculaires de réparation qui les accompagnent. Il existe plusieurs mécanismes moléculaires de réparation, chacun est spécifique à un type de dommage à l'ADN (voir la figure 1.5). Il y a la **réversion directe**, la **réparation des mésappariements** (MMR pour MisMatch Repair), la **réparation par excision de bases** (BER pour Base Excision Repair) et la **réparation par excision de nucléotides** (NER pour Nucleotide Excision Repair), la **recombinaison homologue** (HR pour Homologous Recombination), la **jonction d'extrémités non homologues** (NHEJ pour non-homologous end joining) et la **réparation couplée à la transcription** (TCR pour transcription coupled repair).

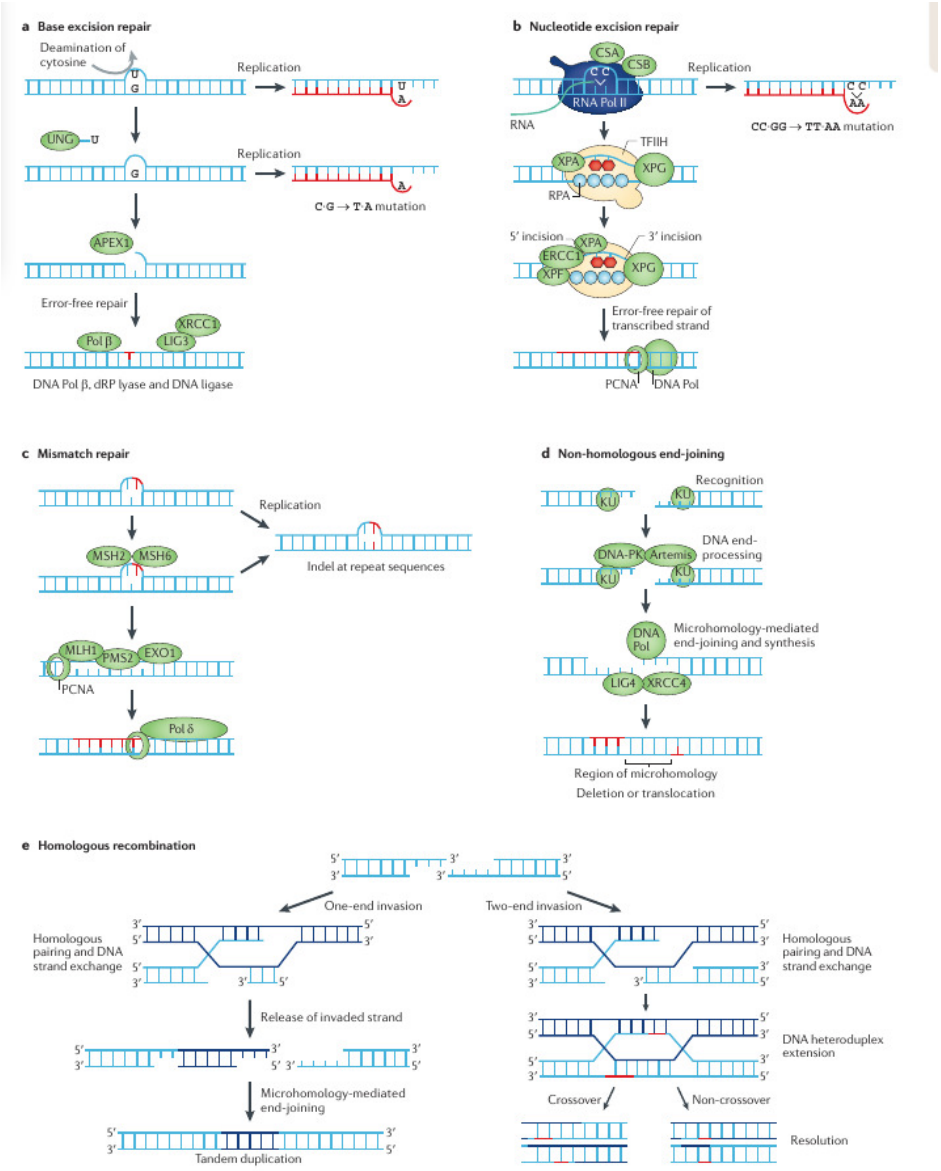

FIGURE 1.5. Processus de réparation (tiré de [Helleday et al., 2014]). Les mécanismes de réparation présentés sont (a) la **réparation par excision de bases** (BER pour Base Excision Repair), (b) la **réparation par excision de nucléotides** (NER pour Nucleotide Excision Repair), (c) la **réparation des mésappariements** (MMR pour MisMatch Repair), (d) la **jonction d'extrémités non homologues** (NHEJ pour non-homologous end joining), (e) la **recombinaison homologue** (HR pour Homologous Recombination).

La réversion directe permet de retrouver la base modifiée dans une seule étape, mais fait référence à un ensemble de mécanismes moléculaires distincts qui permettent d'enlever les groupes alkyles, comme la méthylation de l'oxygène de la guanine en position 6. Dans ce

cas, la réversion directe implique une methyltransferase O(6)-methylguanine-DNA, celle-ci qui permet de transférer le groupe alkyle sur l'oxygène (6) de la guanine à une cystéine de son site actif [Chatterjee and Walker, 2017].

Lors de la réplication chez les Mammifères, les polymérases γ et ϵ génèrent des mutations à un taux par base de 10^{-5} à 10^{-6} et 10^{-6} à 10^{-7} respectivement [Chatterjee and Walker, 2017]. Le mécanisme MMR permet ensuite d'abaisser ce taux par un facteur $\sim$ 100 [Chatterjee and Walker, 2017]. Le mécanisme MMR détecte les bases endommagées, et excise, avec une exonucléase, un fragment d'ADN comprenant la base endommagée et quelques bases adjacentes. Par la suite, une polymérase à ADN comblera l'ADN manquant à partir du brin parent. Une ligase est aussi nécessaire pour fusionner le brin synthétisé. Il est intéressant de noter que le mécanisme de MMR intervient avant que la méthylation du brin nouvellement répliqué ait lieu. Une méthyletransférase permet de répliquer les patrons de méthylation du brin parent au brin fils. La consistance des patrons de méthylation entre les deux brins est en soi un mécanisme de contrôle de qualité [Wang et al., 2016].

Le mécanisme BER permet de réparer les bases altérées par la désamination spontanée, par les lésions oxydatives, par la méthylation ou encore les réactions hydrolytiques. BER assure la réparation par l'excision de la base altérée au moyen d'une ADN glycosylase, ce qui génère un site abasique (apurique ou apyrimidique) hautement mutable puisque toutes les bases peuvent être introduites lors de la réplication d'un site abasique (revue dans [Helleday et al., 2014]). POLB se charge de finaliser la réparation.

NER est un mécanisme qui permet de réparer les dommages causés par les UV, les amines aromatiques (e.g., aflatoxines), et autres molécules qui se fixent à l'ADN. Ce mécanisme de réparation est efficace sur le brin transcrit (3'-5'), alors que le brin anti-sens (5'-3') sera réparé moins efficacement, ce qui génère une asymétrie dans le processus mutationnel. Le mécanisme tire avantage d'une distorsion dans la structure de la double hélice, due à un dommage à l'ADN. La distorsion sera accentuée par la présence de protéines, qui par la suite permettront d'ouvrir la double hélice d'ADN pour l'excision de plusieurs bases, comprenant la base endommagée. L'ADN polymérase bêta, $POLB$, se charge de finaliser la réparation.

Seules les mutations qui affectent les lignées germinales seront étudiées ici, car nous nous intéressons à l'évolution à grande échelle de temps ; ces mutations doivent donc passer d'une génération à l'autre de multiples fois. Chez les mammifères, trois processus mutationnels

sont particulièrement importants (il y en a assurément d'autres) : (1) l'hypermutabilité des m^5C en contexte CpG [Bird, 1980], (2) la conversion génique biaisée vers GC [Duret and Arndt, 2008] qui favorise les allèles $C : G$ à la place des allèles $A : T$, et (3) un ensemble de mutations reliées à l'exposition de l'ADN simple brin durant la transcription [Green et al., 2003], plus particulièrement le brin qui n'est pas transcrit.

L'hypermutabilité des m^5C résulte de leur désamination : soit la perte du groupe amine (NH_3) exocyclique. La désamination des cytosines, des adénines, des guanines et des cytosines méthylées (m^5C) génère des uraciles, des hypoxanthines, des xanthines et des thymines respectivement (revue dans Chatterjee and Walker [2017]). La désamination affecterait plus particulièrement les cytosines et les m^5C. Le taux de désamination par base serait aussi quatre fois plus important pour les m^5C (revue dans Chatterjee and Walker [2017]). Par contre, l'indépendance des mutations résultant de la désamination des m^5C face au mécanisme de la réplication serait controversée. Certains résultats abondent dans le sens d'une indépendance face à la réplication (e.g., Mikkelsen et al. [2005], Thomas et al. [2018]), alors que d'autres études (e.g., Jonsson et al. [2017], Wong et al. [2016]) détectent l'effet de l'âge du père à la reproduction (e.g., Jonsson et al. [2017], Wong et al. [2016]), un proxy du nombre de réplications qui ont lieu durant la spermatogenèse.

La désamination d'une m^5C peut : (1) permettre de retrouver la cytosine (non-méthylée), alors il y a réparation ; (2) générer une transition $C > T$, le cas le plus connu de mutation ; ou (3) générer une transition ou une transversion pour tous les états possibles, $C > N$ (communications personnelles Ignacio Bravo). Dans le premier cas, le mauvais appariement généré ($T : G$) sera transformé en site ab

ique via une thymine glycosylase à ADN qui sera ensuite excisée par une endodeoxyribonuclease apurinique. Avant d'être finalement réparée par $POLB$, le 5'-deoxyribose-phosphate est excisé lui aussi par une lyase. Dans le deuxième cas, si la réplication a lieu avant la formation du site abasique, une thymine sera introduite à la place de la cytosine, menant à la transition $C > T$ habituellement associée à l'hypermutabilité m^5C. Dans le troisième cas, lorsque la réplication se fait en présence du site abasique, donc avant excision de ce site, toutes les mutations sont alors possibles, $C > N$. De son côté, la désamination d'une cytosine non-méthylée génère les mêmes profils de mutation que la désamination d'une cytosine méthylée : $C > T$ ou $C > N$ pour les mêmes conditions face à la réplication. Cependant

la transformation de l'uracile (mauvais appariement $U : G$) en site abasique se fait via une uracile glycosylase à *ADN* qui pourra ensuite être excisée par une endodeoxyribonuclease apyrimidinique avant d'être réparée par *POLB*.

La recombinaison homologue permet la réparation des cassures double-brin qui se forment lors de la méiose chez les eucaryotes (à la phase S ou G2), division cellulaire essentielle à la gamétogenèse. C'est à ce moment que des cassures double-brin sont produites et que les mécanismes de réparation HR et NHEJ interviendront. La HR serait plus efficace que la NHEJ, cette dernière laissant des insertions et des délétions [Tubbs and Nussenzweig, 2017]. Dans le cas où la réparation n'est pas accomplie, la cellule en méiose devra se diriger vers l'apoptose. Les gènes *BRCA1* et *BRCA2* qui contrôlent en partie la HR sont aussi connus pour leur implication dans le développement des cancers du sein ou des ovaires [Helleday et al., 2014]. Les principales étapes consistent à l'appariement des brins d'*ADN* en provenance du chromosome homologue, puis de la synthèse de l'*ADN* et de la séparation des chromosomes homologues par coupure des brin d'*ADN* appariés (revue dans Alberts et al. [2002]).

L'*ADN* simple brin exposé lors de la réplication, de la transcription ou encore de la recombinaison serait plus sensible à la désamination spontanée (revue dans [Chatterjee and Walker, 2017]). La désamination devrait donc affecter particulièrement les gènes fortement exprimés. L'*ADN* simple brin peut aussi être la cible d'endonucléases, ce qui aura pour conséquence de générer des cassures double-brin, ou encore être la cible de désaminases appartenant à la famille des APOBEC [Tubbs and Nussenzweig, 2017]. Le potentiel mutagénique dû au stress de la réplication serait uniforme le long du génome par définition, alors que le potentiel mutagénique lié à la transcription aurait un impact particulièrement important sur les gènes codants pour des protéines fortement exprimées (e.g., protéines ribosomiques dans le soma et les rares gènes fortement exprimés des lignées germinales).

1.1.5. Les mutations germinales

Seules les mutations qui affectent les cellules germinales chez les mammifères peuvent être transmises d'une génération à une autre. Pour être héritées, les mutations doivent se produire dans une fenêtre du développement bien précise, entre la formation du zygote (diploïde), jusqu'à la fécondation, en passant par la gamétogenèse, voir figure 1.6.

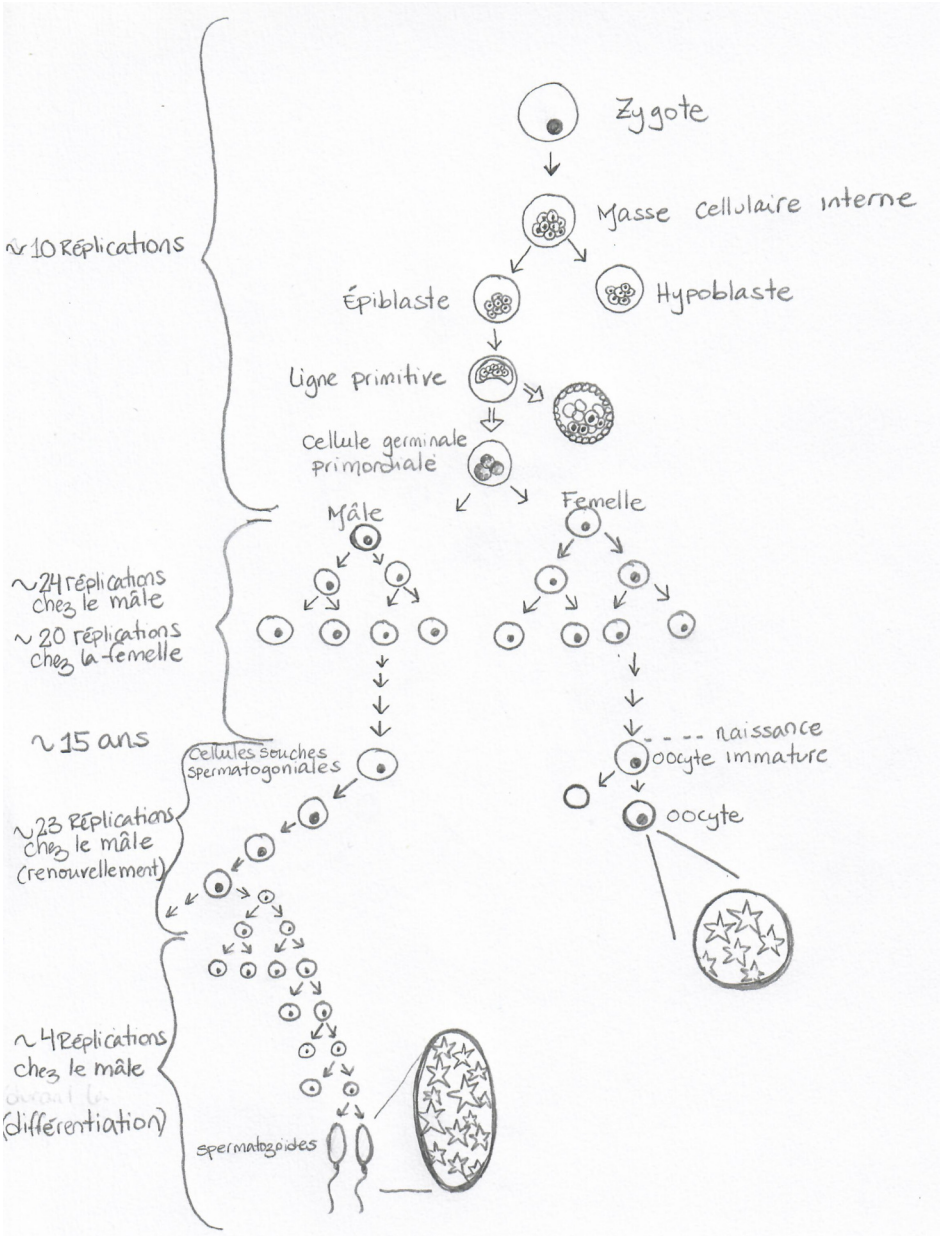

Figure 1.6. Schéma détaillant la gamétogénèse ainsi que le nombre de réplications qui sont nécessaires à chacune des étapes chez le mâle et la femelle *Homo sapiens* (adapté à partir de [Rahbari et al., 2016]). Crédits : Géraldine Philippin

Chez les mammifères, très peu de spermatozoïdes se rendent au site de fertilisation, et un seul participe au bagage génétique de la génération suivante. Chez *Homo sapiens* par exemple, seulement quelques milliers de spermatozoïdes se rendent au site de fécondation, alors qu'il y en avait des millions au site de dépôt [Sakkas et al., 2015]. À cette étape, le potentiel à éliminer les mutations délétères qui affectent les spermatozoïdes (e.g., motilité)

est grand étant donné le nombre d'individus sur lequel peut s'appliquer la sélection. La N_e usuellement estimée intègre toutes les étapes du cycle de vie des systèmes biologiques étudiés ainsi qu'une partie de leur histoire évolutive puisque cette dernière est estimée à partir d'un polymorphisme correspondant à l'histoire évolutive des populations de du système biologique à l'étude (e.g., [Pitt et al., 2019]). C'est-à-dire que pour avoir accès à la N_e la plus récente, il faut s'intéresser aux mutations nouvellement arrivées dans chacune des populations. Ceci dit, chacune des étapes menant à la fécondation pourrait faire l'objet d'une étude afin de quantifier la sélection qui s'y opère.

Une fois rendu au site de fécondation, il ne reste que quelques milliers de spermatozoïdes, la sélection est alors probablement beaucoup moins efficace à discriminer les génotypes/phénotypes ayant la plus grande valeur adaptative avant la fécondation, surtout que la fusion d'un spermatozoïde à l'oocyte nécessite l'action de plusieurs spermatozoïdes (i.e., une sorte de coopération entre individus haploïdes). En absence de sélection pré-fécondation due aux technologies de reproduction assistée, certains auteurs estiment que le fardeau génétique transmis d'une génération à l'autre ne cesse d'augmenter [Lynch, 2016]. Mais aucune étude, à notre connaissance, ne quantifie le fardeau génétique qui grandit par l'absence de sélection lorsqu'une technologie de reproduction assistée est utilisée. Certains praticiens semblent reconnaître l'importance de sélectionner les meilleurs spermatozoïdes [Vaughan and Sakkas, 2019].

Chez *Homo sapiens*, les Hommes transmettent un fardeau génétique plus grand que les Femmes (e.g., [Jonsson et al., 2017]). Cet écart avec les femelles devient plus important en fonction de l'âge : environ deux mutations sont ajoutées au fardeau génétique de la lignée germinale pour chaque année d'âge supplémentaire après la puberté de l'Homme, alors que le fardeau génétique des oocytes est beaucoup plus stable [Rahbari et al., 2016]. La gamétogenèse chez les Femmes est obtenue en un nombre fixe de réplications alors que la spermatogenèse permet de renouveler en continu le stock de spermatozoïdes, ce qui nécessite de nouvelles mitoses tout au long de la vie de l'Homme (figure 1.6). Mais en fait, à notre connaissance, ni l'hypothèse de la réplication [Rahbari et al., 2016] ni l'hypothèse alternative selon laquelle la transcription serait le mécanisme responsable de l'accumulation des mutations n'ont été testées explicitement.

L'environnement cellulaire des gamètes femelles et mâles étant différent, les traits d'histoire de vie qui les caractérisent (tableau 1.1) le sont aussi. En fait, les traits d'histoire de vie correspondent aux **stratégies r et K** [Pianka, 1970] pour les spermatozoïdes et les oocytes respectivement. L'investissement dans les oocytes est beaucoup plus important (stratégie K). Par exemple, chaque spermatozoïde compte environ 100 mitochondries (stratégie r), alors que chaque gamète femelle compte entre des dizaines de milliers à des dizaines de millions mitochondries (stratégie K) chez les mammifères [Dumollard et al., 2007]. Même si le nombre de divisions cellulaires pour générer les gamètes femelles est fixe, contrairement à celui du mâle, le taux de mutation par base par division cellulaire serait plus important chez la femelle : est-ce que ce plus haut taux de mutation détecté chez les gamètes femelles pourrait s'expliquer par une plus grande concentration de dérivés réactifs de l'oxygène due à une plus grande abondance de mitochondries ? D'autre part, il est surprenant de constater que la mitochondrie est d'héritage femelle, alors que la contrainte énergétique que nécessite la motilité des spermatozoïdes aurait permis de sélectionner les mitochondries les plus performantes.

Table 1.1. Comparaison de traits d'histoire de vie entre spermatozoïdes et oocytes à travers les lunettes du modèle des stratégies r et K

traits d'histoire de vie	spermatozoïdes	oocytes
taille de population	grande	petite
investissement parental	petit	grand
âge de la maturité	précoce	tardive
taille	petite	grande
durée de vie	courte	longue
mortalité	élevée	basse

1.1.6. Les mutations somatiques

Rappelons que notre travail n'est pas l'étude des mutations somatiques, mais bien celui des mutations germinales. Dès les années 1970, le domaine de la cancérologie identifie les mutations qui affectent le soma comme une cause au développement des cancers (revue dans [Tubbs and Nussenzweig, 2017]) et plus particulièrement les mutations qui altèrent les fonctions des gènes impliqués dans les mécanismes de réparation de l'ADN (e.g., BRCA1).

Une fois un mécanisme de réparation altérée, la probabilité d'acquérir de nouvelles mutations augmente. La sélection positive pour les mutations qui améliorent localement la valeur adaptative de certaines lignées somatiques détériore la valeur adaptative de l'individu en générant des lignées cellulaires à croissance incontrôlée. Ceci illustre bien le conflit entre les lignées germinales et lignées somatiques (voire la revue suivante [Queller and Strassmann, 2018] sur les conflits dans le domaine de l'évolution). De la même manière, les mutations engendrées par la chimiothérapie et la radiothérapie, moyens utilisés dans la lutte contre le cancer, peuvent provoquer l'apparition de nouveaux sites tumoraux, ou encore, favoriser l'exploration du paysage adaptatif et générer des adaptations qui pourront être sélectionnées et mener au développement de résistances aux thérapies elles-mêmes (revue dans [Tubbs and Nussenzweig, 2017]). Certains génotypes sont plus enclins aux développements de cancer : par exemple les mutations qui affectent les gènes BRCA1 et BRCA2, soit des gènes impliqués dans la réparation des cassures double-brin, sont des prédispositions au cancer du sein et de l'ovaire importantes [Mersch et al., 2015]. Mais de manière générale, les mutations qui affectent BRCA1 et 2 diminuent significativement la valeur adaptative des porteurs, la sélection négative tend alors à les éliminer de la population, car sans la réparation des cassures double-brin qui apparaissent durant la méiose, la cellule entre en apoptose, un contrôle de qualité important.

Dans l'espoir de mieux diagnostiquer les cancers (oncologie de précision), un domaine actif de la cancérologie s'intéresse tout particulièrement à comprendre leur étiologie. Par contre, détecter et identifier les mutations à l'origine des cancers est une tâche particulièrement difficile. Chaque cancer peut posséder jusqu'à quelques milliers de mutations (SNP), sans parler de tous les réarrangements génomiques, comme l'aneuploïdie, aussi impliqués dans le développement des lignées cancéreuses [Helleday et al., 2014]. Accéder à l'étiologie du cancer est d'autant plus difficile que plusieurs lignées cellulaires sont regroupées lors du séquençage haut débit, cachant potentiellement la présence d'une hétérogénéité dans les processus mutationnels ou de sélection qui affectent les différentes lignées cancéreuses d'un même individu. Par contre, il serait aujourd'hui possible de travailler à l'échelle de la cellule unique (single cell sequencing : [Tanay and Regev, 2017]), évitant ainsi de confondre les différents signaux. Une autre difficulté provient de la grande taille du génome humain, soit

plus de $\sim 3 \times 10^9$ bases à séquencer, sans parler de la dimension épigénomique qui devra aussi être étudiée. C'est donc comme chercher une aiguille dans une botte de foin.

Néanmoins, certains chercheurs ont réussi à faire des progrès considérables dans la compréhension de ce que peut être l'étiologie des cancers. Une de celles-ci consiste non pas à travailler avec les mutations prises individuellement, mais plutôt à travailler avec un ensemble de mutations appelé profil mutationnel (mutational signature) [Alexandrov et al., 2013b,c, Alexandrov and Stratton, 2014, Helleday et al., 2014, Nik-Zainal et al., 2012]. Cette méthode utilise les quelques milliers de mutations extraites sous forme de SNP obtenus par le séquençage des exomes/génomes des différents types de cancer, où le génome humain est pris à titre de référence (e.g., GRCh37). Cette démarche cherche à différencier les types de cancer sur la base de leurs profils mutationnels. Le profil mutationnel le plus simple utilise un modèle symétrique en brin, ce qui veut dire que les paires de bases ($A:T$, $T:A$, $G:C$ et $C:G$) sont assumées pour être présentes uniformément sur les brins référence et anti-référence. Un processus mutationnel qui affecte la cytosine ($C > T$) indifféremment du brin référence ou anti-référence devrait alors générer autant de mutations de $C > T$ que de G>A sur le brin référence. Le modèle symétrique en brin possède donc six paramètres ($C:G > T:A$, $C:G > A:T$, $C:G > G:C$, $T:A > C:G$, $T:A > A:T$, $T:A > G:C$). Le stress mutationnel lié à la réplication ou à la transcription n'affecte pas les deux brins de manière équivalente, un modèle mutationnel asymétrique en brin incorporant 12 paramètres mutationnel ($A > C$, $A > G$, $A > T$, $C > A$, $C > G$, $C > T$, $G > A$, $G > T$, $G > C$, $T > A$, $T > C$, $T > G$) serait nécessaire pour caractériser ces deux processus mutationnels. L'utilisation d'un modèle symétrique en brin, mais prenant en compte le contexte des deux bases adjacentes à la base qui mute, pour un contexte trinucléotidique, a permis de discriminer plusieurs types de cancer et même d'identifier les mécanismes à l'origine du **profil mutationnel global** enregistré pour un type de cancer particulier. Ce profil mutationnel global possède 96 entrées, $4 \times 6 \times 4$, dans lesquels les SNP sont répartis. Étant donné la quantité de SNP disponible, la recherche s'est limitée au contexte trinucléotidique pour s'assurer d'un assez grand pouvoir statistique.

L'étape suivante consiste à se demander si le profil mutationnel global retrouvé en prenant en compte le contexte trinucléotidique ne serait pas en fait un assemblage de plusieurs profils mutationnels ayant pour origine différents mécanismes mutationnels (e.g., désamination des

cytosines méthylées et l'exposition aux UV) et cela dans des proportions variables. C'est en partie à cette question que Nik-Zainal et al. [2012] veulent répondre. Ces auteurs proposent d'utiliser un modèle de régression (i.e., factorisation par matrices non-négatives) pour inférer une combinaison linéaire de k profils mutationnels trinucléotidiques, $p_{1 \leq i \leq k}$, dont la participation spécifique au profil global est modulée par un poids, $w_{1 \leq i \leq k}$ spécifique à chacun des k profils dans le but de prédire le profil mutationnel global P construit à partir des SNP contextualisés en 3' et 5'. Ce qui peut être formulé de la manière suivante : $P = \sum_{i=1}^{k} w_i \times p_i$, où P, w et p sont des vecteurs de dimension 96, et les valeurs w_i pour un même vecteur i sont tous égaux.

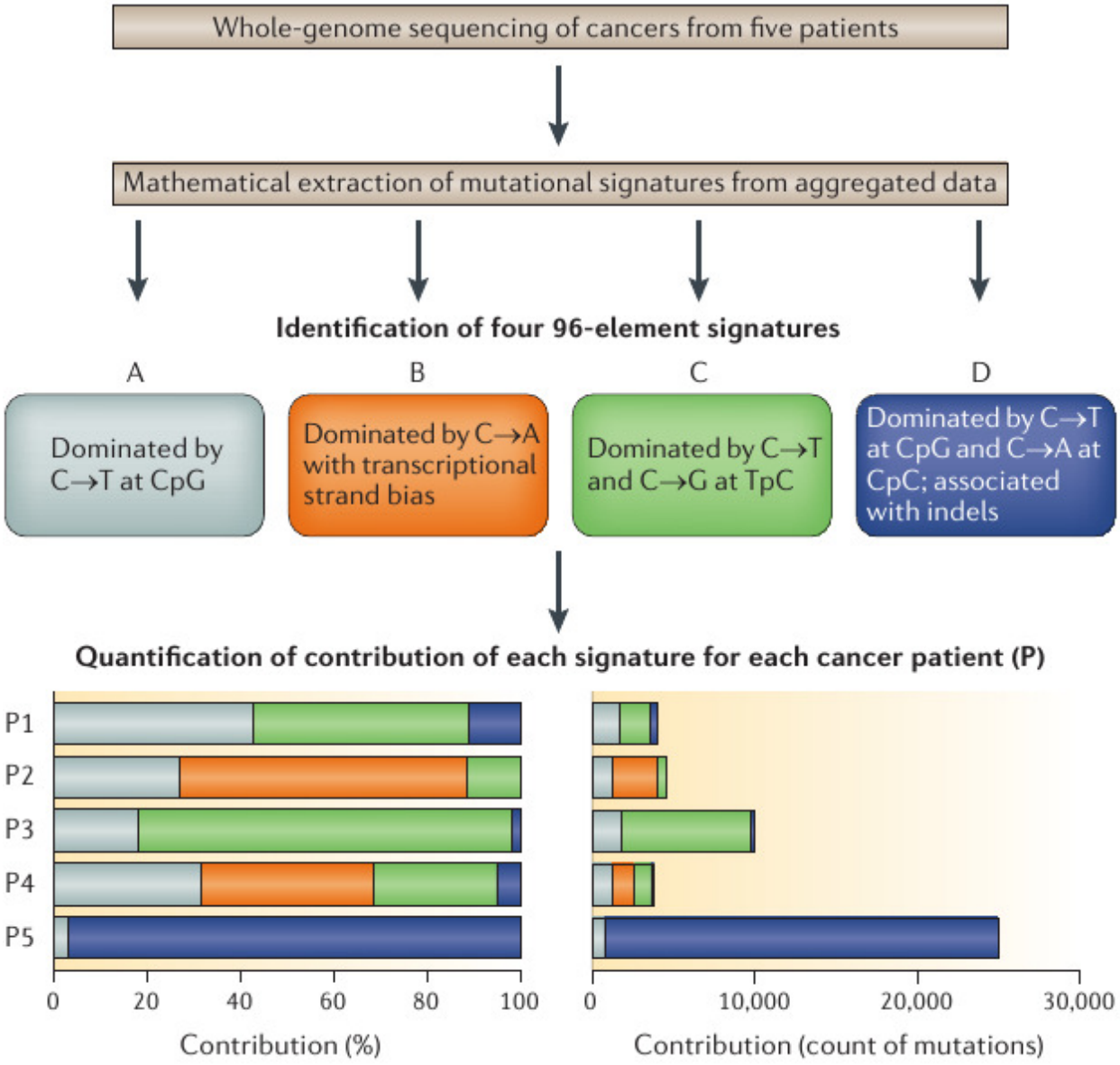

Figure 1.7. Schéma détaillant l'inférence de profils mutationnels à partir de données de séquençage haut débit de cancers (modifié à partir de [Helleday et al., 2014]).

Par exemple, l'hypermutabilité du contexte NpCpG est un profil mutationnel, figure 1.8, qui a été détectée dans tous les types de cancers étudiés Alexandrov et al. [2013a]. Ce qui n'est pas surprenant étant donné l'importance du processus de désamination des cytosines méthylées chez les mammifères. Par contre, d'autres profils mutationnels sont spécifiques à certains types de cancer, comme les profils mutationnels en lien avec une mauvaise réparation des cassures double-brin liées au dysfonctionnement de BRCA1 et BRCA2 (cancers du sein, des ovaires ou du pancréas : figure 1.8).

Cependant, cette approche est hautement phénoménologique puisqu'elle n'est pas contrainte de proposer des profils mutationnels qui auraient pu être causés par des mécanismes moléculaires réels. Plusieurs profils proposés par la méthode ne connaissent pas d'étiologie. Utiliser des profils mutationnels expérimentaux définis *a priori* est une voix intéressante pour contraindre le modèle de régression. Par exemple, Zou et al. [2018] mettent hors d'usage (knock-out) certains gènes appartenant aux mécanismes de réparation (e.g., MMR, BER, HR, NER) via l'utilisation de CRISPR-Cas9, ce qui permet d'obtenir des profils mutationnels spécifiques aux différents mécanismes de réparation, d'autant plus que les auteurs contrôlent pour les processus mutationnels qui ont lieu de toute manière. Une fois la lignée cellulaire générée, le séquençage haut débit permet de caractériser le profil mutationnel contextualisé en trinucléotides. Informée de cette manière, l'approche devient ainsi beaucoup plus mécanistique.

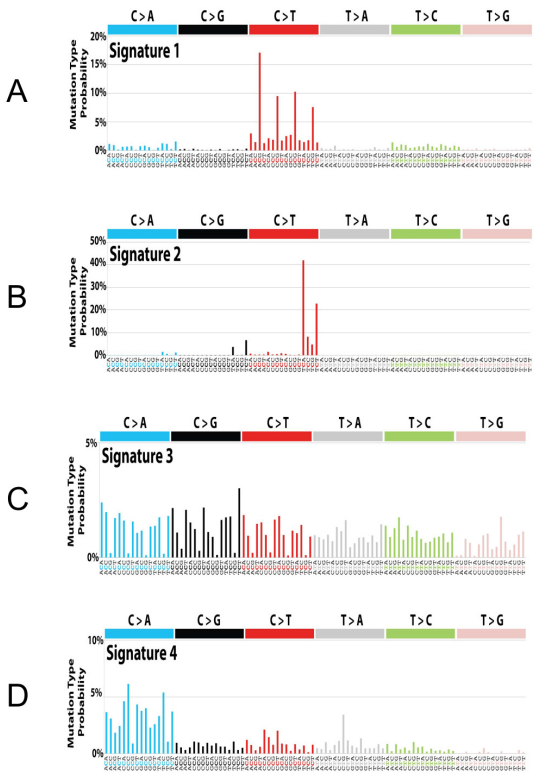

Figure 1.8. (A) Profils mutationnels 1-4 (mutation signature 1A-4). Profil mutationnel 1 (mutation signature 1A) associé à la désamination des cytosine méthylées. (B) Profil mutationnel 2 (mutation signature 2) généré par des désaminases de la famille des AID/APOBEC. (C) Profil mutationnel 3 (mutation signature 3) associé avec une altération du mécanisme moléculaire de réparation des cassures double-brin de l'ADN, la recombinaison homologue. (D) Profil mutationnel 4 (mutation signature 4) associé à l'exposition à la fumée de cigarette, plus particulièrement aux mutagènes comme le benzo[a]pyrène présent dans cette dernière. Adapté à partir de [Alexandrov et al., 2013a])

1.2. Modèles d'évolution des séquences

1.2.1. Phylogénie et histoire de la vie

Historiquement la préoccupation première de la phylogénie était d'inférer les relations de parenté entre systèmes biologiques. Ultimement la phylogénie peut permettre de discuter l'histoire de la vie (arbre de la vie : [Eme et al., 2014]). À savoir comment sont organisés les trois domaines du vivant (i.e., bactéries, eucaryotes, archées). Maintenant, l'histoire de

la vie peut aussi être abordée au niveau de l'évolution des molécules qui font les systèmes biologiques (évolution moléculaire).

Le **modèle d'évolution** qui ressemble le plus à ce qui est connu de la phylogénie actuelle est probablement l'arbre de la vie présenté par Darwin [Darwin, 1859]. Mais au sens mathématique, les premiers modèles d'évolution des caractères n'apparaissent qu'avec le développement de l'informatique. Les modèles d'évolution des caractères permettent d'inférer des **histoires évolutives**, soit des séries de changements d'état de caractères qui ont lieu le long de l'arbre phylogénétique. La structure habituelle sur laquelle les histoires évolutives sont déployées utilise la structure d'un arbre dichotomique ou encore celle d'un réseau. Les paramètres de base de tous les modèles d'évolution sont la topologie et la longueur des branches.

L'algorithme de maximum de parcimonie [Farris, 1970, Fitch, 1971a] permet de trouver l'histoire évolutive qui implique le moins de changements pour un jeu de données, faisant l'hypothèse que l'évolution est parcimonieuse et sans permettre de substitutions multiples. C'est-à-dire que les homoplasies détectées ne sont pas le résultat de l'évolution convergente (e.g., [Rey et al., 2018]), par exemple, mais bel et bien le résultat du partage d'un ancêtre commun.

Avant l'arrivée des données moléculaires, les analyses phylogénétiques se faisaient à partir d'observations morphologiques. Le défaut de la morphologie est qu'elle incorpore une complexité (e.g., dépendances entre caractères) et une diversité de caractères (e.g., présence et absence d'un feuillet embryonnaire versus leur nombre) qui empêche de définir aisément des modèles d'évolution, le nombre de paramètres nécessaires à un modèle un tant soit peu réaliste étant beaucoup plus grand que le nombre d'observations (underfitting). Travailler avec des données de séquençage, soit des séquences codantes pour des protéines ou des gènes d'ARN structural (ARNr 18S), mais aussi avec des séquences non-codantes a permis d'augmenter de beaucoup la qualité des inférences : (1) grâce à la quantité des données disponibles (de quelques dizaines de caractères à des centaines dès la fin des années 1980, puis à des centaines de milliers, voire des millions, dans les années 2010) et (2) dans une moindre mesure, grâce à une meilleure compréhension des mécanismes qui régissent l'évolution des séquences codantes pour des protéines ou des ARN. Par exemple, les analyses phylogénétiques basées sur la morphologie ont proposé de positionner les tortues à la base des Diapsida (figure 1.9 :a)

ou encore de les placer comme groupe frère des Lepidosaures (figure 1.9 :b), alors que la phylogénie moléculaire (figure 1.9 :c) retrouve généralement les tortues avec les Archosauria (récemment confirmé par la phylotranscriptomique [Irisarri et al., 2017]).

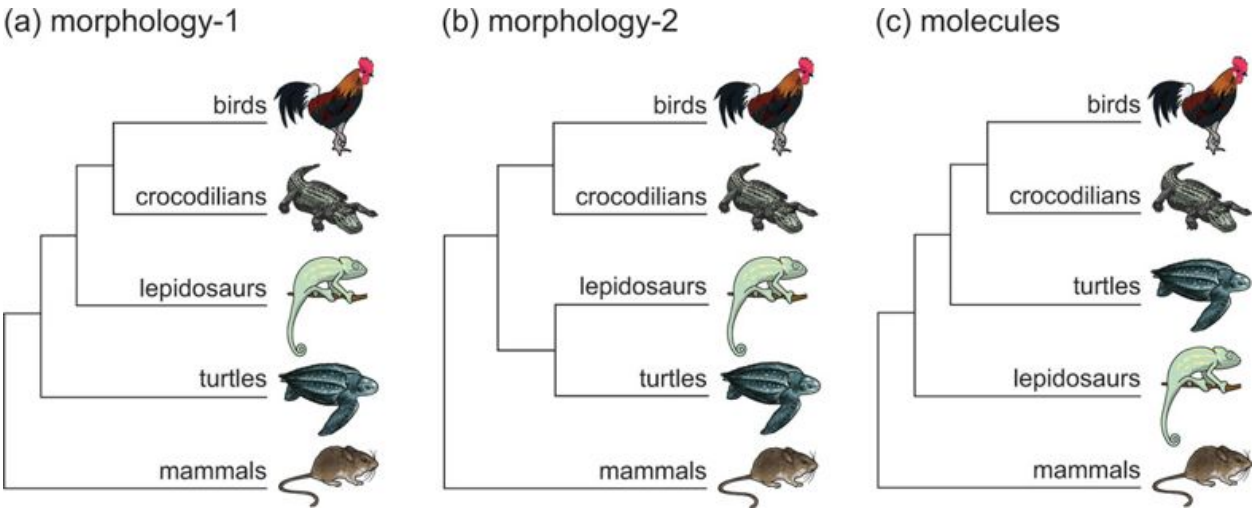

Figure 1.9. Comparaison de la position phylogénétique des tortues selon le type de données utilisées (a, b) données morphologiques et (c) données moléculaires (tiré de [Helleday et al., 2014]).

Le problème d'attraction des longues branches ([Felsenstein, 1978] : LBA), toujours d'actualité, est fréquemment rencontré par les phylogénéticiens. Ce problème a été mis en évidence au moment où l'inférence par maximum de parcimonie prévalait. L'attraction des longues branches est une limitation des modèles à prédire correctement une grande série de substitutions. Ce problème consiste à donner à tort un même ancêtre commun à deux longues branches. L'inférence phylogénétique par maximum de parcimonie est particulièrement sensible à l'attraction des longues branches, certains modèles d'évolution développés dans le cadre probabiliste sont plus en mesure de prédire les substitutions multiples [Yang and Rannala, 2012].

Aujourd'hui l'inférence de la phylogénie est statistique, c'est Felsenstein [1981] qui proposa le premier modèle probabiliste, et sut donc définir la manière de calculer la vraisemblance du modèle. Cette méthode de calcul nécessite cependant de faire l'hypothèse forte d'indépendance des sites, par opposition à une vraisemblance dite site-interdépendante permettant par exemple de prendre en compte la structure des protéines [Robinson et al., 2003]. Les modèles phylogénétiques probabilistes proposent donc des histoires évolutives qui sont

en fait une série de changements d'état (e.g., nucléotides, acides aminés, codons ou autres caractères), le long de l'arbre phylogénétique, de la racine de l'arbre jusqu'aux feuilles. La topologie et les longueurs des branches sont elles-mêmes des paramètres des modèles d'évolution. Les modèles phylogénétiques probabilistes prennent en entrée des alignements de séquences codantes. Différentes méthodes seront utilisées pour conditionner les modèles phylogénétiques probabilistes (voir le chapitre 3), ce qui aura pour but de filtrer l'espace des histoires évolutives proposées par les modèles phylogénétiques probabilistes et garder celles qui sont les plus plausibles (dans le cas où l'inférence bayésienne est utilisée).

Dans le cadre de la phylogénie probabiliste qui prévaut de nos jours, l'attraction des longues branches se manifeste lorsque les substitutions multiples sont si nombreuses que l'état ancestral inféré converge, souvent, vers la distribution stationnaire du modèle d'évolution et sera donc commun aux longues branches. Même si l'emploi de modèles d'évolution prenant en compte une plus grande hétérogénéité, comme la préférence site-spécifique en acides aminés [Lartillot and Philippe, 2004] permet d'inférer plus efficacement une série de substitutions multiples sur une plus grande distance évolutive [Lartillot et al., 2007], la meilleure solution consiste à "briser les longues branches" pour paraphraser Hendy and Penny [1989], en améliorant l'échantillonnage taxonomique. Ainsi, l'ajout d'espèces peut permettre de résoudre des problèmes phylogénétiques.

Par exemple, certains chercheurs se demandaient dans les années 1990 si la capacité d'exuviation (la mue) était un trait phénotypique apparu plusieurs fois indépendamment ou une seule fois ? Les premières analyses phylogénétiques plaçaient les nématodes en dehors du groupe actuel des Ecdysozoa (Nematoda, Nematomorpha, Loricifera, Priapulida, Kinorhyncha, Onychophora, Arthropoda). Une majorité des nématodes évoluent très rapidement, générant par le fait même de longues branches, et potentiellement des problèmes d'attraction des longues branches. Ce qui pouvait expliquer la paraphylie des lignées qui pratiquent l'exuviation précédemment retrouvée [Raff et al., 1994]. L'inclusion et l'exclusion des espèces de nématodes nouvellement séquencées (ARNr 18S) sur la base de leur vitesse d'évolution engendrent des conséquences importantes sur les topologies retrouvées (voir figure 1a et 1b [Aguinaldo et al., 1997]). Dans le premier cas (figure 1a [Aguinaldo et al., 1997]) les nématodes se retrouvent à la base des Bilateria, alors que lorsque seulement l'espèce avec le taux

de substitution le moins élevé, *Trichinella*, est incluse dans l'analyse, la monophylie des espèces pratiquant l'exuviation est retrouvée, créant de ce fait le groupe des Ecdysozoa accepté aujourd'hui [Aguinaldo et al., 1997], ce qui est confirmé plus tard [Delsuc et al., 2005].

L'utilisation de modèles probabilistes confère le grand avantage de permettre d'expliciter les hypothèses et de les tester dans un cadre statistique rigoureux. Ils peuvent être comparés via des méthodologies plus ou moins complexes (e.g., facteurs de Bayes, postérieur prédictif, etc.). Ultimement, cela nous permettra de mieux comprendre certains des processus à l'origine des données génomiques étudiées. Tester chacune de ces hypothèses requiert énormément de ressources computationnelles, d'autant plus que la recherche actuelle a pour objectif de proposer des hypothèses/modèles d'évolution de plus en plus complexes de manière à ce qu'ils soient en mesure de prendre en compte une plus grande diversité de processus moléculaires et modéliser une plus grande part de l'hétérogénéité présente dans les données [Lartillot, 2015].

Les histoires évolutives réelles qui sont à l'origine des données génomiques avec lesquelles nous travaillons sont extrêmement complexes, les modèles d'évolution qui nous permettent de proposer des histoires évolutives sont donc des simplifications outrageuses de ce qui s'est réellement passé. Néanmoins, ces modèles nous permettent de proposer des histoires évolutives qui sont porteuses de savoir autant faut-il que le **signal évolutif** soit assez important par rapport au **non-signal** [Philippe and Roure, 2011]. Identifier le signal évolutif, c'est être capable de retracer l'histoire évolutive. Plus l'histoire évolutive étudiée, la vraie, est hétérogène plus le signal sera difficile à identifier.

Les modèles d'évolution sont donc des hypothèses qui répondent à un ensemble de règles qui les définissent. Par exemple, sous la formulation proposée par Felsenstein [1981] pour le calcul de la vraisemblance, la majorité des modèles d'évolution ne prennent pas en compte de dépendance entre les sites d'un alignement de séquences, car le calcul de la vraisemblance dit site-indépendant est plus simple [Felsenstein, 1973, 1981]. Alors que les processus de mutation et de sélection qui font l'évolution peuvent être site-interdépendants (e.g., hypermutabilité des transitions en contexte CpG et contraintes de sélection sur la structure des protéines). Lorsqu'une seule de ces règles est transgressée, il est alors question de violations

de modèle (model misspecification). À mesure que la quantité de données grandit, si l'estimateur converge sur la vraie valeur de paramètre, l'estimateur est dit consistant. Mais en présence de violations de modèle, la consistance n'est pas forcément assurée (mais ça peut).

Nécessairement, les modèles proposés font face à des violations de modèles (connues et inconnues), qui devront être contrôlées dans la mesure du possible. La corroboration est la démarche consentie pour valider les résultats en faisant varier les conditions expérimentales (données ou pseudo-données) et les hypothèses explicatives (modèles).Deux cas extrêmes sont importants au système de corroboration : (1) les cas où l'effet du processus à isoler est absent (contrôles négatifs), (2) les cas où l'effet du processus à isoler est présent (contrôles positifs).

Les paramètres des modèles d'évolution doivent être conditionnés aux données pour permettre l'inférence phylogénétique : ce qui implique l'utilisation d'une méthode d'inférence (e.g., par maximum de vraisemblance ou encore par le calcul bayésien) et d'une implémentation qui prend la forme d'un programme (e.g., MrBayes [Ronquist and Huelsenbeck, 2003], PhyML[Guindon et al., 2010], Phylobayes MPI [Lartillot et al., 2013a], RAxML [Stamatakis, 2014]). À noter que si aucun modèle d'évolution n'est parfait, aucune implémentation n'est parfaite. Et tout ce dispositif expérimental devient de plus en plus onéreux en ressources computationnelles par la complexification des modèles d'évolution.

Par exemple, avec le développement de l'approche par super matrice (phylogénomique) il a fallu paralléliser le code pour pouvoir analyser des jeux de données de plusieurs millions de positions (e.g., [Irisarri et al., 2017]). Dans le contexte des analyses phylogénétiques, très souvent c'est le calcul de la vraisemblance qui est parallélisé, puisque celui-ci est fait en chacun des sites indépendamment dans un premier temps. Au final il suffit de faire le produit des vraisemblances calculées pour obtenir la vraisemblance de la super matrice pour un modèle phylogénétique donné. Mais la vraisemblance sera différente selon que l'analyse ait été faite sur un seul coeur ou sur plusieurs coeurs en parallèle, et cela dû à la propagation de l'erreur d'arrondie [Darriba et al., 2018]. Les analyses bayésiennes seraient plus robustes à ce type de problèmes par le fait que l'erreur serait répartie sur l'ensemble des valeurs qui forment la distribution *a posteriori*, et non pas cumulée sur une seule valeur comme dans le cas du maximum de vraisemblance [Darriba et al., 2018].

D'autre part, au vu des millions d'heures de calculs nécessaires pour réaliser une seule analyse phylogénomique selon les critères de la science actuelle, l'investissement dans l'optimisation du code reste néanmoins relativement faible par rapport au coût des analyses (millions de dollars canadiens) ; ce sont évidement les programmes les plus utilisés comme Phylobayes MPI [Lartillot et al., 2013a] ou RAxML [Stamatakis, 2014] qui profitent de l'aide au développement [Darriba et al., 2018]. D'un autre côté, la majorité des programmes ont très peu d'aide au développement. Cet état de fait n'est certainement pas profitable à l'avancement du savoir puisque les risques de bogue sont plus élevés. L'utilisation de pipe-lines se généralise en bio-informatique. Or la plupart des programmes intégrés dans ces pipe-lines recevront très peu d'aide au développement leur permettant d'atteindre les niveaux du génie logiciel, la probabilité d'avoir un bogue est d'autant plus importante. Pour limiter ce risque, doit-on développer indépendamment le même programme plusieurs fois par des personnes différentes, comme l'avait fait Yang and Nielsen [2008] lors du développement des modèles FMutSel et FMutSel0 disponibles dans CodeML de la suite PAML [Yang, 2007b]. Un autre exemple significatif sur les risques d'utilisation de programmes multiples dans des pipe-lines est celui de [Smith et al., 2011]. Après avoir filtré les données génomiques servant à la construction d'une super matrice au moyen d'un programme publié [Smith and Dunn, 2008], les auteurs ont obtenu une matrice dans laquelle six acides aminés (E, F, I, L, P et Q) ont été remplacés par des données manquantes [Smith et al., 2012]. Dans ce contexte Phylobayes MPI, le programme d'inférence phylogénétique utilisé, reconnaissait les alignements comme étant des alignements d'ADN avec des états ambigus (e.g., Y pour les pyrimidines et R pour les purines). Les auteurs ont alors modifié le code de Phylobayes MPI, pensant avoir découvert un bogue, pour forcer le programme à reconnaître les alignements étudiés comme des alignements de séquences d'acides aminés. Il est navrant de noter que moins de 10% des scientifiques ayant cité l'article initial [Smith et al., 2011] auront également cité le *corrigendum* [Smith et al., 2012].

Nous avons fait un bref survol de la phylogénie. L'avantage de travailler dans un cadre probabiliste est qu'il permet de tester des hypothèses dans un cadre statistique formel. La systématique est le premier domaine d'utilisation de la phylogénie au sens de son importance. Les modèles d'évolution déployés en systématique sont dits phénoménologiques, ils n'ont pas l'ambition de modéliser le processus mutationnel, le processus de sélection, ni la fixation des

mutations dans les populations [Rodrigue and Philippe, 2010] ; au contraire des modèles de type mutation-sélection. Cependant, les modèles de type mutation-sélection ne sont pour l'instant utilisés que dans le cadre de la détection des facteurs déterminants l'évolution des séquences codantes, et non pas pour déterminer la systématique des systèmes biologiques.

1.2.2. Modèles utilisés en phylogénie

L'inférence phylogénétique a pour but de retrouver l'ordre des évènements de spéciation ainsi que les distances qui les séparent. L'ordre des évènements de spéciation correspond à la topologie de l'arbre phylogénétique alors que les distances correspondent aux nombres de changements. C'est la partie la plus mécanistique des modèles d'évolution utilisés en phylogénie actuellement. Alors que la paramétrisation du modèle d'évolution qui suppose une **matrice de substitution** est phénoménologique, car la fixation des mutations dans les populations (i.e., les substitutions) n'est pas explicitement modélisée.

La matrice de taux, Q, représente les taux de changement dans l'espace des états. De plus, étant donné que les histoires évolutives proposées sont en fait une série de changements d'états le long de l'arbre phylogénétique, ces histoires sont modélisées tel un processus continu (processus de Markov). Cette astuce permet de calculer la probabilité de passer d'un état l à un état m sans avoir a prendre en compte ce qui s'est passé précédemment. Quand le processus est supposé indépendant du temps, il est alors question d'un **processus non-stationnaire**. Un modèle stationnaire peut être réversible, ce qui implique que les taux de changement entre états soient réversibles en temps $l > m = m > l$, une hypothèse fort utile au calcul. Ce système permet de calculer la probabilité de passer de l'état l à l'état m pour un temps d'attente donné, t sans avoir à se soucier de l'emplacement de la racine, puisque celle-ci pourra être définie *a posteriori* grâce au positionnement du groupe extérieur (out-group).

Comme nous l'avons vu plus tôt, les longues branches peuvent être problématiques, faisant référence au problème de l'attraction des longues branches. Mais les évènements de spéciation anciens séparés par de courtes branches (e.g., [Laurin-Lemay et al., 2012]), dus à des évènements rapides de diversification, par exemple [Esselstyn et al., 2017], sont aussi une source de travail pour les phylogénéticiens, car très difficiles à résoudre. Ceci est dû au fait que très peu de signal évolutif est identifié par les modèles d'évolutions utilisés. Ce qui

se traduit par de faibles supports dans les noeuds en question, car l'incertitude sur les séquences ancestrales est trop importante pour permettre de distinguer l'ordre des évènements de spéciation.

Trois aspects, en lien avec l'identification du signal évolutif, pourront alors améliorer l'inférence phylogénétique, voire dénouer l'incertitude : (1) la quantité de données (nombre de positions), (2) la qualité de ces dernières (sans contamination et avec le moins de données manquantes possible), et (3) l'utilisation des meilleurs modèles d'évolution disponibles afin d'identifier le maximum de signal évolutif (sensibilité) et réduire les risques de violations de modèles (spécificité) [Philippe et al., 2011, Philippe and Roure, 2011, Roure et al., 2013]. Mais encore là, il n'existe aucune garantie sur l'issue de l'inférence phylogénétique (résolution ou pas du problème phylogénétique), tout simplement par l'absence d'un signal évolutif assez important. Le défi actuel de l'inférence phylogénétique est d'obtenir des jeux de données propres et de proposer des modèles d'évolution qui permettent d'identifier un signal évolutif qui va s'avérer représentatif de l'histoire évolutive moyenne des taxons étudiés. Si le modèle d'évolution n'est pas une hypothèse qui s'applique à l'ensemble des taxons étudiés lors de l'inférence phylogénétique, il y a des risques de violations de modèle.

En phylogénie, les modèles d'évolution diffèrent essentiellement par leur paramétrisation. Par exemple, le modèle JC [Jukes and Cantor, 1969] n'a que la topologie et les longueurs de branche comme paramètres. Quant à lui, le modèle K80 prend en compte la différence dans le taux de transition par rapport au taux de transversion [Kimura, 1980] au moyen d'un seul paramètre, κ. Le modèle F81, permet de prendre en compte une préférence globale au niveau des nucléotides, en ajoutant trois paramètres libres [Felsenstein, 1981], les fréquences stationnaires, notées $\{\pi_A, \pi_C, \pi_G, \pi_T\}$, et où $\pi = (\pi_n)_{1 \leq n \leq 4}$ avec $\sum_{n=1}^{4} \pi_n = 1$. Les modèles HKY [Hasegawa et al., 1985] et GTR [Lanave et al., 1984, Tavare et al., 1997, Rodriguez et al., 1990], sont probablement les modèles les plus connus pour les nucléotides. Le modèle HKY, comme le modèle K80, prend en compte la différence dans le taux de transition par rapport au taux de transversion, tout en permettant une préférence spécifique à chacun des nucléotides comme le modèle F81. Le modèle GTR [Lanave et al., 1984, Tavare et al., 1997] permet quant à lui de modéliser les douze types de substitutions simples avec huit paramètres libres. La fréquence des nucléotides, trois paramètres, puis cinq autres

paramètres libres sont utilisés pour modéliser l'ensemble des taux de transitions et de transversions : $\{\rho_{AC}, \rho_{AG}, \rho_{AT}, \rho_{CG}, \rho_{CT}, \rho_{GT}\}$ qui se définissent comme suit $\rho = (\rho_{lm})_{1 \leq l,m \leq 4}$, avec $\sum_{1 \leq l < m \leq 4} \rho_{lm} = 1$.

Sous le modèle GTR, dans un contexte nucléotidique, la matrice de taux (figure 1.2.1) est dénotée de la manière suivante :

$$Q = \begin{pmatrix} & A & C & G & T \\ A & -\pi_A(\sum_{l \neq m} \rho_{lm}) & \pi_A \rho_{AC} & \pi_A \rho_{AG} & \pi_A \rho_{AT} \\ C & \pi_C \rho_{CA} & -\pi_c(\sum_{l \neq m} \rho_{lm}) & \pi_C \rho_{CG} & \pi_C \rho_{CT} \\ G & \pi_G \rho_{GA} & \pi_G \rho_{GA} & -\pi_G(\sum_{l \neq m} \rho_{lm}) & \pi_G \rho_{GT} \\ T & \pi_T \rho_{TA} & \pi_T \rho_{TC} & \pi_T \rho_{TG} & -\pi_T(\sum_{l \neq m} \rho_{lm}) \end{pmatrix}, \quad (1.2.1)$$

où chaque taux s'obtient de la manière suivante $Q_{lm} = \pi_l \rho_{lm}, l \neq m$, et où les valeurs de la diagonale sont obtenues suivant $Q_{ll} = -\sum_{l \neq m} Q_{lm}$, où $1 \leq l,m \leq 4$. L'exponentiation de la matrice de taux permet de calculer la probabilité de passer de l'état l à l'état m après un temps d'attente t, dénoté par $P_{l \to m}(t) = e^{tQ_{lm}}$. Les longueurs de branche doivent être normalisées au moyen d'un facteur de normalisation, Z, de manière à s'assurer qu'elles soient représentatives de tous les types de substitutions. Ce qui est dénoté par $Z_Q = 2 \times \sum_{1 \leq l,m \leq 4} \rho_{lm} \pi_l \pi_m$.

Les modèles d'évolution sont conçus pour identifier certaines hétérogénéités. De manière générale l'hétérogénéité présente dans les données peut se décliner sur le plan d'une hétérogénéité (1) le long de l'alignement ou (2) au cours du temps, mais j'ajouterais la classe des hétérogénéités dues à (3) des interdépendances entre sites qui possèdent leurs spécificités d'implémentation et qui en font une classe à part. L'hétérogénéité à identifier peut aussi être causée par la présence de (4) transfert horizontaux.

L'**hétérotachie** est une hétérogénéité connue des phylogénéticiens depuis plus de 40 ans [Fitch, 1971b, Fitch and Markowitz, 1970] qui consiste en une variation du taux d'évolution au cours du temps. L'hétérotachie a mené au développement de modèles de type covarion (e.g., Galtier and Gouy [1995], Huelsenbeck [2002], Zhou et al. [2010]), où le taux d'évolution varie au cours du temps, en d'autres mots le taux évolue. Mais l'hétérogénéité peut aussi être dans la nature des changements au cours du temps, ou par la préférence en acides aminés, il est alors question d'**hétéropécillie** [Roure and Philippe, 2011], ou encore par une variation du taux de GC [Gueguen and Duret, 2018]. Le modèle BP (Break-Point) permet de prendre

en compte une variation dans la préférence des acides aminés au cours du temps de manière globale et non-stationnaire [Blanquart and Lartillot, 2008]. Des changements dans la préférence des acides aminés au cours du temps suggèrent un changement au niveau du processus mutationnel et/ou de sélection au cours du temps. Mais, ces modèles phénoménologiques ne permettent pas de tester explicitement des hypothèses sur les processus qui pourraient expliquer cette variation de préférence d'acides aminés au cours du temps ; il faut alors se rabattre sur les modèles d'évolution moléculaire, plus mécanistiques. Lors de transferts horizontaux, l'arbre de certains gènes ne correspondra pas à l'arbre d'espèces. La manière la plus simple pour prendre en compte cette hétérogénéité de l'évolution des génomes dans un contexte d'évolution des séquences codantes consiste à détecter les séquences issues des transferts horizontaux, pour ensuite les retirer des alignements des gènes étudiés. Mais souvent les supports pour les arbres de gènes sont faibles, ce qui rend la tâche d'identification des transferts horizontaux un peu plus difficile lorsque l'arbre simple gène est comparé à l'arbre d'espèce. Par contre, il est possible aussi de modéliser les transferts horizontaux (e.g.,[Daubin and Szoellosi, 2016]), ce qui peut être utile pour corroborer des évènements de spéciation [Davin et al., 2018]. À l'échelle des Vertébrés les transferts horizontaux sont peu nombreux, mais jouent un rôle certain dans la détermination des caractéristiques de ces derniers. Par exemple, à la base des mammifères placentaires, un virus syncytial aurait transmis un gène, ce qui aurait permis l'apparition du placenta empêchant le système immunitaire d'attaquer le foetus [Haig, 2012]. Sur une plus grande échelle, celle des eucaryotes, la mitochondrie est un organelle lui-même issu d'une endosymbiose, un type de transfert horizontal [Roger et al., 2017].

Au niveau de la modélisation, identifier des aspects de l'hétérogénéité qui se présentent au long de l'alignement est sans nul doute ce qui a fait le plus avancer l'inférence phylogénétique dans le contexte des séquences codantes : pour cause le rôle central de la structure primaire dans la détermination des structures protéiques/ARN de plus hauts niveaux (e.g., ARN structural [Meyer et al., 2019]). Dès le début des années 1990, une distribution gamma discrétisée en catégories (2-16) est utilisée pour modéliser des taux de changement variables entre sites [Yang, 1993]. Les variations dans le taux de changement entre sites identifiées par ce modèle phénoménologique peuvent être dues à des variations dans les contraintes de sélection négative. Dans ce cas, le taux de changement sera inversement proportionnel à

l'importance de la sélection négative. Plus récemment, le modèle CAT [Lartillot and Philippe, 2004] propose d'utiliser un processus de mélange basé sur une distribution multivariée de Dirichlet pour modéliser la préférence site-spécifique en acides aminés. Cette modélisation phénoménologique rend non seulement compte des contraintes biochimiques qui affectent les protéines de manière site-spécifique mais permet aussi de moduler le taux de changement en chacun des sites. Un site extrêmement conservé aura un profil très piqué pour un acide aminé, n'acceptant de ce fait que très peu de types de changements au cours de l'évolution de la protéine.

La phylogénie s'intéresse aux branches internes de l'arbre phylogénétique, alors que les modèles d'évolution moléculaire s'intéressent à l'arbre phylogénétique dans son entièreté (e.g., prédire des séquences ancestrales, mais aussi les états aux feuilles). Les modèles proposés dans le domaine de l'évolution moléculaire cherchent à reconstruire ce qui est observé aux feuilles, pour identifier les processus biologiques en jeu. Ils peuvent aussi servir à dresser un portrait auquel pouvait ressembler l'ancêtre des espèces étudiées. Les modèles de type mutation-sélection sont tous dessinés pour cette tâche. Ce sont des modèles dits mécanistiques, car ils ont pour objectif de tester des hypothèses quant au rôle des processus mutationnels et de sélection dans la détermination de ce qui est observé dans les données génomiques.

1.2.3. Modèles mutation-sélection utilisés en évolution moléculaire

Les modèles mutation-sélection font partie d'une grande famille de modèles regroupés sous l'appellation "origin-fixation" qui comprend des modèles de génétique des populations et des modèles phylogénétiques (revue dans [McCandlish and Stoltzfus, 2014]). Dans notre cas, ce sont des modèles phylogénétiques d'évolution moléculaire qui sont utilisés. L'appellation "origin-fixation" provient du fait que la paramétrisation de ces modèles s'appuie sur les principes de la génétique des populations où la sélection s'applique sur les mutations qui apparaissent au cours de l'évolution. La paramétrisation du processus mutationnel détermine la manière dont le paysage adaptatif peut être exploré. Les modèles mutation-sélection permettent de tester explicitement des hypothèses concernant les processus mutationnels et de sélection à l'oeuvre (revue dans [McCandlish and Stoltzfus, 2014]). Les modèles mutation-sélection intègrent les processus mutationnels et de sélection via un processus de substitution qui nécessite le calcul des probabilités de fixation des mutations qui prend la forme d'une

matrice de substitution. Dans ce cas la séquence codante intègre certains processus de mutation et de sélection qui ont pu se produire dans les populations des espèces étudiées. Détecter la sélection est une tâche statistique qui consiste à identifier une déviation par rapport à ce qui est attendu sous le modèle mutationnel. De plus, l'avantage de ce type de modèle est qu'à tout instant, c'est-à-dire le long de l'arbre phylogénétique, il est possible de calculer le coefficient de sélection moyen de la séquence codante ou encore d'un codon spécifique. Les modèles mutation-sélection se distinguent entre eux principalement par les stratégies utilisées pour modéliser la sélection. Les modèles de type mutation-sélection sont dits de sélection négative ou purificatrice, au sens de la théorie quasi-neutre de l'évolution.

Les modèles à codon de type mutation-sélection prennent en compte la structure du code génétique et donc la nature des substitutions, synonymes et non-synonymes, pour inférer l'évolution des protéines. Afin de réduire le temps calcul, les modèles d'évolution moléculaire utilisent habituellement une topologie inférée dans une étape précédente. Mais rien n'empêche d'inférer la topologie de manière jointe aux autres paramètres du modèle. Dans le cas des modèles à codon qui nous intéressent, la matrice de substitution possède généralement une dimension de 61×61 (code génétique universel) ; les codons d'arrêt ont une probabilité nulle d'exister selon ces modèles d'évolution, mais cela pourrait être autrement, par exemple pour tenir compte de l'édition des ARNm.

La paramétrisation du processus mutationnel a très peu évolué au cours des 30 dernières années (e.g., [Muse and Gaut, 1994b, Yang and Nielsen, 2008], elle sera identifiée de la manière suivante : M[parametrisation]. La majorité des modèles d'évolution n'acceptent que des substitutions simples ; mais quelques articles s'intéressent aux mutations doubles et triples ainsi que leurs effets sur les modèles prenant en compte que les mutations simples [Doron-Faigenboim and Pupko, 2007, Kosiol et al., 2007, De Maio et al., 2013, Miyazawa, 2011, Pouyet et al., 2016, Zoller and Schneider, 2012, Jones et al., 2018, Dunn et al., 2019, Venkat et al., 2018]. La plupart des analyses publiées ont été faites avec la paramétrisation M[HKY] tirée du modèle d'évolution HKY [Hasegawa et al., 1985]. L'avantage principal du modèle M[HKY] est d'avoir un seul paramètre d'échangeabilité, κ, rendant compte de l'hétérogénéité principale du processus mutationnel, à savoir un taux de mutation plus grand pour les transitions que les transversions. Le terme échangeabilité est utilisé ici pour identifier les paramètres relatifs aux changements entre états (e.g., nucléotides). Ce modèle prend aussi en

compte le biais mutationnel global via les propensions nucléotidiques qui sont définies de la manière suivante : $\{\varphi_A, \varphi_C, \varphi_G, \varphi_T\}$, où $\varphi = (\varphi_n)_{1 \leq n \leq 4}$ avec $\sum_{n=1}^{4} \varphi_n = 1$. Les propensions nucléotidiques permettent de capturer l'enrichissement "universel" en AT [Keightley et al., 2009, Lynch et al., 2008, Hershberg and Petrov, 2010, Hildebrand et al., 2010, Ossowski et al., 2010]. C'est probablement la seconde hétérogénéité la plus importante à prendre en compte. Alors que la paramétrisation M[GTR], tirée du modèle GTR [Lanave et al., 1984, Tavare et al., 1997], est plus riche mais beaucoup moins utilisée. Les propensions nucléotidiques sont communes aux deux paramétrisations M[HKY] et M[GTR]. Contrairement à M[HKY], M[GTR] possède plusieurs paramètres d'échangeabilité, $\{\varrho_{AC}, \varrho_{AG}, \varrho_{AT}, \varrho_{CG}, \varrho_{CT}, \varrho_{GT}\}$ qui se définissent comme $\varrho = (\varrho_{lm})_{1 \leq l,m \leq 4}$, avec $\sum_{1 \leq l < m \leq 4} \varrho_{lm} = 1$. Au final 8 paramètres libres sont disponibles pour modéliser un processus mutationnel qui se décline en douze échanges possibles $A > C$, $A > G$, $A > T$, $C > A$, $C > G$, $C > T$, $G > A$, $G > T$, $G > C$, $T > A$, $T > C$, $T > G$. La paramétrisation M[HKY] est mise en l'avant dans la suite PAML [Yang, 2007b]. Alors que la paramétrisation M[GTR] est implémentée par défaut dans les modèles de la suite Phylobayes MPI [Lartillot et al., 2013a]. Les longueurs de branches sont en nombre de mutations attendues sous le modèle mutationnel. Les longueurs de branche doivent être normalisées au moyen d'un facteur de normalisation, Z, de manière à s'assurer qu'elles soient représentatives de tous les types de mutations. Ce qui est dénoté par $Z_{mut} = 2 \times \sum_{1 \leq l,m \leq 4} \varrho_{lm} \varphi_l \varphi_m$. La sélection négative aura pour conséquence d'abaisser le taux de substitions, alors que la sélection positive aura pour conséquence d'augmenter le taux de substitutions.

Le développement des modèles mutation-sélection s'est principalement fait sur la paramétrisation de la sélection. Cette paramétrisation sera identifiée de la manière suivante : S[parametrisation]. Une première génération de modèles a vu le jour dans les années 1990 [Muse and Gaut, 1994a, Goldman and Yang, 1994]. Ces travaux ont mené à des modèles utilisant exclusivement le paramètre ω, soit le rapport du taux de substitutions non-synonymes au taux de substitutions synonymes (dN/dS), pour identifier la sélection, négative ($\omega < 1$) ou positive ($\omega > 1$). Tous les modèles n'utilisant que le paramètre ω pour modéliser la sélection sont des modèles phénoménologiques puisque la matrice de substitution est en fait une matrice de taux d'échange comme celles utilisées en phylogénie, aucune probabilité de fixation

n'est calculée. Le modèle M[GTR]-S[ω], inspiré de [Muse and Gaut, 1994a], est implémenté dans Phylobayes MPI et se définit de la manière suivante :

$$Q_{ab} = \begin{cases} \varrho_{a_c b_c} \varphi_{b_c}, & \text{si synonyme,} \\ \varrho_{a_c b_c} \varphi_{b_c} \omega, & \text{si non-synonyme,} \end{cases} \tag{1.2.2}$$

où b_c renvoie un indice, de 1 à 4, du nucléotide trouvé à la $c$ième position du codon b.

Sachant que la sélection est un processus extrêmement hétérogène en raison de la diversité des contraintes biochimiques, biophysiques et structurelles qui peuvent agir sur les protéines [Liberles et al., 2012], les chercheurs ont développé des modèles d'évolution toujours de plus en plus riches. Les articles fondateurs [Muse and Gaut, 1994a, Goldman and Yang, 1994] ont inspiré de nombreux développements. Par exemple, parmi les modèles les plus connus, M1a et M2a, sont utilisés pour détecter l'importance de la sélection positive sur la base d'un test de rapport de vraisemblance [Yang, 2007b]. Le modèle M1a modélise la proportion des sites qui seront sous forte contrainte de sélection négative ($\omega_0 < 1$) ainsi que la proportion des sites qui sont dans un régime neutre ($\omega_1 = 1$). Deux paramètres libres sont alors requis : ω_0 et la proportion des sites sous contrainte de sélection négative p_0, puisque p_1, la proportion des sites sous régime de sélection neutre est déduite avec $p_1 = 1 - p_0$. Le modèle M2a permet la détection de sélection positive. Il possède quatre paramètres libres. Il y a la proportion de sites sous régime de sélection négative (p_0) qui seront associés à un paramètre $\omega_0 < 1$. Une certaine proportion de sites (p_1) seront associés à un régime neutre ($\omega_1 = 1$), ainsi qu'une proportion de sites $p_2 = 1 - p_0 - p_1$ associés au régime de sélection positive ($\omega_2 > 1$). Pour chacun des sites de l'alignement, $i = 1 \leq i \leq N$, où N est la longueur du gène (en codons), le logarithme de la vraisemblance d'une séquence est obtenu de la manière suivante dans le cas du modèle M1a :

$$\mathcal{L} = \sum_{1 \leq i \leq N} log(p_0 P(D_i|\omega_0,\theta) + p_1 P(D_i|\omega_1,\theta)), \tag{1.2.3}$$

où D_i correspond au site (en codons) i de l'alignement, et où θ inclus les paramètres mutationnels, les longueurs de branches et la topologie de l'arbre. De la manière suivante dans le cas du modèle M2a :

$$\mathcal{L} = \sum_{1 \leq i \leq N} log(p_0 P(D_i|\omega_0,\theta) + p_1 P(D_i|\omega_1,\theta) + p_2 P(D_i|\omega_2,\theta)). \tag{1.2.4}$$

À remarquer que cette vraisemblance est en fait une moyenne pondérée de la vraisemblance sous chaque valeur ω. Les valeurs optimales des paramètres (ω et proportions) sont typiquement obtenues par maximum de vraisemblance. Les mêmes auteurs [Yang, 2007b] ont aussi proposé un modèle de mélange avec de plus grands nombres de composantes, K, mais cette fois tous les paramètres ω sont libres de modéliser tous les régimes de sélection possibles, ce qui est une grande amélioration par rapport aux implémentations précédentes. Les K paramètres sont requis pour modéliser l'ensemble des paramètres ω et $K-1$ paramètres sont requis pour modéliser les proportions, tous comme les paramètres M1a et M2a. Ce qui intéresse le plus les chercheurs est de pouvoir connaître quels sont les sites sous régime de sélection négative, neutre et positive et de corréler cette information à des processus biologiques. Il est possible d'estimer la probabilité *a posteriori* qu'un site i appartienne à la classe K via estimation bayésienne empirique [Yang, 2007b]. D'autres modèles de sélection existent pour détecter la sélection positive, par exemple, dans une branche spécifique de l'arbre phylogénétique [Yang, 2007b].

Par la suite, Yang and Nielsen [2008] ont proposé de modéliser la préférence globale en acides aminés avec un profil de valeurs adaptatives de dimension 20. Le modèle se différencie des précédents, car il permet de calculer la probabilité de fixation des mutations. Le vecteur de préférence en acides aminés est dénoté par $\psi = (\psi_l)_{1 \leq l \leq 20}$, où le profil somme à l'unité, $1 = \sum_{1 \leq l \leq 20} \psi_l$, et 19 paramètres sont libres. Le modèle incorpore aussi un paramètre ω pour identifier la sélection négative ou positive en conjoncture avec le profil de préférence en acides aminés. Pour les changements non-synonymes, du codon a à b, des coefficients de sélection (voir [Rodrigue et al., 2010]) sont alors calculés comme suit :

$$S_{ab} = \ln(\frac{\psi_{f(b)}}{\psi_{f(a)}}), \tag{1.2.5}$$

où $f(a)$ retourne un indice, de 1 à 20, de l'acide aminé codé par le codon a. La valeur S_{ab}, à son tour, définit un facteur de fixation, désigné $h(S_{ab})$, et calculé comme suit :

$$h(S_{ab}) = \frac{S_{ab}}{1 - e^{-S_{ab}}}. \tag{1.2.6}$$

Le modèle M[GTR]-S[1CatAA*], inspiré de [Yang and Nielsen, 2008], est implémenté dans Phylobayes MPI et se définit de la manière suivante :

$$Q_{ab} = \begin{cases} \varrho_{a_c b_c} \varphi_{b_c}, & \text{si synonyme,} \\ \varrho_{a_c b_c} \varphi_{b_c} \omega_* h(S_{ab}), & \text{si non-synonyme} \quad . \end{cases} \tag{1.2.7}$$

Mais, la grande limitation de ce modèle est de ne pas prendre en compte la préférence site-spécifique en acides aminés. C'est en fait un raccourci phénoménologique, car quel est le sens biologique d'un usage global des acides aminés ? Les mêmes auteurs Yang and Nielsen [2008] proposent un deuxième modèle afin d'identifier la sélection sur l'usage des codons via un profil de valeurs adaptatives, de dimension 61, donc sans les codons d'arrêt. Le vecteur de préférence des codons est donc une extension du vecteur ψ, maintenant dénoté $\psi = (\psi_l)_{1 \leq l \leq 61}$, où le profil somme à l'unité, $1 = \sum_{1 \leq l \leq 61} \psi_l$, 60 paramètres sont donc libres. Le modèle incorpore aussi un paramètre ω_* pour capturer potentiellement la sélection positive dans un contexte où la sélection négative aurait été identifiée par le profil de préférence d'usage des codons. L'astérisque apparait afin de faire distinction avec un paramètre qui représente le rapport dN/dS (ω).

Le modèle M[GTR]-S[1CatCodon*], inspiré de [Yang and Nielsen, 2008], est implémenté dans Phylobayes MPI et se définit de la manière suivante :

$$Q_{ab} = \begin{cases} \varrho_{a_c b_c} \varphi_{b_c} h(S_{ab}), & \text{si synonyme,} \\ \varrho_{a_c b_c} \varphi_{b_c} \omega_* h(S_{ab}), & \text{si non-synonyme} \quad . \end{cases} \tag{1.2.8}$$

Par contre, utiliser le même vecteur de paramètres de valeurs adaptatives pour identifier le signal évolutif relatif à la sélection sur l'usage des acides aminés ainsi que la sélection sur l'usage des codons est potentiellement limitatif. C'est dans l'optique de mieux identifier ces deux aspects du signal évolutif, et surtout celui en lien avec la sélection sur l'usage des codons, que Pouyet et al. [2016] proposent de modéliser la sélection sur les évènements non-synonymes à partir d'un vecteur de valeurs adaptatives de préférence des acides aminés de dimension 20 ainsi que la sélection sur les évènements synonymes à partir d'un vecteur de valeurs adaptatives de préférence des codons de dimension 61. Le vecteur de préférence en acides aminés est dénoté par $\psi = (\psi_l)_{1 \leq l \leq 20}$, où le profil somme à l'unité, $1 = \sum_{1 \leq l \leq 20} \psi_l$, 19 paramètres sont donc libres. Cette paramétrisation du profil de préférence globale des acides aminés est équivalente au modèle M[GTR]-S[1CatAA*]. Puis la paramétrisation de

préférence des codons dénotés par $\psi = (\psi_l)_{1 \leq l \leq 61}$ ne somme plus à un comme pour M[GTR]-S[1CatCodon*], mais à 20, $20 = \sum_{1 \leq l \leq 61} \psi_l$, 41 paramètres sont donc libres. Ceci est dû au fait que pour chacun des acides aminés un profil de valeurs adaptatives est dédié à ses codons synonymes respectifs. Par exemple, $\psi^{Arg} = (\psi_l^{Arg})_{1 \leq l \leq 6}$, où le profil ψ^{Arg} somme à l'unité, $1 = \sum_{1 \leq l \leq 6} \psi_l^{Arg}$, 5 paramètres sont donc libres pour ce profil de valeurs adaptatives des codons synonymes de l'arginine.

Une seconde génération de modèles à codon a été développée à partir du travail de Halpern and Bruno [1998], permettant de modéliser la préférence en acides aminés de manière site-spécifique afin de retrouver les contraintes de sélection. Cette paramétrisation de la sélection est notée par S[NCatAA]. Mais comme dans le modèle CAT [Lartillot and Philippe, 2004], la préférence en acides aminés s'appuie sur un mélange de profils de préférence généré à partir d'un processus de Dirichlet. Il a fallu attendre plus de 10 ans avant de retrouver des implémentations efficaces des idées de Halpern and Bruno [1998], et cela principalement parce que ces modèles demandent d'importantes ressources computationnelles (e.g., [Rodrigue et al., 2010, Tamuri et al., 2012]). Rodrigue and Lartillot [2017] ont par la suite proposé d'ajouter un paramètre ω_* dans le but d'identifié toute déviation au rapport dN/dS alors que la sélection négative était déjà identifiée par la paramétrisation S[NCatAA]. Dans ces conditions, le paramètre ω_* peut plus facilement mener à la détection de régimes de sélection particulier, comme celui de la course aux armements (hypothèse de la Reine-Rouge) qui engendre des patrons de sélection positive [Rodrigue and Lartillot, 2017]. L'hypothèse de la Reine-Rouge décrit une dynamique de coévolution entre hôtes et parasites ou entre proies et prédateurs (e.g.,[Raffel et al., 2008]) ou encore au niveau intra-génome (e.g., [Latrille et al., 2017]). La sélection qu'introduit le parasite sur la population de son hôte aura pour conséquence d'augmenter la résistance de l'hôte à son parasite à la génération suivante, à son tour cette nouvelle résistance de l'hôte aura pour conséquence de sélectionner les parasites les plus virulents.

La paramétrisation S[NCatAA*] implique un ensemble de K vecteurs, avec chacun 20 entrées correspondant aux préférences en acides aminés (aussi appelés **profils**), dénotés $\psi = (\psi_l^{(k)})_{1 \leq l \leq 20, 1 \leq k \leq K}$ et où chacun des profils somme a l'unité, $1 = \sum_{1 \leq l \leq 20} \psi_l^{(k)}$. L'implémentation du modèle implique un algorithme utilisant une variable d'allocation notée $z = (z_i)_{1 \leq i \leq N}$, où N est la longueur du gène (en codons) ; pour un site donné i, z_i retourne

un index de 1 à K, précisant le profil en acides aminés opérant sur ce site. Pour les changements non-synonymes, du codon a à b au site i, des coefficients de sélection (voir [Rodrigue et al., 2010]) sont calculés comme suit :

$$S_{ab}^{(i)} = \ln(\frac{\psi_{f(b)}^{(z_i)}}{\psi_{f(a)}^{(z_i)}}), \tag{1.2.9}$$

où $f(a)$ retourne un indice, de 1 à 20, de l'acide aminé codé par le codon a. La valeur $S_{ab}^{(i)}$, à son tour, définit un facteur de fixation, désigné $h(S_{ab}^{(i)})$, et calculé comme suit :

$$h(S_{ab}^{(i)}) = \frac{S_{ab}^{(i)}}{1 - e^{-S_{ab}^{(i)}}}, \tag{1.2.10}$$

qui sera ensuite utilisé directement dans la matrice de substitution. L'ensemble des profils d'acides aminés K, la variable d'allocation z, et la valeur de K elle-même, sont des variables aléatoires du processus de Dirichlet [Rodrigue et al., 2010]. C'est le modèle de référence utilisé dans cette thèse, M[GTR]-S[NCatAA*], qui est implémenté dans Phylobayes MPI et se définit de la manière suivante :

$$Q_{ab}^{(i)} = \begin{cases} \varrho_{a_c b_c} \varphi_{b_c}, & \text{si synonyme,} \\ \varrho_{a_c b_c} \varphi_{b_c} h(S_{ab}^{(i)}) \omega_*, & \text{si non-synonyme.} \end{cases} \tag{1.2.11}$$

Certains auteurs ont aussi cherché à identifier l'hétérogénéité liée à des interdépendances entre sites, en lien par exemple avec la sélection sur la structure des protéines [Robinson et al., 2003]. Prendre en compte des interdépendances entre sites augmente non seulement l'espace des paramètres à échantillonner, mais nécessite aussi de développer une alternative au calcul de la vraisemblance usuellement utilisé en phylogénie, car le calcul de la vraisemblance se fait de manière site indépendant [Felsenstein, 1973, 1981]. Il existe des stratégies pour calculer la vraisemblance en prenant en compte des interdépendances dans les données, mais elles sont difficiles à mettre en place.

Par exemple, les auteurs de [Rodrigue et al., 2009] proposent un modèle de substitution qui prend en compte les contraintes liées à la structure des protéines (e.g., contact et accessibilité au solvant) en utilisant un potentiel pseudo-énergétique $G(s)$ développé par [Kleinman et al., 2006]. Ce potentiel $G(s)$ d'une séquence codante $s = (s_i) 1 \leq i \leq N$, où N est la longueur de la séquence (en codons), est obtenue par l'équation suivante :

$$G(s) = \sum_{1 \leq i < j \leq N} \Delta_{ij} \epsilon_{f(s_i) f(s_j)} + \sum_{1 \leq i \leq N} \Xi^{w_i}_{f(s_i)} + \sum_{\sum_{1 \leq i \leq N}} \mu_{f(s_i)}, \tag{1.2.12}$$

où $f(s_i)$ retourne le codon a à la position i de la séquence s. Le symbole Δ du premier terme correspond à une matrice de dimension $N \times N$, qui définit la présence et l'absence de contact au niveau de la structure tertiaire de la protéine entre les acides aminés en position i et j de la séquence s. Lorsqu'il y a contact, $\Delta_{ij} = 1$ et sinon $\Delta_{ij} = 0$. Le symbole ϵ du premier terme correspond à une matrice de dimension 20×20 qui permet de retourner le potentiel pseudo-énergétique de chacune des paires d'acides aminés possibles. Le symbole Ξ du deuxième terme correspond à une matrice de dimension 14×20 qui permet de retourner le potentiel d'accessibilité au solvant d'un acide aminé l à la position i et où w_i correspond à l'index d'une des 14 classes accessibilité au solvant. Le symbole μ du dernier terme correspond à un vecteur de dimension 20 qui permet de définir le potentiel pseudo-énergétique de chacun des acides aminés. La matrice de substitution $20^N \times 20^N$ est formulée de la manière suivante lorsque la paramétrisation M[GTR] est utilisée :

$$R_{ss'} = \begin{cases} \varrho_{s_{i_c} s'_{i_c}} \varphi_{s'_{i^c}}, & \text{si synonyme,} \\ \varrho_{s_{i_c} s'_{i_c}} \varphi_{s'_{i^c}} \omega e^{\beta(G(s) - G(s'))}, & \text{si non-synonyme,} \end{cases} \tag{1.2.13}$$

où s_{i_c} renvoie le nucléotide à la position c du codon i de la séquence s. Le symbole β correspond à un modulateur du potentiel pseudo-énergétique.

Mais comme ce n'est pas possible d'estimer la matrice de substitution de dimension $20^N \times 20^N$ efficacement, il est possible d'utiliser l'augmentation de données (stochastic mapping [Nielsen, 2002]). Cette stratégie consiste à conditionner des paramètres site-interdépendants, ϕ, à partir des histoires substitutionnelles L (variables latentes), elles-mêmes conditionnées aux valeurs de paramètres du modèle site-indépendant, θ [Robinson et al., 2003] : ce qui revient à calculer $p(\phi|\theta,D,M)$. Où D et M correspondent aux données et au modèle respectivement. L'algorithme Metropolis-Hastings est utilisé pour échantillonner ϕ. Un récent article propose d'utiliser cette méthode phylogénétique d'augmentation de donnée modéliser une hétérogénéité au cours du temps en GC et ainsi quantifier l'impacte de ce processus sur les rapports de taux de substitutions non-synonymes aux taux de substitutions synonymes [Gueguen and Duret, 2018].

Comme nous l'avons vu dans les deux dernières sections, les modèles phylogénétiques probabilistes sont des hypothèses pour expliquer ce qui est observé dans les données génomiques. Les modèles sont plus ou moins mécanistiques, c'est-à-dire qu'ils incorporent, à des degrés différents, des processus déterminants de l'évolution des séquences codantes. Ils sont aussi plus ou moins sophistiqués. Les modèles de type mutation-sélection sont donc les modèles qui se veulent les plus mécanistiques de par le fait qu'ils possèdent une paramétrisation explicite à des hypothèses sur les processus de mutation et de sélection. Ces hypothèses définissent la manière dont le paysage adaptatif sera exploré et la manière dont les mutations seront sélectionnées dans une perspective de génétique des populations. Ce type de modèle ne modélise pas explicitement l'évolution des populations, mais intègre cette information au niveau du processus de substitution. En d'autres mots, ce genre de modèle ne modélise pas explicitement le polymorphisme, puisque les mutations sont instantanément fixées ou éliminées par la sélection.

1.3. Les systèmes d'inférence

1.3.1. L'inférence par méthode de maximum de vraisemblance

L'inférence par calcul de la vraisemblance consiste à optimiser la fonction de vraisemblance, $p(D|\theta,M)$. Où D représente les données observées, θ le vecteur de paramètres et M représente la manière dont les paramètres du modèle phylogénétique sont agencés. Il existe un ensemble de méthodes d'optimisation (e.g., expectation-maximisation [Dempster et al., 1977], descente de gradient [Robbins and Monro, 1951], chaîne de Markov Monte-Carlo avec recuit simulé [Kirkpatrick et al., 1983]). En maximum de vraisemblance, l'incertitude face aux données est souvent évaluée par bootstrap non-paramétrique [Efron and Tibshirani, 1994]. Cette technique permet de construire les intervalles de confiance pour chacun des paramètres du modèle. Les bootstraps paramétriques sont utilisés pour générer des simulations à partir des valeurs de paramètre inférées par le calcul de la vraisemblance maximale [Efron and Tibshirani, 1994]. Il est alors possible d'évaluer la qualité du modèle à reproduire les données, et des barres d'erreur sur les prédictions sont construites. Les modèles de la suite PAML [Yang, 2007b] sont implémentés en maximum de vraisemblance. Nous utilisons deux modèles de cette suite dans le chapitre suivant.

1.3.2. L'inférence par méthode de calcul bayésien

L'inférence bayésienne ne cherche pas à trouver la valeur optimale de chacun des paramètres du modèle comme dans le cas d'une inférence par maximum de vraisemblance, mais plutôt à obtenir la distribution *a posteriori*. La distribution *a posteriori* se note de la manière suivante $p(\theta|D,M) \propto p(D|\theta,M)p(\theta|M)$, où $p(\theta|M)$ est la distribution *a priori*. Une prémisse de l'inférence bayésienne est de représenter l'état de connaissance sur les paramètres à partir d'une distribution. Le calcul bayésien permet de pondérer la vraisemblance du modèle par des probabilités *a priori*. Celles-ci représentent notre état de connaissance sur les paramètres avant même d'avoir considéré les données. L'inférence bayésienne s'appuie sur le théorème de Bayes (équation 1.3.2) pour obtenir une distribution inconnue, la distribution *a posteriori* :

$$p(\theta|D,M) = \frac{p(D|\theta,M)p(\theta|M)}{p(D|M)}.$$

Les distributions *a priori* non-informatives sont utiles pour tester des hypothèses puisqu'elles permettent d'éviter toute forme de connaissance subjective liée au choix des distributions *a priori*, ce qui facilite l'acceptation et le rejet des hypothèses à tester. Par exemple, dans le cas où nous voulons détecter de la sélection positive ($\omega_* > 1$), il suffit de calculer les intervalles de crébidibilté pour un seuil alpha de 5%, et vérifier que la valeur 1 n'est pas comprise dans cet intervalle.

1.3.3. L'échantillonnage par méthode de Monte-Carlo par chaîne de Markov

Les méthodes de Monte-Carlo par chaîne de Markov (MCMC) permettent d'échantillonner efficacement la distribution *a posteriori*. Parmi les méthodes MCMC, l'algorithme de Metropolis-Hastings [Metropolis et al., 1953, Hastings, 1970] est fréquemment utilisé. Cet algorithme procède en deux étapes qui consistent à simuler des valeurs de paramètres θ^* à partir d'un noyau de perturbation $q(\theta^*|\theta_{t-1})$, où θ_{t-1} est donné en entrée. Puis θ^* est accepté selon le rapport Metropolis-Hastings $\frac{p(D|\theta^*,M)p(\theta^*)q(\theta^*|\theta_{t-1})}{p(D|\theta_{t-1},M)p(\theta_{t-1})q(\theta_{t-1}|\theta^*)}$ tout en assurant une exploration stochastique de la distribution *a posteriori*. À cette dernière étape, lorsque θ^* est accepté, θ_t prend alors la valeur de θ^*, soit la nouvelle valeur.

Algorithm 1 Algorithme Metropolis-Hastings

for t=1 to N-1 **do**
  if t=1 **then**
    Simuler $\theta_{t=1}$ depuis $p(\theta|M)$
  else
    Simuler θ^* depuis $q(\theta^*|\theta_{t-1})$
    Simuler u depuis $Uniforme(0,1)$
    if $u < \frac{p(D|\theta^*,M)q(\theta^*|\theta_{t-1})}{p(D|\theta_{t-1},M)q(\theta_{t-1}|\theta^*)}$ **then**
      $\theta_t = \theta^*$
    else
      $\theta_t = \theta_{t-1}$
    end if
  end if
end for

À long terme la distribution de paramètres visités par l'algorithme correspond à la distribution *a posteriori*. En d'autres mots, la fréquence relative à laquelle l'algorithme visite un intervalle de paramètres quelconque est proportionnelle à la probabilité *a posteriori* contenue dans cet intervalle.

1.3.4. L'échantillonnage par méthode de calcul bayésien approché

Le calcul bayésien approché (Approximate Bayesian Computation) permet d'obtenir la distribution *a posteriori* d'un modèle en contournant le calcul de la vraisemblance, lorsque celle-ci n'est pas accessible analytiquement. Dans ce contexte, l'algorithme d'Acceptation-Rejet consiste à générer des valeurs de paramètre, θ_i^*, à partir de la distribution *a priori*, $p(\theta|M)$, pour ensuite générer N simulations ($Z = z_1, z_2, ..., z_N$) à partir d'un modèle génératif, $p(z|\theta_i^*,M)$. Seules les simulations qui sont identiques aux observations, D, seront acceptées. Cette méthode a pour prémisse que les observations auraient pu être simulées sous le modèle génératif proposé.

Algorithm 2 Algorithme Acceptation-Rejet ABC

Simuler θ_i^* depuis $p(\theta|M)$ pour $1 \leq$ i $\leq$ N, où N est le nombre de simulations

Simuler z_i^* depuis $p(z|\theta_i^*,M)$ pour $1 \leq$ i $\leq$ N

Accepter θ_i^* lorsque $z_i^* = D$ pour $1 \leq$ i $\leq$ N

Il est peut-être un peu trompeur de référer à cet algorithme comme étant approximatif, puisque dans la limite où la taille de l'échantillon tend vers l'infini l'approximation sera exacte (approximatif quand fini). Mais lorsque les données sont complexes, comme dans le cas des alignements de séquence, il est pratiquement impossible de reproduire exactement les données. Le calcul bayésien approché permet de relâcher le critère d'acceptation de l'algorithme en calculant une distance euclidienne, ou pas, entre les pseudo-données simulées obtenues sous le modèle génératif, $p(z|\theta_i^*,M)$, et les données observées, $d(z_i^*, D)$. Les valeurs de paramètre pour lesquelles les distances calculées sont plus petites ou égales à un seuil ϵ, souvent défini *a posteriori* en pratique, seront acceptées pour faire l'approximation de la distribution *a posteriori*, $p(\theta|M,(d(z_i^*,D) \leq \epsilon))$. Lorsque le seuil ϵ tend vers zéro, l'erreur liée à l'approximation tend elle aussi vers zéro. Une autre stratégie utilisée pour relâcher le critère d'acceptation consiste à utiliser des statistiques descriptives (SD) pour n'aborder que certaines dimensions des données observées. Le vecteur de SD est obtenu en appliquant une fonction sur les données observées, $S(D)$, et sur les simulations $S(z_i^*)$. De la même manière, une distance est obtenue $d(S(D)-S(z_i^*))$ pour faire l'approximation de la distribution *a posteriori*, $p(\theta|M,(d(S(D)-S(z_i^*)) \leq \epsilon))$. En fait, cet algorithme fait référence à un algorithme des k-Plus Proches Voisins libre du calcul de la vraisemblance (k-Nearest-Neighbors Likelihood-Free : kNN). Dans un premier temps une table de référence est construite, comprenant l'ensemble des paramètres $\theta = \theta_1, \theta_2,..., \theta_{N_\theta}$, où N_θ est le nombre de paramètres du modèle, ainsi que des statistiques descriptives calculées à partir des simulations $S = s_1, s_2,..., s_{N_s}$, où N_s est le nombre de statistiques descriptives extraites.

$$\begin{pmatrix} \theta_{11} & \theta_{21} & ... & \theta_{N_\theta 1} & s_{11} & s_{21} & ... & s_{N_s 1} \\ \theta_{12} & \theta_{22} & ... & \theta_{N_\theta 2} & s_{12} & s_{22} & ... & s_{N_s 2} \\ ... & ... & ... & ... & ... & ... & ... & ... \\ \theta_{1N} & \theta_{2N} & ... & \theta_{N_\theta N} & s_{1N} & s_{2N} & ... & s_{N_s N} \end{pmatrix}, \tag{1.3.1}$$

où N est le nombre de lignes (simulations) dans la table de référence. L'algorithme k-Plus Proches Voisins libre du calcul de la vraisemblance prend la forme suivante :

Algorithm 3 Algorithme k-Plus Proches Voisins libre du calcul de la vraisemblance

Simuler θ_i^* depuis $p(\theta|M)$ pour $1 \leq i \leq N$, où N est le nombre de simulations

Simuler z_i^* depuis $p(z|\theta_i^*,M)$ pour $1 \leq i \leq N$

Calculer $d_i^* = d(S(z_i^*),S(D))$ pour $1 \leq i \leq N$

Ordonner $d_1^*, ..., d_N^*$

Retourner les θ_i^* qui correspondent aux k plus petites distances

1.3.5. Les modèles de régression

Les modèles de régression servent à prédire les **variables réponses** via une fonction, Y, à partir des **variables explicatives**, $f(X)$. Deux types de paramètres peuvent être utilisés dans les régressions : des poids w et des biais b. Les poids sont multipliés aux variables explicatives, alors que les biais sont additionnés aux précédents produits, comme dans le cas d'une régression linéaire $Y = wX + b$. Différents algorithmes peuvent être utilisés pour calculer les valeurs des poids et des biais (e.g., moindre carré, descente de gradient) à partir d'une fonction de coût qui peut prendre la forme de l'erreur sur la prédiction ou de la vraisemblance du modèle de régression $p(D,X|Y)$. Le modèle de régression le plus simple, lorsque l'intérêt n'est porté que sur Y permet de prédire la moyenne ($Y = b$) : dans ce cas les poids prennent la valeur zéro et le biais reste libre pour inférer la moyenne, ce qui donne $Y = b$.

Parmi les modèles de régression, il y a les modèles de régression linéaire simple (une seule variable explicative) et multiple (plusieurs variables explicatives), les analyses en composantes principales, les arbres de régression (e.g., forêts aléatoires [Breiman, 2001]). Les modèles de régression supervisés permettent de détecter des relations entre variables explicatives et variables réponses. La **classification** étant un cas particulier de régression où les variables réponses prennent des valeurs discrètes (étiquettes). Alors que l'apprentissage non-supervisé permet de mettre à jour des structures dans les données, ultimement cela peut servir à réduire la dimensionnalité des données étudiées en choisissant les dimensions où la variance entre les objets est maximale. De par les distances qui séparent les objets dans

l'espace réduit, il est potentiellement possible de classer les objets. Cela fait référence au concept de **similarité**, très important en bio-informatique (e.g., [Altschul et al., 1990]).

Nous avons travaillé avec l'algorithme de forêts aléatoires sous sa forme de modèle de régression, et de modèle de classification. Dans la section suivante, nous décrirons les grandes lignes de cet algorithme. Il est déjà intéressant de noter que l'efficacité de l'algorithme repose essentiellement sur des preuves empiriques, le contexte théorique n'étant pas complètement établi. C'est donc dans cette optique que sera présenté l'algorithme de forêts aléatoires dans la section suivante. Auparavant, nous introduirons l'utilisation des modèles de régression pour corriger les distributions *a posteriori* approximées avec les méthodes de calcul bayésien approché.

1.3.6. Correction de la distribution *a posteriori* au moyen de modèles de régression

Il est possible d'utiliser un modèle de régression pour améliorer l'approximation de la distribution *a posteriori* faite sous l'algorithme kNN [Beaumont et al., 2002, Csilléry et al., 2012, Blum and Francois, 2010, Saulnier et al., 2017]. L'approche est relativement simple, elle consiste dans un premier temps à modéliser la relation entre statistiques descriptives (variables explicatives) et les valeurs de paramètres (variables réponses) à l'aide d'un modèle de régression (e.g., régression linéaire multiple, réseaux de neurones, forêts aléatoires). Puis, dans un deuxième temps, l'objectif est d'appliquer une correction à l'ensemble des échantillons décrivant la distribution *a posteriori* approximée. À partir du modèle de régression et des données réelles, des valeurs de paramètres sont prédites, puis la correction ou l'ajustement est obtenu en rapportant l'erreur sur la prédiction de chacun des échantillons décrivant la distribution *a posteriori* de chacune des valeurs de paramètre spécifiquement. Donc plus le modèle de régression est habile à modéliser la relation entre statistiques descriptives et valeurs des paramètres, plus l'ajustement devrait réduire l'erreur sur la prédiction faite par le modèle.

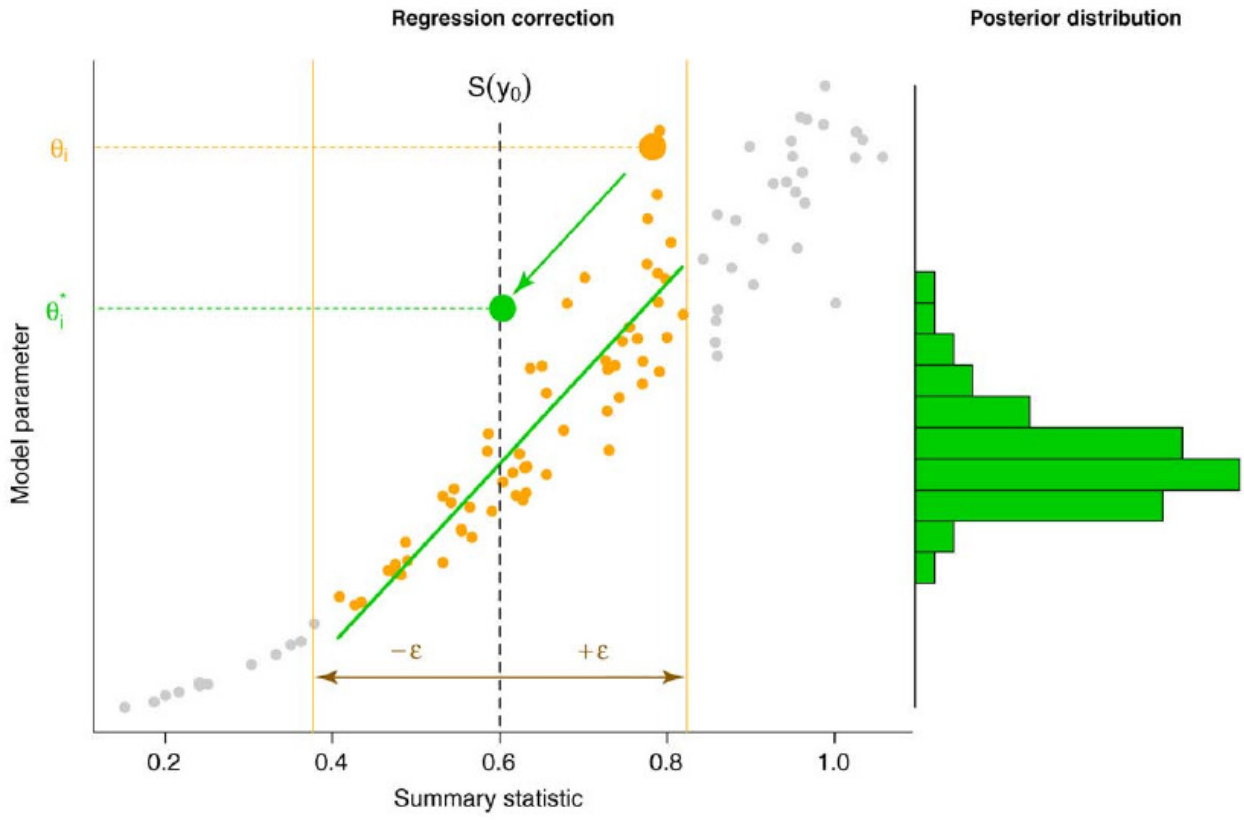

FIGURE 1.10. Ajustement par modèle de régression (tiré de [Csilléry et al., 2012]). $S(y_0)$ correspond à la valeur réelle, ϵ au seuil d'acceptation. Les points jaunes correspondent aux k plus proches voisins conservés lors de l'approximation de la distribution *a posteriori* sur lesquels un modèle de régression multiple est appliqué. Le point vert précédé d'une flèche et d'un gros point jaune illustre la correction apportée par le modèle de régression.

1.3.7. Présentation de l'algorithme de forêts aléatoires

L'algorithme de forêts aléatoires [Breiman, 2001] est un type d'arbre de décision particulier. C'est une heuristique qui incorpore des aspects aléatoires comme le suggère son nom. Cet algorithme permet de déployer des modèles de régression non-supervisée, de régression supervisée et de classification supervisée utilisant une méthodologie de type **arbre de décision**. Le développement de l'algorithme est guidé par son efficacité, et non pas sur la base de fondements théoriques [Louppe, 2014]. Les propriétés de l'algorithme sont actuellement étudiées [Tang et al., 2018]. Le code est facilement parallélisable sur la base du calcul de chacun des arbres de l'algorithme.

L'algorithme utilise deux aspects aléatoires. Ces aspects sont très importants, car ils assurent que les arbres générés sont non-corrélés entre eux, permettant ainsi d'estimer une moyenne à travers l'ensemble des arbres produits par l'algorithme. Pour chaque arbre que l'algorithme de forêts aléatoires construit, (1) un échantillonnage avec remise (bootstrap) est réalisé à partir du jeu de données. Dans notre cas, les données sont les simulations qui peuplent la table de référence. À chaque division binaire de l'arbre, (2) un ensemble de variables explicatives est déterminé au hasard, le nombre de variables varie entre une seule et l'ensemble de celles-ci (déterminé a priori). De cette manière, l'algorithme est capable de gérer un très grand nombre de variables explicatives potentiellement corrélées entre elles et de nature différente (e.g., continue versus discrète). Cette étape permettrait aussi de rendre l'analyse robuste aux valeurs marginales (outliers) et au fléau de la dimensionnalité (curse of dimensionality), n'ayant qu'à prendre des décisions sur un sous-ensemble des variables explicatives. À chaque noeud de l'arbre, la valeur (Gini ou entropie) est calculée pour chacune des variables explicatives tirées et pour chacune des valeurs présentes dans les jeux de données échantillonnées de manière à identifier le seuil et la variable explicative sur lesquels scinder les données en deux groupes (figure 1.11).

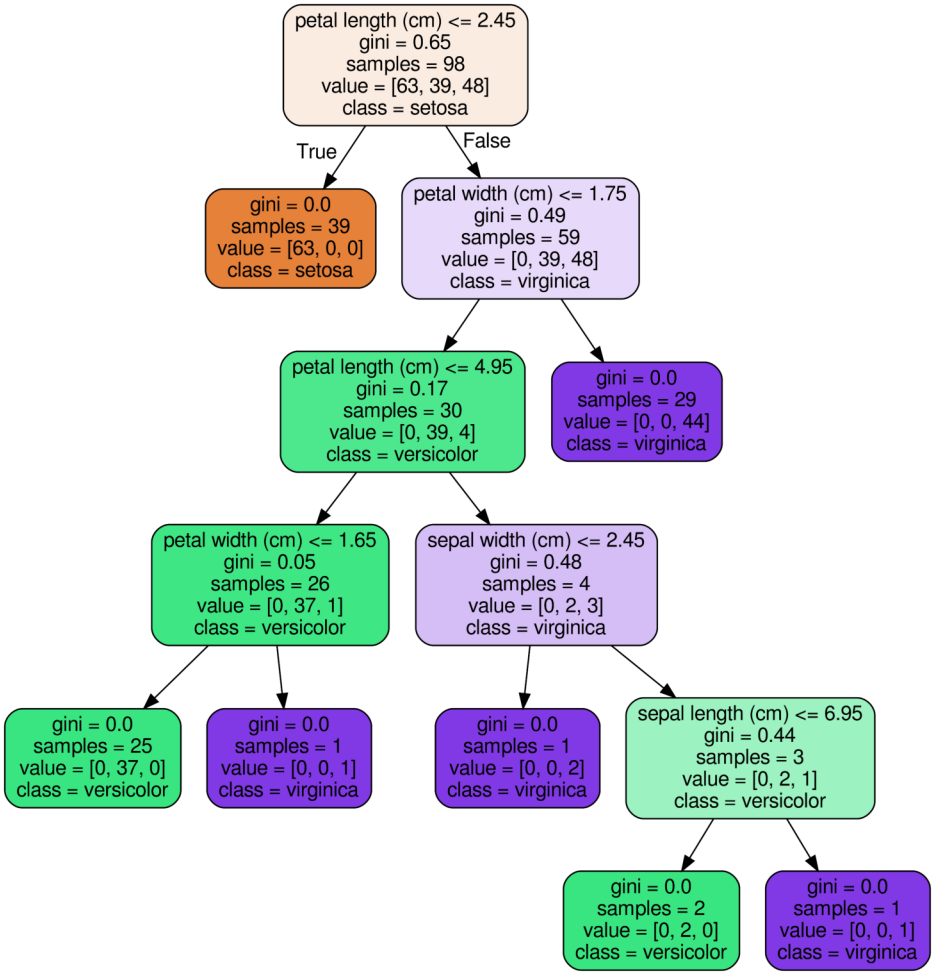

FIGURE 1.11. Exemple de visualisation d'un arbre de décision tiré d'un ensemble de 100 arbres obtenus avec l'algorithme de forêts aléatoires sur le jeu de données Iris (caractéristiques de fleurs [FISHER, 1936]). L'arbre de décision est réalisé avec le programme graphviz [Ellson et al., 2003].Le jeu de données comprend 150 échantillons pour lesquels la longueur des sépales (sepal length), la largeur de sépales (sepal width), la longueur des pétales (petal length) et la largeur des pétales (petal width) ont été caractérisées. Pour chaque noeud la variable explicative optimale, selon le critère (mesure de Gini de l'impureté ou gain d'information), et la valeur optimale sur laquelle la décision de scinder le jeu de données en deux est prise sont affichées ainsi que la mesure de Gini de l'impureté (gini), le nombre d'échantillons (samples), la proportion de chacune des classes (value) ainsi que la classe la plus abondante (class). Dans cet arbre de décision, l'espèce *I. setosa* est classée à la première décision, sur la base de la longueur des pétales (≤ 2.45 cm). Les deux autres espèces, *I. versicolor* et *I. virginica* vont être essentiellement classées au noeud suivant sur la base de la largeur des pétales (≤ 1.75 cm) ; seulement quelques individus, 6%, de *I. virginica* et *I. versicolor* demanderont jusqu'à trois étages de décision de plus.

Le nombre d'arbres à construire ainsi que les hyperparamètres de l'algorithme (e.g., nombre d'échantillons par noeuds terminaux) peuvent être évalués par validation croisée. Tout particulièrement, la validation croisée de type hors-du-sac (out-of-bag : OOB) prend avantage du fait que chacun des arbres est construit avec un sous-ensemble des données. Habituellement la validation croisée se fait avec un jeu de données de test, alors que les paramètres du modèle sont calculés sur un jeu d'entraînement, mais l'astuce de la validation croisée de type OOB permet d'éviter de scinder les données en jeux de test et d'entraînement. Il est donc possible de calculer le coefficient de détermination (R^2) sur la prédiction OOB à partir des données qui n'ont pas servi à la construction des arbres, et ainsi éviter de regarder deux fois les données.

1.3.8. Comparaison de modèles

Identifier le meilleur modèle parmi un ensemble de modèles est une tâche statistique fondamentale de la modélisation probabiliste. C'est un domaine de recherche en soi des statistiques où se rencontre toute la complexité que peut générer l'utilisation de modèles probabilistes (e.g., [Rodrigue and Aris-Brosou, 2011, Pudlo et al., 2016]).

En maximum de vraisemblance, les rapports de la vraisemblance (Likelihood ratio test : LRT) et dans le cadre bayésien, les facteurs de bayes, sont fréquemment utilisés pour ordonner les modèles sur leur capacité à expliquer les données étudiées. Les rapports de vraisemblance requièrent des modèles imbriqués les uns dans les autres, en d'autres mots, qu'un modèle soit une forme plus simple de l'autre modèle (e.g., HKY versus GTR). Quand ce n'est pas le cas, il est possible d'utiliser des critères comme AIC, (Akaike Information Criterion : AIC [Akaike, 1974]), l'AIC corrigé (Corrected Akaike Information Criterion : AICc [EC, 1997]) ou encore le critère d'information bayésien (Bayesian Information Criterion : BIC [Schwarz, 1978]). De leur côté, les facteurs de Bayes sont coûteux en temps calcul, car il faut échantillonner suffisamment la distribution de la vraisemblance marginale (distribution prédictive *a priori*) des deux modèles avant d'en faire le rapport [Jeffreys, 1935]. Plus les modèles à comparer sont riches en matière de nombre de paramètres plus le calcul des facteurs de Bayes sera coûteux en temps calcul [Rodrigue et al., 2008a].

Dans le contexte où la fonction de vraisemblance n'est pas accessible, comme dans le cas du calcul bayésien approché, les facteurs de Bayes ne peuvent être calculés pour permettre

la comparaison de modèles. Pudlo et al. [2016] proposent d'utiliser (1) un modèle de classification basé sur l'algorithme de forêts aléatoires pour déterminer le modèle préféré et (2) un second modèle de régression basé sur l'algorithme de forêts aléatoires pour évaluer l'erreur sur la prédiction construite à partir de la distribution prédictive a priori, ce qui permet d'obtenir la probabilité *a posteriori* de préférer un modèle plus qu'un autre.

Dans un premier temps, le modèle de forêts aléatoires (classificateur) est conditionné à partir d'une table de comparaison construite en joignant les tables de référence (distributions prédictives *a priori*) des deux modèles à comparer. En fait cette méthode est spécialement conçue pour les inférences par calcul bayésien approché, donc où la table de comparaison est obtenue en joignant la partie correspondant aux statistiques descriptives utilisées dans les tables de références des deux modèles et en ajoutant une colonne identifiant chacune des entrées avec une étiquette correspondant à l'identité des modèles (dans le but de réaliser une tâche de classification). La table de comparaison prend la forme suivante :

$$\begin{pmatrix} M1 & s_{11} & s_{21} & ... & s_{N_s1} \\ M1 & s_{12} & s_{22} & ... & s_{N_s2} \\ ... & ... & ... & ... & ... \\ M1 & s_{1N} & s_{2N} & ... & s_{N_sN} \\ M2 & s_{11} & s_{21} & ... & s_{N_s1} \\ M2 & s_{12} & s_{22} & ... & s_{N_s2} \\ ... & ... & ... & ... & ... \\ M2 & s_{1N} & s_{2N} & ... & s_{N_sN} \end{pmatrix}, \tag{1.3.2}$$

où chacune des entrées de la table de comparaison comprend N_s statistiques descriptives. La table est balancée, elle comprend le même nombre d'entrées (N) pour chacun des modèles $M1$ et $M2$. Une fois le modèle conditionné à partir de la table de comparaison, il est possible de prédire la classe à laquelle appartient l'alignement du gène étudié. Un seuil (e.g., $>$50%) est alors utilisé pour déterminer l'issue du vote auquel chacun des arbres construits par l'algorithme participe, et donc déterminer à quelle classe l'alignement du gène étudié appartient. En fait, en faisant varier ce seuil, et en calculant le taux de vrais positifs, soit la fréquence à laquelle les deux modèles sont bien classés (sensibilité), ainsi qu'en calculant le taux de faux positifs (spécificité), il est possible de construire la fonction d'efficacité du récepteur (receiver operating characteristic : ROC). À mesure que le seuil augmente pour

décider de l'issue de la votation, la spécificité du test augmente, mais la sensibilité diminue. L'étude de la courbe ROC est donc une manière d'identifier les limites de ce qui est inféré par l'algorithme de forêts aléatoires.

Les auteurs de [Pudlo et al., 2016] ont par la suite mis en place un système astucieux pour évaluer l'incertitude sur la classification. Elle est obtenue en modélisant l'erreur sur la prédiction à l'aide d'un second modèle de forêts aléatoires, mais cette fois un modèle de type régression. Une table de comparaison trafiquée est utilisée, où la variable réponse correspond à l'erreur binaire sur la prédiction de la classification (0 lorsque la prédiction moyenne est plus petite ou égale à 0,5 et 1 lorsque la prédiction est plus grande que 0,5). L'erreur, binaire, sur les prédictions de la classification est évaluée au moyen de la validation croisée OOB [Pudlo et al., 2016]. Les variables explicatives sont donc les statistiques descriptives. La table d'erreur binaire prend la forme suivante :

$$\begin{pmatrix} 1 & s_{11} & s_{21} & ... & s_{N_s 1} \\ 1 & s_{12} & s_{22} & ... & s_{N_s 2} \\ ... & ... & ... & ... & ... \\ 0 & s_{1N} & s_{2N} & ... & s_{N_s N} \\ 1 & s_{11} & s_{21} & ... & s_{N_s 1} \\ 0 & s_{12} & s_{22} & ... & s_{N_s 2} \\ ... & ... & ... & ... & ... \\ 1 & s_{1N} & s_{2N} & ... & s_{N_s N} \end{pmatrix}, \tag{1.3.3}$$

où chacune des entrées de la table de comparaison comprend N_s statistiques descriptives. Une fois le modèle de régression de type forêts aléatoires conditionné avec la table de comparaison trafiquée (TCtraffique), il est possible de calculer la probabilité a posteriori de la classification (figure 1.12).

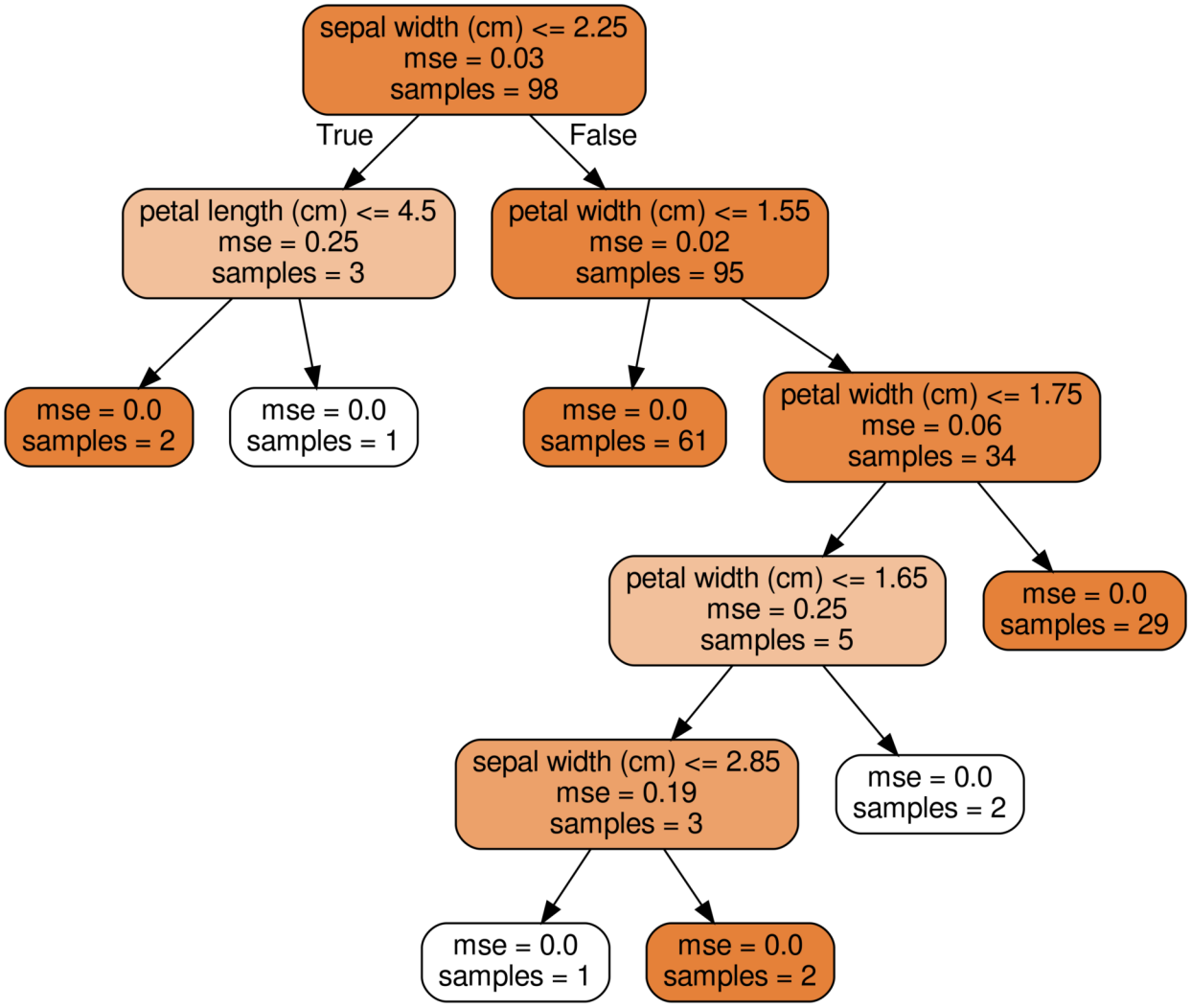

Figure 1.12. Visualisation d'un arbre de décision obtenu avec l'algorithme de régression par forêts aléatoires à partir d'un jeu de données où les variables explicatives sont celles du jeu de données *iris* (voir la figure 1.11) et la variable réponse est l'erreur sur la prédiction obtenue avec le même algorithme, mais sur le jeu *iris* (voir la figure 1.11). Pour chaque noeud de l'arbre, la variable explicative optimale, selon le critère de l'erreur au carré moyenne (mse), ainsi que la valeur optimale sur laquelle la décision de scinder le jeu de données en deux est prise sont affichées. Le nombre d'échantillons restants (samples) est aussi présenté pour chacune des étapes de l'arbre de régression. *Iris setosa* est identifiée dans cet arbre de décision à la première décision sur la base de la longueur des pétales (≤ 2.45 cm). Les deux autres espèces, *Iris versicolor* et *Iris virginica* vont être majoritairement classées au noeud suivant sur la base de la largeur des pétales (≤ 1.75 cm) ; seulement quelques individus, 6%, de *Iris virginica* et *I. versicolor* demanderont jusqu'à trois étages de décision de plus.

Lorsque la probabilité a posteriori est de 0, cela signifie que la classification est fausse, alors que lorsque la probabilité a posteriori est de 1, cela signifie que la classification est vraie à 100% (dans la limite de la présence d'effets confondants). Une valeur de probabilité a posteriori entre 0 et 1 demande une interprétation plus soignée. L'étude de la sensibilité et de la spécificité prend alors tout son sens. Il faut noter que l'identification du signal tient au choix des statistiques descriptives utilisées, à l'information qu'elles contiennent, mais aussi plus en amont, à la quantité de signal disponible dans l'absolu sur lequel l'algorithme peut s'appuyer pour faire sa classification (e.g., le nombre de transversions G<>T sont habituellement rares chez les Mammifères et il est donc probablement difficile de classer les modèles sur cette base).

Chapitre 2

Multiples facteurs confondant la détection phylogénétique de la sélection sur l'usage des codons

2.1. Information

Multiple factors confounding phylogenetic detection of selection on codon usage

Molecular Biology and Evolution, Volume 35, Issue 6, 1 June 2018, Pages 1463–1472, https ://doi.org/10.1093/molbev/msy047

Published : 27 March 2018

Simon Laurin-Lemay[1], Hervé Philippe[1,2], Nicolas Rodrigue[*,3]

[1]Robert-Cedergren Center for Bioinformatics and Genomics, Department of Biochemistry and Molecular Medicine, Faculty of Medicine, Université de Montréal, Montréal, Québec, Canada

[2]Centre de Théorisation et de Modélisation de la Biodiversité, Station d'Écologie Théorique et Expérimentale, UMR CNRS 5321, Moulis, Ariège, France

[3]Department of Biology, Institute of Biochemistry, and School of Mathematics and Statistics, Carleton University, Ottawa, Canada

Running head : Confounding factors to uneven codon usage

Keywords : synonymous substitution, nonsynonymous substitution, CpG hypermutability, model violation.

*Correspondence : Nicolas Rodrigue
209 Nesbitt Biology Building,
1125 Colonel By Drive Ottawa, Ontario, CANADA
K1A 0C6
nicolas.rodrigue@carleton.ca
tel : +1 613 520 2600 x 4194

2.2. Résumé

La détection de la sélection sur l'usage des codons est une tâche difficile, car l'usage des codons peut être façonné à la fois par le processus mutationnel et par des contraintes de sélection opérant au niveau de l'ADN, de l'ARN et des protéines. Yang and Nielsen [2008] ont développé un test, que nous appelons CUYN, pour détecter la sélection sur l'usage des codons au moyen de deux modèles à codon de substitution de type mutation-sélection. Le modèle nul suppose que l'usage des codons est déterminée uniquement par le processus mutationnel, tandis que le modèle alternatif suppose que l'usage des codons est déterminée par le processus mutationnel et/ou par des contraintes de sélection globables sur l'usage des codons. La plupart des alignements de gène de mammifères obtiennent un test CUYN positif pour la sélection sur l'usage des codons. Cela est d'autant plus surprenant, étant donné la petite N_e des mammifères, ce qui nous a incités à évaluer la robustesse du test CUYN face à des violations de modèle potentielles. Un niveau élevé de faux positifs a été obtenu en utilisant le test sur des alignements simulés générés de manière à imiter des aspects importants de l'évolution moléculaire. En particulier, les simulations utilisant un niveau modeste d'hypermutabilité CpG trompent complètement le test, avec 100 % de faux positifs. Étonnamment, un niveau élevé de faux positifs (56,1 %) résulte simplement de l'utilisation de la paramétrisation moins riche HKY lors du test CUYN sur des simulations effectuées avec une paramétrisation plus riche GTR du processus mutationnel. Enfin, en utilisant une procédure d'optimisation crue sur le paramètre contrôlant le taux d'hypermutabilité du context CpG, nous trouvons que cette propriété mutationnelle pourrait expliquer une partie importante de la l'usage des codons observée chez les mammifères. Dans l'ensemble, nos travaux mettent en perspective la nécessité d'évaluer l'impact potentiel des violations des modèles sur les tests statistiques dans le domaine de l'évolution moléculaire. Le code source du simulateur et des gènes de mammifères utilisés sont disponibles sous forme de dépôt GitHub (https ://github.com/Simonll/LikelihoodFreePhylogenetics.git).

2.3. Abstract

Detecting selection on codon usage (CU) is a difficult task, since CU can be shaped by both the mutational process and selective constraints operating at the DNA, RNA and protein levels. Yang and Nielsen [2008] developed a test (which we call CUYN) for detecting selection on CU using two competing mutation-selection models of codon substitution. The null model assumes that CU is determined by the mutation bias alone, whereas the alternative model assumes that CU is determined by mutation bias and/or selection on CU. In applications on mammalian-scale alignments, the CUYN test detects selection on CU for numerous genes. This is surprising, given the small effective population size of mammals, and prompted us to evaluate the robustness of the CUYN test to model violations. A high level of false positives was obtained when using the test on simulated alignments generated in a manner mimicking important aspects of molecular evolution. In particular, simulations using a modest level of CpG hypermutability completely mislead the test, with 100% false positives. Surprisingly, a high level of false positives (56.1%) resulted simply from using the HKY mutation-level parameterization within the CUYN test on simulations conducted with a GTR mutation-level parameterization. Finally, by using a crude optimization procedure on a parameter controlling the CpG hypermutability rate, we find that this mutational property could explain a very large part of the observed mammalian CU. Altogether, our work emphasizes the need to evaluate the potential impact of model violations on statistical tests in the field of molecular phylogenetic analysis. The source code of the simulator and the mammalian genes used are available as a GitHub repository (https ://github.com/Simonll/LikelihoodFreePhylogenetics.git).

2.4. Introduction

Elucidating the evolutionary factors influencing codon usage (CU) in protein-coding genes is a challenging endeavor. One difficulty resides in the complex nature of the CU genotype-phenotype relation. The nature of this relationship is affected by features of the mutational process, the degeneracy structure of the genetic code, synonymous codon recognition by tRNA molecules, and multiple other constraints operating at the DNA, RNA and protein levels.

Unequivocally, the most important feature highlighted since the beginning of CU studies (e.g., Grantham et al. 1980), is the positive correlation found between the CU of highly expressed genes and the isoaccepting-tRNA pool composition [Ikemura, 1985, Bulmer, 1987, Akashi, 1995, Sharp et al., 1995, Duret, 2002a]. Presumably, the co-evolution of CU and tRNA pool composition ensures the accuracy and efficiency of translation of highly expressed genes (reviewed in Quax et al. 2015).

The CU of highly expressed genes is usually postulated to result from a selection process on synonymous mutations, as investigated in large comparative frameworks across bacteria [Sharp et al., 2005] and eukaryotes [dos Reis and Wernisch, 2009], and several more specific contexts, such as in *Drosophila* [Shields et al., 1988, Akashi, 1994, Bierne and Eyre-Walker, 2006, Lawrie et al., 2013], in *Caenorhabditis elegans* [Duret and Mouchiroud, 1999], in *Daphnia pulex* [Burge et al., 1992], in vertebrates [Doherty and McInerney, 2013], in the Euarchotonglires from mammals [Yang and Nielsen, 2008], as well as in humans [Lavner and Kotlar, 2005]. While there might be mutational features that could partially account for the match between CU and the tRNA pool composition, "mutational pressures alone cannot explain why the more frequent codons [...] are those that are recognized by more abundant tRNA molecules" [Hershberg and Petrov, 2008], probably because selection might be more effective at modulating the tRNA pool composition. Nonetheless, until a plausible mechanism is proposed, there generally remains "an understandable reluctance to accept selection at synonymous sites" [Chamary et al., 2006].

Exogenous gene expression is one area in which these issues have practical ramifications. Indeed, practitioners in this field remain without a clear rationalization of the methods for their activity : "A future challenge in studying the relation between coding sequences and protein production is to perform a thorough comparative analysis of all currently known,

and yet to be discovered, features of coding sequences that influence the translation process." [Quax et al., 2015]. To date, two strategies are available to them for increasing the efficiency of exogenous genes expression : the first consists of expressing additional tRNA genes in the host cell to match the CU of the exogenous gene to be expressed ; and the second consists of modifying the CU of the exogenous genes (reviewed in Quax et al. 2015). For instance, when using a bacterial system, one cannot achieve a worthwhile production of some desired human protein without modifying the codons used for encoding the desired amino acid sequence [Doble and Gummadi, 2007]. Nonetheless, there are still many instances in which the CU modifications for a particular sequence proceed by trial-and-error [Webster et al., 2017], without discernible reasons for the final CU that optimizes protein production.

Probabilistic models of molecular evolution have the potential to tease apart the determining factors of CU. The *mutation-selection* models (reviewed by McCandlish and Stoltzfus 2014), and particularly the phylogenetic ones, are well suited for testing hypotheses related to coding sequence evolution. The distinguishing features of these models is that they specify a substitution process with distinct parameterizations for the manner in which genetic variation is generated and for the fixation probability of genetic variants. In some cases, the models have specifically included considerations of selection on CU [McVean and Vieira, 2001, Nielsen et al., 2007, Rodrigue et al., 2008b, Yang and Nielsen, 2008, Rodrigue and Philippe, 2010, Pouyet et al., 2016].

The most well known of these models of codon substitution are those of Yang and Nielsen [2008]. In their work, they propose a likelihood ratio test (LRT) to detect selection on CU, which we refer to as the CUYN test. The LRT is performed using two competing mutation-selection models : the null model is built with selection acting only on amino acid usage, assigning the same fitness to each degenerate codon encoding a particular amino acid ; and the alternative model assigns a distinct fitness to each codon, and thus accounts for selection acting on both amino acid usage (by assigning a higher/lower fitness overall to the codons that encode a particular amino acid) and codon usage (by assigning a distinct fitness to each codon of an amino acid).

Yang and Nielsen [2008] found that much of the mammalian genes they tested rejected the null model, suggesting pervasive selection on CU. However, population genetics principles suggest that for organisms with small effective population sizes, like mammals, selection is

too inefficient to distinguish small effects conferred by certain synonymous mutations (reviewed in Chamary et al. 2006, Charlesworth 2009, Lynch et al. 2016). In contrast, selection is expected to be efficient within highly expressed proteins and in groups of fast-growing organisms (e.g., Sharp et al. 2010). Consequently, mammalian CU is expected to be mainly determined by mutation bias (reviewed in Chamary et al. 2006).

Several important features of the mutational process are unaccounted for in most codon substitution models. Among these, a phenomenon known as *CpG hypermutability*, whereby certain mutations occur at higher rates in the context of the states at adjacent positions in the sequence, is considered pervasive in mammalian genomes (reviewed in Hodgkinson and Eyre-Walker 2011). The resultants of CpG hypermutability are numerous. For instance, the most used synonymous codon for Alanine in human, GCC, is four times more represented than GCG [Coleman et al., 2008], probably because the latter codon includes a highly mutable context (i.e., CpG), and is therefore short-lasting. Also, on the basis of the *relative synonymous codon usage* (RSCU) metric, the most frequent adjacent codon pair for Alanine-Glutamine is expected to be GCC-GAR. In fact, however, this pair is the most underrepresented in humans [Coleman et al., 2008], probably because of the instability of the CpG context found at the interface of these two sequential codons. Another important feature that is known to bias the generation of genetic variation within mammalian genomes is *GC-biased gene conversion* [Duret and Galtier, 2009b] : a phenomenon produced from meiotic recombination favoring transmission of G :C alleles over A :T alleles. Hypermutability of CpG contexts and GC-biased gene conversion are considered responsible for the isochore structures found in vertebrates [Duret and Galtier, 2009b, Munch et al., 2014, Mugal et al., 2015]. These phenomena may have an impact on the test proposed by Yang and Nielsen [2008], in a manner that is very difficult to foresee. Indeed, in light of their surprising results, Yang and Nielsen [2008] point out that "the sensitivity of the LRT to violations of the assumed mutation model is not well understood and merits further research."

In this work, we use simulations to evaluate the effect of model violations on the accuracy of the CUYN test. We explore several types of violations at the mutation level, including one with dependence across codons, related to CpG hypermutability. We also evaluate how site-heterogeneous preferences on amino acids can affect the test. Numerous false positives are obtained under these model violations. While not excluding other potential features that

can affect CU bias, our findings suggest that CpG hypermutability alone could explain the results of the CUYN test on mammalian genes.

2.5. Results and Discussion

2.5.1. Phylogenetics Mutation-Selection Models

The branch lengths are free parameters for all models used in this work, while the tree topology is fixed. The models we use herein assume a point-mutation process, going from one codon a to another b, which differ at the cth position. In other words, c is an index of value 1, 2, or 3, indicating which of the three nucleotide positions is different between codons a and b. Stop codons are also disallowed from the state space, leading to a 61×61 rate matrix Q. The rate matrix is rather sparse, however, since entries corresponding to nucleotide doublet or triplet events are set to 0. All non-null, non-diagonal entries of the matrix are specified from two overall sets of parameters : those controlling a mutation level, and those controlling a selection level.

At the mutation level, nucleotide propensity parameters are invoked, defined as $\varphi = (\varphi_n)_{1 \leq n \leq 4}$, with $\sum_{n=1}^{4} \varphi_n = 1$. When using the M[HKY] settings, one parameter is introduced (i.e., $\kappa > 0$) to account for unequal rates between transitions and transversions. When using the M[GTR] settings, the exchangeabilities of each unique pair of nucleotides, m and n, are defined as $\varrho = (\varrho_{mn})_{1 \leq m,n \leq 4}$, with $\sum_{1 \leq m < n \leq 4} \varrho_{mn} = 1$. For some models, the transition rates within the CpG context (i.e., C to T and G to A) are modulated via a multiplicative parameter, λ_{CpG}.

At the level of selection, in the most elaborate settings, the specification of the model involves a set of K vectors, with each having 20 entries corresponding to amino acid preferences (also called *profiles*), denoted $\psi = (\psi_l^{(k)})_{1 \leq l \leq 20, 1 \leq k \leq K}$. The specification also involves an allocation variable denoted $z = (z_i)_{1 \leq i \leq N}$, where N is the length of the gene (in codons) ; for a given site i, z_i returns an index from 1 to K, specifying the amino acid profile operating at that site. The scaled selection coefficient [see Yang and Nielsen, 2008] associated to a nonsynonymous change from codon a to b at site i is given as

$$S_{ab}^{(i)} = \ln\left(\frac{\psi_{f(b)}^{(z_i)}}{\psi_{f(a)}^{(z_i)}}\right), \tag{2.5.1}$$

where $f(a)$ returns an index, from 1 to 20, of the amino acid encoded by codon a. The value of $S_{ab}^{(i)}$, in turn, defines a fixation factor, denoted $h(S_{ab}^{(i)})$, and calculated as

$$h(S_{ab}^{(i)}) = \frac{S_{ab}^{(i)}}{1 - e^{-S_{ab}^{(i)}}}, \tag{2.5.2}$$

which will then be used directly in the rate matrix. The set of K amino acid profiles, the allocation variable z, and the value of K itself, are random variables within a Dirichlet process device [Rodrigue et al., 2010]. We refer to this setting as S[NCatAA]. Of course, one can dispense with the Dirichlet process device, and simply use a single category of amino acid preferences, in which case we drop the index on sites, i, from the notation, and refer to it as S[1CatAA].

Another setting at the level of selection replaces a single 20-dimensional vector with one that includes 61 values instead, one for each codon. The scaled selection coefficients and fixation factors are computed as before, although in this case, they are required for both synonymous and nonsynonymous events. We refer to this as the S[1CatCodon] setting.

Finally, we include a parameter (i.e., $\omega_* > 1$) on nonsynonymous rates, aimed at capturing deviations from the mutation-selection balance [Rodrigue and Lartillot, 2017]. The different models obtained from the combinations of mutation and selection parts are thus as follows :

— M[HKY]-S[1CatAA] : the mutation part of this model has four parameters (3 degrees of freedom) controlling nucleotide propensities, as well as a parameter (1 df) to distinguish transitions from transversions. The selection part of this model has a single category of amino acid preference (20 parameters and 19 df). The model was first described by Yang and Nielsen [2008] as FMutSel0, and is detailed below :

$$Q_{ab} = \begin{cases} \varphi_{b_c}, & \text{if syn. tr.,} \\ \varphi_{b_c}\kappa, & \text{if syn. ts.,} \\ \varphi_{b_c}\omega_* h(S_{ab}), & \text{if nonsyn. tr.,} \\ \varphi_{b_c}\kappa\omega_* h(S_{ab}), & \text{if nonsyn. ts.,} \end{cases} \tag{2.5.3}$$

where “syn.” and “nonsyn.” are short for “synonymous” and “nonsynonymous”, “tr.” and “ts.” are short for “transversion” and “transition”, and b_c returns an index, from 1 to 4, of the nucleotide found at the cth position in codon b.

— M[GTR]-S[1CatAA] : This model is nearly the same as M[HKY]-S[1CatAA], but has 6 distinct parameters (5 df) controlling the relative rate of each (un-ordered) pair of nucleotides. The model was also first described in Yang and Nielsen [2008], and is detailed in equation 2.5.4 :

$$Q_{ab} = \begin{cases} \varrho_{a_c b_c}\varphi_{b_c}, & \text{if syn.,} \\ \varrho_{a_c b_c}\varphi_{b_c}\omega_* h(S_{ab}), & \text{if nonsyn.} \end{cases} \tag{2.5.4}$$

— M[HKY+λ_{CpG}]-S[1CatAA] : Similarly as before, this model only differs from equation 2.5.3 in having an additional parameter, λ_{CpG}, controlling the mutation rate of transitions in the CpG context. The model is detailed below :

$$Q_{ab} = \begin{cases} \varphi_{b_c}, & \text{if syn. tr.,} \\ \varphi_{b_c}\kappa, & \text{if syn. ts. non-CpG,} \\ \varphi_{b_c}\kappa\lambda_{CpG}, & \text{if syn. ts. CpG,} \\ \varphi_{b_c}\omega_* h(S_{ab}), & \text{if nonsyn. tr.,} \\ \varphi_{b_c}\kappa\omega_* h(S_{ab}), & \text{if nonsyn. ts. non-CpG,} \\ \varphi_{b_c}\kappa\omega_* h(S_{ab})\lambda_{CpG}, & \text{if nonsyn. ts. CpG.,} \end{cases} \tag{2.5.5}$$

where "CpG" refers to a hypermutability context (assuming $\lambda_{CpG} > 1$) type of event.

— M[GTR+λ_{CpG}]-S[1CatAA] : Similarly as before, this model only differs from equation 2.5.4 in having one parameter, λ_{CpG}, controlling the mutation rate of the CpG context. The model is detailed in equation 2.5.6 :

$$Q_{ab} = \begin{cases} \varrho_{a_c b_c}\varphi_{b_c}, & \text{if syn. tr.,} \\ \varrho_{a_c b_c}\varphi_{b_c}, & \text{if syn. ts. non-CpG,} \\ \varrho_{a_c b_c}\varphi_{b_c}\lambda_{CpG}, & \text{if syn. ts. CpG,} \\ \varrho_{a_c b_c}\varphi_{b_c}\omega_* h(S_{ab}), & \text{if nonsyn. tr.,} \\ \varrho_{a_c b_c}\varphi_{b_c}\omega_* h(S_{ab}), & \text{if nonsyn. ts. non-CpG,} \\ \varrho_{a_c b_c}\varphi_{b_c}\omega_* h(S_{ab})\lambda_{CpG}, & \text{if nonsyn. ts. CpG.} \end{cases} \tag{2.5.6}$$

— M[HKY]-S[1CatCodon] : In contrast to M[HKY]-S[1CatAA], which in effect assigns all codons encoding a particular amino acid the same preference, this model has as distinct parameter for each codon (60 df). The model was first described by Yang and Nielsen [2008] as FMutSel, and is detailed in equation 2.5.7 :

$$Q_{ab} = \begin{cases} \varphi_{b_c} h(S_{ab}), & \text{if syn. tr.,} \\ \varphi_{b_c} \kappa h(S_{ab}), & \text{if syn. ts.,} \\ \varphi_{b_c} \omega_* h(S_{ab}), & \text{if nonsyn. tr.,} \\ \varphi_{b_c} \kappa \omega_* h(S_{ab}), & \text{if nonsyn. ts.} \end{cases} \tag{2.5.7}$$

— M[GTR]-S[1CatCodon] : Similarly as before, this model only differs from the previous one in having 6 parameters controlling relative rates of nucleotide exchange, as described in Yang and Nielsen [2008], and equation 2.5.8 :

$$Q_{ab} = \begin{cases} \varrho_{a_c b_c} \varphi_{b_c} h(S_{ab}), & \text{if syn.,} \\ \varrho_{a_c b_c} \varphi_{b_c} \omega_* h(S_{ab}), & \text{if nonsyn.} \end{cases} \tag{2.5.8}$$

— M[HKY]-S[NCatAA] : Of the same form as M[HKY]-S[1CatAA], this model allows for heterogeneity across sites for the amino acid preference by including multiple categories of amino acid preference, as described by Rodrigue et al. [2010], and equation 2.5.9 :

$$Q_{ab}^{(i)} = \begin{cases} \varphi_{b_c}, & \text{if syn. tr.,} \\ \varphi_{b_c} \kappa, & \text{if syn. ts.,} \\ \varphi_{b_c} \omega_* h(S_{ab}^{(i)}), & \text{if nonsyn. tr.,} \\ \varphi_{b_c} \kappa \omega_* h(S_{ab}^{(i)}), & \text{if nonsyn. ts.} \end{cases} \tag{2.5.9}$$

— M[GTR]-S[NCatAA] : Finally, this model extends the previous one, to 6 parameters controlling relative rates of nucleotide exchange. This is the model studied in Rodrigue et al. [2010], and given in equation 2.5.10 :

$$Q_{ab}^{(i)} = \begin{cases} \varrho_{a_c b_c} \varphi_{b_c}, & \text{if syn.,} \\ \varrho_{a_c b_c} \varphi_{b_c} h(S_{ab}^{(i)}) \omega_*, & \text{if nonsyn.} \end{cases} \tag{2.5.10}$$

2.5.2. CUYN test on observed data

We applied the CUYN test on 137 placental mammalian gene alignments (see Materials and Methods). At a level of 5%, all of the 137 genes analyzed showed a significant LRT (fig. 1A). Yang and Nielsen [2008] and Kessler and Dean [2014] also found a large proportion of genes with significant LRTs, but not 100%. This is likely due to the fact that the genes we study are longer and include more species, resulting in more evolutionary signal (i.e., substitutions) available for the models to learn their parameters, and therefore increases the statistical power of the test.

2.5.3. Protocols and validations of the CUYN test

In order to perform a verification of the CUYN test, we conducted several sets of simulations using parameter values obtained from the use of various mutation-selection models on 16 genes among the 137 previously subjected to the CUYN (see Materials and Methods). Our simulations involve numerous variants of mutation-selection models, and so we make use of a descriptive nomenclature of the models to refer to the different parameterizations. For instance, the mutation-level part of a model could consist of a set of parameters controlling nucleotide frequencies along with a transition rate parameter, and we denote this part as M[HKY], in reference to Blanquart and Lartillot [2008]. The selection-level parameterization could consist of a single set, or category, of parameters controlling amino acid fitness, and we denote this part as S[1CatAA]. The resulting mutation-selection model would be referred to as M[HKY]-S[1CatAA], corresponding to the basic null model in Yang and Nielsen [2008]. The alternative model, in this case, would be denoted M[HKY]-S[1CatCodon], where 1CatCodon refers to a single set of parameters controlling codon fitness. Other models we explored include those with *general time reversible* mutation-level specifications [Lanave et al., 1984], as in M[GTR]-S[1CatAA] or M[GTR]-S[1CatCodon], also studied by Yang and Nielsen [2008], and models that treat amino acid fitness profiles across sites as random effects, captured using a Dirichlet process prior [Rodrigue et al., 2010], denoted M[HKY]-S[NCatAA] and M[GTR]-S[NCatAA]. We also used two intermediate parameterizations to the M[HKY] and the M[GTR] settings. The first of these is referred to as M[Tr], since it introduces distinct parameters to transversion exchangeabilities (i.e., where exchangeability parameters for A

and C, A and T, C and G, and G and T are distinct, but where a single exchangeability parameter is shared for transition events involving A and G and those involving C and T : Posada 2008), and the second is denoted M[Ts], since it includes distinct parameters for transition exchangeabilities (where a single exchangeability parameter is shared for events involving A and C, A and T, C and G, and G and T, whereas those involving A and G, as well as C and T, have distinct exchangeability parameters : Tamura and Nei 1993). We also explore the impact of introducing a multiplicative parameter, λ_{CpG}, to events involving a CpG context ; we denote such a model as, for instance, M[HKY+λ_{CpG}]-S[1CatAA]. Importantly, models invoking λ_{CpG} imply a dependence across sites, since they consider CpG contexts both within codons and across adjacent codons. All models are detailed in the Materials and Methods section.

As a negative control, we generated a set of simulated alignments without selection on CU using parameter values obtained under the null models (i.e, M[HKY]-S[1CatAA] and M[GTR]-S[1CatAA]). The rate of false positives, at a significance level of 5%, is close to expectations for the analyses with M[HKY] mutation-level parameterization (fig. 1B : 5.2%). Unexpectedly, almost twice the amount of false positives (Table 1 : 8.8%) is obtained when using M[GTR] parameterization. A similar result is also obtained when using M[GTR] over M[HKY] (Table 1 : 8.4), suggesting non-standard behavior of the LRT under low signal content. To explore this surprising result, we increased the lengths of the branches of the tree over which our simulations were conducted by multiplying their lengths with a parameter $\lambda_{TBL} = 50$. In these conditions, the rate of false positive approaches the expectation (Table 1 : 5.7%), suggesting that the overall signal present in the data when $\lambda_{TBL} = 1$ is weak, even if we used a richer taxon sampling than Yang and Nielsen [2008], and hence limiting the statistical accuracy of the CUYN test when using the M[GTR] setting.

As a positive control, we generated a set of simulated alignments with selection on CU using the parameter values obtained under the alternative models (i.e., M[HKY]-S[1CatCodon] and M[GTR]-S[1CatCodon]). In this case, all the simulated alignments tested show significant evidence for selection on CU. Altogether, for these control simulations, the CUYN test generally performs correctly, by detecting signatures of selection on CU when they are indeed present in the data, and not rejecting the null models in the absence of selection on CU.

2.5.4. Model violations at the mutation level

We next explored simulations done without selection on CU, using the parameter values obtained from analyses of the same 16 genes previously used. We found that alignments simulated with M[GTR] and analyzed with M[HKY] mutation-level settings generated 56.1% false positives (fig. 1C). This is a surprising result, since it involves one of the simplest model violations one could test. Moreover, it does not support the frequent use of the M[HKY] in the analyses conducted with the PAML package [Yang, 2007a]. When using the intermediate settings to M[HKY] and M[GTR], where transition exchangeabilities are set equal and transversion exchangeabilities are kept distinct, the M[Tr] setting, or where transition exchangeabilities are kept distinct and transversion exchangeabilities are set equal, the M[Ts] setting, we find that the heterogeneity of transversion rates has more impact than the heterogeneity of transition rates (i.e., 43.1% of false positives versus 19.4% respectively).

To study the properties of the CUYN test when dealing with increasingly information-rich data, we again increased the lengths of the branches of the tree over which our simulations were conducted with a multiplicative parameter λ_{TBL}, taking on values 5, 10, and 50 (Table 1). In one set of simulations, we increased the tree length when using M[Tr] and M[Ts] mutation-level specifications. We observed that even if the false positive rate increases with dilated branch lengths (e.g., 33% for $\lambda_{TBL} = 50$ versus 19.4% for $\lambda_{TBL} = 1$ when using the M[Ts]) it does not reach the false positive rate detected when analyzing simulations made with the M[GTR] (fig. 1C : 56.1%), suggesting that the main factor determining the number of false positives is the relative complexity of the model used to generate the simulated alignments compared to the model used for its analysis. This is confirmed by the fact that the accuracy of the test does not change for trees of greater length when M[HKY]-S[1CatCodon] is used against M[HKY]-S[1CatAA] (Table 1 : 4.2%-5.7%).

Next, we simulated data with a new model that further modulates the mutation rates of the CpG context found within and across codon boundaries, using the multiplicative parameter λ_{CpG} (see Material and Methods). This new CpG context-dependent mutability, coupled with the null selection setting (i.e., M[HKY+λ_{CpG}]-S[1CatAA]), generates ∼100% of false positives when λ_{CpG} reaches a value of only 5 (fig. 1D), even when CpG hypermutability only explains ∼10-20% of the substitutions. We note that a value of $\lambda_{CpG} = 5$ is probably an underestimate in mammals [Hodgkinson and Eyre-Walker, 2011]. On the other hand,

hypomutability of CpG (i.e., $\lambda_{CpG} < 1.0$) generates a similar amount of false positives, suggesting that any context-dependent mutation pattern may generate a high level of false positives for the CUYN test.

2.5.5. Impacts of model violations at the level of selection

A low rate of false positives (Table 1 : 20.6%) was recovered when analyzing alignments generated with heterogeneous amino acid fitness across sites. However, our alignments are highly conserved, with a small number of multiple amino acid substitutions at a given position (i.e., up to 68% of the positions have less than 3 nucleotide substitutions), leaving little opportunity to exhibit site-specific amino acid preferences. Increasing the total tree length by 5, 10 and 50 using the multiplicative parameter λ_{TBL} leads to an increase in the number of false positives (Table 1 : 42.4%, 65.9% and 97.2%, respectively). Thus, site-heterogeneous amino acid preference can mislead results of the CUYN test, particularly at deep evolutionary scales or for fast evolving proteins.

2.5.6. CpG hypermutability can largely explain CU

The hypermutability of CpGs appears to be the model violation having the greatest impact on the rate of false positives when using the CUYN test. Given that CpG hypermutability is significant in mammals [Hodgkinson and Eyre-Walker, 2011], we suspect that the selection detected on CU by Yang and Nielsen [2008] is largely due to this phenomenon. We designed a crude experimental protocol to explore the ability of the null model, M[GTR]-S[1CatAA], and the alternative model, M[GTR]-S[1CatCodon], to reproduce the codon frequencies of the 16 gene alignments. To do so, we computed a distance between the mean RSCU retrieved from batches of sequences generated under the stationarity of the different models and that same statistic computed from the real sequences of our alignments (refer to Material and Methods for details on the procedure). As expected, the alternative model performed much better than the null model at predicting the codon frequencies, by rendering an RSCU closer to that of the true alignment (fig. 2).

We also explored the impact of CpG hypermutability on CU by generating sequences using parameter values from the stationarity of the null model obtained for each of the same 16 genes along with various values of λ_{CpG}, ranging from 0.1 to 20 (i.e., M[GTR+λ_{CpG}]-S[1CatAA]). Focusing on the RSCU retrieved from the batches of sequences generated under

the latter conditions, we sought to find the values of λ_{CpG} that minimized the distance with the observed CU, leading to a rough approximation of the maximum likelihood value of λ_{CpG} itself. We then compare the resulting distance with those retrieved from the null and the alternative models (fig. 2).

The distance to the true data greatly decreases when invoking λ_{CpG} within the null model to account for CpG hypermutability. Interestingly, the CU induced by M[GTR+λ_{CpG}]-S[1CatAA] model appears to be close to that of the M[GTR]-S[1CatCodon] model (fig. 2). In two cases, the rough optimization of λ_{CpG} brings the sequences drawn from the stationary distribution closer to the observed RSCU. This is particularly significant, since the M[GTR]-S[1CatCodon] model involves 41 additional jointly optimized parameters (relative to the null), whereas the M[GTR+λ_{CpG}]-S[1CatAA] involves only a single additional parameter, estimated crudely.

The batches of sequences simulated from the stationary distribution are independent one from the other, and thus computing the mean RSCU on them is justified. The sequences of the true alignments, however, are not independent of each other, such that averaging RSCU over them ignores their phylogenetic inertia. Therefore, investigated how the rough estimation procedure for λ_{CpG} using sub-sets of sequences from the true alignment when computing the mean RSCU, to the point of using a single sequence. As expected, using only one sequence picked from the true alignment leads to a high variance of the minimized distance between the true data and the simulated data (fig. S1). This is obviously a result of the low ratio between the degrees of freedom involved in the computation of the RSCU statistics (i.e., 41 df) and the number of codon states available in a single sequence from drawn from the true alignment. In spite of this high variance, the general tendency of the results is similar : the distance between true and simulated data is comparatively low when invoking λ_{CpG}. Results obtained when averaging with 1/3 and 2/3 of the sequences present in the true alignments (fig. S2 and S3) show progressively decreasing variance, as expected, and lead to very similar values of λ_{CpG} when compared to the results obtained with the entire alignment.

Figure 2 suggests that the selection detected on CU in mammals could be due to CpG hypermutability. The λ_{CpG} estimated from our rough procedure leads to values greater than 1, as expected, and ranges between 4 and 12. These are probably underestimates, as the

procedure is conditional on other parameter values obtained from the plain null model (i.e., M[GTR]-S[1CatAA]) rather than a proper joint estimation under the model with dependence across sites. The number of transition substitutions occurring in the CpG context is potentially limited, as the null model is probably attributing aspects of the hypermutability related to CpG contexts to the ω_* parameter and to the transition rates of the GTR mutation model. The CU of few genes (e.g., EDEM2) were not significantly altered by introducing the λ_{CpG} parameter, suggesting that CpG hypermutability is not among the most important factor determining CU for those genes. Other potential model violations should be further investigated, such as GC-biased gene conversion, and its potential interaction with CpG hypermutability.

2.6. Conclusions and future directions

In spite of being formulated from mechanistic principles to account for CU, the parameters introduced in the alternative model of the CUYN test appear to be absorbing other features of the evolutionary process than those intended, in a manner that is a statistically significant departure from the values expected under the null model. We have shown that violations on both the mutation and the selection aspects of the models can greatly impact the accuracy of the CUYN test, to a point where it may not be useful for the analysis of mammalian genes. We have shown that the frequent use of the M[HKY] setting should be avoided when the M[GTR] is available, since this oversimplified mutational parameterization alone can generate an important amount of false positives. We have also shown that two important aspects of evolution, CpG hypermutability, and, to a much lesser extent, site-heterogeneous preferences on amino acids, can mislead the CUYN test.

Simulation studies are key to evaluating the robustness of probabilistic inferences with respect to model violations that are well known to be pervasive in the data at hand. If the model-based test appears to be robust to violations (e.g., in some instances, site-heterogeneous selection on amino acids), its use is reassuring. If not (e.g., CpG hypermutability), the reliability of the test is in doubt, and a more complex model should be considered.

In the case of CpG hypermutability, the most obvious modeling expansion from the work we've conducted here would be to include the λ_{CpG} parameter within the overall inference.

While Monte Carlo methods for implementing such site-dependent models have been developed (e.g., Rodrigue and Lartillot 2012), our crude approximation based on simulations also suggests the use of *Approximate Bayesian computation* [Csilléry et al., 2010]. In any case, with such a formulation, it is hoped, the parameters introduced by the alternative model would no longer need to absorb CpG effects, and would presumably be "freed" for their intended purpose. Or, the introduced parameters could absorb other model violations, still. Indeed, the issue is much more difficult when model violations are poorly characterized (e.g., selective constraints on mRNA secondary structure) or even unknown, but the results we present here suggest that this should be a major part of research efforts.

Table 2.1. Comparison of the proportion of false positives detected when computing CUYN test on simulated alignments generated using parameter values inferred from various mutation-selection models.

models used to generate the simulated alignments	mutational model used to compute CUYN test	CUYN test (% positives)
M[GTR]-S[1CatAA]	M[GTR]	8.8
M[GTR]-S[1CatAA]	M[HKY]	56.1
M[HKY]-S[1CatAA]	M[GTR]	8.4
M[HKY]-S[1CatAA]	M[HKY]	5.2
M[HKY]-S[1CatAA]+($\lambda_{TBL} = 5$)	M[HKY]	4.2
M[HKY]-S[1CatAA]+($\lambda_{TBL} = 10$)	M[HKY]	5.3
M[HKY]-S[1CatAA]+($\lambda_{TBL} = 50$)	M[HKY]	5.7
M[GTR]-S[1CatAA]+($\lambda_{TBL} = 50$)	M[GTR]	5.7
M[GTR]-S[1CatCodon]	M[GTR]	100
M[GTR]-S[1CatCodon]	M[HKY]	100
M[HKY]-S[1CatCodon]	M[GTR]	100
M[HKY]-S[1CatCodon]	M[HKY]	100

λ_{TBL} corresponds to a multiplicative parameter used to dilate the branches length. λ_{CpG} corresponds to a multiplicative parameter used to modulate the transition rates of the CpG context.

Table 2.1. Comparison of the proportion of false positives detected when computing CUYN test on simulated alignments generated using parameter values inferred from various mutation-selection models.(*continued*)

models used to generate the simulated alignments	mutational model used to compute CUYN test	CUYN test (% positives)
M[Ts]-S[1CatAA]	M[HKY]	19.4
M[Tr]-S[1CatAA]	M[HKY]	43.1
M[Ts]-S[1CatAA]+($\lambda_{TBL} = 5$)	M[HKY]	24.8
M[Tr]-S[1CatAA]+($\lambda_{TBL} = 5$)	M[HKY]	74.1
M[Ts]-S[1CatAA]+($\lambda_{TBL} = 10$)	M[HKY]	27.0
M[Tr]-S[1CatAA]+($\lambda_{TBL} = 10$)	M[HKY]	77.1
M[Ts]-S[1CatAA]+($\lambda_{TBL} = 50$)	M[HKY]	33
M[Tr]-S[1CatAA]+($\lambda_{TBL} = 50$)	M[HKY]	85.3
M[HKY+$\lambda_{CpG} = 0.1$]-S[1CatAA]	M[HKY]	99.9
M[HKY+$\lambda_{CpG} = 0.5$]-S[1CatAA]	M[HKY]	67.1
M[HKY+$\lambda_{CpG} = 2$]-S[1CatAA]	M[HKY]	76.9
M[HKY+$\lambda_{CpG} = 4$]-S[1CatAA]	M[HKY]	99.8
M[HKY+$\lambda_{CpG} = 5$]-S[1CatAA]	M[HKY]	100
M[HKY+$\lambda_{CpG} = 8$]-S[1CatAA]	M[HKY]	100
M[HKY+$\lambda_{CpG} = 16$]-S[1CatAA]	M[HKY]	100
M[HKY]-S[NCatAA]	M[HKY]	20.6
M[HKY]-S[NCatAA]+($\lambda_{TBL} = 5$)	M[HKY]	42.4
M[HKY]-S[NCatAA]+($\lambda_{TBL} = 10$)	M[HKY]	65.9
M[HKY]-S[NCatAA]+($\lambda_{TBL} = 50$)	M[HKY]	97.2

λ_{TBL} corresponds to a multiplicative parameter used to dilate the branches length. λ_{CpG} corresponds to a multiplicative parameter used to modulate the transition rates of the CpG context.

2.7. Materials and Methods

2.7.1. Data preparation and tree inference

We queried the OrthoMaM database, version 9 [Douzery et al., 2014], to retrieve all gene alignments where all of the 43 species available in the database were present, which leads to a collection of 137 mammalian alignments. We then arbitrarily removed from these the 4 non-placental mammals that are part of OrthoMaM, in order to focus our study on the 39 placentals mammals. Each alignment was treated with HmmCleaner [Amemiya et al., 2013] and Gblocks [Talavera and Castresana, 2007] to remove structural annotation errors and poorly aligned regions, respectively. Gblocks was used under a non-stringent setting, resulting in very few positions being removed ($<$1.5%). Our analyses were conducted with a fixed tree topology, as obtained under the CAT model [Lartillot and Philippe, 2004] implemented in PhyloBayes-MPI [Lartillot et al., 2013b] on the amino acid concatenation of the 137 alignments (fig. S4). When compared to a recent review on the topic[Foley et al., 2016], our topology is almost identical, with the only difference being in the relationships at the base of Laurasiatheria, which precisely corresponds to very short internal branches. The impact of such topology variation is therefore expected to be negligible.

2.7.2. Inferring model parameters

Parameters were estimated on the 137 genes with PhyloBayes-MPI [Rodrigue and Lartillot, 2014] for the M[HKY]-S[1CatAA], M[GTR]-S[1CatAA], M[HKY]-S[1CatCodon], M[GTR]-S[1CatCodon], M[HKY]-S[NCatAA] and M[GTR]-S[NCatAA] models. Parameters were estimated separately on the 137 genes with CodeML for the M[HKY]-S[1CatAA], M[GTR]-S[1CatAA], M[HKY]-S[1CatCodon] and M[GTR]-S[1CatCodon] models. With CodeML, we found that 121 genes gave extreme parameter values when applying the M[GTR]-S[1CatCodon], for instance, with one of the nucleotide propensity parameters at a value close to 0 or 1. Yang and Nielsen [2008] mentioned that in comparing M[HKY]-S[1CatCodon] and M[GTR]-S[1CatCodon] "[...] the estimates of codon-fitness parameters for the concatenated data under the 2 mutation models are very different (results not shown). This is the case even though both mutation models predicted very similar codon frequency parameters, which closely match the observed frequencies. Our estimates of the selection coefficients are affected by the mutation model. Thus, we found that the LRT is somewhat insensitive to the assumed

mutation model but the estimates of codon fitness parameters are." Obviously, the mutation-level parameterization had to compensate in order to obtain similar codon frequencies. In order to keep computation-time of our simulation study manageable, we chose to work with the parameter values obtained from a sub-sample of alignments, and decided to work the 16 genes that did not exhibit this issue (this still implies 100×16 simulations for each condition, and several times as many ML inferences and Bayesian MCMC runs). Since the results based on CodeML and PhyloBayes-MPI for the M[HKY]-S[1CatAA], M[GTR]-S[1CatAA], M[HKY]-S[1CatCodon], and M[GTR]-S[1CatCodon] models were similar, we only present results based on CodeML (see Tables S1 and S2).

2.7.3. Simulation program

We wrote a simulation program that evolves sequences along a phylogenetic tree, generating DNA alignments with various mutation-selection models (see the previous section). Our simulation software can take as input the outputs of CodeML or of PhyloBayes-MPI when used with mutation-selection models.

In our implementation, sequences are evolved in a jump-chain manner, substitution by substitution, starting from the root and traversing all branches to the tips of the tree. We first sample a sequence from the stationary distribution as it would be under the site-independent model (with $\lambda_{CpG} = 1$). This sequence is then evolved for a large number of events, to reach stationarity under the site-interdependent framework, before being set as the starting state at the root of the tree. The simulation along each branch of the tree proceeds by first drawing a waiting time from an exponential distribution with a parameter corresponding to the sum of rates to all nearest-neighbors of the current sequence. If the waiting time drawn does not go beyond the length of the current branch, the nature of the event is drawn, with a probability proportional to the rate to each nearest-neighbor sequence. From this new state, and time-point along the branch, another waiting time is drawn as before, until the waiting time sampled goes beyond the length of the branch ; once it does, the state at the descendant node is set to the current sequence, and the simulation procedure splits and continues independently along the daughter branches. This continues until sampling waiting times beyond the final terminal branches, thereby yielding the states in the simulated alignment set.

2.7.4. Simulated datasets

We simulated datasets by using the parametric bootstrapping and the posterior predictive procedures. The different experimental conditions were defined by the models used. Experimental conditions were replicated at two levels : 100 simulated alignments were generated from the parameter values retrieved under the use of the different models on the 16 genes. The root position was set at 90% of the branch from the Afrotheria to the (Xenarthra + Laurasiatheria + Euarchontoglires). In some cases, we manipulated the values of the model parameters prior to the simulation procedures : (1) we set the transversion rates to the average of nucleotide exchangeability parameters of transversions obtained from a GTR-based analysis (i.e., M[Ts]) ; (2) we did as in (1), but for the transition rates rather than transversion rates (i.e., M[Tr]) ; (3) we increased the total tree length by 5, 10 and 50 times using a multiplicative parameter, λ_{TBL} ; and (4) we modulated the mutation rate in the CpG context with a multiplicative parameter, λ_{CpG}, using different values (i.e., 0.1, 0.5, 1, 2, 4, 5, 8, 16).

2.7.5. Approximating the observed CU

Here, we developed a rough methodology to help build our intuition about how CpG hypermutability could impact CU. We drew sequences from the stationary distribution (i.e., 10,000 sequences per set of parameter values) to study how the different models predict the codon frequencies of the true alignments. The comparison of the CU obtained under the different models is achieved by retrieving the squared Euclidean distance (i.e., equation 2.7.1) between the mean RSCU computed from 1,000 batches of 10 sequences drawn from each model stationary distribution and the mean RSCU obtained from the true alignment, or subsets thereof :

$$CU_{\text{dist}} = \sum_{a=1}^{61} (\log_2(z_a) - \log_2(y_a))^2, \tag{2.7.1}$$

where y and z are the mean codon state frequencies normalized by amino acid (i.e., RSCU) obtained from the batches of simulated sequences and the true sequence(s) respectively.

We investigated the effect of computing the distances when including different numbers of sequences randomly picked from the true alignment. This was achieved by randomly sampling

various proportion of sequences (i.e., 1/39, 1/3 and 2/3) from each alignment studied ; note that 39 sequences are present in each of the alignments studied. Our expectation was that the number of codon states present in a sequence, even if the sequences contain more than 500 codons, will limit the accuracy of the rough procedure we designed here, as the RSCU involves 41 degrees of freedom. We compared the ability of the null and the alternative models (i.e., M[GTR]-S[1CatAA] and M[GTR]-S[1CatCodon]) to predict the CU for each of the 16 genes, but also performed the comparison with the CpG context model M[GTR+λ_{CpG}]-S[1CatAA], incorporating different λ_{CpG} values (i.e., 0.25, 0.5, 0.75, 1, 2, 4, 8, 10, 12, 14, 16, 18, 20). We also tracked the proportion of CpG context substitutions occurring during the simulations.

2.8. Acknowledgments

We would like to thank both reviewers and the editor for their insightful comments. Also, we like to heartily thank Nicolas Lartillot for all the help provided during the project and commentary on the manuscript. This work was supported by the French Laboratory of Excellence project entitled TULIP (ANR-10-LABX- 41 ;ANR-11-IDEX-0002-02), and by the Natural Sciences and Engineering Research Council of Canada. Computations were made on the supercomputer Mammouth-série from Université de Sherbrooke, managed by Calcul Québec and Compute Canada. The operation of this supercomputer is funded by the Canada Foundation for Innovation, the ministère de l'Économie, de la science et de l'innovation du Québec and the Fonds de recherche du Québec - Santé.

2.9. Supplementary material

Supplementary tables S1.1-1.2 are available at Molecular Biology and Evolution.

— Table S1 : parameter values obtained under the different models implemented in CodeML : the null models (i.e., M[HKY]-S[1CatAA], M[HKY]-S[1CatCodon]) and the alternative models (i.e., M[GTR]-S[1CatAA] and M[GTR]-S[1CatCodon]) on 137 mammalian genes.

— Table S2 : mean parameter values (100 replicates per set of parameter values) obtained under the different models implemented in CodeML (i.e., M[HKY]-S[1CatAA], M[HKY]-S[1CatCodon], M[GTR]-S[1CatAA] and M[GTR]-S[1CatCodon]) applied on simulated alignments generated using parameter values obtained under various mutation-selection models on 16 mammalian genes.

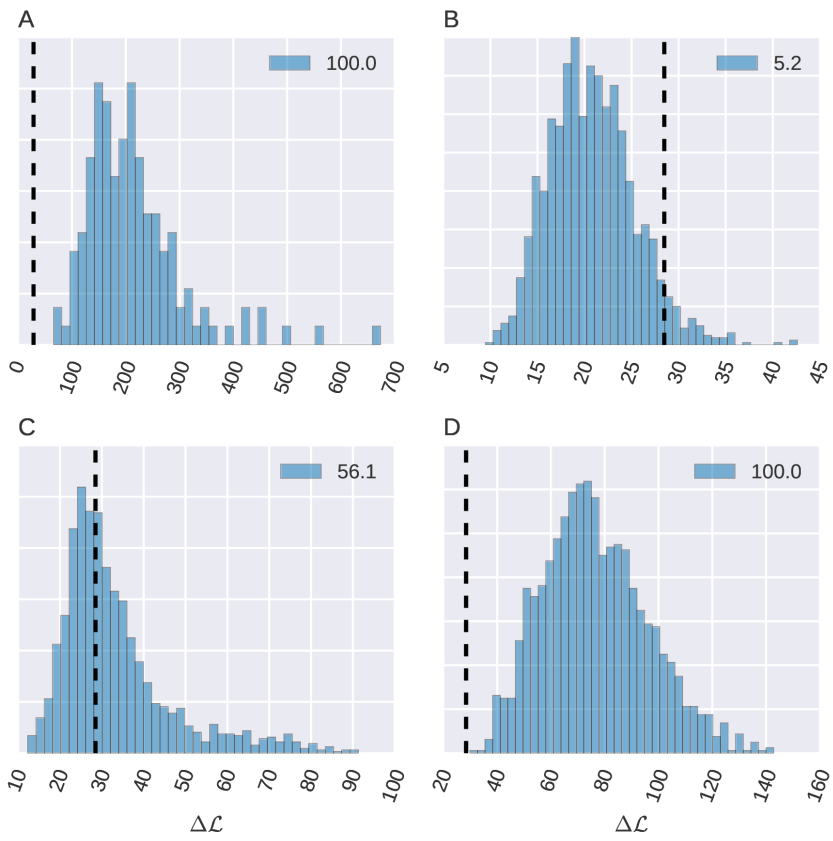

Figure 2.1. Histograms of the log-likelihood differences ($\Delta\mathcal{L}$) computed using the null hypothesis and the alternative hypothesis (i.e., M[HKY]-S[1CatAA] and M[HKY]-S[1CatCodon] respectively) on (A) 137 mammalian genes and on (B,C,D) simulated alignments, generated to mimic important aspects of mammalian evolution. (B) Distribution of log-likelihood differences computed on the simulated alignments (100 replicates per set of parameter values), generated from parameter values obtained under M[HKY]-S[1CatAA] on 16 genes. (C) Distribution of log-likelihood differences computed on the simulated alignment (100 replicates per set of parameter values), generated from parameter values obtained under M[GTR]-S[1CatAA] on 16 genes. (D) Distribution of log-likelihood differences computed on the simulated alignments (100 replicates per set of parameter values), generated from the parameter values obtained under M[HKY+$\lambda_{CpG} = 5$]-S[1CatAA] on 16 genes. The vertical line corresponds to the threshold of significance (i.e., 28.47) at 5% with 41 degrees of freedom (i.e., 60-19 parameters) according to the χ^2 distribution. The proportion of significant analyses is shown at top right.

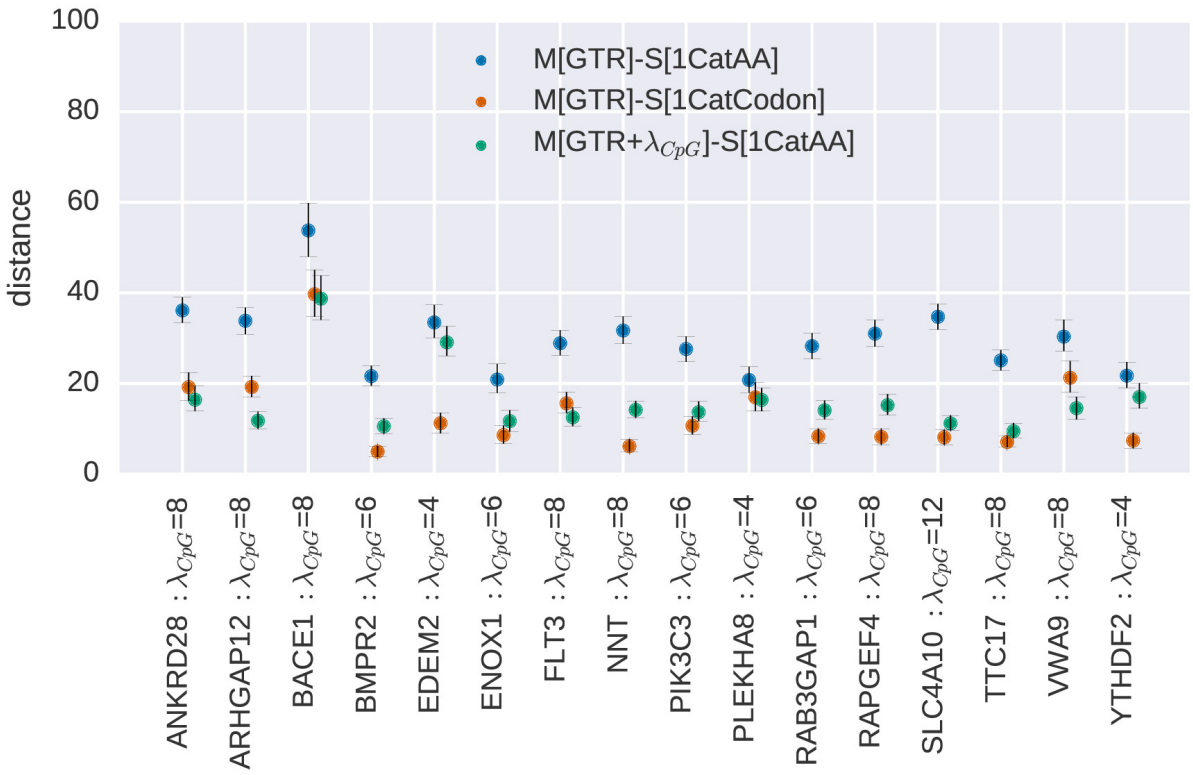

Figure 2.2. Models comparison on the basis of there ability to predict CU (i.e., blue : M[GTR]-S[1CatAA], red : M[GTR]-S[1CatCodon] and green from the rough approximation procedure : [GTR+λ_{CpG}]-S[1CatAA]). The mean distances (i.e., dots) obtained for each specific analysis of the 16 mammalian genes are plot along with their associated error bars, that corresponds to 95% intervals, where the closest distances (2.5%) and the farthest distances (2.5%) were removed. The distances are computed between the mean RSCU retrieved independently from batches of sequences (i.e., 1,000 batches of 10 sequences) all generated under the stationarity of each specific model used and the RSCU recovered from the true alignment. Only the results rendering the minus mean distance obtained from the rough optimization procedure are presented (i.e., green). The label of each gene analyzed is added as well as the values of λ_{CpG} that minimized the mean distance.

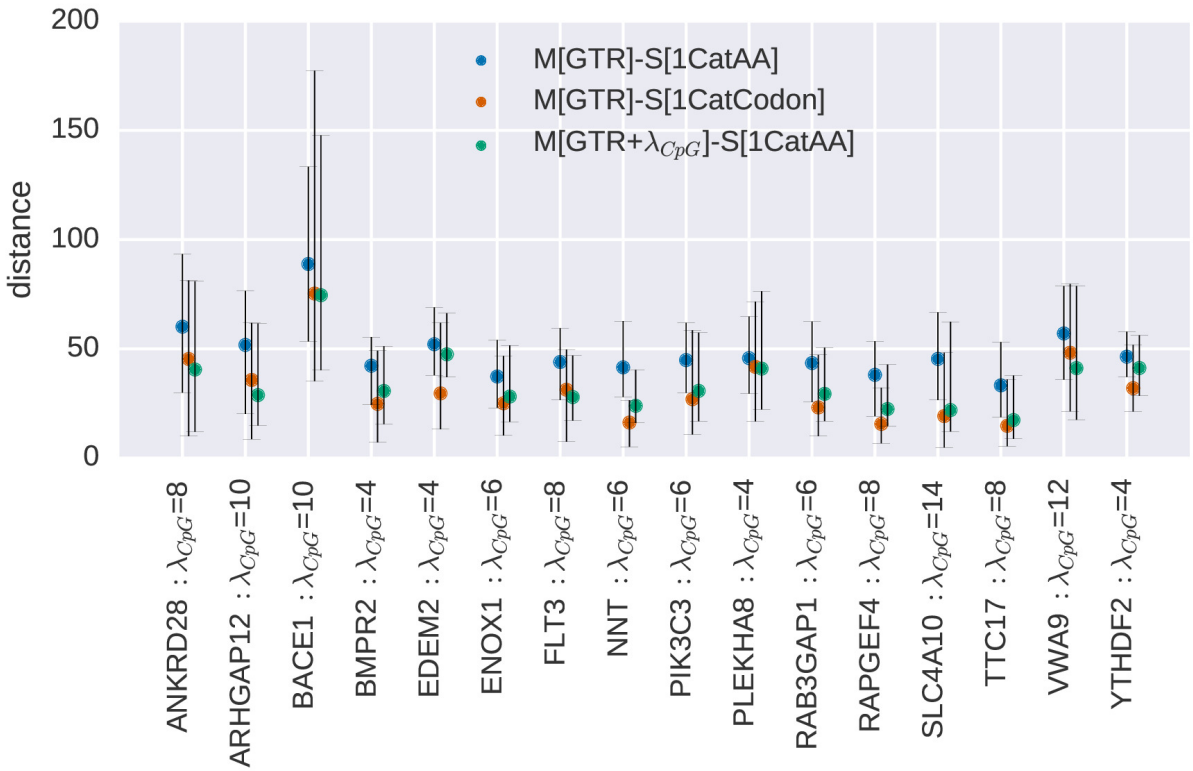

Figure S2.1. Models comparison on the basis of there ability to predict CU (i.e., blue : M[GTR]-S[1CatAA], red : M[GTR]-S[1CatCodon] and green from the rough approximation procedure : [GTR+λ_{CpG}]-S[1CatAA]). The mean distances (i.e., dots) obtained for each specific analysis of the 16 mammalian genes are plot along with their associated error bars, that corresponds to 95% intervals, where the closest distances (2.5%) and the farthest distances (2.5%) were removed. For each gene, the distances are computed between the mean RSCU retrieved independently from batches of sequences (i.e., 1,000 batches of 10 sequences) all generated under the stationarity of each specific model used and the RSCU recovered from one sequence randomly picked from the true alignment. Only the results rendering the minus mean distance obtained from the rough optimization procedure are presented (i.e., green). The label of each genes analyzed is added as well as the values of λ_{CpG} that minimized the mean distance.

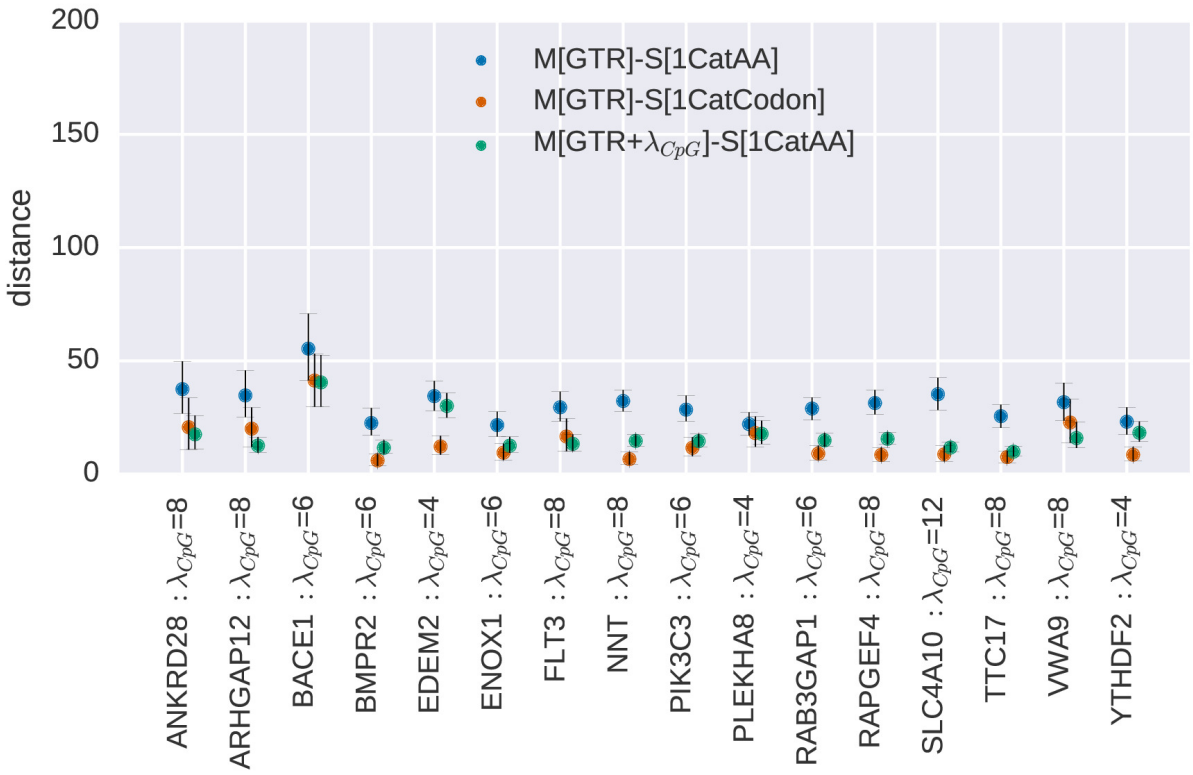

Figure S2.2. Models comparison on the basis of there ability to predict CU (i.e., blue : M[GTR]-S[1CatAA], red : M[GTR]-S[1CatCodon] and green from the rough approximation procedure : [GTR+λ_{CpG}]-S[1CatAA]). The mean distances (i.e., dots) obtained for each specific analysis of the 16 mammalian genes are plot along with their associated error bars, that corresponds to 95% intervals, where the closest distances (2.5%) and the farthest distances (2.5%) were removed. For each gene, the distances are computed between the mean RSCU retrieved independently from batches of sequences (i.e., 1,000 batches of 10 sequences) all generated under the stationarity of each specific model used and the RSCU recovered from 13 sequence randomly picked from the true alignment (i.e., 2/3 of the sequences). Only the results rendering the minus mean distance obtained from the rough optimization procedure are presented (i.e., green). The label of each genes analyzed is added as well as the values of λ_{CpG} that minimized the mean distance.

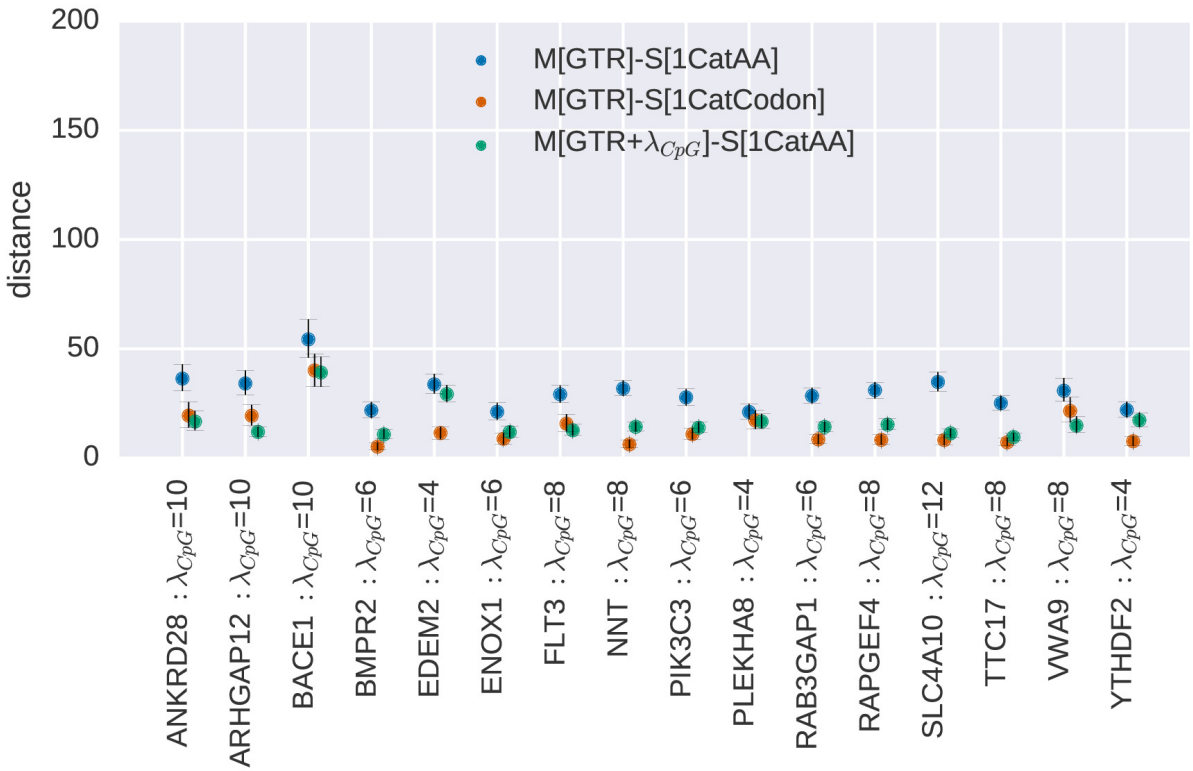

Figure S2.3. Models comparison on the basis of there ability to predict CU (i.e., blue : M[GTR]-S[1CatAA], red : M[GTR]-S[1CatCodon] and green from the rough approximation procedure : [GTR+λ_{CpG}]-S[1CatAA]). For each genes, The mean distances (i.e., dots) obtained for each specific analysis of the 16 mammalian genes are plot along with their associated error bars, that corresponds to 95% intervals, where the closest distances (2.5%) and the farthest distances (2.5%) were removed. The distances are computed between the mean RSCU retrieved independently from batches of sequences (i.e., 1,000 batches of 10 sequences) all generated under the stationarity of each specific model used and the RSCU recovered from 26 sequences randomly picked from the true alignment (i.e., 2/3 of the sequences). Only the results rendering the minus mean distance obtained from the rough optimization procedure are presented (i.e., green). The label of each genes analyzed is added as well as the values of λ_{CpG} that minimized the mean distance.

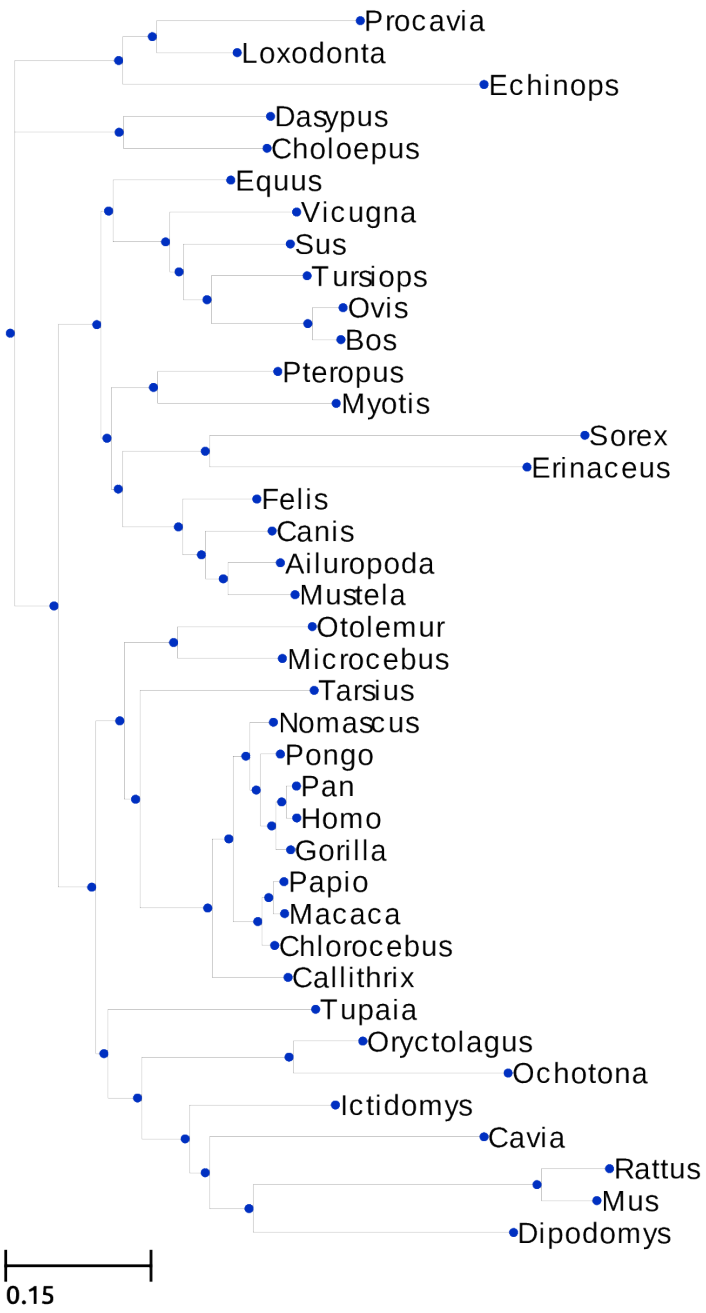

Figure S2.4. Consensus tree obtained with the CAT model on the amino acid concatenation of the 137 gene alignments (121,441 amino acid positions). The scale bar represents the expected mean number of substitutions per site.

Chapitre 3

Méthode de calcul bayésien approché conditionnel : une nouvelle approche pour les modèles de type mutation-sélection site-interdépendents et de haute dimensionalité

Molecular Biology and Evolution, Volume 35, Issue 11, November 2018, Pages 2819–2834, https ://doi.org/10.1093/molbev/msy173 Published : 07 September 2018

3.1. Information

Conditional Approximate Bayesian Computation, a new approach for across-site dependency in high-dimensional mutation-selection models

Simon Laurin-Lemay*[1], Nicolas Rodrigue[2], Nicolas Lartillot[3], Hervé Philippe*[1,4]

[1]Robert-Cedergren Center for Bioinformatics and Genomics, Department of Biochemistry and Molecular Medicine, Faculty of Medicine, Université de Montréal, Montréal, Québec, Canada

[2]Department of Biology, Institute of Biochemistry, and School of Mathematics and Statistics, Carleton University, Ottawa, ON, Canada

[3]Laboratoire de Biométrie et Biologie Évolutive, UMR CNRS 5558, Université Lyon 1, Lyon, France

[4]Centre de Théorisation et de Modélisation de la Biodiversité, Station d'Écologie Théorique et Expérimentale, UMR CNRS 5321, Moulis, France

Running head : Approximate Bayesian Computation for modeling CpG hypermutability

Keywords : Markov chain Monte Carlo, synonymous substitution, non-synonymous substitution, posterior predictive, phylogenetics

*Correspondence : simon.laurin-lemay@umontreal.ca and herve.philippe@sete.cnrs.fr

3.2. Résumé

Une question centrale au domaine de la biologie moléculaire et évolutive concerne les rôles déterminants des processus mutationnels et des contraintes de sélection sur l'organisation des génomes. Les processus mutationnels et les contraintes de sélection sont hétérogènes le long du génome et dans le temps. Les modèles de substitution des codons de type mutation-sélection sont des approches mécanistiques prometteuses pour identifier l'effet du processus mutationnel et de sélection. Dans la pratique, cependant, plusieurs complications surviennent, puisque la prise en compte de ces hétérogénéités impliquent souvent l'utilisation de modèles possédant une paramétrisation de haute dimensionnalité (e.g., préférences site-spécifique pour les acides aminés) ou devant prendre en compte des interdépendances dans les données (e.g., hypermutabilité CpG), rendant la fonction de vraisemblance insoluble. Le calcul bayésien approximatif (ABC) permet de contourner ce problème. Nous proposons ici une nouvelle approche, appelée ABC conditionnel (CABC), qui combine l'efficacité d'échantillonnage du MCMC et la flexibilité de l'ABC. Pour illustrer le potentiel de l'approche CABC, nous l'appliquons à l'étude de l'hypermutabilité du contexte CpG chez les mammifères par l'utilisation d'un paramètre mutationnel impliquant une dépendance entre sites adjacents, combinée à une sélection purificatrice de préférence site-spécifique sur les acides aminés modélisé au moyen d'un processus de Dirichlet. Notre démonstration d'efficacité de la méthodologie CABC ouvre de nouvelles perspectives de modélisation. De plus, notre application de la méthode révèle un niveau élevé d'hétérogénéité de l'hypermutabilité CpG entre loci et une légère hétérogénéité entre les groupes taxonomiques étudié ; et enfin, nous montrons que l'hypermutabilité CpG est un facteur déterminant de l'évolution de l'usage des codons synonymes. Le code source est disponible sous forme de dépôt GitHub (https ://github.com/Simonll/LikelihoodFreePhylogenetics.git).

3.3. Abstract

A key question in molecular evolutionary biology concerns the relative roles of mutation and selection in shaping genomic data. Moreover, features of mutation and selection are heterogeneous along the genome and over time. Mechanistic codon substitution models based on the mutation–selection framework are promising approaches to separating these effects. In practice, however, several complications arise, since accounting for such heterogeneities often implies handling models of high dimensionality (e.g., amino acid preferences), or leads to across-site dependence (e.g., CpG hypermutability), making the likelihood function intractable. Approximate Bayesian Computation (ABC) could address this latter issue. Here, we propose a new approach, named Conditional ABC (CABC), which combines the sampling efficiency of MCMC and the flexibility of ABC. To illustrate the potential of the CABC approach, we apply it to the study of mammalian CpG hypermutability based on a new mutation-level parameter implying dependence across adjacent sites, combined with site-specific purifying selection on amino-acids captured by a Dirichlet process. Our proof-of-concept of the CABC methodology opens new modeling perspectives. Our application of the method reveals a high level of heterogeneity of CpG hypermutability across loci and mild heterogeneity across taxonomic groups ; and finally, we show that CpG hypermutability is an important evolutionary factor in rendering relative synonymous codon usage. All source code is available as a GitHub repository (https ://github.com/Simonll/LikelihoodFreePhylogenetics.git).

Keywords : Markov chain Monte Carlo, synonymous substitution, nonsynonymous substitution, posterior predictive, phylogenetics Issue Section : METHODS Associate Editor : Rasmus Nielsen

3.4. Introduction

The mutational process, a main basis of genetic variability, itself varies according to the environment (e.g., abiotic : Maharjan and Ferenci 2017 ; biotic Krasovec et al. 2017) and along the genome [Hodgkinson and Eyre-Walker, 2011]. One example of this mutational heterogeneity is the case of cytosine being much more mutable when followed by guanine in the genomes of vertebrates [Bird, 1980], a phenomenon known as CpG hypermutability. As a result, a biased variability is subjected to selective processes, leaving a signal that seems clear in the cases of parallel adaptation [Stoltzfus and McCandlish, 2017]. Features of selection are probably even more heterogeneous along the genome and over time than those of mutation. Selection acts at multiple levels (e.g., DNA, RNA, protein, cell, tissue, organism, population, community, and ecosystem), and conflicts can exist between levels or because of fluctuations in the environment. The heterogeneity of selection is obvious when examined at a fine scale, for instance within a protein, where each site typically displays a strong preference for a small sub-set of amino acids [Halpern and Bruno, 1998, Lartillot and Philippe, 2004, Rodrigue and Lartillot, 2012, Tamuri et al., 2012, Rodrigue, 2013, Rodrigue and Lartillot, 2014, Tamuri et al., 2014, Echave et al., 2016, Hilton et al., 2017, Rodrigue and Lartillot, 2017, Wang et al., 2018].

In comparative genomics, these complexities make it difficult to separate the effects of selection from the bias induced by mutational features. Codon usage (CU) in mammals provides a good illustration of this problem. Some authors argue that selection is acting on CU [Yang and Nielsen, 2008, Kessler and Dean, 2014] to favor efficiency of translation [Drummond and Wilke, 2008, Cannarrozzi et al., 2010, Tuller et al., 2010], while others argue that population sizes are too small to allow selection of such a minor advantage, particularly in Primates [Duret, 2002b, Pouyet et al., 2016, Laurin-Lemay et al., 2018a, Galtier et al., 2018], and therefore that CU is the result of neutral evolution [Ohta, 1973]. In agreement with the latter view, CU in mammals mostly reflects GC3 content [Sueoka, 1961, 1962, Muto and Osawa, 1987, Ermolaeva, 2001, Knight et al., 2001, Chen et al., 2004, Li et al., 2015] or isochore structure [Filipski et al., 1973, Bernardi, 2000], suggesting that it is determined by the mutational pressure and by fixation biases likely related to GC-biased gene conversion [Duret, 2002b, Duret and Galtier, 2009a, Katzman et al., 2011, Glemin et al., 2015].

A promising solution to tease apart mutation and selection in coding sequences is to develop mechanistic codon substitution models [Rodrigue and Philippe, 2010] that operate in a mutation-selection framework. Such mutation-selection models have previously been developed to study the role of protein structure [Robinson et al., 2003, Rodrigue et al., 2006, 2005, 2009, Kleinman et al., 2010], codon preference [McVean and Vieira, 2001, Nielsen et al., 2007, Rodrigue et al., 2008b, Yang and Nielsen, 2008, Rodrigue and Philippe, 2010, Pouyet et al., 2016], or site-specific amino acid preferences [Halpern and Bruno, 1998, Rodrigue et al., 2010, Tamuri et al., 2012, Rodrigue, 2013, Tamuri et al., 2014]. However, thus far, the main focus has been on the modeling of complex features of selection, whereas simple, homogeneous, parameterization were used for the mutational aspects of the model, often the very simple HKY model [Blanquart and Lartillot, 2008]. Yet, violations in the mutational part of the model can easily lead to erroneous detection of selection [e.g., Lartillot, 2013, Van den Eynden and Larsson, 2017, Laurin-Lemay et al., 2018a]. In particular, the latter study shows erroneously inferred selection on codon usage when using simple models on sequence alignments simulated with mild CpG hypermutability, but without any selection on codon usage.

To take full advantage of the mutation-selection models, it may be necessary to incorporate more complexity (i.e., natural heterogeneity) in both mutation-level and selection-level specifications of the model. However, heterogeneity often implies handling parameter vectors of high dimensionality and across-site dependency, both of which create computational difficulties. High dimensionality can lead to over-fitting in a maximum likelihood framework. As for across-site dependency, it leads to intractable likelihood calculations [precluding the use of the pruning algorithm ; Felsenstein, 1973, 1981]. The Bayesian framework, thanks to the use of Markov chain Monte Carlo (MCMC ; Metropolis et al. 1953, Hastings 1970), enables the study of rich models accounting for across-site heterogeneity of amino acid profiles, as previously shown in the case of site-specific amino acid preferences [Rodrigue et al., 2010]. Approximate Bayesian Computation (ABC) avoids the computation of the likelihood [Pritchard et al., 1999, Beaumont et al., 2002, Marjoram et al., 2003, Sisson et al., 2007], and could be a means of addressing the across-site dependency issue, whether at the level of mutation (e.g., CpG contexts ; Pedersen et al. 1998, Jensen and Pedersen 2000, Christensen et al. 2005, Arndt et al. 2003, Hwang and Green 2004, Huttley 2004, Siepel and Haussler

2004, Arndt and Hwa 2005, Christensen 2006, Hobolth et al. 2006, Hobolth 2008, Duret and Arndt 2008, Lindsay et al. 2008, Misawa and Kikuno 2009, Suzuki et al. 2009, Misawa 2011, Keightley et al. 2011, Ying and Huttley 2011, Berard and Gueguen 2012, Lee et al. 2015, 2016, Huttley and Yap 2012) or selection (e.g., on protein structure ; Robinson et al. 2003, Rodrigue et al. 2006, 2005, 2009, Kleinman et al. 2010). Unfortunately, the classical rejection sampling ABC cannot deal with complex models involving parameter vectors of high dimensionality [Kousathanas et al., 2016]. Here we propose a new approach, named Conditional ABC, which combines the advantages of MCMC and ABC. As a proof-of-concept, we study the across-site dependent hypermutability of CpG, while modelling the high dimensionality of site-specific amino acid selection.

3.5. New approaches : Conditional ABC

We consider the general situation where we have a model with parameters (λ, θ) and a data set D under study. The parameter θ represents the (potentially high-dimensional) nuisances. The parameter λ, on the other hand, is our parameter of interest. Computationally, the model is assumed to be intractable by classical MCMC in the generic case, except under a reference value for λ (e.g., $\lambda = 1$). Here, we can think of λ as the relative rate of CpG mutation—a feature that implies across-site dependency—such that, for $\lambda = 1$, the model reduces to the usual site-independent model.

Ideally, we would like to sample from the joint posterior :

$$(\lambda, \theta) \sim p(\lambda, \theta|D), \tag{3.5.1}$$

and then conduct inference on λ (e.g., by visualising the marginal posterior distribution of λ and computing the mean and 95% credible interval). Noting that the joint posterior can be factorized as follows :

$$p(\lambda,\theta|D) = p(\lambda|D,\theta)p(\theta|D), \tag{3.5.2}$$

the sampling procedure denoted by equation (3.5.1) could equivalently be done in two steps :

$$\theta \sim p(\theta|D), \tag{3.5.3a}$$

$$\lambda \sim p(\lambda|D, \theta), \tag{3.5.3b}$$

i.e., by first sampling θ from its marginal posterior (marginal over λ) and then sampling λ from its conditional posterior (conditional on θ).

Neither of the two sampling steps described by 3.5.3a and 3.5.3b can be performed exactly. Accordingly, the Conditional ABC (CABC) approach proposed here relies on two main approximations. First, the marginal posterior (3.5.3a) is approximated by the posterior on θ under the reference model, i.e., $p(\theta|D, \lambda = 1)$, using MCMC ; we denote this approximated posterior distribution as $p_{MCMC}(\theta|D, \lambda = 1)$. Second, sampling λ conditional on θ (3.5.3b) is done by classical ABC, denoted $p_{ABC}(\lambda|D, \theta)$. Provided that the nuisance parameters and λ are weakly correlated under the true posterior, these approximations should be relatively accurate.

In summary, the approach proceeds in two steps :

$$\theta \sim p_{MCMC}(\theta|D, \lambda = 1), \tag{3.5.4a}$$

$$\lambda \sim p_{ABC}(\lambda|D, \theta), \tag{3.5.4b}$$

or equivalently :

$$(\lambda, \theta) \sim p_{CABC}(\lambda, \theta|D), \tag{3.5.5}$$

where :

$$p_{CABC}(\lambda, \theta|D) = p_{ABC}(\lambda|D, \theta)p_{MCMC}(\theta|D, \lambda = 1). \tag{3.5.6}$$

Comparing (3.5.2) and (3.5.6), the two approximations invoked by the CABC are the use of ABC, instead of exact Bayesian inference on λ conditional on θ, and the fact that θ is not from its marginal posterior (marginal on λ), but is instead from its reference posterior (with $\lambda = 1$).

In practice, some of the nuisance parameters collectively denoted by θ might be strongly correlated with λ, in which case the approach will be inaccurate. Let us further subdivide the parameterization, by defining :

$$\theta = (\theta_{sc}, \theta_{wc}), \tag{3.5.7}$$

where θ_{sc} is strongly correlated, and θ_{wc} weakly correlated, with λ under the joint posterior. Provided that θ_{sc} is sufficiently low-dimensional, we can re-sample it by ABC, along with λ :

$$\theta_{wc} \sim p_{MCMC}(\theta_{wc}|D, \lambda = 1), \tag{3.5.8a}$$

$$(\lambda, \theta_{sc}) \sim p_{ABC}(\lambda, \theta_{sc}|D, \theta_{wc}), \tag{3.5.8b}$$

or equivalently :

$$(\lambda, \theta_{sc}, \theta_{wc}) \sim p\prime_{CABC}(\lambda, \theta_{sc}, \theta_{wc}|D), \tag{3.5.9}$$

where :

$$p\prime_{CABC}(\lambda, \theta_{sc}, \theta_{wc}|D) = p_{ABC}(\lambda, \theta_{sc}|D, \theta_{wc}) p_{MCMC}(\theta_{wc}|D, \lambda = 1). \tag{3.5.10}$$

Working with (3.5.10) instead of (3.5.6) will decrease the impact of the approximation implied by using the reference marginal posterior, as opposed to the true marginal posterior, on a smaller component of the parameter vector, although at the cost of an increase in the impact of the approximation entailed by conducting the ABC on a higher dimensional parameter (λ, θ_{sc}). Note that we do not have a theoretical basis from which to establish which nuisance parameters are to be considered as weakly or strongly correlated to λ_{CpG}. This problem is to be addressed empirically, exploiting our knowledge of the underlying biology and modeling system, and ultimately studied through simulations.

To illustrate the potential of the CABC approach, we apply it to the estimation of the well-established hypermutability at CpG sites—which involves dependence across sites—in the context of a complex reference model combining site-independent mutation, handled by a general-time-reversible nucleotide level parameterization, [Lanave et al., 1984], denoted M[GTR], along with purifying selection on amino-acids (i.e., site-specific amino-acid preferences) captured by a Dirichlet process prior [Rodrigue et al., 2010], denoted S[NCatAA*]. In this specific application of CABC, the parameter of interest (denoted above as λ) is λ_{CpG}, the ratio of the mutation rate of transitions at CpG sites to the mutation rate of transitions at non-CpG sites. The reference model (without CpG hypermutation, or equivalently, with $\lambda_{CpG} = 1$) is denoted by M[GTR]-S[NCatAA*], while the complete model (with CpG hypermutation) is referred to as M[GTR+ts-CpG]-S[NCatAA*].

The high-dimensional parameter vector of the reference model was partitioned into strongly and weakly correlated components, as discussed above, by reasoning as follows. On one hand, the estimation of the site-specific fitness profiles and of relative branch lengths should be robust to the specific model used for the mutation process (whether or not CpG hypermutation is included). On the other hand, the context-independent component of the mutation process (the GTR process) is expected to be strongly correlated with λ_{CpG} under

the true posterior distribution. Accordingly, the high-dimensional amino-acid profiles and the branch lengths were estimated by MCMC under the reference posterior distribution, with $\lambda_{CpG} = 1$ (i.e., were included in θ_{wc}), while the 10 GTR parameters (8 degrees of freedom), as well as three modulator parameters (meant as correcting factors for total tree length, mean non-synonymous/synonymous rate deviation, and relative position of the root along the branch separating the in- and the out-group, see Materials and Methods for details) were re-estimated at the ABC step, along with λ_{CpG} (i.e., were therefore included in θ_{sc}). We study the approach using simulations, and apply it to 137 real protein-coding genes from 39 mammals (see Materials and Methods for details).

3.6. Results and Discussion

3.6.1. Validation of the CABC procedure

We validated the CABC approach using simulations. We simulated 5,000 alignments, using various values for λ_{CpG} (ranging from 0.5 to 8) combined with empirically estimated parameter values for the reference model. Then, we applied the CABC approach to these simulated alignments, and evaluated the relative mean square error (RMSE) of the approximated posterior and the coverage properties of the posterior credible intervals. For the ABC step, we used two alternative approaches : either a simple ABC rejection sampling (RS) algorithm [Pritchard et al., 1999], or a more sophisticated approach based on the use of a linear regression model (LRM) for getting closer to the true posterior distribution [Blum and Francois, 2010]. Note that the ABC step itself relies on simulations, with numerous summary statistics computed and compared between this step's simulated data sets and the data set under analysis. Only a small percentages of these simulations are retained by the procedure, as controlled by the *tolerance* level. We used 13 summary statistics related to the frequency of certain states and the counts of specific pairwise differences between sequences (see Materials and Methods for details). We explored empirically different sample sizes and tolerance levels.

All the approximate posterior distributions obtained by running the CABC procedure with the RS algorithm alone were inaccurate : the global RMSE ranged from $\sim$4 to $\sim$34 depending on the value of λ_{CpG} (tables 1-2). The global RMSE decreases when the tolerance level is decreased (from 1% to 0.1% of the simulated samples), but remains high even under

the most stringent settings, suggesting that much smaller tolerance levels (implying a much larger total number of simulated samples) would still be needed in order for the simple RS approach to yield a reasonable approximation to the true posterior distribution [Barber et al., 2015].

Table 3.1. Global relative mean square error (without λ_{ROOT}) computed for different λ_{CpG} values (1000 replicates per λ_{CpG} value) under two tolerance levels, 10% and 1%, and over 10^5 simulations.

Method	Tolerance level	$\lambda_{CpG} = 0.5$	$\lambda_{CpG} = 1$	$\lambda_{CpG} = 2$	$\lambda_{CpG} = 4$	$\lambda_{CpG} = 8$
RS	10%	34.30±3.09	15.06±3.13	10.25±3.6	9.49±3.89	9.87±4.10
RS	1%	18.20±4.03	10.02±1.73	6.78±1.84	6.06±2.20	6.53±2.55
RS+LRM	10%	0.86±0.19	0.86±0.14	0.89±0.13	0.89±0.12	0.86±0.16
RS+LRM	1%	0.70±0.19	0.69±0.14	0.71±0.13	0.72±0.12	0.71±0.13

RS : rejection sampling algorithm

LRM : linear regression model

Table 3.2. Global relative mean square error (without λ_{ROOT}) computed for different λ_{CpG} values (1000 replicates per λ_{CpG} value) under two tolerance levels, 1% and 0.1%, and over 10^6 simulations.

Method	Tolerance level	$\lambda_{CpG} = 0.5$	$\lambda_{CpG} = 1$	$\lambda_{CpG} = 2$	$\lambda_{CpG} = 4$	$\lambda_{CpG} = 8$
RS	1%	18.21±3.86	10.01±1.66	6.78±1.84	6.05±2.18	6.54±2.56
RS	0.1%	8.22±2.98	6.27±1.35	4.50±0.76	3.73±0.95	4.07±1.28
RS+LRM	1%	0.69±0.18	0.69±0.14	0.71±0.13	0.71±0.11	0.70±0.12
RS+LRM	0.1%	0.53±0.17	0.52±0.14	0.54±0.13	0.54±0.11	0.54±0.11

RS : rejection sampling algorithm

LRM : linear regression model

In contrast, the global RMSE obtained when using the LRM of Blum and Francois [2010] fall under 1 (tables 1-2). The accuracy of CABC with LRM rose when the sampling effort is increased and the tolerance level is reduced (tables 1-2). The global RMSE remained very similar, around 0.70, when using the best 1% of the 10^5 simulations compared to the best 1% of 10^6 simulations. A reduction of the tolerance level (0.1%), however, decrease RMSE, by ∼20% (table 2). We note that, in spite of performing well here, the behavior of LRM in presence of model violations has been shown to be potentially misleading [Frazier et al., 2017, unpublished data]. Given our simulation results, however, all rejection sampling results were corrected using the LRM unless stated otherwise.

The RMSE associated with each parameter (fig. 1 and table S1) appears to be strongly linked to the amount of signal relevant to that parameter. For instance, the RMSE for the transition exchangeabilities (ϱ_{AG} and ϱ_{CT}) were 4 times lower than for the transversion exchangeabilities (ϱ_{AC}, ϱ_{AT}, ϱ_{CG} and ϱ_{GT}). Similarly, the four nucleotide propensities have the smallest RMSE, being in fact the smallest for the two most frequent nucleotides (C and G). The RMSE for λ_{ROOT} (the correcting factor for the relative position of the root along the branch separating the in- and the out-group) is the highest (> 0.30); indeed the posterior distribution of this parameter is almost identical to its prior distribution, demonstrating that the signal provided by the non-reversibility of the context-dependent mutation process is too tenuous to be captured when analyses are conducted on single genes.

The improvement brought by a sample size of 10^6 and a tolerance level of 0.1% applied mainly to λ_{TBL} and λ_{ω_*} (the correcting factors for total tree length and dN/dS deviation), as well as the transversion exchangeabilities. In contrast, the improvement was minor for λ_{CpG}. Of note, the RMSE for λ_{CpG} is smaller under high rates of CpG hypermutability, reflecting the more abundant empirical signal (i.e., a higher number of CpG hypermutation events) in this regime; thus, when the true λ_{CpG} is equal to 8, the corresponding RMSE (0.0362) is below that observed for transversion exchangeabilities and close to the one obtained for transition exchangeabilities. To explore the idea that more evolutionary signal leads to a decreased in RMSE, we plotted the relation between the total number of expected substitutions and the RMSE computed on λ_{CpG} (fig. S1). As expected, we found a negative relationship between the amount of evolutionary signal and the RMSE on λ_{CpG}. Moreover, as the evolutionary signal for λ_{CpG} becomes more prominent (panels S1 : A-E), the fit of the regression become

higher : the r^2 values go from 0.249 (lowest value of λ_{CpG}, 0.5) to 0.797 (highest value of λ_{CpG}, 8). As a result, when applied to real data, CABC will be precise if there is a high rate of transition in the CpG context. In conclusion, the computational burden of 10^6 simulations is mainly useful if one wants to study the effect of CpG on the transversion rates. Otherwise, a less intense sampling effort (10^5), combined with a moderate tolerance level (1%), gives reasonably accurate inference.

The coverage properties (i.e., the frequency at which credible intervals cover the true value) provide another interesting perspective on the statistical properties of CABC. Here, Probability-Probability (P-P) plots are used to investigate the coverage properties for several parameters of interest. On these plots, a straight line along the diagonal indicates that the nominal and true coverage coincide, that is, $1 - \alpha$ credible intervals cover the true value at a frequency equal to $1 - \alpha$. If this is the case, then credible intervals are true frequentist confidence intervals. Coverage is not necessarily expected to be perfect for all aspects of the model (i.e., nuisance parameters), but is an important property when the intention is to test a null hypothesis (e.g., $\lambda_{CpG} = 1$), with a frequentist control of the type I error (rate of false positive).

The coverage properties were poor for all parameters when using RS alone (fig. S2). As for RMSE, the use of LRM greatly improved the concordance between nominal and true coverage (fig. S3), while the increase in sample size from 10^5 to 10^6 allowed a minor improvement (fig. S4). The coverage properties were good for all parameters but λ_{ω_*}, λ_{TBL} and λ_{ROOT}, and to a lesser extent for nucleotide exchangeabilities when $\lambda_{CpG} = 8$. The poor coverage of ϱ_{AG} and ϱ_{CT} when λ_{CpG} is greater than 1 (fig. S4) could be explained by the rise of the uncertainty since a great amount of mutational signal related to the GTR component is transferred to the λ_{CpG}. Importantly, our parameter of interest, λ_{CpG}, had excellent coverage properties (fig. 2), which is of prime importance to test the hypothesis that λ_{CpG} is greater than 1.

We further characterized the properties of CABC by transferring the parameters of the GTR mutation model from θ_{sc} to θ_{wc} at the ABC step. The expectation is that CABC will be inaccurate, because the hypermutability of CpG will lead to an artifactual increase in the transition/transversion ratio and the A+T content inferred under the reference model. To investigate this point, we used simulations made with a $\lambda_{CpG} = 8$ and a sample size of 10^5. Indeed, under these new settings, the relative mean square error on λ_{CpG} was much

increased (with a 2-fold increase of the RMSE). Similarly, coverage was poor for λ_{CpG}, as well as for all the GTR parameters (fig. S5). This is in sharp contrast to the case where the GTR parameters are re-estimated (i.e., within θ_{sc}, fig. S3) : in this case, λ_{CpG} and nucleotide propensities (except φ_G) are well estimated. The estimation of relative nucleotide exchangeabilities is equally poor in the two cases, suggesting that these parameters might not be strongly correlated with λ_{CpG} (see below), but probably just impacted by the lack of signal under $\lambda_{CpG} = 8$ for the GTR component, as previously explained. The correcting factors, λ_{TBL} and λ_{ω_*}, are more are more accurately inferred when the GTR parameters are not themselves re-estimated (fig. S3, fig. S5).

Finally, we evaluated the impact of the estimation of the large number of nuisance parameters represented by branch lengths and site-specific amino-acids profiles on the overall accuracy of CABC, by running the entire procedure with all these parameters fixed to their true values. Granting perfect knowledge about these nuisances is expected to improve the accuracy of the estimation of all other parameters. However, if the improvement turns out to be minor, this will show that (i) in itself, uncertainty about these nuisance parameters is not detrimental, and (ii) our approximation based on estimating these nuisances under the reference model (and not under the target model) does not compromise the overall quality of the inference. We used simulations made with a $\lambda_{CpG} = 8$ and a sample size of 10^5.

The RMSE for all parameters were very similar to the results obtained under the standard validation procedure (table S1). For instance, the estimation of λ_{CpG} was only weakly impacted by the use of the true branch lengths and the true site-specific amino acid preferences. The parameter most impacted was λ_{ω_*} : its RMSE decreased from 0.100 to 0.053 (table S1), accounting for 67% of the reduction of the global RMSE. The P-P plots (fig. S6) are in agreement with RSME and are very similar to the case where we drew θ_{wc} from $p(\theta_{wc}|D, \lambda_{CpG} = 1)$ (fig. S3).

In conclusion, the CABC procedure is reasonably accurate as long as the parameters included in θ_{wc} are indeed weakly influenced by λ_{CpG}. In particular, the accuracy suggested by our simulation study is largely sufficient to test the hypothesis that λ_{CpG} is equal to 1, with a good control of type I error, and even to study the impact of the CpG hypermutability on the GTR parameters (with a somewhat greater uncertainty concerning the four transversion

exchangeabilities). From here, all the results we present below are obtained with the LRM (see Materials and Methods) made on the 0.1% best of 10^6 simulations.

3.6.2. Estimation of the mutation rate in the CpG context using CABC

We applied the CABC to approximate the posterior distribution of λ_{CpG} for a sample of 137 mammalian genes from 39 species (fig. 3). In agreement with previous observations [Hodgkinson and Eyre-Walker, 2011], CABC always inferred a posterior mean transition rate in the CpG context greater than one, with an average value of 7.45 ; none of the 137 genes included $\lambda_{CpG} = 1$ within their 99% credible intervals (table S2). The bimodal shape of the marginal distribution (fig. 3) is due to two genes (*ARNT* and *KIAA0100*) for which the transition rate in the CpG context obtains a posterior mean of 18.8 and 19.5, respectively (table S2). A total of 16 genes displayed posterior mean values for λ_{CpG} greater than 10, i.e., outside the prior belief [1/10,10]. Values outside the prior were reached through the use of the LRM approach. To further explore this result, new CABC analyses were conducted over the 137 genes with a broader prior (log-uniform over [1/50,50]), using the same sampling scheme and tolerance level. The impact of the prior on the estimation of λ_{CpG} was minor (fig. S7), as indicated by the fact that the posterior means are highly correlated between the two alternative prior settings ($R^2 = 0.98$). Of note, the use of a narrow prior (over [1/10,10]), leads to an underestimation of λ_{CpG}, making our approach conservative in the evaluation of hypermutability in the CpG context.

We then applied CABC to investigate whether the value of λ_{CpG} is homogeneous across the placental tree. We subdivided our data sets into three clades : Glires (7 species), Laurasiatheria (14 species), and Primates (12 species). The gene-specific estimates of λ_{CpG} obtained for each of these three clades (fig. S8) are well correlated with the ones obtained for placentals (r^2=0.94, 0.96 and 0.96, respectively), indicating that the hypermutability in the CpG context is relatively well conserved along the placental tree. However, the slope of the regression (passing through the origin) is below one (0.96 and 0.91) for Glires and Laurasiatheria, respectively, and greater than one (1.14) for Primates. The higher level of CpG transition rate in Primates is congruent with the results of Keightley et al. [2011], although heterogeneity across clades is less marked here. This could be due to the fact that the analysis of

Keightley et al. [2011] is based on pairs of species, whereas the present analysis relies on the information contributed by 39 placental species considered simultaneously.

Finally, we looked at the effect of taking into account CpG hypermutability on the other parameters of the mutation process of the model, M[GTR+ts-CpG]. Overall, these parameters were slightly affected by the inclusion of the λ_{CpG} parameter, which is understandable given the relative rarity of CpG in mammalian protein coding sequences (with the mean observed/expected CpG ratio of 0.41). However, comparison of the values of $\varphi_G + \varphi_C$ (fig. 4) shows that the CpG hypermutability has a complex effect, strongly dependent on the gene. This is congruent with our assumption that the GTR parameters are strongly correlated with λ_{CpG} and should be re-estimated at the ABC step. On average, a tenuous increase in $\varphi_G + \varphi_C$ is observed (fig. 4), which is expected since the hypermutability of CpG tends to decrease G+C content.

3.6.3. Posterior predictive checks to analyze the effect of CpG hypermutability on some sequence characteristics

Instead of looking at the GTR parameters, a more sensible approach is to examine the predictions made by both models, including and not including CpG hypermutability. First, we compared the GC content observed at the third codon positions (GC3) in empirical data to the GC3 predicted by the two models (fig. 5). The model not including CpG hypermutability over-predicts GC3, as previously noticed by [Mugal et al., 2015]. The model including CpG hypermutability gets closer to the observed GC3, but with a small under-prediction especially for high values of GC3. The inclusion of the CpG hypermutability by CABC therefore allows to improve the prediction of GC3, a widely used measure to estimate mutational pressure [Sueoka, 1961, 1962, Muto and Osawa, 1987, Ermolaeva, 2001, Knight et al., 2001, Chen et al., 2004, Li et al., 2015]. Second, during the simulations conducted to generate the posterior predictive alignments, we computed statistics on the substitution histories over the tree (table 3). The number of substitutions is higher for the model including CpG hypermutability compared to the reference model (6342±3202 versus 5214±2511). Focusing on the relative frequencies of substitution types, the model including CpG hypermutability predicted more transitions relative to transversions (77.3% versus 71.3% for the reference model). The C->T and G->A transition rates show the sharpest increase (+14.2% and +10.7%), in agreement

with the increased transition rate at CpG sites implied by CpG hypermutability, but the T->C (+7.1%) and A->G (+1.0%) transition rates also show a non-negligible increase. The pattern is similarly complex for transversions, with an important decrease for G->T (-54.5%), G->C (-61.6%), C->G (-67.3%) and C->A (-69.4%), the other types of transversions being relatively unaffected. The complexity of the impact of the CpG hypermutability on the relative frequencies of the 12 types of substitutions is difficult to interpret, being the result of an interplay between the mutation process, the genetic code and selection on amino-acids.

Table 3.3. Comparison of the proportions of substitution types recovered from the posterior predictive simulations (mean over 137 mammalian gene analyses).

Type of substitutions	without CpG	with CpG
ts	71.33±3.38	77.29±2.91
A > G	17.16±2.1	17.33±2.05
G > A	17.04±2.15	18.87±2.05
C > T	18.49±2.01	21.12±2.25
T > C	18.64±2.08	19.97±2.17
tv	28.67±3.38	22.71±2.91
A > C	4.70±0.93	4.68±0.88
C > A	4.74±0.92	3.29±0.77
A > T	2.53±0.7	2.53±0.7
T > A	2.59±0.72	2.46±0.68
C > G	3.70±1.05	2.49±0.75
G > C	3.62±1.06	2.23±0.71
G > T	3.52±1.08	1.92±0.49
T > G	3.28±0.69	3.10±0.59

ts for transitions ; tv for transversions ;

Third, we exclusively looked at the substitutions in the CpG context (table 4), which should be easier to interpret. Unsurprisingly, the number of CpG->TpG or ->CpA transitions among all substitutions were much more frequent (from 234±133 to 584±318) than other substitution types. When analyzed with respect to the position of CpG within codons, it appears that only CpG->CpA at positions 2-3 and CpG->TpG at positions 3-1 drastically increase under the model with CpG hypermutability. This is entirely expected since most of these substitutions are synonymous. In fact, the proportion of non-synonymous substitutions (transitions) at CpG sites only increases from 1.9% to 2.5% while the synonymous transitions jump from 7.5% to 14.6%. This is congruent with the analysis of thousands of genes between human and chimpanzee showing that about 14% of the substitutions (synonymous or non-synonymous) are related to CpG hypermutability [Misawa and Kikuno, 2009]. Table 4 shows that selection at the amino-acid level severely filters the effect of CpG hypermutability [Stoltzfus and McCandlish, 2017], but suggests that codon usage might be affected (see below).

Table 3.4. Comparison of the proportion of transition substitutions within CpG context recovered from the posterior predictive simulations (mean over 137 mammalian gene analyses)

Codon position	Substitution types	Without CpG	With CpG
1-2	CG > TG	0.14±0.14	0.23±0.22
2-3	CG > TG	0.54±0.32	0.65±0.42
3-1	CG > TG	4.33±1.05	8.27±2.14
1-2	CG > CA	0.45±0.36	0.75±0.57
2-3	CG > CA	3.19±0.79	6.36±1.44
3-1	CG > CA	0.78±0.42	0.88±0.54
synonymous	CG > TG + CG > CA	7.52±1.47	14.63±2.85
non-synonymous	CG > TG + CG > CA	1.91±1.02	2.51±1.47

Fourth, we investigated the dinucleotide frequencies related to CpG hypermutability (CpG, TpG and CpA) and, as negative controls, the other dinucleotides involving the same pairs of nucleotides (GpC, GpT and ApC). For all codon positions (i.e., 1-2, 2-3, 3-1) the negative controls are similarly, and accurately, predicted by both models, with or without CpG hypermutability (fig. S9-11, D-F). In contrast, introducing CpG hypermutability severely impacted the prediction of CpG, TpG and CpA dinucleotide frequencies (fig. S10-11, A-C), except at codon position 1-2 (fig. S9, A-C). This is expected because almost all substitutions at these positions are non-synonymous, hence almost exclusively predicted by the selection part of the model, which is identical between the two models. At codon positions 2-3 and 3-1, the CpG frequency is always better predicted by the model that includes CpG hypermutability (fig. S10-11, A) and are in fact very close to the observed values (mean Z-score of -0.58 and 0.01, respectively). The frequency of TpG at codon position 3-1, and of CpA at position 2-3, are both better predicted by the model with CpG hypermutability (fig. S11C and S10B). As noticed for the predicted substitutions (table 4), the mutational results of CpG hypermutability are synonymous events at the codon level. For TpG (CpA) frequency at codon position 3-1 (2-3), the predictions of the two models are virtually identical because these products of CpG hypermutability are non-synonymous. Including the CpG hypermutability therefore allows to improve the prediction of dinucleotide frequencies almost exclusively in a synonymous context. In contrast, in a non-synonymous context, the model without CpG hypermutability appears to yield globally correct predictions.

Fifth, we compared the amino acid frequencies predicted by the two models. We did not observe major differences (fig. S12), again, probably because this characteristic is mainly modelled by the selection part, which is shared by the two models. However, it is known that the mutational process has an impact on amino acid frequencies, through variation in GC content in mitogenomes [Foster et al., 1997], or differences between the leading and lagging strands [Rocha et al., 1999]. The case of arginine constitutes a good illustration of this specific point. The frequency of arginine is over-predicted by the reference model, and under-predicted by the model including CpG hypermutability. Strikingly, arginine is the only amino acid encoded by codons having a CpG at position 1-2 (CGN, and also by codons AGR). If the selective advantage of arginine at a given position is not sufficiently strong, the high mutational pressure away from CpG can easily lead to the replacement of arginine

by a less favourable amino acid. The case of arginine also demonstrates that site-specific amino-acid preferences might in fact be correlated with λ_{CpG} under the posterior, something which was ignored in our analysis, by pre-estimating amino-acid fitness parameters under the reference model (without CpG) without any subsequent correction. In this respect, one possible improvement of our approach would be to globally modulate the site-specific amino acid fitness profiles using a vector of 20 correcting factors that would be estimated at the ABC step. However, the pattern shown in figure S12 is complex. For instance, it is not clear why the model including CpG hypermutability over-predicts the frequencies of isoleucine (codons AUH) and methionine (codon AUG).

3.6.4. Posterior predictive checks and the codon usage bias in mammals

We have shown that the modelling of CpG hypermutability has a major impact on the ability to predict synonymous aspects of mammalian protein coding gene evolution. It is therefore particularly interesting to examine its effect on codon usage. We used posterior predictive checks to study the entropy of relative synonymous codon usage (RSCU, fig. 6) and of relative codon frequencies (RFC, fig. S13). The results obtained with these two alternative statistics were similar and we will focus on RSCU, which is commonly used in empirical analyses of codon usage (e.g., Pouyet et al. 2017). The model with CpG hypermutability more accurately predicts the codon usage entropy observed on the empirical alignments, compared to the reference model (fig. 6) : the mean Z-scores are -4.84 and -3.16, respectively. Since the entropy is maximal under equal use of each synonymous codon, the predicted RSCU are generally more homogeneous than the observed RSCU. A large proportion of the genes (41.6% and 20.4% for the models without and with λ_{CpG} respectively) yields very poor predictions of the entropy of RSCU with Z-scores under -5 (fig. 6). This suggests that other important determinants, such as splicing enhancer, or mRNA structure, are still missing to our modeling strategy.

To better understand the mutational or selective forces determining the small entropy of RSCU observed in mammalian protein coding genes, we performed a principal component analysis of the RSCU predicted by the two models, along with the RSCU observed in the empirical alignments (fig. 7). The first axis of the PCA explains most of the variance (59.2%), and is related to the GC3 content (the r^2 between the first axis and the GC3 of the real

alignments is equal to 0.984). This is congruent with similar analyses based on a larger number of genes but restricted to *Homo sapiens* (e.g., Pouyet et al. 2017). The model without CpG hypermutation is slightly shifted to the right (fig. 7), in agreement with its GC3 over-prediction (fig. 5). In contrast, the predictions of the model including CpG are comparable to real data on the first axis. The second axis explains 14.3% of the variance and strongly discriminates the real data from the predictions of the reference model, the predictions of the model with CpG hypermutability being in-between. The model that includes CpG hypermutability is, as expected from previous results, closer to the real data.

All the G/C ending codons (in red) but TTG and AGG are located to the right, in agreement with the correlation between the first axis and GC3 content. The second axis, driven by the difference between observed and predicted data, is more complex to interpret. The codons ending by CpG are all located in the lower right corner, indicating that CpG hypermutability contributes to this axis. Including λ_{CpG} is indeed necessary to explain why codons TCG, GCG, CCG and ACG are un-preferred in humans with a RSCU of 0.05, 0.11, 0.11, 0.11 respectively [Nakamura et al., 2000], whereas G ending codons are always otherwise preferred. However, the synonymous products of transitions of these codons (NCA) do not meaningfully contribute to the second axis. In contrast, three codons ending by G (GTG, CTG and CAG, up right corner) heavily contribute to this axis but do not seem to be linked to CpG. Codons CTA (Leucine), ATA (Isoleucine) and GTA (Valine) are also major drivers of the second axis. They are over-predicted by both models. A deficit of TpA could be due to the hypermutability of this dinucleotide [Milholland et al., 2017] or to selection against the attachment to transcription or termination factors [Burge et al., 1992]. Codons for arginine are also separated on the second axis, AGR clearly on the upper part and CGN on the lower one (very weakly for CGG). This is likely related to CpG hypermutability, which will erode CGN codons towards TGN and CAN. The possibly complicated evolutionary path between CGN and AGR codons to conserve functionally-important arginines could be responsible for the surprising position of codon AGG along the first axis. In summary, the PCA of RSCU (fig. 7) demonstrates that including CpG hypermutability into the mutation-selection model leads to an improved prediction of codon usage but that other characteristics (e.g., TpA) are poorly predicted, requiring future additions in the mutation and/or selection part(s) of the model.

3.7. Conclusions and future directions

We have proposed a new approach, CABC, that combines MCMC and ABC to simultaneously handle high-dimensional parameter vectors and site-interdependent substitution processes. We have shown that this approach allows accurate estimation of the level of transition hypermutability in the CpG context. Our analysis confirms that CpG hypermutability is prevalent in mammals and variable among loci. This proof of concept of the CABC methodology opens new perspectives towards improved mutation-selection models better able to tease apart the relative role of these two evolutionary forces.

We used a simple implementation for ABC, where summary statistics were manually selected and the posterior distribution was approximated with the rejection sampling algorithm followed by the use of a LRM. This appears to be sufficient to accurately estimate the rate of CpG hypermutation, λ_{CpG}, although some biases and/or inaccuracies were observed for other parameters (e.g. λ_{TBL} and λ_{ω_*}). From there, the method could be improved in several respects. For instance, rejection sampling could be replaced by MCMC [Marjoram et al., 2003] or by sequential Monte Carlo [Sisson et al., 2007]. Similarly, LRM could be replaced by other regression models [e.g., random forest ; Raynal et al., 2017, unpublished work], in the hope of getting closer to the true posterior and potentially reducing the computation burden. The choice of summary statistics could also be reconsidered, for instance by computing the number of substitutions by maximum parsimony instead of simply counting the number of observed pairwise differences, which might improve the estimation of λ_{TBL}. Perhaps more importantly, the choice of summary statistics could be performed automatically [Prangle et al., 2014b]. The Random Forest ABC [Pudlo et al., 2016] may be particularly well suited for sequence data, for which hundreds of summary statistics can in principle be contemplated. Finally, one specific aspect of strategy that was adopted here, i.e., introducing modulator parameters, which are estimated at the ABC step, to correct for the fact that most nuisance parameters are sampled under the reference posterior distribution (i.e. under $\lambda_{CpG} = 1$), could be generalized to other aspects of the model, in particular, to amino-acid frequencies across the proteome (as illustrated by the case of arginine, fig. S12).

The CABC approach will make it possible to develop complex mutation-selection models handling several of the well identified and complex features of mutation and selection processes. Concerning mutation, context-dependent effects are clearly understudied in molecular

evolution, mostly for computational reasons, and despite the fact that the prevalence of such effects is widely recognized [Siepel and Haussler, 2004, Nevarez et al., 2010, Seplyarskiy et al., 2017, Guo et al., 2018]. In addition to CpG hypermutability, TpA hypermutation [Milholland et al., 2017] or more complex context-dependent mutational pattern, such as inferred from the large number of de novo mutations discovered through the sequencing of trios (e.g., Francioli et al. 2015, Wong et al. 2016, Jonsson et al. 2017), could be further investigated. Concerning selection, the perspectives are broader, including, among other things, selection against mono-nucleotide repeats, mRNA secondary structure, motif for RNA binding proteins (e.g., splicing enhancers) and obviously protein structure. Such improved models should have a broad applicability in molecular evolutionary studies, by making it possible to tease apart the role of mutation, purifying and diversifying selection in the evolution of genomic sequences.

3.8. Materials and Methods

3.8.1. Data sets and tree topology

All the 137 mammalian gene alignments used in this work as well as the mammalian tree were recovered from , and both are available via the GitHub repository (https ://github.com/Simonll/LikelihoodFreePhylogenetics/ ; last accessed July 24, 2018).

3.8.2. Codon substitution models

To mechanistically disentangle mutation from selection processes, we used the codon substitution model of Rodrigue et al. [2010] with the modification of Rodrigue and Lartillot [2017], as implemented in Phylobayes-MPI [Lartillot et al., 2013b, Rodrigue and Lartillot, 2014]. Let us briefly recall the parameterization of this reference model, denoted as M[GTR]-S[NCatAA*]. Branch lengths are free parameters, while the tree topology is kept fixed. The mutational part of the model, M[GTR], is modelled with the general-time-reversible approach [Lanave et al., 1984] using 10 parameters (8 degrees of freedom) and assumes a point mutation process between codon a to codon b. Stop codons are prohibited (i.e., have zero probability). The mutational process act identically on all codon positions (1, 2 and 3), whereas codons a and b differ at the cth position. The nucleotide propensities, are defined as $\varphi = (\varphi_n)_{1 \leq n \leq 4}$, with $\sum_{n=1}^{4} \varphi_n = 1$ and the nucleotide exchangeabilities are defined as

$\varrho = (\varrho_{mn})_{1\leq m,n\leq 4}$, with $\sum_{1\leq m<n\leq 4} \varrho_{mn} = 1$. The selective part of the model, S[NCatAA*], acts at the amino acid level. The amino acid relative scaled fitness *profiles*, or NCatAA from S[NCatAA*], are elements of a Dirichlet process [Rodrigue et al., 2010]. The Dirichlet process is a nonparametric method, controlled by a few hyper-parameters, which allows to approximate any unknown distribution [Ferguson, 1973]. Here the dimensionality of the latent variables is huge (e.g., the number of profiles times 20 amino acids) noting that there is in average about 70.80±22.42 profiles to deal with when working with the mammalian gene alignments studied here. The K profiles are defined as vectors $\psi = (\psi_l^{(k)})_{1\leq l\leq 20, 1\leq k\leq K}$. Site specific allocation of the K profiles is specified for the length of the gene, N, via the vector $z = (z_i)_{1\leq i\leq N}$. Therefore, the scaled selection coefficient for non-synonymous events, $S_{ab}^{(i)}$, is obtained as in Yang and Nielsen 2008 :

$$S_{ab}^{(i)} = \ln(\frac{\psi_{f(b)}^{(z_i)}}{\psi_{f(a)}^{(z_i)}}), \tag{3.8.1}$$

where $f(a)$ returns an index, from 1 to 20, of the amino acid encoded by codon a. The value of $S_{ab}^{(i)}$ in turn defines a fixation factor, denoted $h(S_{ab}^{(i)})$, and calculated as

$$h(S_{ab}^{(i)}) = \frac{S_{ab}^{(i)}}{1 - e^{-S_{ab}^{(i)}}}. \tag{3.8.2}$$

A deviation parameter, ω_*, or $*$ from S[NCatAA*], was recently introduced by [Rodrigue and Lartillot, 2017] to capture the excess or the deficit of non-synonymous rates with respect to the purifying selection modelled by the amino acid fitness profiles, corresponding to Darwinian selection or other forms of purifying selection (e.g., the secondary structure of mRNA or the 3D structure of protein), respectively. The substitution rate matrix Q of the reference model has entries of the form :

$$Q_{ab}^{(i)} = \begin{cases} \varrho_{a_c b_c}\varphi_{b_c}, & \text{if syn.,} \\ \varrho_{a_c b_c}\varphi_{b_c} h(S_{ab}^{(i)})\omega_*, & \text{if non-syn.} \end{cases} \tag{3.8.3}$$

To capture the transition mutation rate in the CpG context (i.e., CpG > TpG or CpG > CpA), we extended the mutation component M[GTR] of the reference model by including an across-site dependent parameter, λ_{CpG}. The Q matrix has now entries of the form :

$$Q_{ab}^{(i)} = \begin{cases} \varrho_{a_c b_c} \varphi_{b_c}, & \text{if syn. tr. or ts. non-CpG,} \\ \varrho_{a_c b_c} \varphi_{b_c} \lambda_{CpG}, & \text{if syn. ts. CpG,} \\ \varrho_{a_c b_c} \varphi_{b_c} \omega_* h(S_{ab}^{(i)}), & \text{if non-syn. tr. or ts. non-CpG,} \\ \varrho_{a_c b_c} \varphi_{b_c} \omega_* h(S_{ab}^{(i)}) \lambda_{CpG}, & \text{if non-syn. ts. CpG.} \end{cases} \tag{3.8.4}$$

3.8.3. Overview of CABC

The hypermutability of C in the CpG context introduces across-site dependency, making the computation of the likelihood intractable. CABC eschews this difficulty by inferring the high-dimensional parameter vectors using a standard MCMC with the model M[GTR]-S[NCatAA*], i.e., with $\lambda_{CpG} = 1$, and then by using ABC to infer the posterior distribution of the model with CpG hypermutation, assuming the values of the high-dimensional parameter vectors previously estimated. More precisely, the parameter vectors θ_{sc} and θ_{wc} of equation 9 consist of the propensities and exchangeabilities of the GTR matrix (φ and ϱ) plus three modulators (λ_{ω_*}, λ_{TBL} and λ_{ROOT}) and of the branch lengths plus the amino acid fitness profiles, respectively. As explained above, the parameters of θ_{wc} are of high dimensionality and cannot be accurately inferred by ABC, whereas the parameters of θ_{sc} are strongly correlated with the site-interdependent parameter λ_{CpG} and therefore cannot be inferred by MCMC. As a result, θ_{wc} is first obtained using MCMC assuming $\lambda_{CpG} = 1$ and λ_{CpG} and θ_{sc} are obtained using ABC conditional on θ_{wc} as formulated CABC equation 3.5.9. Priors are defined for λ_{CpG} and θ_{sc} before running the ABC step.

3.8.4. MCMC part of CABC

We applied the reference model implemented in Phylobayes-MPI on the 137 alignments, composed of 39 placentals, available from Laurin-Lemay et al. [2018a]. All the analyses were carried under fixed topology previously obtained by Laurin-Lemay et al. [2018a]. The analyses was also conducted on subparts of the mammalian tree : Glires (7 species), Laurasiatheria (14 species) and Primates (12 species). Convergence was first visually assessed using two independent chains, and then by computing the effective size of the parameters. The priors used under the reference model are listed in Rodrigue and Lartillot [2014]. Parameters from

θ_{wc} are drawn from the posterior distribution estimated under the reference model using MCMC (i.e., assuming $\lambda_{CpG} = 1$).

3.8.5. ABC part of CABC

In addition to the 10 GTR parameters, which are expected to be strongly correlated to λ_{CpG}, θ_{sc} includes λ_{TBL}, a modulator serving as a multiplicative parameter for every branch lengths. As the mutation-selection equilibrium was disrupted by the new parameterization, the tree length (Total Branch Length or TBL), which is measured in the number of mutations per site, will likely increase because of the additional mutations proposed at CpG sites (when $\lambda_{CpG} > 1$). Also included in θ_{sc} is λ_{ω_*}, a multiplicative modulator to ω_*, to respond to changes in non-synonymous rates that might emerge when accommodating CpG hypermutability. It is difficult to anticipate the value of λ_{ω_*}, given the potentially complex interplay between parameter values that could be produced from modeling CpG hypermutation. Finally, θ_{sc} includes λ_{ROOT}, a multiplicative parameter to fix the exact position of the root of the tree between the in- and out-group. Since we used a model, M[GTR+ts-CpG]-S[NCatAA*], that makes the process non-time-reversible, the position of the root, in our case on the branch separating Afrotheria (set as the out-group from Xenarthra + Euarchontoglires + Laurasiatheria), influences the output. It is difficult to know whether the phylogenetic signal will be sufficient to precisely estimate λ_{ROOT}.

Priors used are non-informative except for the nucleotide exchangeabilities, or ϱ. We informed the model that the transition rates are in average two times higher than transversions [Wakeley, 1996] to reduce the dimensionality of the ABC search. The use of non-informative priors makes the rejection of the null hypothesis more reliable at the expense of computational time.

$$\lambda_{CpG} \sim \log_{10} \text{Uniform}[0.1,\ 10]$$

$$\varrho_{ts} \sim \text{Gamma}[\alpha = 1,\ \beta = 1]$$

$$\varrho_{tr} \sim \text{Gamma}[\alpha = 2,\ \beta = 1]$$

$$\varphi \sim \text{Dirichlet}[1,\ 1,\ 1,\ 1]$$

$$\lambda_{\omega_*} \sim \log_2 \text{Uniform}[0.5,\ 2]$$

$$\lambda_{TBL} \sim \log_2 \text{Uniform}[0.5,\ 2]$$

$$\lambda_{ROOT} \sim \text{Uniform}[0,\ 1]$$

The simulator developed in Laurin-Lemay et al. [2018a] allows one to generate sequence alignments from the model with CpG across-site dependency along a phylogenetic tree. It was modified to work in parallel and to compute distances between the vectors of summary statistics (SS) recovered from simulated and true alignments. Concretely, the simulator program generates a reference table of SS along with the parameter values (SS, λ_{CpG}, θ_{sc}, θ_{wc}), ordered by increasing distance values. The ABC rejection sampling algorithm [RS ; Pritchard et al., 1999] was implemented into the simulation package. One can run the RS for a defined number of simulations (i.e., sampling size) and select the best simulations (i.e., tolerance level) on the basis of the distances computed for each simulation (see below). The selected simulations correspond to the *RS table* used to approximate the posterior distribution. The program is accessible via the GitHub repository (https ://github.com/Simonll/LikelihoodFreePhylogenetics/ ; last accessed July 24, 2018). The two steps procedure (MCMC followed by CABC) developed here takes for a single gene analysis approximately 10 hours on an AMD Opteron 6172 using 12 cores (for 100,000 simulations).

The SS are key to capturing the relevant information in the data [Fu and Li, 1997, Tavare et al., 1997, Weiss and von Haeseler, 1998, Pritchard et al., 1999]. Preliminary analyses were performed to select among >200 possible SS those that are the most useful in discriminating different values of λ_{CpG} and θ_{sc}. Thirteen SS were selected to summarize the alignments. First, we used the relative dinucleotide frequency of CpG, TpG, and CpA (SS_{C3pG1}, SS_{T3pG1}, SS_{C3pA1}) at the third and first positions of two adjacent codons, mainly in order to fit the λ_{CpG} parameter. Second, the frequency of four nucleotides at the third codon positions (SS_{A3}, SS_{C3}, SS_{G3}, SS_{T3}) was considered, mainly to fit the nucleotide propensities, or φ. Third, the sum over all the possible pairs of sequences of the absolute numbers of differences were computed at the nucleotide level for each possible unordered pair of nucleotides, leading to six summary statistics ($SS_{A<>C}$, $SS_{A<>G}$, $SS_{A<>T}$, $SS_{C<>G}$, $SS_{C<>T}$, $SS_{G<>T}$) ; they should mainly allow to fit the exchangeability parameters (ϱ), but also λ_{TBL}. Fourth, we also computed the sum over all the possible pairs of sequences of the absolute number of all non-synonymous differences indiscriminately (SS_{NS}), with the aim to fit λ_{TBL} and λ_{ω_*}. We did not find any summary statistics informative for λ_{ROOT}. In this study, the ordering of simulations was achieved by using the squared Euclidean distance. All the 13 summary

statistics were log base 2 transformed to avoid over representing SS with large values (e.g., $SS_{A<>C} \sim 10^5$ while SS_{C3pG1} are $\sim 10^{-2}$) when applying the distance function.

Two sampling sizes (10^5 or 10^6 simulations) were investigated under the RS algorithm. To approximate the posterior distribution of λ_{CpG} and θ_{sc}, we selected the best simulations following different tolerance levels : we kept the best 10% or 1% for the sampling size of 10^5 or the best 1% or 0.1% for 10^6. Given the large combinatorics of parameter values for λ_{CpG} and θ_{sc}, it is likely that the RS algorithm would require a much larger sampling size to accurately infer the posterior distribution [Barber et al., 2015]. To get closer to the true posterior distribution, we modeled the relationship between the parameter values sampled during the CABC (i.e., λ_{CpG}, θ_{sc}) as the response variables and the SS present in the RS table of the best simulations as the predictors of a regression model as introduced by Beaumont et al. [2002]. More specifically, we applied the nonparametric weighted multiple linear regression model (previously identified as LRM), which also accounts for heteroskedasticity, as proposed by Blum and Francois [2010] and available in the ABC package [Csilléry et al., 2012] from R CRAN [R Core Team, 2017]. The weighted scheme, done for each entry of the RS table, are obtained by applying an Epanechnikov kernel to the Euclidean distances computed. In other words, the weights are maximal for the entries with the smallest distances, and minimal for the biggest distances. This ensures that the LRM optimizes its parameters from the best samples present in the RS table.

3.8.6. Validation of CABC

To validate the new CABC method we analyzed alignments simulated using known parameter values. To ensure the realism of the simulated alignments, we drew the parameter values from the posteriors obtained under the reference model (10 genes) along with five values of λ_{CpG} (0.5, 1, 2, 4, and 8). The same mammalian tree topology was used for the validation and the analyses, taken from Laurin-Lemay et al. 2018a. The 10 genes (see table S3 for details) were selected among the 137 used in Laurin-Lemay et al. 2018a to represent the variation of the GC content found within mammalian genomes (table S3) and to have a sequence length of $\sim$ 1000 codons (a compromise between the amount of evolutionary signal and the computational burden). More specifically, for each gene, 100 sets of parameter values were drawn from the posterior distribution and used for 5 simulation sets (i.e., the

five values of λ_{CpG}). This leads to a total of 5,000 ($5 \times 10 \times 100$) DNA sequence alignments to benchmark the CABC.

This validation framework enables us to investigate the reliability of inferences conducted in this study, as a function of the various settings of our approximation methods. Specifically, we explored the number of simulations (i.e., 10^5 or 10^6), the tolerance level to be applied (10%, 1% or 0.1%), as well as the use of regression models. The tolerance levels were chosen to have RS table of at least 1000 points.

Two standard methods were used to evaluate the accuracy of CABC. First, we quantified estimation error for each parameter fitted under the CABC procedure by using the relative mean square error as used by Beaumont et al. [2002] :

$$RMSE_i = \frac{1}{N}\sum_{j=1}^{N}(\frac{\hat{\theta}_i - \theta_{ij}}{\hat{\theta}_i})^2, \tag{3.8.6}$$

where $RMSE_i$ corresponds to the average error computed for the parameter i (e.g., λ_{CpG}). The RMSE is obtained by averaging the relative squared discrepancy between the true parameter value ($\hat{\theta}_i$) used for generating the simulated alignment and the N parameter values (θ_{ij}) from the approximated posterior recovered under the CABC procedure when analyzing that very same simulated alignment. Note that the error is calculated relative to the scale of the true parameter value. A global RMSE can be obtained by averaging the total error computed independently for each parameter ($RMSE_i$) over all the analyses of a validation set (i.e., defined upon the five λ_{CpG} values).

We also investigated the coverage property of each parameter [Cook et al., 2006, Prangle et al., 2014a, Fearnhead and Prangle, 2012] fitted under the CABC using the Probability-Probability (P-P) plots. The coverage was investigated with a set of 99 credibility intervals (1-α), where α is ranging from 1 to 99%, and increased by steps of 1%. More precisely we computed the frequency for which the true parameter value was found within each credibility interval (1000 replicates per λ_{CpG} values) and compared those frequencies to the expected ones (1-α). In other words, when a 95% credibility interval ($1 - 0.05$) is used, we should recover the true value within this credibility interval 95% of the times. Conformity between the coverage recovered was assessed using a two sided Kolmogorov-Smirnov test available from SciPy [Eric et al., 2001–]. The rejection of the null hypothesis (i.e., the coverage is

expected to be uniform along all credible intervals tested) would demonstrate that there is a bias in the CABC analyses.

We also evaluated the impact of the two approximations of CABC on its accuracy. To study the choice of the parameters to be included in θ_{sc}, we transferred the GTR parameters into θ_{wc}. The strongly correlated (to λ_{CpG}) parameter vector, θ_{sc}, is now only composed of λ_{TBL}, λ_{omega_*} and λ_{ROOT}. The second approximation (that the parameters contained in θ_{wc} might not be uncorrelated to λ_{CpG}) was investigated by fixing the values of the θ_{wc} parameters to the true values. In other words, instead of drawing θ_{wc} from the posterior distribution of the simulated alignments under the reference model, we took the θ_{wc} values that have been used to generate the simulated alignments.

3.8.7. Application to mammalian protein coding genes

We would like to evaluate the ability of CABC to estimate hypermutability in the CpG context in the cases of mammalian protein coding genes. All the 137 genes of Laurin-Lemay et al. [2018a] were analyzed with CABC using a sampling size of 10^6 and a tolerance level of 0.1%. The topology of Laurin-Lemay et al. [2018a] was used for all the genes. We then carried out the hypothesis testing related to CpG hypermutability (i.e., $\lambda_{CpG} > 1$), for credibility intervals of 95% and 99%. We further investigated the impact of the prior on λ_{CpG} parameter by comparing our results to the ones obtained by using a broader prior on λ_{CpG} (i.e., [1/50,50]). We also explored the heterogeneity of CpG hypermutability over the placental tree by analyzing three clades independently. For each analysis (i.e., Glires, Laurasiatheria and Primates) we sampled the root position using λ_{ROOT} parameter on the branch connecting *Dipodomys* and the rest of the Glires (7 species), on the branch connecting *Mustela* and the rest of Laurasiatheria (14 species) and on the branch connecting *Callithrix* and the rest of the Primates (12 species).

3.8.8. Posterior predictive checks

Posterior predictive analysis is a powerful framework to evaluate model properties [Gelman et al., 2013]. Ten replicates were generated per posterior sample under the model without and with CpG hypermutability. First we compared the model predictions on the basis of substitution histories generated over simulations. Specifically, we quantified the total number of substitutions and the proportion of each substitution types as defined by each unique pair of

nucleotide substitution (e.g., A to C) or by their effect at the amino acid level (synonymous versus non-synonymous). We also tracked substitutions related to the CpG context (i.e., CpG to TpG and CpG to CpA) from all codon position (1-2, 2-3 and 3-1). We computed various SS from simulated alignments to compare model fit using Z-scores. Among the key features investigated, we looked at the GC3 content, the entropy of the RSCU (relative synonymous codon usage), the entropy of the RCF (relative codon frequencies), the relative dinucleotide frequencies for codon positions 1-2, 2-3 and 3-1, as well as the amino acid frequencies. We also performed a principal component analysis using the VEGAN package [Oksanen et al., 2017] from R CRAN [R Core Team, 2017] on the matrix of the RSCU recovered from the true alignments and from alignments generated by both models.

3.9. Acknowledgments

This work was supported by the French Laboratory of Excellence project entitled TULIP (ANR-10-LABX- 41 ;ANR-11-IDEX-0002-02), and by the Natural Sciences and Engineering Research Council of Canada. Computations were made on the supercomputer Mammouth-parallel from Université de Sherbrooke, managed by Calcul Québec and Compute Canada. The operation of this supercomputer is funded by the Canada Foundation for Innovation (CFI), the Ministère de l'Économie, de la Science et de l'Innovation du Québec - Nature et technologies (FRQ-NT). S.L.L is the recipient of a Fonds de la Recherche en Santé du Québec (FRSQ) Graduate Scholarship.

3.10. Supplementary materials

Supplementary tables (tables S1-S3) are available at Molecular Biology and Evolution.

— Table S1 : Comparison of relative mean square error computed over the different validation conditions.

— Table S2 : λ_{CpG} posterior mean and credible intervals (95% and 99%) estimates computed over the 137 genes analysis.

— Table S3 : GC content and sequence length of the 137 mammalian genes.

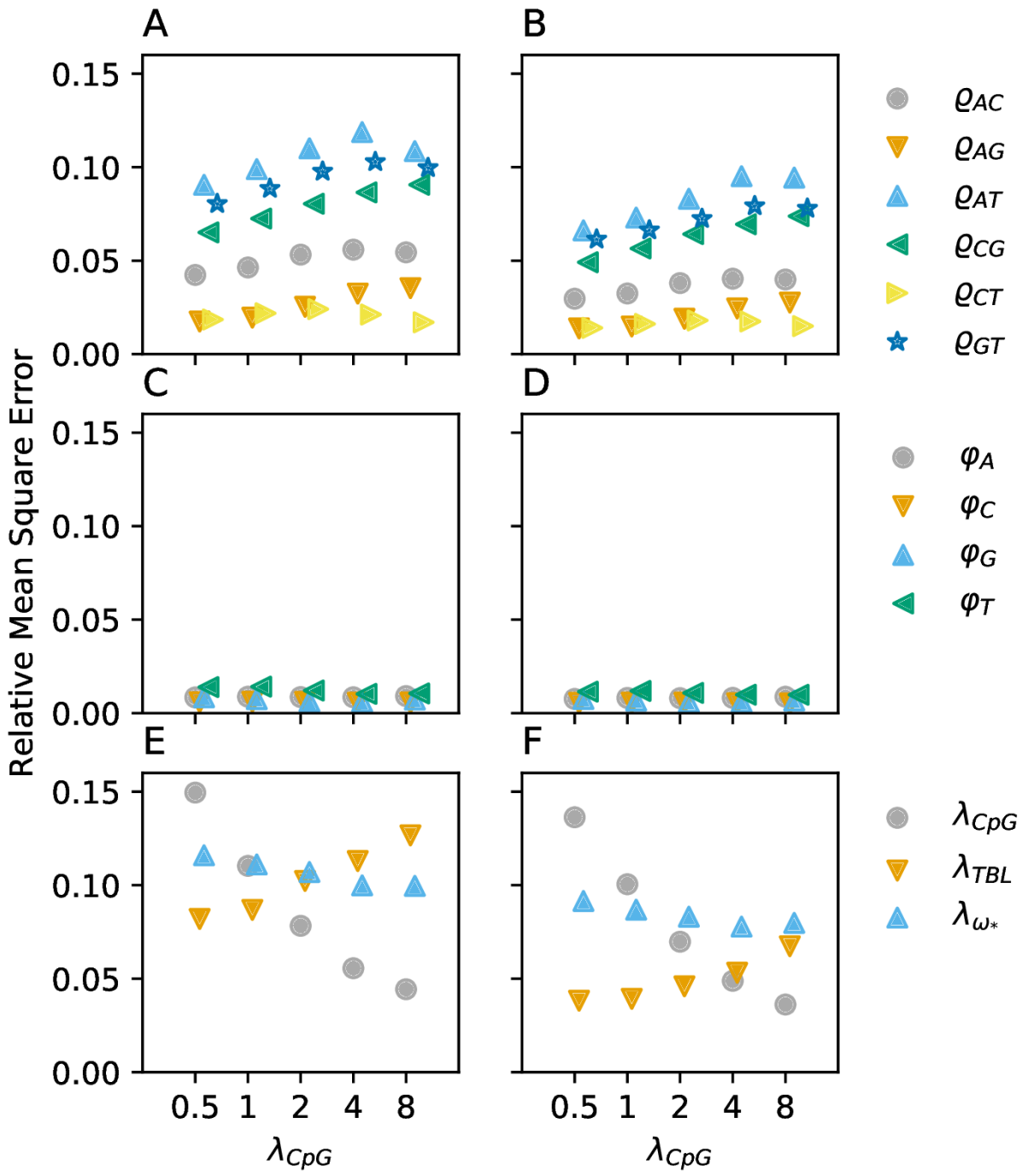

FIGURE 3.1. Relative mean square error (mean over 1000 replicates) under different λ_{CpG} values (x axes). Two tolerance levels, 1% (left panels) and 0.1% (right panels) over 10^6 simulations were used. Parameter values were corrected using linear regression model. (A-B) Mean RMSE of the six nucleotide exchangeabilities (ϱ_{AC}, ϱ_{AG}, ϱ_{AT}, ϱ_{CG}, ϱ_{CT} and ϱ_{GT}). (C-D) Mean RMSE of the four nucleotide propensities (φ_A, φ_C, φ_G and φ_T). (E-F) Mean RMSE of λ_{CpG}, λ_{TBL} and λ_{ω_*}.

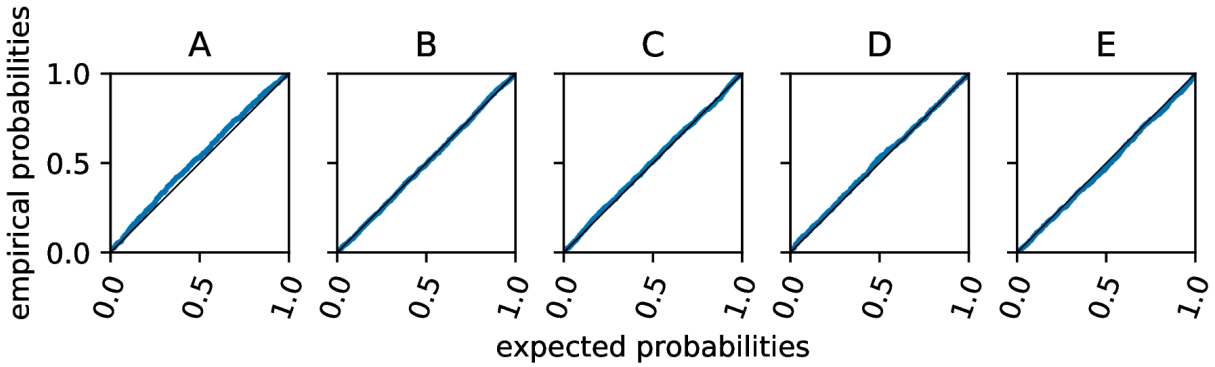

FIGURE 3.2. P-P plots of the λ_{CpG} recovered from the analyses of simulated alignments generated under λ_{CpG} values (0.5, 1.0, 2.0, 4.0, 8.0), corresponding respectively to (A-E). Empirical probabilities were obtained using rejection sampling (the best 0.1% of 10^6 simulations) corrected with a linear regression model. The frequency at which the true values of λ_{CpG} within each credibility intervals is uniformly distributed (two sided Kolmogorov-Smirnov test : p= 0.848, p= 1, p= 0.999, p= 0.996 and p= 1 respectively). A diagonal line is added (black) to appreciate any deviation between the expectations and the results.

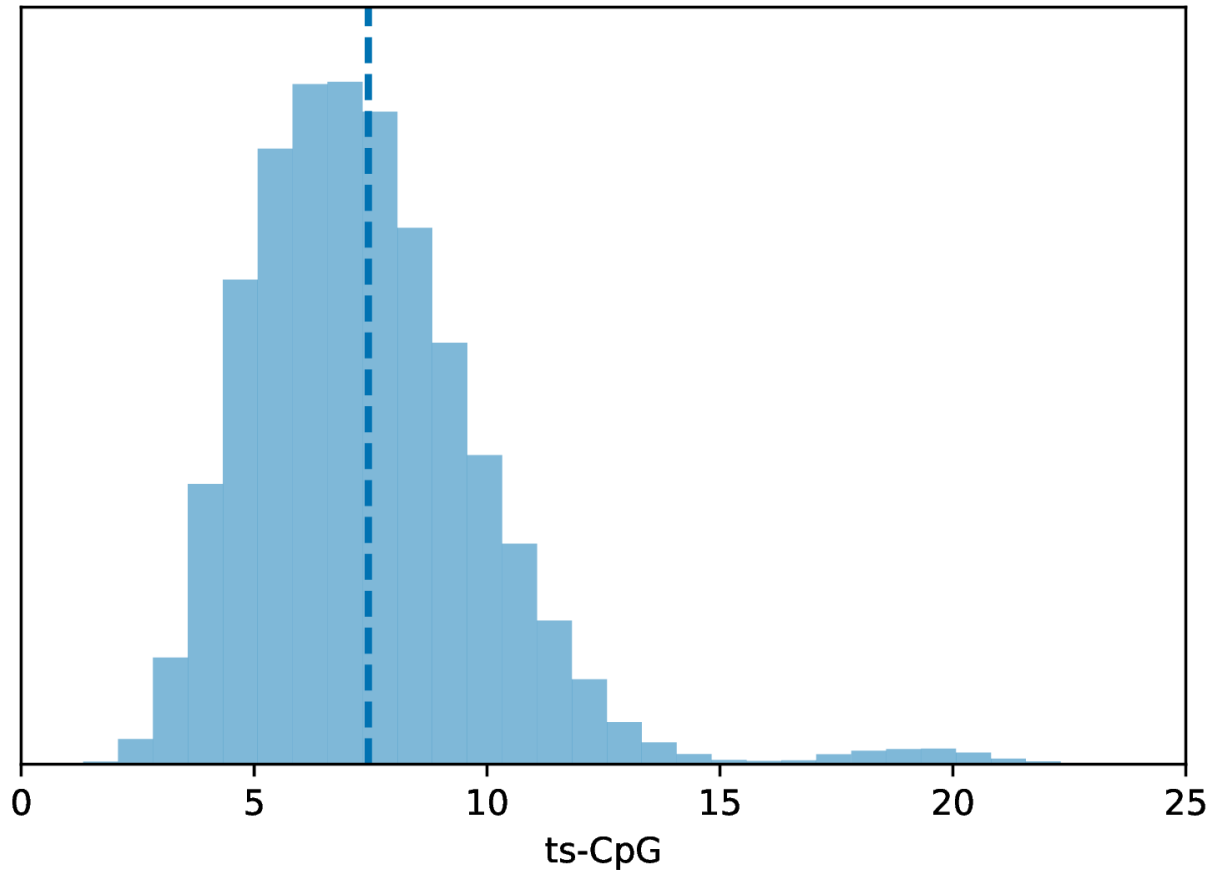

FIGURE 3.3. Aggregation of posterior distributions of λ_{CpG} recovered from 137 mammalian genes using the CABC methodology. Rejection sampling (the best 0.1% of 10^6 simulations) with linear regression model were used to approximate posteriors. The vertical blue dash line represents the mean λ_{CpG} value (7.45) over all posterior values pooled.

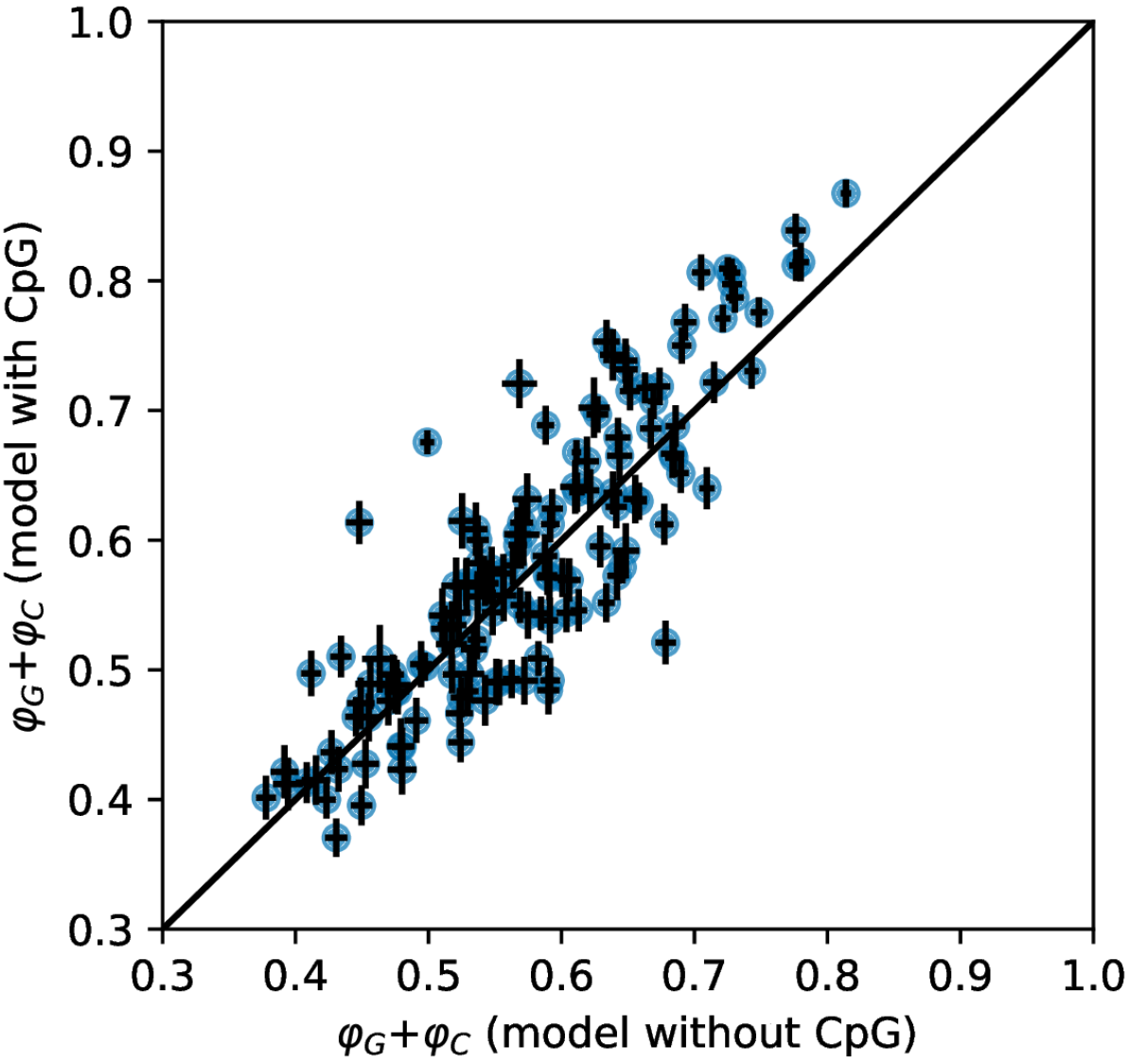

Figure 3.4. Comparison of the $\varphi_G + \varphi_C$ posterior mean estimates under the models without (x axis) and with CpG hypermutation (y axis) recovered from the analysis of the 137 mammalian gene alignments. A diagonal line is added (black) to appreciate any deviation between both models estimate. The error bars correspond to the standard deviations computed from each posterior.

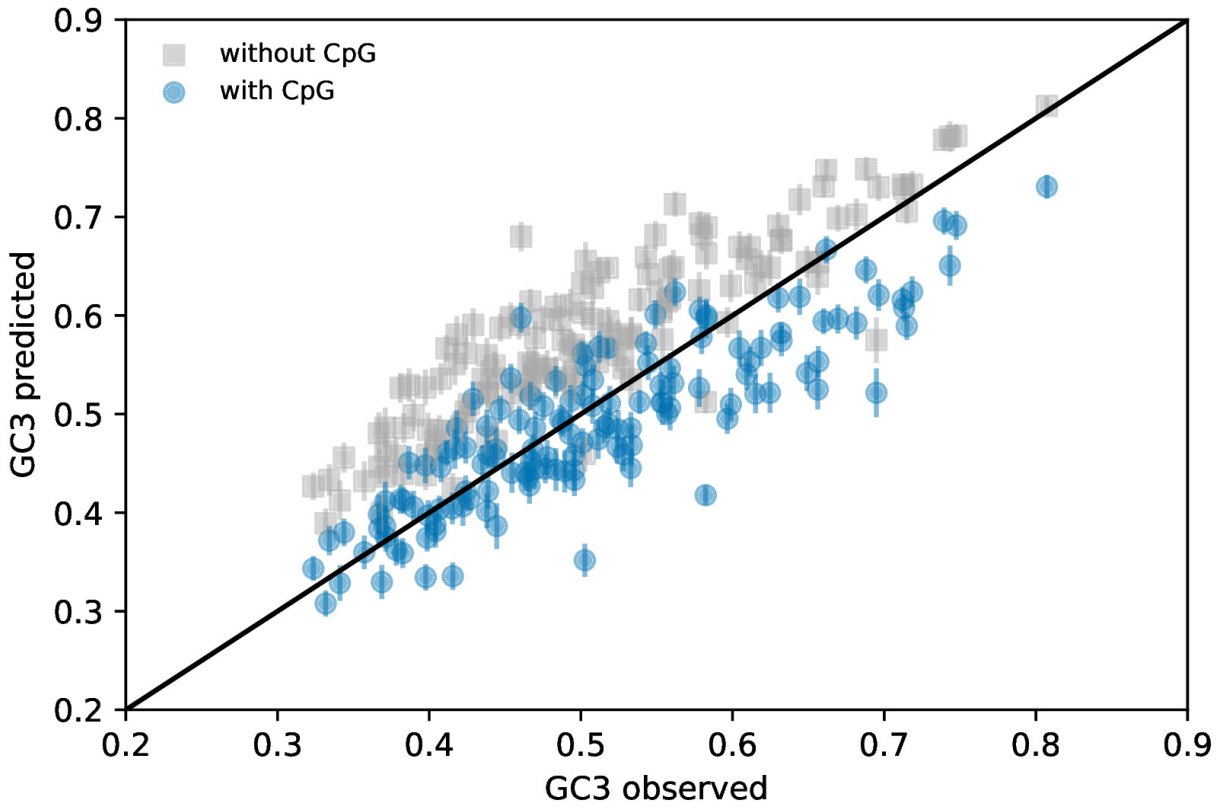

Figure 3.5. Comparison of the ability of the models without (gray squares) and with CpG hypermutation (blue circles) to predict the GC3 content of the 137 mammalian gene alignments using posterior predictive simulations. The observed GC3 is plot against the mean predictions (y axis) from both models. A diagonal line is added (black) to appreciate any deviation between observations and the predictions. The error bars correspond to the standard deviations computed from models predictions.

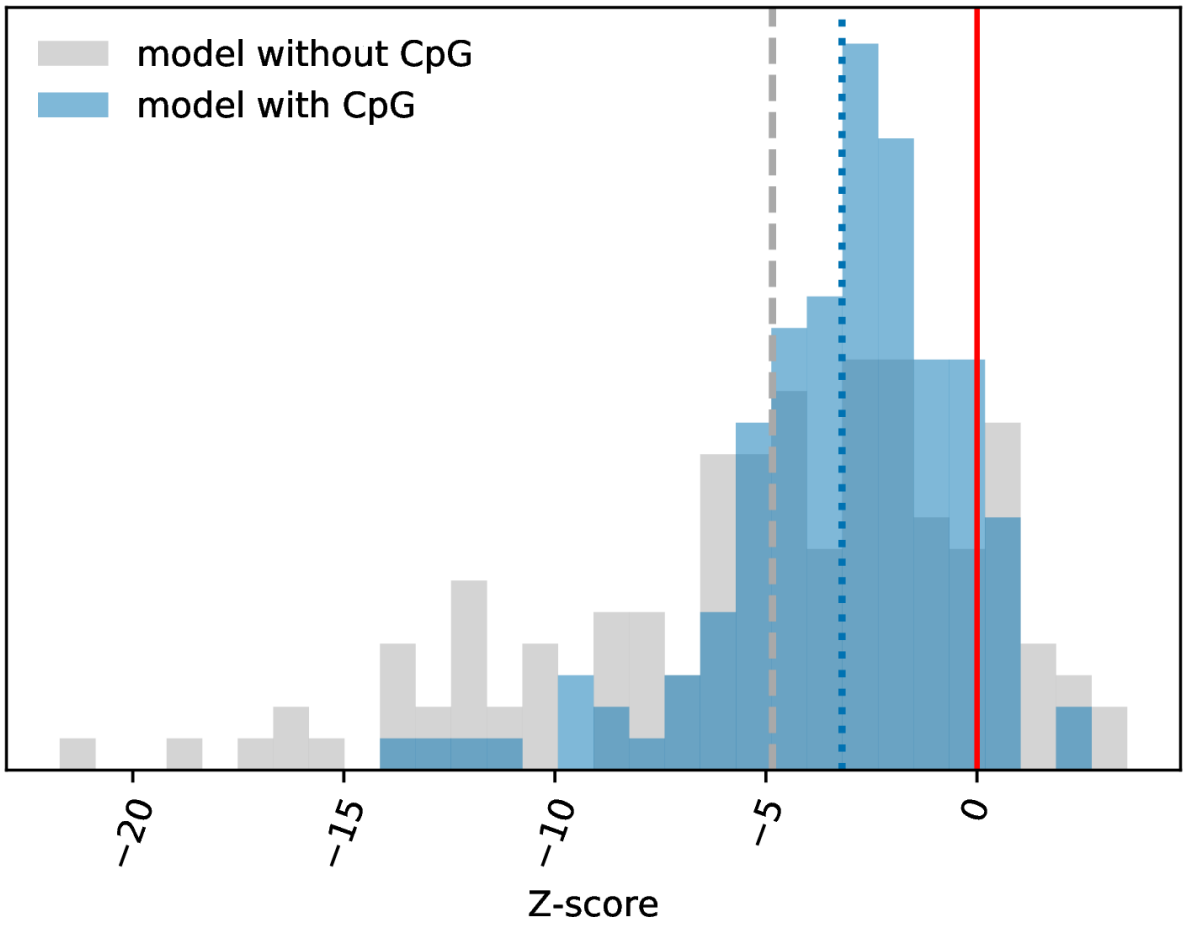

Figure 3.6. Distribution of Z-scores computed from RSCU (without stop, methionine and tryptophan codons) entropy predicted under the models without (gray) and with (blue) CpG hypermutation. The vertical dashed (gray) and dotted (blue) lines represent the mean Z-scores obtained under each model respectively (i.e., without and with CpG hypermutation). The vertical solid line (red) represents the zero value.

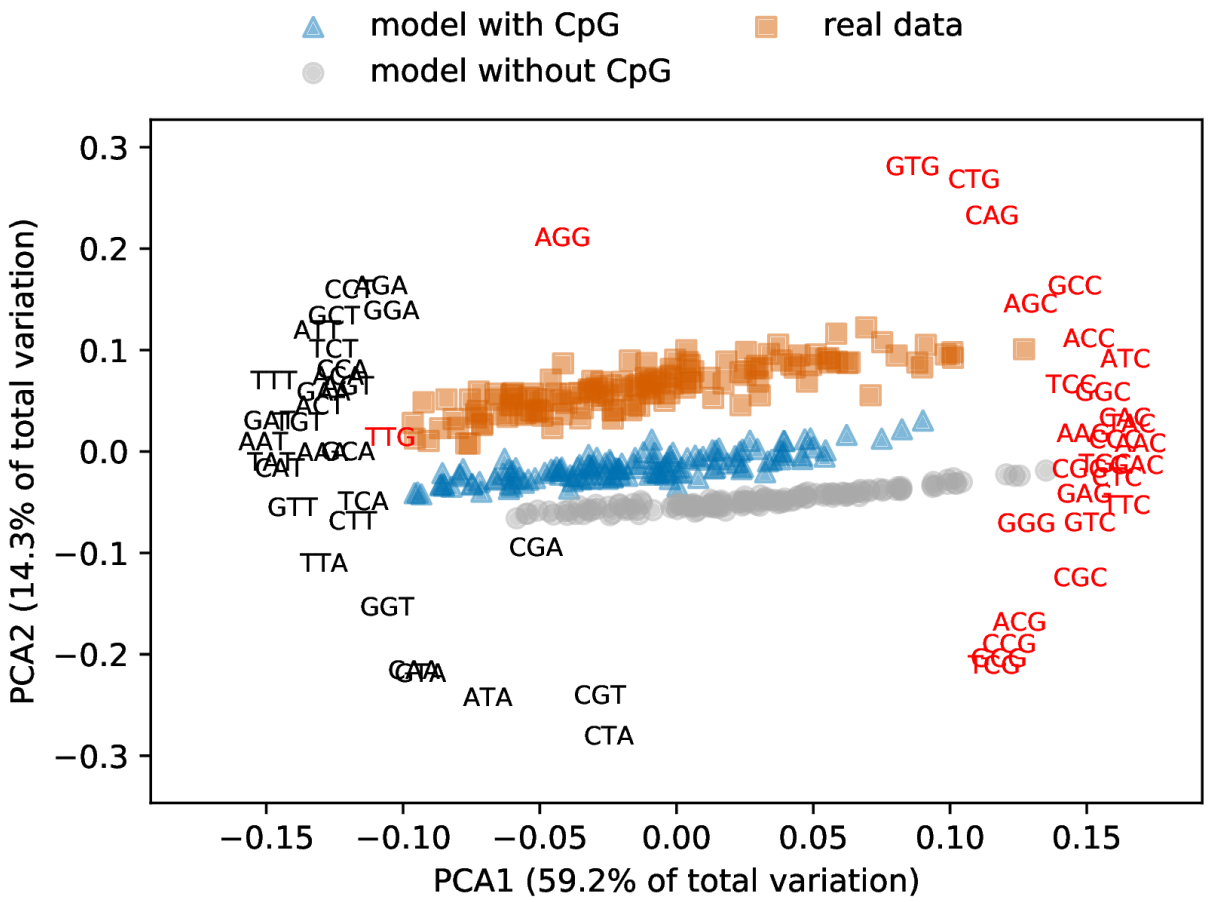

Figure 3.7. Principal component analysis of the RSCU (without stop, methionine and tryptophan codons) recovered from the 137 mammalian gene alignments and from the mean RSCU predicted under models without and with CpG hypermutation. G/C-ending codons are annotated in red while A/T-ending codons are annotated in black.

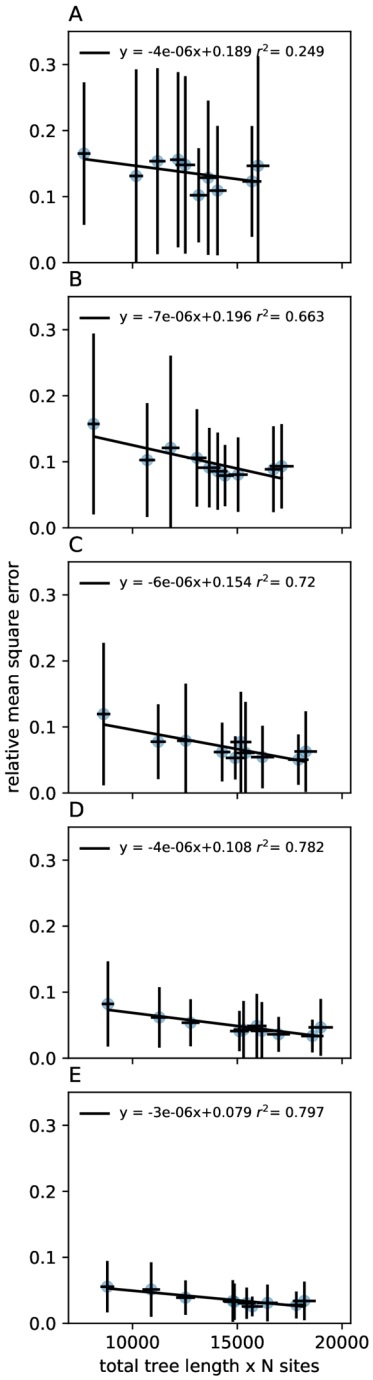

Figure S3.1. Ability to predict the relative mean square error of λ_{CpG} parameter (y axes) from the amount of evolutionary signal (x axes ; expected number of substitutions) present in the simulated alignment used for validation purpose. Dots and error bars correspond to the means and the standard deviations respectively. Means and standard deviations are computed by pooling validation replicates over the different mammalian genes (10) used to generated the simulated alignments. (A-E) correspond to the λ_{CpG} values (0.5, 1, 2, 4 and 8) used respectively to generate the validation data sets (see Materials and Methods for details). The regression equations are added as well as theirs r-squared.

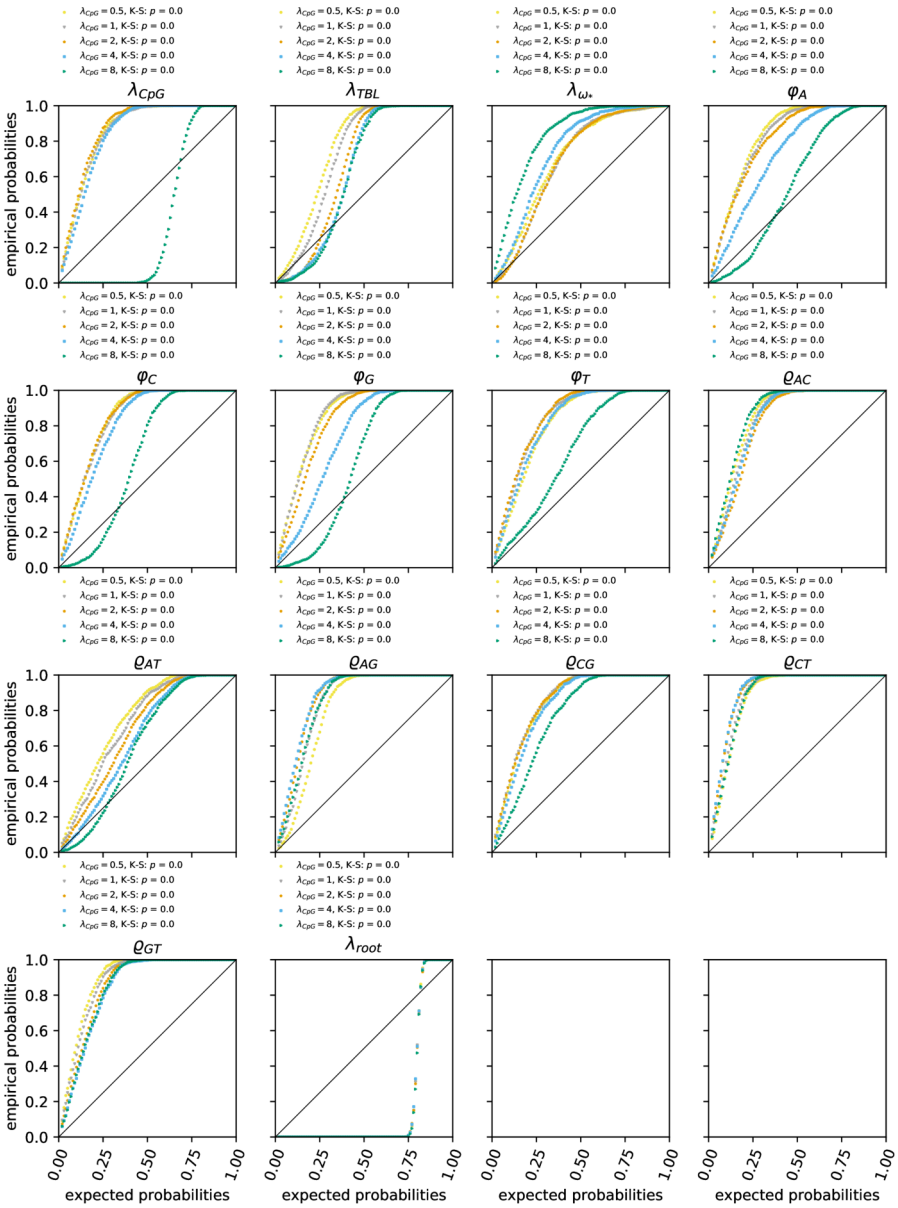

FIGURE S3.2. P-P plots of all parameters (λ_{CpG}, λ_{TBL}, $\lambda\omega_*$, φ_A, φ_C, φ_G, φ_T, ϱ_{AC}, ϱ_{AG}, ϱ_{AT}, ϱ_{CG}, ϱ_{CT}, ϱ_{GT}, λ_{ROOT}) recovered from the analysis of simulated alignments generated under different λ_{CpG} values (0.5, 1.0, 2.0, 4.0, 8.0). Rejection sampling (the best 0.1% of 10^5 simulations) alone was used to approximate posteriors. Coverage results are shown for each λ_{CpG} value (yellow, gray, orange, blue and green respectively). A two sided Kolmogorov-Smirnov test is applied, p=values are shown for each set of simulated alignments. A diagonal line is added (black) to appreciate any deviation between the expectations and the results.

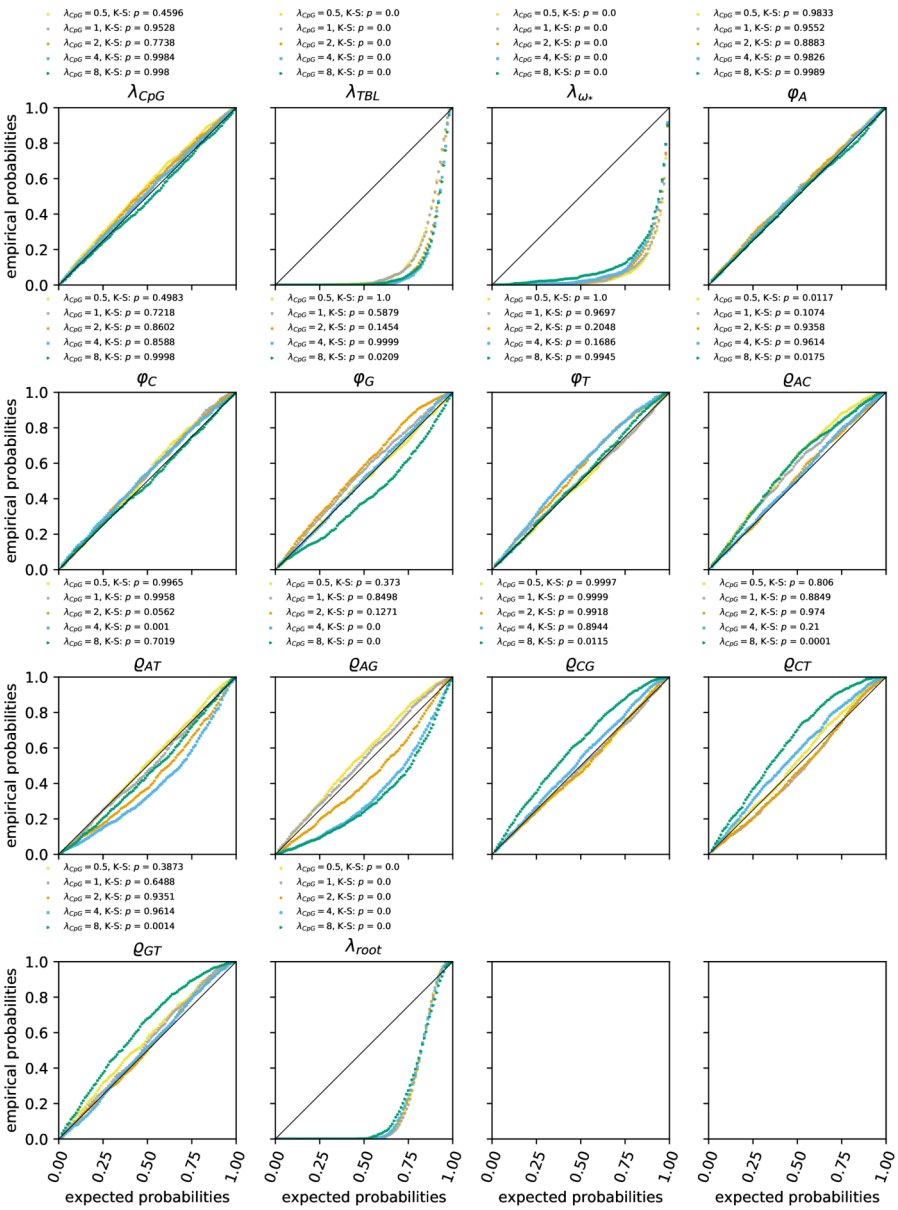

FIGURE S3.3. P-P plots of all parameters (λ_{CpG}, λ_{TBL}, λ_{ω_*}, φ_A, φ_C, φ_G, φ_T, ϱ_{AC}, ϱ_{AG}, ϱ_{AT}, ϱ_{CG}, ϱ_{CT}, ϱ_{GT}, λ_{ROOT}) recovered from the analysis of simulated alignments generated under different λ_{CpG} values (0.5, 1.0, 2.0, 4.0, 8.0). Rejection sampling (the best 1% of 10^5 simulations) with linear regression model were used to approximate posteriors. Coverage results are shown for each λ_{CpG} value (yellow, gray, orange, blue and green respectively). A two sided Kolmogorov-Smirnov test is applied, p=values are shown for each set of simulated alignments. A diagonal line is added (black) to appreciate any deviation between the expectations and the results.

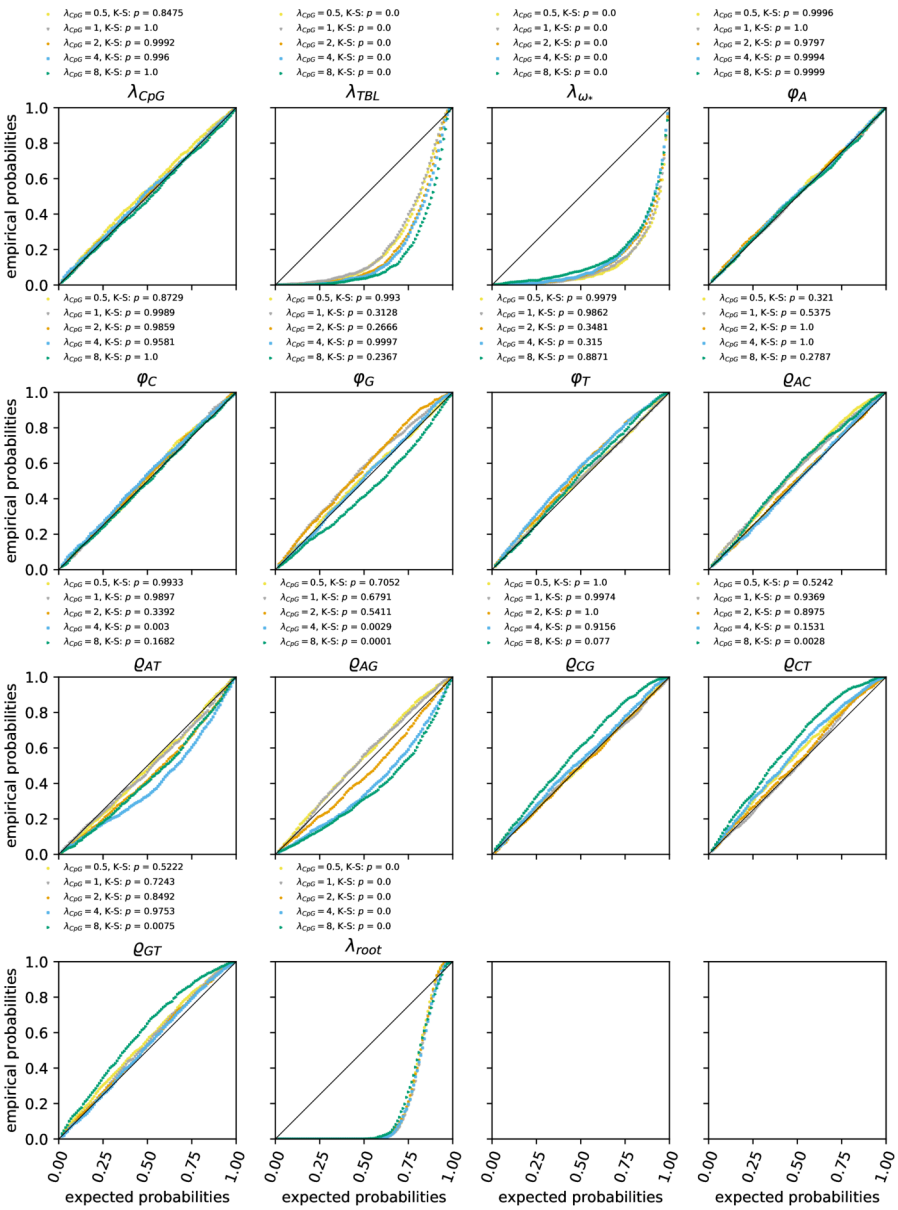

FIGURE S3.4. P-P plots of all parameters (λ_{CpG}, λ_{TBL}, λ_{ω_*}, φ_A, φ_C, φ_G, φ_T, ϱ_{AC}, ϱ_{AG}, ϱ_{AT}, ϱ_{CG}, ϱ_{CT}, ϱ_{GT}, λ_{ROOT}) recovered from the analysis of simulated alignments generated under different λ_{CpG} values (0.5, 1.0, 2.0, 4.0, 8.0). Rejection sampling (the best 0.1% of 10^6 simulations) with linear regression model were used to approximate posteriors. Coverage results are shown for each λ_{CpG} value (yellow, gray, orange, blue and green respectively). A two sided Kolmogorov-Smirnov test is applied, p=values are shown for each set of simulated alignments. A diagonal line is added (black) to appreciate any deviation between the expectations and the results.

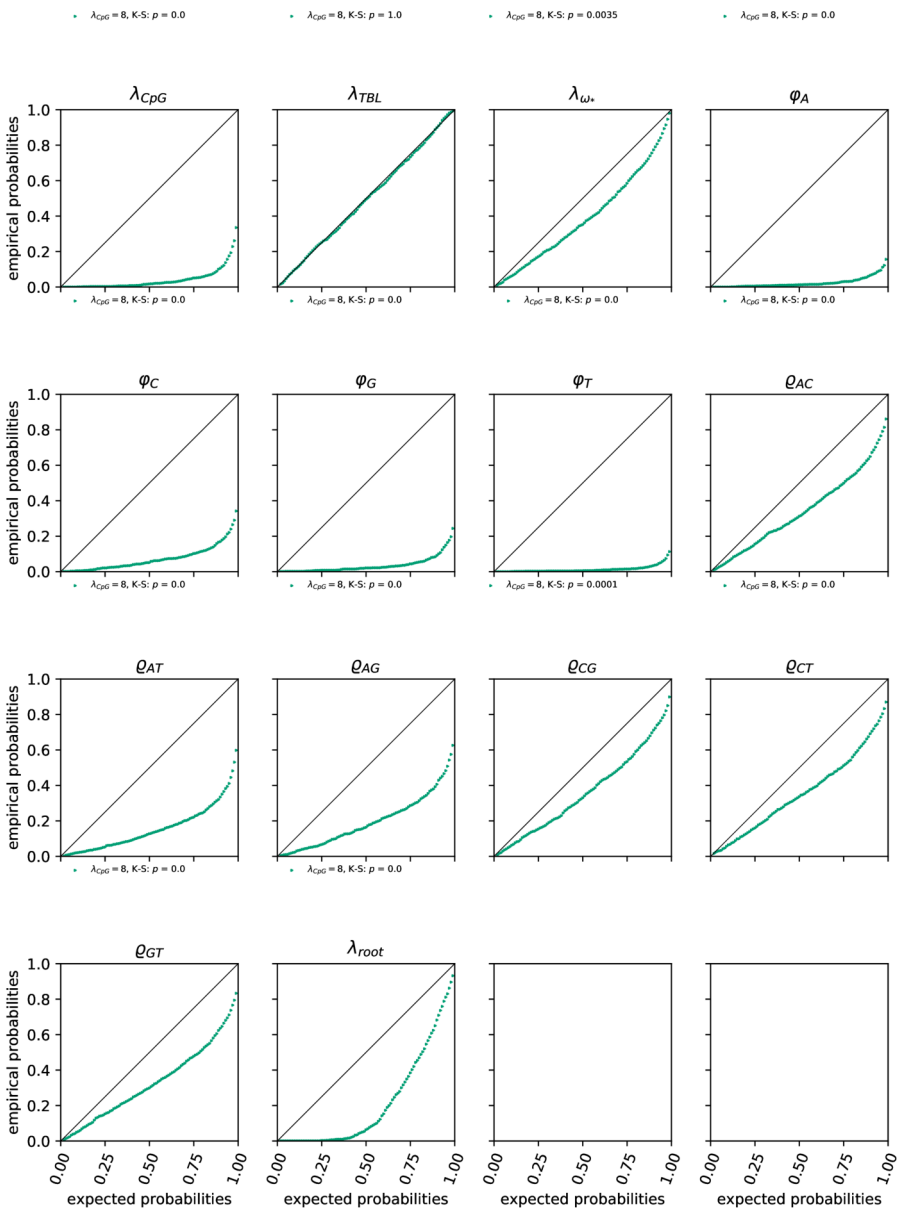

FIGURE S3.5. P-P plots of all parameters (λ_{CpG}, λ_{TBL}, λ_{ω_*}, φ_A, φ_C, φ_G, φ_T, ϱ_{AC}, ϱ_{AG}, ϱ_{AT}, ϱ_{CG}, ϱ_{CT}, ϱ_{GT}, λ_{ROOT}) recovered from the analysis of simulated alignments generated under λ_{CpG} set to 8 when the GTR parameters are considered to belong to θ_{wc}. Rejection sampling (the best 1% of 10^5 simulations) with linear regression model were used to approximate posteriors. Coverage results are shown for each λ_{CpG} value (yellow, gray, orange, blue and green respectively). A two sided Kolmogorov-Smirnov test is applied, p=values are shown for each set of simulated alignments. A diagonal line is added (black) to appreciate any deviation between the expectations and the results.

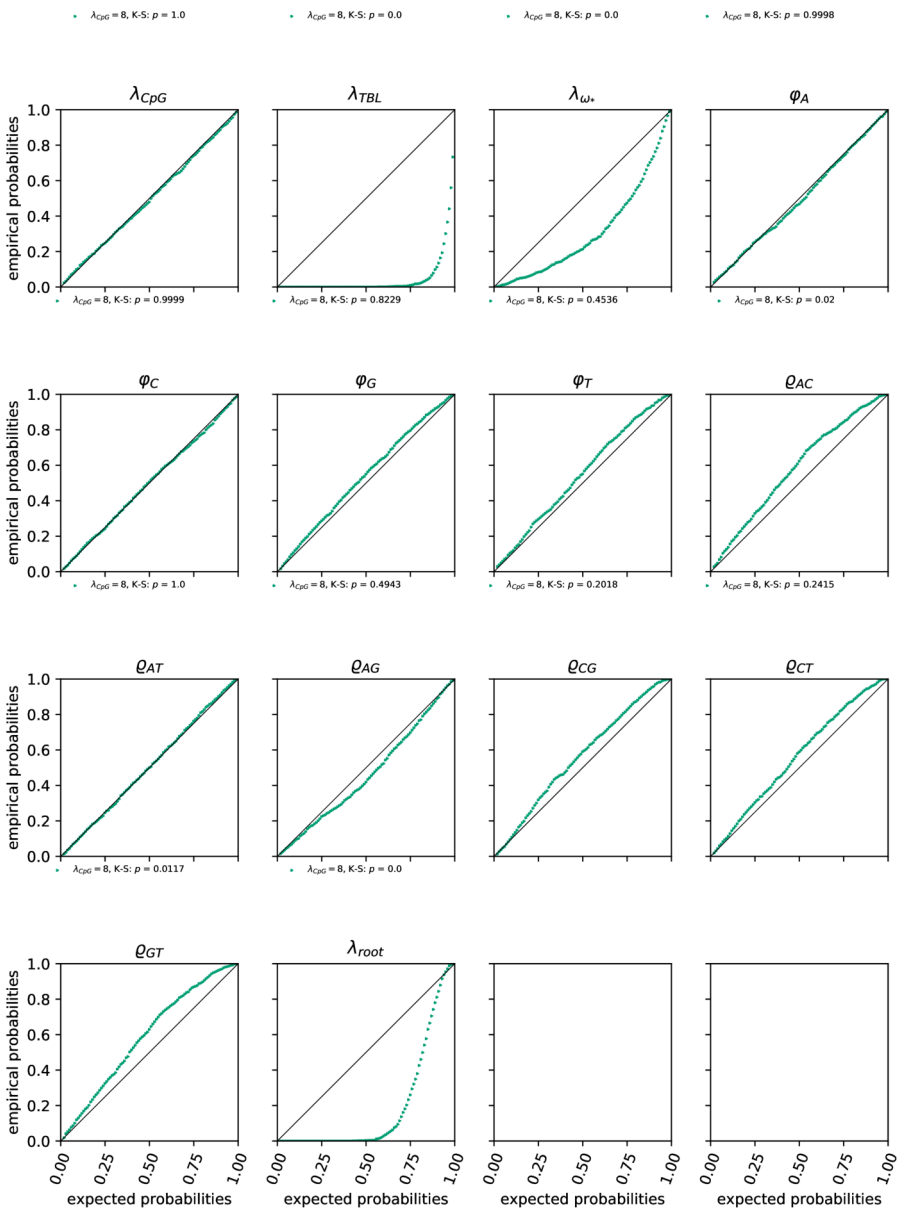

FIGURE S3.6. P-P plots of all parameters (λ_{CpG}, λ_{TBL}, λ_{ω_*}, φ_A, φ_C, φ_G, φ_T, ϱ_{AC}, ϱ_{AG}, ϱ_{AT}, ϱ_{CG}, ϱ_{CT}, ϱ_{GT}, λ_{ROOT}) recovered from the analysis of simulated alignments generated under λ_{CpG} set to 8.0 when θ_{wc} are set to the true values (as used for generating the simulated alignments). Rejection sampling (the best 1% of 10^5 simulations) with linear regression model were used to approximate posteriors. Coverage results are shown for each λ_{CpG} value (yellow, gray, orange, blue and green respectively). A two sided Kolmogorov-Smirnov test is applied, p=values are shown for each set of simulated alignments. A diagonal line is added (black) to appreciate any deviation between the expectations and the results.

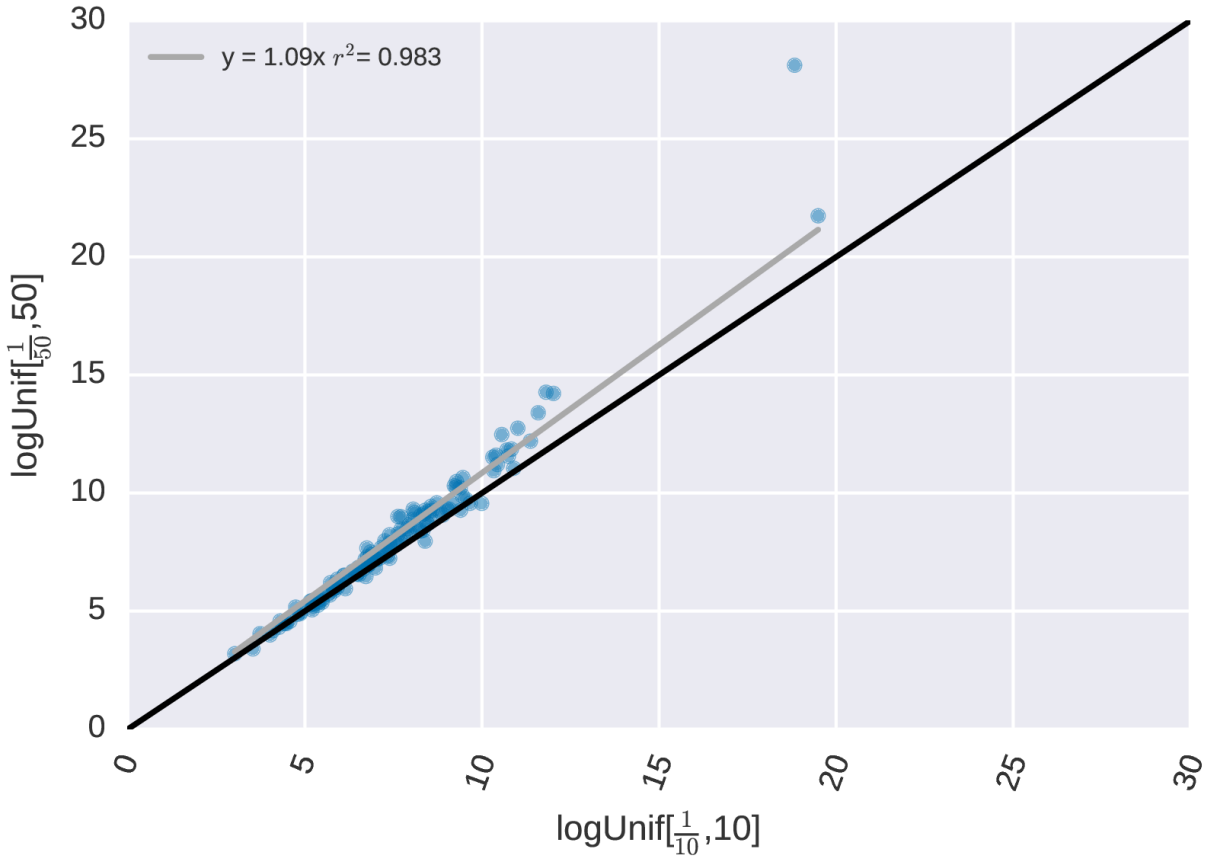

FIGURE S3.7. Comparison of λ_{CpG} parameter estimation (posterior mean) obtained under the M[GTR+ts-CpG]-S[NCatAA*] model from 137 mammalian gene alignments analyzed when using two different prior beliefs : logUniform[1/10,10], x axis, and the logUniform[1/50,50], (y axis). Rejection sampling (the best 0.1% of 10^6 simulations) with linear regression model were used to approximate posteriors. The slope and the r-squared of the regression passing trough the origin is added to the plot. A diagonal line is added (black).

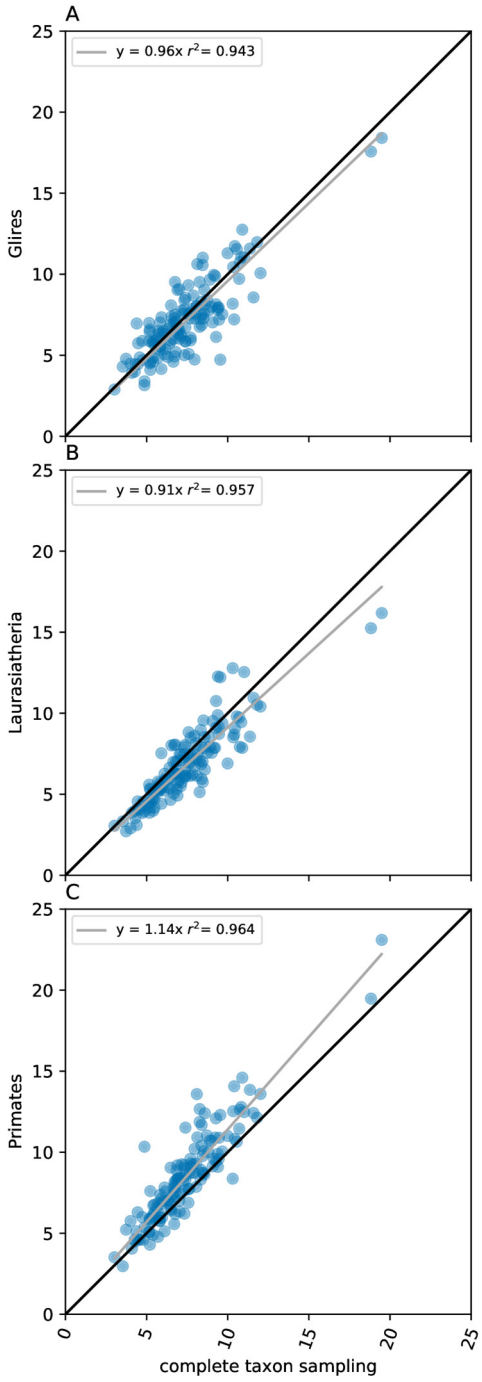

FIGURE S3.8. Comparison of λ_{CpG} parameter estimation (posterior mean) obtained under the M[GTR+ts-CpG]-S[NCatAA*] model from 137 mammalian gene alignments analyzed when using the complete taxon sampling (x axes) and composed of (A) Glires, (B) Laurasiatheria and (C) Primates only (y axes respectively). Rejection sampling (the best 0.1% of 10^6 simulations) with linear regression model were used to approximate posteriors. The slope and the r-squared of the regression passing trough the origin is added to the plot. A diagonal line is added (black).

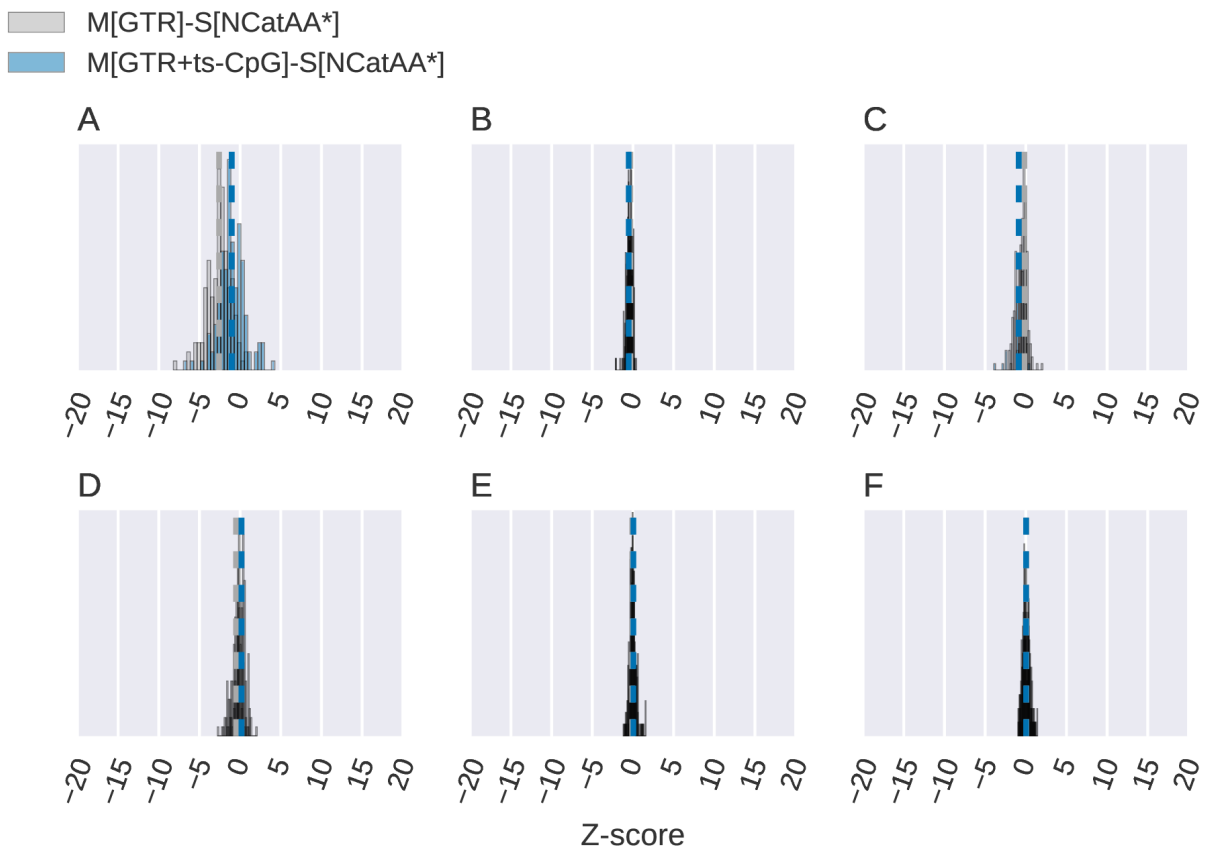

Figure S3.9. Comparison of the ability of the models without and with CpG hypermutation to predict the dinucleotide frequencies at codon position 1-2 when using posterior predictive simulations from the analysis of 137 mammalian gene alignments. Z-scores are computed from dinucleotides frequencies directly affected by CpG hypermutability (i.e., A : CpG ; B : TpG ; C : CpA) and not directly affected by CpG hypermutability (i.e., D : GpC ; E : GpT ; F : ApC). The distribution represent the mean Z-scores computed for each gene analysis. The vertical gray and blue dash lines represent the mean Z-scores retrieved under both models respectively (without and with CpG hypermutation)

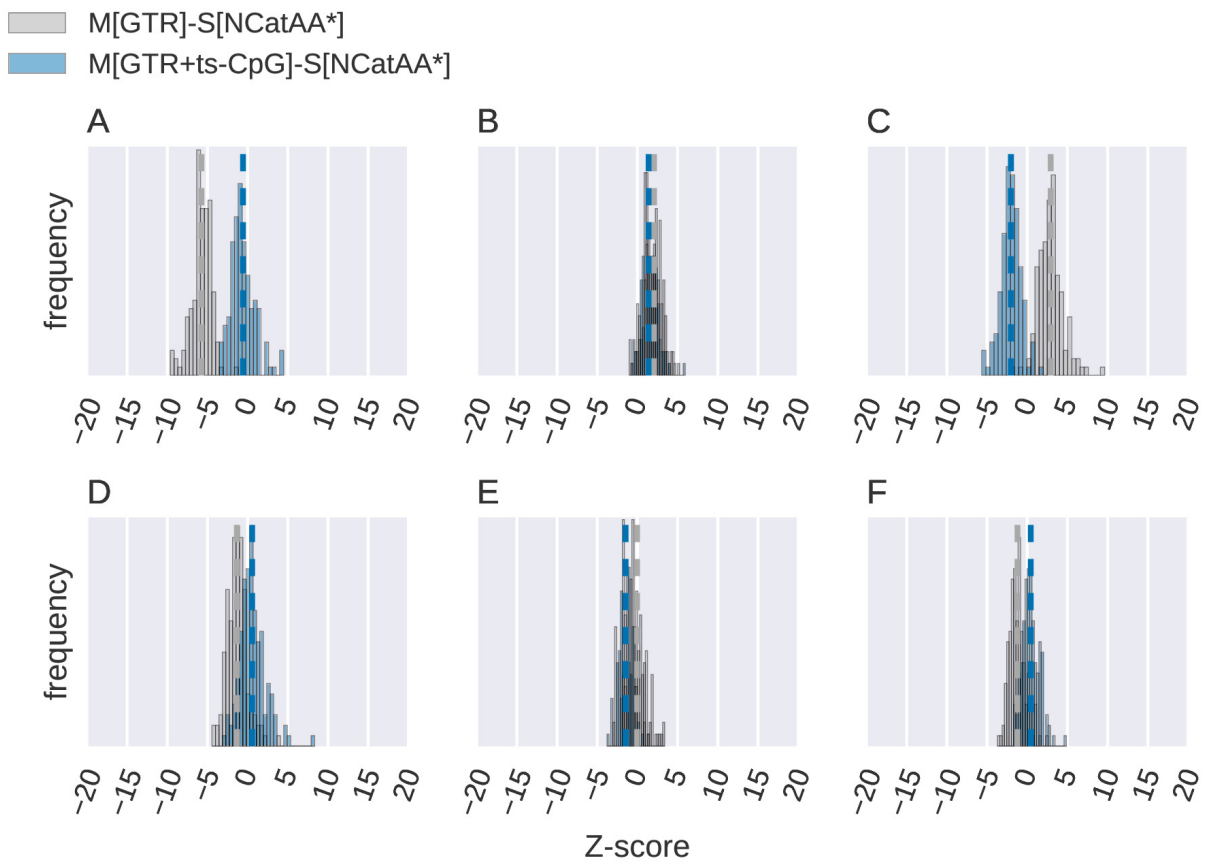

Figure S3.10. Comparison of the ability of the models without and with CpG hypermutation to predict the dinucleotide frequencies at codon position 2-3 when using posterior predictive simulations from the analysis of 137 mammalian gene alignments. Z-scores are computed from dinucleotides frequencies directly affected by CpG hypermutability (i.e., A : CpG ; B : TpG ; C : CpA) and not directly affected by CpG hypermutability (i.e., D : GpC ; E : GpT ; F : ApC). The distribution represent the mean Z-scores computed for each gene analysis. The vertical gray and blue dash lines represent the mean Z-scores retrieved under both models respectively (without and with CpG hypermutation)

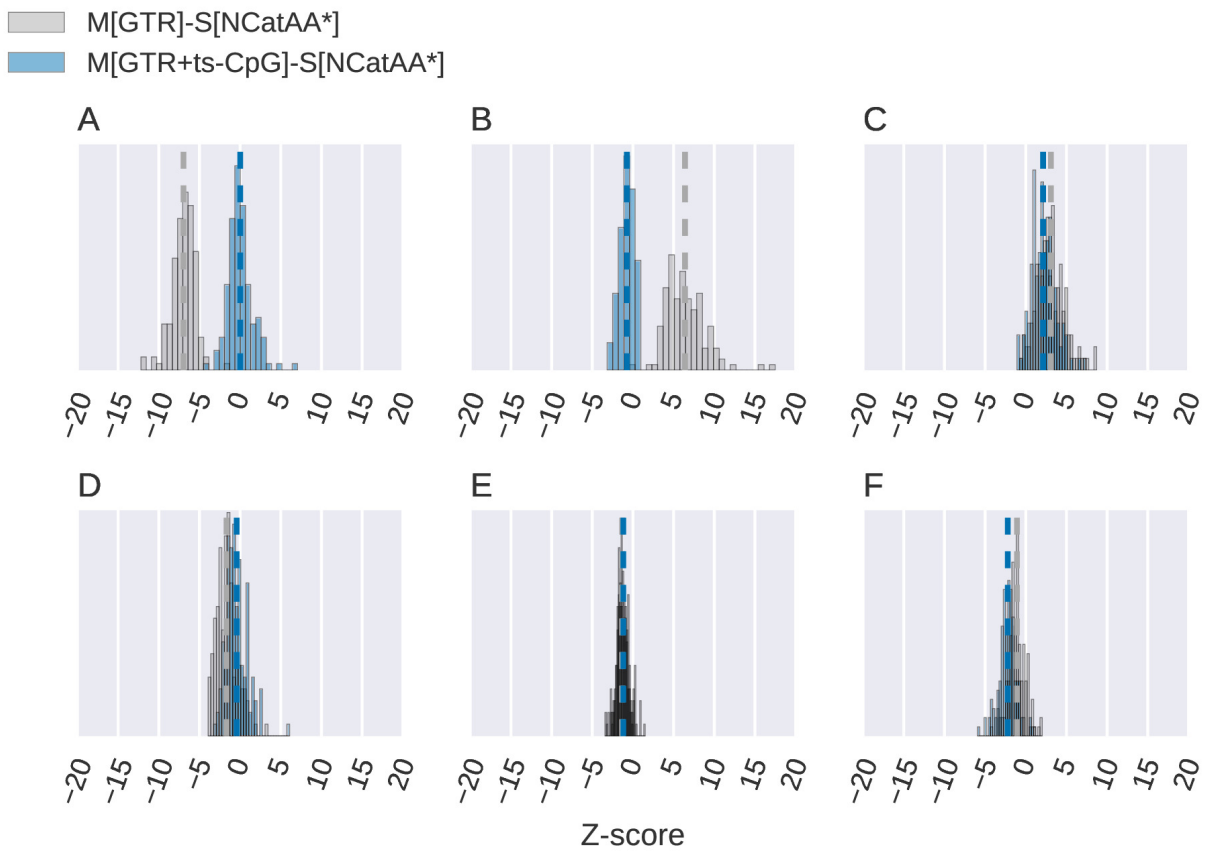

Figure S3.11. Comparison of the ability of the models without and with CpG hypermutation to predict the dinucleotide frequencies at codon position 3-1 when using posterior predictive simulations from the analysis of 137 mammalian gene alignments. Z-scores are computed from dinucleotides frequencies directly affected by CpG hypermutability (i.e., A : CpG ; B : TpG ; C : CpA) and not directly affected by CpG hypermutability (i.e., D : GpC ; E : GpT ; F : ApC). The distribution represent the mean Z-score computed for each gene analysis. The vertical gray and blue dash lines represent the mean Z-scores retrieved under both models respectively (without and with CpG hypermutation)

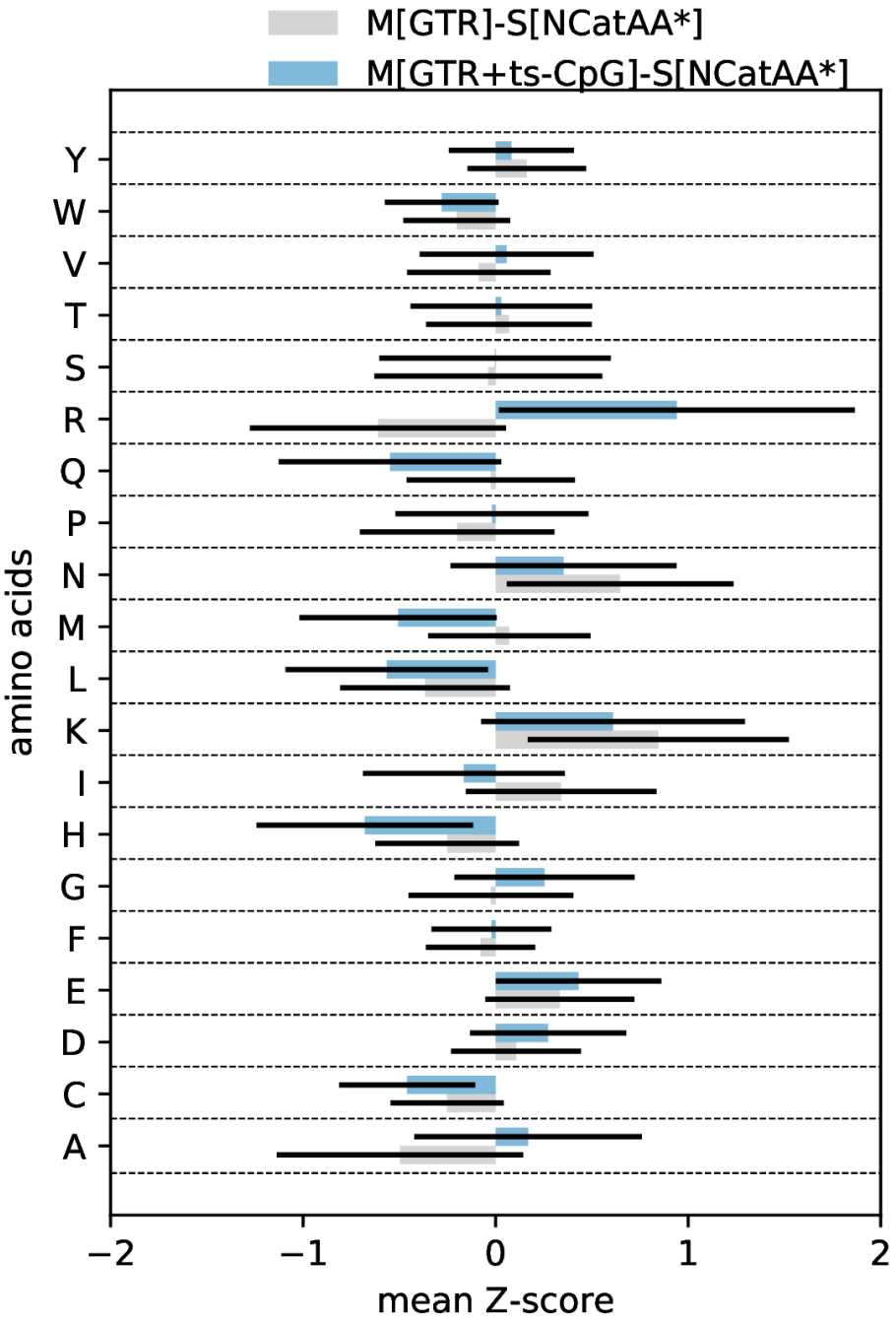

Figure S3.12. Comparison of the ability of the models without and with CpG hypermutation to predict the amino acid frequencies when using posterior predictive simulations from the analysis of 137 mammalian gene alignments. The width of the bars represents the mean Z-scores and the error bars correspond to the standard deviations both computed from all simulations.

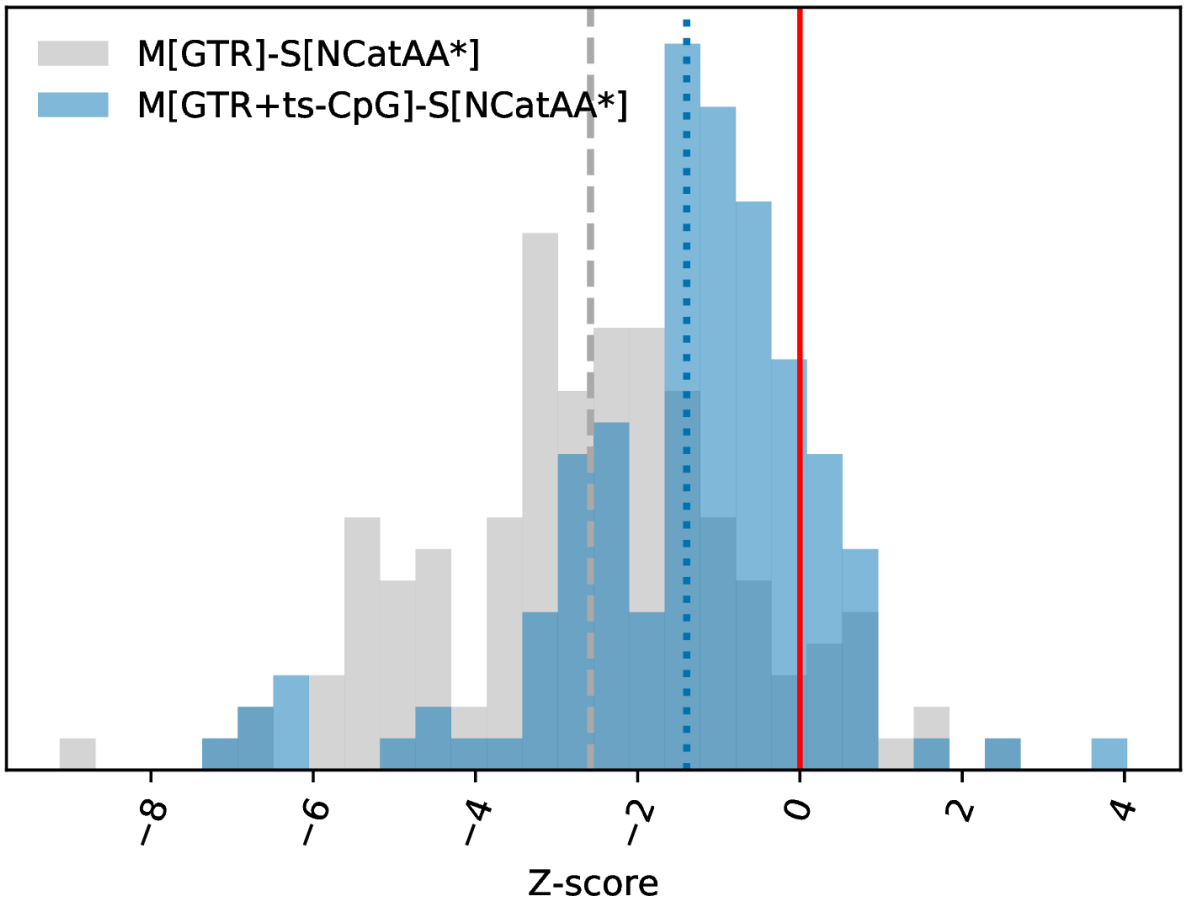

Figure S3.13. Distribution of Z-scores computed from the relative codon frequency (without stop codons) entropy predicted under the models without (gray) and with (blue) CpG hypermutation. The vertical dashed (gray) and dotted (blue) lines represents the mean Z-scores obtained under each models (without and with CpG hypermutation respectively). The vertical solid line (red) represents the zero value.

Chapitre 4

Conclusion

4.1. Retour sur le travail de thèse

4.1.1. La tâche difficile de détecter la sélection sur l'usage des codons

Bien que la sélection négative sur l'usage des codons soit connue pour affecter les bactéries à croissance rapide, chez les systèmes biologiques possédant de plus petites N_e, comme les mammifères, la sélection négative sur l'usage des codons est controversée. Effectivement, selon les principes de génétique des populations, les mutations faiblement délétères ou avantageuses, comme les mutations synonymes, ne peuvent être efficacement sélectionnées étant donné l'importance sélective des mutations non-synonymes. Par contre, la sélection sur l'usage des codons pourrait être très localisée le long des séquences codantes, à des sites précis, relevant de contraintes de sélection particulières dues à des motifs utilisés par la machinerie d'épissage ou par les facteurs de transcriptions, par exemple.

Détecter l'importance de la sélection, négative ou positive, est une tâche difficile qui nécessite d'identifier un écart statistique par rapport à ce qui est attendu sous le modèle mutationnel, par rapport à ce qui est obtenu en modélisant certaines contraintes de sélection. La tâche est d'autant plus difficile lorsque le signal est faible, car ce dernier peut être confondu par la présence de d'autres processus de mutations qui ne sont pas pris en compte par le modèle. De la même manière, lorsque les branches qui séparent les évènements de spéciations sont courtes, il est très difficile d'en inférer l'ordre, sans parler des biais systématiques qui peuvent confondre l'inférence phylogénétique.

Les modèles phylogénétiques de type mutation-sélection sont les outils de prédilection pour tester explicitement des hypothèses en lien avec les processus mutationnels ainsi que

les contraintes de sélection qui agissent sur les séquences codantes. Ces modèles sont complexes, ainsi que les questions qu'ils permettent d'aborder. Lorsque la controverse surgit, il est important de corroborer les résultats à l'aide d'un design expérimental bien conçu qui permettra d'évaluer la réponse du modèle face à différentes situations. Les études par simulations sont une stratégie efficace pour évaluer le comportement des modèles d'évolution face à de potentielles violations de modèle. Toutes les hétérogénéités testées avec les modèles mutation-sélection de Yang and Nielsen [2008] ont généré des taux de faux positifs allant de 20%, en présence de sélection site-spécifique, à 100%, lorsqu'en présence d'hypermutabilité des transitions en contexte CpG [Laurin-Lemay et al., 2018b].

L'hypermutabilité des transitions du contexte CpG peut donc à elle seule expliquer la sélection détectée par [Yang and Nielsen, 2008] sur l'usage des codons. Il faut retenir aussi que d'autres processus mutationnels sont probablement en cause, comme l'hypermutabilité du contexte TpA (communications personnelles Laurent Duret), ou encore la conversion génique biaisée [Duret, 2002b, Duret and Galtier, 2009a, Katzman et al., 2011, Glemin et al., 2015], un processus mutationnel qui change au cours du temps (e.g., [Lartillot, 2013]).

4.1.2. Nouvelle méthode d'inférence : calcul bayésien approché conditionnel

Modéliser des hétérogénéités qui résultent d'interdépendances dans les données (par exemple l'hypermutabilité du contexte CpG) augmente de beaucoup la complexité des fonctions de vraisemblance. À certains moments la fonction de vraisemblance peut devenir impossible à calculer, et des astuces doivent être utilisées. Par exemple, en phylogénie, prendre en compte les contraintes dues à la structure tertiaire des protéines augmente significativement la grandeur de la matrice de substitution, et faire l'exponentiation de la matrice de substitution n'est tout simplement plus abordable, alors, certains auteurs on eu l'idée d'utiliser une stratégie d'augmentation de données (e.g., [Rodrigue et al., 2009]). De plus, le niveau de sophistication des modèles fait en sorte que des vecteurs de paramètres de haute dimensionnalité sont nécessaires pour modéliser l'hétérogénéité des processus étudiés, dans notre cas de contraintes de sélection sur la protéine.

Le calcul bayésien approché (Approximate Bayesian Computation ou ABC, voir la récente revue sur cette méthode d'inférence [Beaumont, 2019]) est une astuce qui permet de contourner le calcul de la vraisemblance basée sur la capacité de générer des données à partir

de la fonction de vraisemblance. Mais cette approche n'est pas aussi efficace, à notre connaissance, que l'échantillonnage par MCMC communément utilisé pour faire l'approximation de la distribution *a posteriori* dans le contexte où la fonction de vraisemblance est accessible. Dans certaines conditions, lorsque les paramètres de haute dimensionnalité sont faiblement corrélés aux paramètres d'intérêt, il est possible, dans un premier temps, d'inférer ces mêmes paramètres de haute dimensionnalité avec la méthode plus efficace, par MCMC, et puis, dans un deuxième temps, les paramètres d'intérêt au moyen de l'ABC. Autrement, l'ensemble des paramètres devraient être échantillonnés conjointement. Cette nouvelle approche se nomme CABC [Laurin-Lemay et al., 2018a], pour calcul bayésien approché conditionnel (Conditional Approximate Bayesian Computation : CABC).

Nous avons validé extensivement la méthode en étudiant des simulations pour lesquelles nous connaissons les valeurs de paramètres utilisées pour les générer, de cette manière nous pouvons calculer des métriques quantifiant l'erreur sur les paramètres conditionnés ou encore les propriétés de couvertures. Nous avons pu ensuite vérifier l'efficacité de la méthode CABC en étudiant un cas d'école, soit celui de l'hypermutabilité des transitions en contexte CpG chez les Eutheria [Laurin-Lemay et al., 2018c]. Nous trouvons que 100% des 137 gènes testés possèdent une hypermutabilité des transitions dans un intervalle de crédibilité à 95% ne contenant pas 1 ; autrement dit, la probabilité que le paramètre d'hypermutabilité des transitions en contexte CpG soit plus grand que 1 est plus grande que 0,975 [p(λ_{CpG}>1) $\geq$0,975]. Nous avons aussi montré que les modèles incorporant l'hypermutabilité des transitions en contexte CpG prédisent un usage des codons plus proche de celui des gènes étudiés. Ceci suggère qu'une partie importante de l'usage des codons peut être expliquée à lui seul par les processus mutationnels et non pas par la sélection, mais qu'une autre partie de l'usage des codons n'est toujours pas expliqué, laissant place à la modélisation d'autres processus mutationnels.

Dans ce contexte, nous avons donc décidé d'étendre l'étude de l'hypermutabilité des transitions en contexte CpG à plusieurs groupes de Vertébrés et sur un plus grand nombre de gènes ainsi que d'étendre la gamme des types hypermutabilités à tester. Parmi les hypermutabilités que nous testons, il y a entre autres l'importance des transversions dans le contexte

CpG, mais aussi celui des transitions et des transversions dans le contexte TpA (communication personnelle avec Laurent Duret). Nous présentons ci-dessous des résultats préliminaires sur ces questions.

4.2. Application du CABC

4.2.1. Évolution de l'hypermutabilité des transitions en contexte CpG chez les Vertébrés

L'hypermutabilité des transitions en contexte CpG est un phénomène connu chez les Vertébrés et à notre connaissance, personne n'a étudié ce type d'hypermutabilité sur un grand nombre de groupes taxonomiques représentatifs de la diversité des Vertébrés. Chez la plupart des études, seulement deux groupes taxonomiques sont utilisés pour représenter les Vertébrés (e.g., [Mugal et al., 2015]). De plus, personne n'a étudié l'hypermutabilité des transitions en contexte CpG tout en prenant en compte la préférence site-spécifique en acides aminés comme notre modèle M[GTR+ts-CpG]-S[NCatAA*] le permet.

Ici, nous utilisons 300 alignements de séquences codantes échantillonnés dans 14 groupes de Tétrapodes, contenant de 5 à 42 espèces différentes (tableau 4.1). Ils ont été obtenus à partir des >7000 alignements de l'article d'Irisarri et al. [2017] et complétés par des données transcriptomiques du laboratoire de Miguel Vences (non publiées). Les 300 gènes retenus devaient avoir au moins 80% des espèces dans chacun des 14 groupes étudiés. Pour chaque groupe, un arbre phylogénétique de référence est inféré avec le modèle GTR+G par maximum de vraisemblance (RAxML [Stamatakis, 2014]) à partir d'une concaténation des 1689 alignements qui contiennent au moins 50% des espèces de chacun des groupes de Vertébrés étudiés.

Table 4.1. Nombre d'espèces présentes chez les 14 groupes de Vertébrés étudiés

OTU	N d'espèces
Afrotheria	5
Gekkota	7
Glires	22
Hyloidea	13
Iguania	22
Lacertibaenia	24
Laurasiatheria	38
Microhyloidae	10
Neognathae	41
Primates	23
Salamandrininae	42
Scinciformata	6
Serpentes	11
Testudines	11

Nous avons utilisé le même protocole de CABC que dans [Laurin-Lemay et al., 2018c], dans le but d'inférer le taux de transitions (λ_{tsCpG}) du M[GTR+ts-CpG]-S[NCatAA*]. La première étape consiste à analyser les 4200 alignements (300 gènes X 14 groupes taxonomiques) avec le modèle de référence, M[GTR]-S[NCatAA*], implémenté dans Phylobayes MPI [Lartillot et al., 2013a] pour obtenir les paramètres de haute dimensionnalité que sont les longueurs de branches et les préférences site-spécifiques en acides aminés. Nous avons écarté les mille premières itérations du MCMC, correspondant au burnin, et considéré les mille itérations suivantes comme un échantillonnage représentatif de la distribution *a posteriori*. Pour la partie ABC, 10^5 simulations ont été réalisées avant d'appliquer l'algorithme des k-Plus Proches Voisins libre du calcul de la vraisemblance (kNN-CABC) pour sélectionner les mille entrées de la table de référence qui minimisent la distance euclidienne au carrée. Nous utilisons le même ensemble de statistiques descriptives (que nous appelons *ss-CABC2018*) ainsi que la correction, soit un modèle de régression linéaire (LRM) est utilisé comme proposé dans l'article original de la méthode CABC [Laurin-Lemay et al., 2018c].

Entre 94.3 et 100% des gènes testés chez tous les groupes de Vertébrés étudiés obtiennent une probabilité du paramètre $p(\lambda_{tsCpG} > 1) \geq 0{,}975$ (tableau 4.2). Cet échantillonnage représente $\sim 1\%$ des séquences codantes des Vertébrés et est biaisé en faveur des gènes fortement exprimés, car essentiellement basé sur le transcriptome, il n'est donc pas certain que l'on puisse extrapoler les résultats pour l'ensemble du génome. Le taux moyen d'hypermutabilité des transitions en contexte CpG pour les 14 groupes taxonomiques varie entre 4,6 et 8,4. Les Primates, les Scinciformata et les Testudines sont les systèmes biologiques avec les plus hautes hypermutabilités des transitions en contexte CpG ($\lambda_{tsCpG} > 8$), alors que les Anura (Microhyloidae et Hyloidea) sont les systèmes biologiques avec les plus petites hypermutabilités ($\lambda_{tsCpG} < 5{,}5$). La structuration phylogénétique du λ_{tsCpG} moyen n'est pas claire, car des Amphibiens (Salamandrininae), des Mammifères (Afrotheria) et des Squamates (Serpentes) ont des taux d'hypermutabilité très proches (6,34, 6,51 et 6,74, respectivement), mais pas pour les mêmes gènes comme nous allons le voir.

OTU	proportion des gènes	moy. λ_{tsCpG}	σ
Afrotheria	1	6,513	3,143
Gekkota	0,993	7,008	3,849
Glires	1	6,613	2,952
Hyloidea	0,943	4,597	2,434
Iguania	1	6,539	3,004
Lacertibaenia	0,993	6,554	3,348
Laurasiatheria	1	5,956	3,083
Microhylidae	0,967	5,478	2,908
Neognathae	0,997	7,343	3,593
Primates	0,997	8,337	4,171
Salamandrininae	0,987	6,342	2,966
Scinciformata	0,997	8,223	4,125
Serpentes	0,967	6,739	4,270
Testudines	1	8,072	3,936

Table 4.2. Proportion des 300 gènes de Vertébrés ayant obtenues une probabilité de paramètre $p(\lambda_{tsCpG} > 1) \geq 0.975$ avec la méthodologie CABC

Le taux d'hypermutabilité des transitions en contexte CpG d'un gène est structuré phylogénétiquement sur une échelle de plusieurs dizaines de millions d'années. Par exemple, la figure 4.1 illustre la corrélation de l'hypermutabilité entre les Afrotheria et les Laurasiatheria ($R^2 = 0{,}5$), les Lacertibaenia ($R^2 = 0{,}4$), les Neognathae ($R^2 = 0{,}2$) et Microhyloidea ($R^2 = 0{,}1$), voir la figure 4.2 pour la distribution des R^2 à l'ensemble des groupe de Vertébrés. La corrélation est la plus forte entre Mammifères et la plus faible entre Mammifères et Amphibiens, et intermédiaire entre Mammifères et Sauropsides. Cependant, deux groupes de Sauropsides qui sont aussi éloignés phylogénétiquement des Afrotheria ont des corrélations assez différentes (Lacertibaenia et Neognathae).

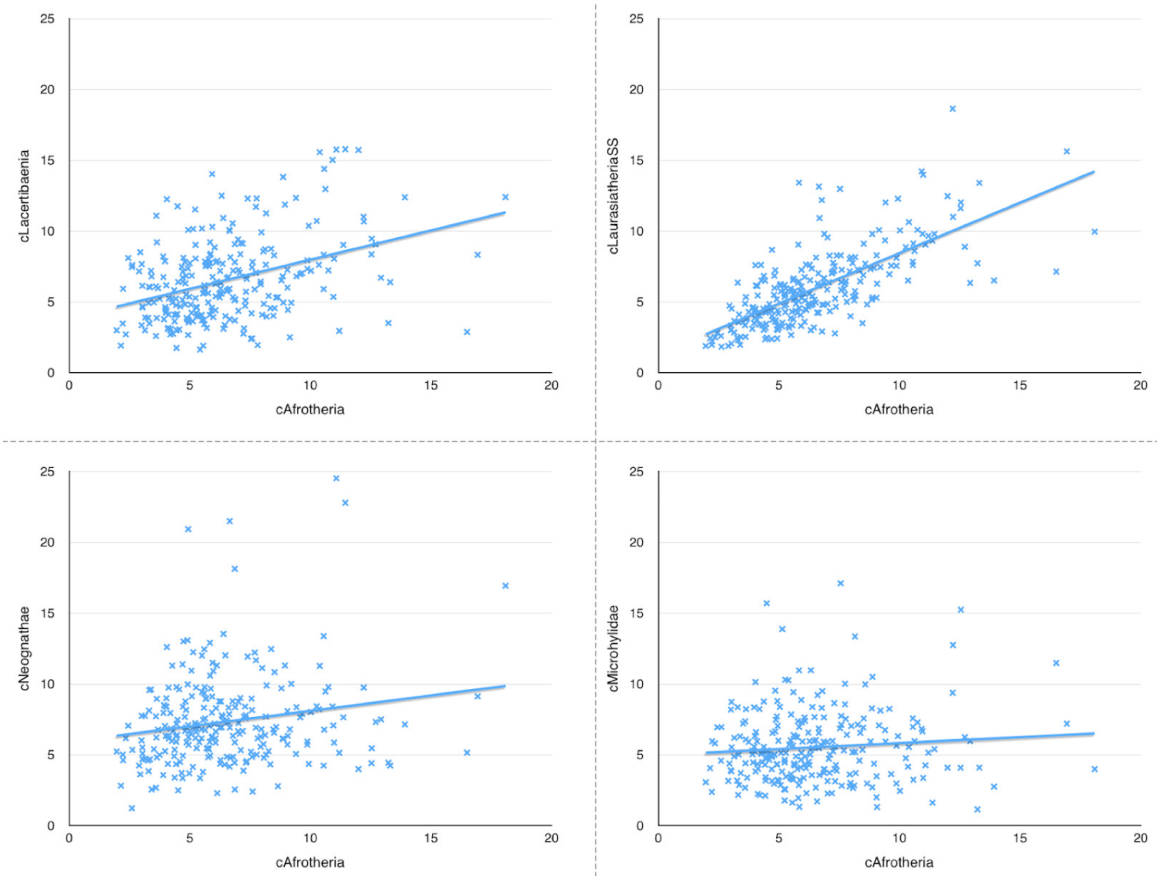

Figure 4.1. Évalulation de l'hétérogénéité d'hypermutabilité des transitions en contexte CpG entre gènes (300) et entre groupes de Vertébrés (14) détectés au moyen du modèle M[GTR+ts-CpG]-S[NCatAA*]. Une régression linéaire passant par l'origine est utilisée pour calculer le R^2.

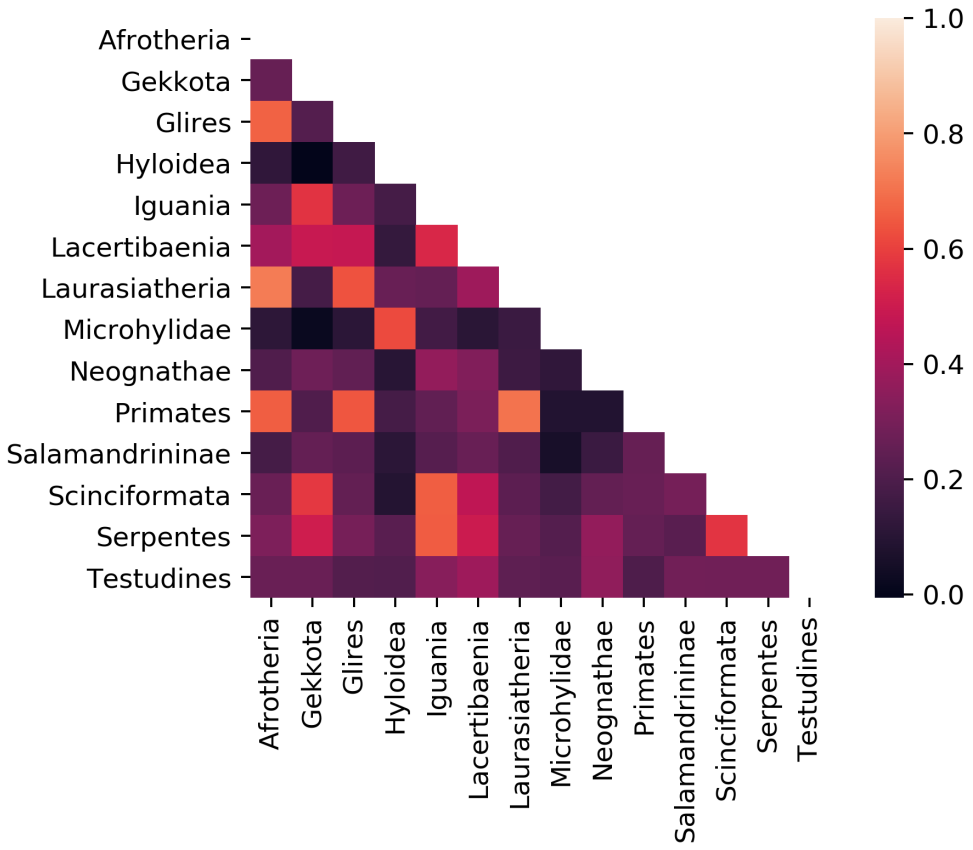

Figure 4.2. Carte de chaleur des R^2 obtenus à partir de régressions linéaires entre taux d'hypermutabilité moyens des 300 gènes de vertébrés étudiés, pour toutes les paires uniques de groupes de vertébrés.

L'ensemble des corrélations (figure 4.3) confirme l'existence d'une certaine conservation évolutive de l'hypermutabilité des transitions en contexte CpG, mais avec de multiples fluctuations. Elles peuvent être dues à un échantillon trop petit (300 gènes), mais il est difficile de l'augmenter si on souhaite que le niveau de données manquantes demeure faible pour les 14 groupes. Néanmoins, un dendrogramme calculé sur la base des R^2 représentant la conservation de l'hypermutabilité moyenne des transitions en contexte CpG entre groupes, et cela pour les 300 gènes étudiés. Il est intéressant de noter que le dendrogramme (Figure 3) retrouve la monophylie des Lepidosauria (Serpentes, Iguania, Scinciformata, Gekkota, Lacertibaenia), des Archolosauria (Neognathae, Testudines), des Eutheria (Laurasiatheria, Afrotheria, Primates, Glires), et des Anura (Microhyloidae et Hyloidea). Cela correspond à des événements de spéciation datant de 100 à 200 millions d'années. Les groupes plus anciens

(Amphibia, Amniota) ne sont pas retrouvés, montrant l'érosion du signal évolutif ainsi que la limite d'un simple modèle de régression linéaire à identifier un signal évolutif.

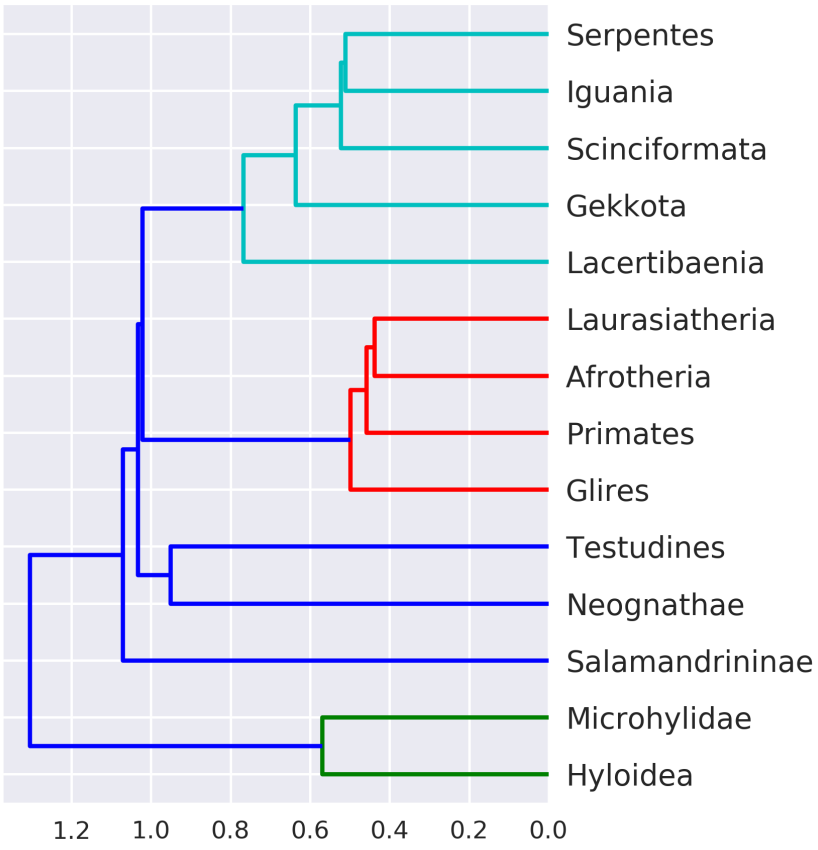

Figure 4.3. Dendrogramme représentant un groupement hiérarchisé calculé au moyen de l'algorithme des points les plus proches (Nearest Point : [Eric et al., 2001–]) à partir d'une matrice de R^2. Les R^2 sont eux-mêmes calculés au moyen d'une régression linéaire entre niveaux d'hypermutabilité des transitions en contexte CpG des 300 gènes de Vertébrés, et cela pour chacune des paires uniques tirées à partir de l'ensemble des groupes de vertébrés.

L'utilisation d'un cadre phylogénétique formel devrait grandement améliorer l'identification du signal évolutif en lien avec cette hétérogénéité de taux de transition en contexte CpG au cours du temps. Un tel modèle d'évolution n'existe pas à notre connaissance, mais pourrait potentiellement contribuer à la connaissance dans le domaine de l'inférence phylogénétique. Rappelons que l'implémentation de modèles d'évolution prenant en compte des aspects d'interdépendances entre sites de l'alignement n'est pas une tâche simple (e.g., Rodrigue et al 2009). Et sans prendre en compte la structure tertiaire des protéines, ce modèle

devrait au moins modéliser les préférences site-spécifiques en acides aminés, une hétérogénéité maintes fois reconnue pour être importante à l'étude des séquences codantes. De plus, cette variation au cours du temps de l'hypermutabilité des transitions en contexte CpG suggère aussi que certaines datations basées sur l'horloge moléculaire [dos Reis et al., 2016] sont à revoir pour deux raisons : (1) l'hypermutabilité des transitions en contexte CpG n'est pas constante au cours du temps et (2) l'hypermutabilité est potentiellement dépendante du taux de réplication qui opère dans les lignées germinales (voir l'introduction).

Autrement, sans faire l'exploration de la topologie conjointe aux autres paramètres du modèle, il serait possible d'étudier l'hypermutabilité des transitions en contexte CpG comme un trait phénotypique qui évolue le long de l'arbre des Vertébrés. Le cadre d'évolution moléculaire Coevol [Lartillot and Poujol, 2011] est tout à fait dessiné pour cette tâche. L'attrait de Coevol est de permettre de modéliser des traits qui varient au cours du temps à l'aide d'un mouvement brownien et d'étudier le niveau de corrélation moyen qu'ils entretiennent au cours de cette même évolution avec d'autres aspects de l'évolution des séquences codantes ou de manière plus générique, n'importe quels traits phénotypiques. Il serait intéressant d'essayer d'identifier avec quels autres traits l'hypermutabilité détectée peut coévoluer, par exemple à la masse corporelle ou encore à la taille du génome et à l'abondance d'éléments transposables. Les Vertébrés avec les plus grands génomes (Salamandrininae, Microphylidae, Hyloidea) et qui possèdent le plus d'éléments transposables sont aussi les Vertébrés pour lesquels les niveaux d'hypermutabilité détectés sont les plus faibles.

4.2.2. Caractérisation d'autres types d'hypermutabilité chez les Mammifères

L'hypermutabilité des transitions en contexte CpG est censée être le résultat de la désamination des cytosines méthylées. Mais nous savons aussi que la désamination des cytosines méthylées peut générer des transversions lorsque la réplication se fait sur un site abasique avant qu'il ne soit excisé par une endodeoxyribonuclease apurinique [Tubbs and Nussenzweig, 2017]. Existe-t-il une hypermutabilité des transversions en contexte CpG ?

Parmi les dinucléotides qui sont mal prédits par le modèle de référence, M[GTR]-S[NCatAA*], il y a les dinucléotides en lien avec l'hypermutabilité des CpG (CpG, CpA et TpG), mais aussi ceux en lien avec le contexte TpA (résultats non montrés). Les deux contextes CpG et TpA sont rares chez les Vertébrés [Beutler et al., 1989]. Mais contrairement

au contexte CpG, l'origine de cette sous-représentation du contexte TpA est controversée. Certains auteurs avancent que la sous-représentation du contexte TpA est le résultat des contraintes de sélection purificatrice afin d'éviter les risques liés à la similarité du contexte avec les codons d'arrêts et autres éléments de régulation comme les boîtes TATA [Beutler et al., 1989]. Les contextes TpA seraient aussi contre-sélectionnés pour éviter de déstabiliser la structure des ARNm [Beutler et al., 1989]. D'autre part, la sous-représentation du contexte TpA chez les mammifères a déjà été identifiée comme étant le résultat indirect de l'hypermutabilité CpG [Duret and Galtier, 2000]. Existe-t-il une hypermutabilité des transitions et/ou des transversions en contexte TpA ?

Pour répondre à ces questions, nous avons utilisé le même jeu de données que [Laurin-Lemay et al., 2018a] (137 alignements de gènes de 39 espèces de mammifères placentaires), qui sera appelé Eutheria39, et un nouveau jeu de données similaires, mais contenant potentiellement plus de signal évolutif (plus d'espèces et des gènes plus longs). Ce deuxième jeu d'alignements Eutheria est construit à partir de Orthomam V10 [Scornavacca et al., 2019] et comprend 100 alignements de gènes de 97 à 111 espèces et de 908 à 3055 codons (moyenne de 1484), qui sera appelé Eutheria111. Les gènes qui composent les deux jeux d'alignements Eutheria ne se chevauchent pas.

4.2.3. Modèles de substitution à codon

Plusieurs modèles de substitution à codon de type mutation-sélection ont été utilisés dans cette étude. Les modèles diffèrent par leur paramétrisation du processus mutationnel. Le modèle de référence reste M[GTR]-S[NCatAA*] comme décrit dans [Laurin-Lemay et al., 2018c]. Nous avons également étudié la possibilité d'utiliser un modèle plus simple, M[HKY]-S[NCatAA*]. Parmi les nouveaux modèles développés dans ce travail préliminaire, trois modèles explorent différents aspects de l'hypermutabilité du contexte CpG : (1) le modèle classique M[GTR+ts-CpG]-S[NCatAA*], avec un paramètre libre pour le taux de transition, λ_{tsCpG}, (2) le modèle M[GTR+tstv-CpG]-S[NCatAA*], avec un paramètre libre pour le taux de transition et de transversion ,$\lambda_{tstvCpG}$, (3) le modèle M[GTR+ts-CpG+tv-CpG]-S[NCatAA*], avec deux paramètres libres pour chacun des taux de transition et de transversion ,λ_{tsCpG} et λ_{tvCpG} respectivement.

Trois autres modèles explorent différents aspects de l'hypermutabilité du contexte TpA : (1) le modèle M[GTR+ts-TpA]-S[NCatAA*], avec un paramètre libre pour le taux de transition ,λ_{tsTpA}, (2) le modèle M[GTR+tstv-TpA]-S[NCatAA*], avec un paramètre libre pour le taux de transition et de transversion ,$\lambda_{tstvTpA}$, (3) le modèle M[GTR+ts-TpA+tv-TpA]-S[NCatAA*], avec deux paramètres libres pour chacun des taux de transition et de transversion (λ_{tsTpA} et λ_{tvTpA} respectivement). Puisque les produits de l'hypermutabilité des transitions en contextes CpG et TpA sont communs à ces deux hypermutabilités (TpG et CpA) il est intéressant de modéliser de manière jointe les deux hypermutabilités des transitions en contexte CpG et TpA. Nous avons donc aussi testé l'hypermutabilité en contextes CpG et TpA conjointement avec les modèles M[GTR+tstv-CpG+tstv-TpA]-S[NCatAA*]. Les différents modèles sont listés dans le tableau 3, ainsi que le nombre de paramètres libres de chacun des modèles mutationnels.

4.2.4. Paramétrisation de la méthodologie CABC

Pour le jeu de données Eutheria39, nous avons récupéré les chaînes MCMC déjà calculées. Pour l'autre jeu de données Eutheria111, des chaînes ont été obtenues par Phylobayes MPI [Lartillot et al., 2013a] avec le modèle M[GTR]-S[NCatAA*] (3000 cycles, burnin de 2000).

La deuxième étape consiste à construire les tables de référence pour chacune des analyses CABC à produire. Les tables de références comportent 10^5 simulations. Ces tables de référence correspondent aux distributions prédictives *a priori* des modèles. Notons que chaque table de référence est spécifique à un gène et un modèle mutation-sélection, ce qui contraint grandement la manière dont nous pouvons adresser la validation des inférences. L'approximation de la distribution *a posteriori* sera faite en conservant les 10^3 simulations qui minimisent la distance euclidienne au carré calculée à partir d'un ensemble de statistiques descriptives, comme dans l'article original du CABC [Laurin-Lemay et al., 2018c].

Différents ensembles de statistiques descriptives sont nécessaires pour permettre aux paramètres de chacun des modèles d'identifier le signal recherché. Nous allons commencer par décrire les statistiques descriptives qui sont communes à l'ensemble des modèles. Nous avons calculé les fréquences relatives des quatre bases à la troisième position des codons (SS_{A3} ; SS_{C3} ; SS_{G3} ; SS_{T3}) afin d'identifier le signal des propensions nucléotidiques, φ. Nous avons calculé la somme absolue des différences entre toutes les paires uniques non ordonnées de

séquences, et cela pour toutes les combinaisons de nucléotide ($SS_{A<>C}$; $SS_{A<>G}$; $SS_{A<>T}$; $SS_{C<>G}$; $SS_{C<>T}$; $SS_{G<>T}$) afin identifier le signal des échangeabilités, ϱ, et de l'ajustement de la grandeur de l'arbre $lambda_{TBL}$. Nous avons également calculé la somme absolue du nombre de différences non-synonymes entre toutes les paires uniques non ordonnées de séquences (SS_{NS}) afin d'identifier le signal de $lambda_{omega_*}$ et $lambda_{TBL}$. Rappelons que $lambda_{omega_*}$ permet un ajustement sur le paramètre omega*, une mesure de l'écart du taux de substitutions non-synonymes par rapport à ce qui serait attendu à l'équilibre mutation-sélection.

Par contre, nous avons différents ensembles de statistiques descriptives pour identifier le signal correspondant aux différentes paramétrisations des hypermutabilités en contextes CpG et TpA. Nous avons choisi d'utiliser les fréquences relatives des dinucléotides en position 3-1 des codons des contextes CpG, TpG et CpA (SS_{C3pG1}, SS_{T3pG1} et SS_{C3pA1}) pour permettre d'identifier le signal évolutif en lien avec le paramètre λ_{tsCpG}. Pour l'hypermutabilité des transitions en contexte TpA (λ_{tsTpA}) nous remplaçons SS_{C3pG1} par SS_{T3pA1}, puisque SS_{T3pG1} et SS_{C3pA1} sont communs.

4.2.5. Hypermutabilité des transitions et des transversions en contexte CpG chez les Eutheria

Le modèle tstv-CpG identifie une hypermutabilité des transitions et des transversions moyenne plus petite (tableau 4.4) que le modèle ts-CpG (tableau 4.3) : soit 3,73±2,34 et 5,96±2,32 versus 8,76±3,61 et 7,74±3,11, pour les jeux de données à 111 et 39 espèces respectivement. Le paramètre $\lambda_{tstvCpG}$ doit nécessairement faire un compromis entre l'hypermutabilité des transitions et l'hypermutabilité des transversions pour en moyenne expliquer le mieux les données, sachant que le taux est potentiellement beaucoup plus faible pour les transversions. Les hypermutabilités des transversions et des transitions sont en lien avec deux mécanismes moléculaires différents, soit la réplication sur le site abasique ou après son excision. Ces deux mécanismes peuvent survenir à des rythmes différents, mais lesquels ?

La proportion de gènes de Vertébrés avec une hypermutabilité des transitions et des transversions, $p(\lambda_{tstvCpG} > 1) >= 0{,}975$, est élevée : soit 0,821 pour Eutheria39 et 1 pour Eutheria111 (tableau 4.4). D'autre part, lorsque les deux hypermutabilités sont inférées séparément avec le modèle M[GTR+ts-CpG+tv-CpG], l'hypermutabilité des transitions se

rapproche du niveau obtenu avec le modèle M[GTR+ts-CpG], mais reste en dessous de la valeur moyenne obtenue avec ce modèle : 7,86 versus 8,76 pour Eutheria111 et 7,29 versus 7,74 pour Eutheria39 (tableau 4.5 et 4.3 respectivement). Le paramètre λ_{tvCpG} est souvent plus grand que 1, mais rarement de façon significative (tableau 4.5 : 16,8% et 3,6% pour Eutheria111 et Eutheria39, respectivement). Le fait qu'il y ait plus de cas significatifs pour les gènes ayant le plus de signal évolutif (Eutheria111) suggère que l'hypermutabilité des transversions existe bien, mais constitue un signal ténu. Le niveau d'hypermutabilité des transversions seules est donc difficile à estimer, mais semble être bien plus faible que celui des transitions ($\lambda_{tvCpG} = 4{,}25$ versus $\lambda_{tsCpG} = 7{,}86$ chez Eutheria111 et $\lambda_{tvCpG} = 2{,}51$ versus $\lambda_{tsCpG} = 7{,}29$ pour Eutheria39 : tableau 4.5). Quand les taux de transitions et de transversions en contexte CpG sont contraints d'être les mêmes (modèle M[GTR+tstv-CpG]), les taux inférés sont sans surprise intermédiaires pour Eutheria39 ($\lambda_{tstvCpG} = 5{,}96$) mais plus petits que les deux taux pris séparément pour Eutheria111 ($\lambda_{tstvCpG} = 3{,}73$).

Table 4.3. Proportion des gènes des 2 jeux de données Eutheria ayant $p(\lambda_{tsCpG} > 1) \geq 0.975$ obtenues au moyen de la méthodologie CABC.

OTU	proportion des gènes	moy.	σ
Eutheria111	1	8,758	3,608
Eutheria39	1	7,743	3,105

Table 4.4. Proportion des gènes des 2 jeux de données Eutheria ayant $p(\lambda_{tstvCpG} > 1) \geq 0.975$ obtenues au moyen de la méthodologie CABC.

OTU	proportion des gènes	moy.	σ
Eutheria111	0,821	3,733	2,336
Eutheria39	1	5,964	2,315

Table 4.5. Proportions des gènes des 2 jeux de données Eutheria ayant $p(\lambda_{tsCpG} > 1) \geq 0.975$ et $p(\lambda_{tvCpG} > 1) \geq 0.975$ obtenues au moyen de la méthodologie CABC.

OTU	proportion des gènes (λ_{tsCpG})	moy.	σ
Eutheria111	0,989	7,861	3,183
Eutheria39	0,993	7,291	2,896
OTU	proportion des gènes (λ_{tvCpG})	moy.	σ
Eutheria111	0,168	4,249	7,821
Eutheria39	0,036	2,505	3,605

4.2.6. Hypermutabilité des transitions et des transversions en contexte TpA chez les Eutheria

Nous avons caractérisé l'hypermutabilité des transitions et des transversions en contexte TpA. Nous trouvons que la probabilité $p(\lambda_{tsTpA} > 1) \geq 0.975$ dans 100% des gènes testés (tableau 4.6). La valeur moyenne est de 3,95±1,38 (tableau 4.6), donc environ deux fois moins importantes que pour l'hypermutabilité des transitions en contexte CpG (tableau 4.3). Des valeurs très similaires sont obtenues avec le paramètres $\lambda_{tstvTpA}$, pour l'hypermutabilité des transitions et des transversions, soit 3,84±1,15 (tableau 4.7). Par contre, en utilisant deux paramètres, λ_{tsTpA} augmente à 4,23±1.83, et reste significativement plus grand que 1 dans 98,5% des gènes testés (tableau 4.8). Au contraire TpA ne montre aucun signe d'hypermutabilité des transversions : λ_{tvTpA} a en moyenne de 1.21±1.55 et n'est pas significativement plus grand que 1. Un seul gène, CASR, est significativement plus grand que 1 (tableau 4.8).

Table 4.6. Proportion des gènes des 2 jeux de données Eutheria ayant $p(\lambda_{tsTpA} > 1) \geq 0.975$ obtenues au moyen de la méthodologie CABC.

OTU	proportion des gènes	moy.	σ
Eutheria39	1	3,943	1,378

Table 4.7. Proportion des gènes des 2 jeux de données Eutheria ayant $p(\lambda_{tstvTpA} > 1) \geq 0.975$ obtenues au moyen de la méthodologie CABC.

OTU	proportion des gènes	moy.	σ
Eutheria39	1,000	3,840	1,146

Table 4.8. Proportions des gènes des 2 jeux de données Eutheria ayant $p(\lambda_{tsTpA} > 1) \geq 0.975$ et $p(\lambda_{tvTpA} > 1) \geq 0.975$ obtenues au moyen de la méthodologie CABC.

OTU	proportion des gènes (λ_{tsTpA})	moy.	σ
Eutheria39	0,985	4,228	1,831
OTU	proportion des gènes (λ_{tvTpA})	moy.	σ
Eutheria39	0,007	1,207	1,546

Cependant, nous avons toutes les raisons de penser qu'il y a une forte interaction dans l'inférence de l'hypermutabilité des CpG et de celle des TpA : par exemple, une transition produit TpG ou CpA dans les deux cas. Nous avons donc commencé à explorer cette complication en étudiant le modèle M[GTR+tstv+CpG+tstv-TpA] (tableau 4.9). L'hypermutabilité des CpG domine (4,94±2,26 versus 2,02±2,02). Il est important de noter que l'hypermutabilité des TpA, bien que faible, est souvent significative (dans 84% des cas). L'hypermutabilité des CpG est plus faible quand TpA est pris en compte que lorsque l'on en tient pas compte (4,94±2,26 versus 5,96±2,32). Cela suggère que le processus d'hypermutabilité dépend du contexte de manière complexe. Il est donc nécessaire de poursuivre l'exploration des différentes hypermutabilités et de leurs combinaisons.

Table 4.9. Proportions des gènes des 2 jeux de données Eutheria ayant $p(\lambda_{tstvCpG} > 1) \geq 0.975$ et $p(\lambda_{tstvTpA} > 1) \geq 0.975$ obtenues au moyen de la méthodologie CABC.

OTU	proportion des gènes ($\lambda_{tstvCpG}$)	moy.	σ
Eutheria39	1	4.943	2.258
OTU	proportion des gènes ($\lambda_{tstvTpA}$)	moy.	σ
Eutheria39	0.839	2.016	0.641

4.2.7. Utilisation de la distribution prédictive *a posteriori* pour comparer les modèles

La comparaison de modèles est un problème statistique complexe, particulièrement dans un cadre ABC (voir l'introduction de la thèse). Il est néanmoins possible de comparer les différents modèles par leur habilité à prédire l'usage des codons en calculant une distance euclidienne au carré entre l'usage des codons observés de chaque alignement à celui de l'usage des codons calculé sur les simulations générées à partir de la distribution prédictive *a posteriori*. C'est ce que nous avons fait dans [Laurin-Lemay et al., 2018b]. Mille tirages (simulations) sont réalisés pour chacune des conditions (137 gènes X 8 modèles). Les distances sont calculées sur le RSCU (Relative Synonynous Codon Usage) excluant la méthionine et le tryptophane (tableau 4.10). Dans le cas du RSCU standardisé (stand.), avant d'effectuer le calcul de la distance, les valeurs sont centrées réduites.

Le modèle qui a les meilleures performances est le modèle qui possède un seul paramètre d'hypermutabilité combinant les transitions et des transversions pour les contextes CpG ($\lambda_{tstvCpG}$) et TpA ($\lambda_{tstvTpA}$) respectivement, donc M[GTR+tstv-CpG+tstv-TpA]. Par contre, étant donné les barres d'erreurs, toutes les prédictions se chevauchent, ce qui montre que cette statistique n'est pas suffisamment puissante. Ceci a motivé notre intérêt pour l'utilisation de l'algorithme de forêts aléatoires (voir ci-dessous) qui pourrait permettre des comparaisons de modèles dans un cadre ABC [Raynal et al., 2017].

Table 4.10. Moyenne des distances calculées entre l'usage des codons (RSCU) réel et l'usage des codons prédit par tirage à partir de la distribution prédictive *a posteriori* pour chacun des 137 gènes Eutheria39.

modèles mutationnels	RSCU moy.	RSCU σ	RSCU stand. moy.	RSCU stand. σ
M[GTR]	0,973	0,283	101,911	24,679
M[HKY]	0,984	0,286	102,98	25,053
M[GTR+ts-CpG]	0,669	0,255	78,36	25,132
M[GTR+ts-TpA]	1,17	0,365	113,108	31,305
M[GTR+ts-CpG+tv-CpG]	0,667	0,251	77,295	24,63
M[GTR+tstv-CpG]	0,645	0,252	73,042	24,428
M[GTR+tstv-CpG+tstv-TpA]	0,632	0,256	67,945	23,749
M[GTR+tstv-TpA]	1,212	0,373	117,348	30,559

Nous avons ensuite effectué une analyse en composante principale pour essayer de mieux comprendre ce qui était, ou non, expliqué par les différents modèles. Pour chaque gène et chaque modèle, on a utilisé la valeur moyenne de l'usage des codons (RSCU) obtenu pour les mille simulations. L'axe des abscisses (PCA1) explique la plus grande variance du RSCU, soit 60% (Figure 4), et permet de séparer les codons qui se termine par A/T et G/C, sauf pour TTG (leucine) et AGG (arginine), tel que décrit dans [Laurin-Lemay et al., 2018c]. De très loin, sur le premier axe, seul le modèle de référence semble être différent des données réelles (trop à droite), tous les autres modèles semblent être aussi proches des données réelles.

Le deuxième axe qui explique le plus de variance (11%), positionne en bas les prédictions de RSCU faites par le modèle de référence (vert) et en haut le RSCU observés sur les données réelles (gris). De nouveau, le modèle de référence apparaît bien séparer les données réelles. Par après, il est plus difficile de différencier les RSCU prédits par les différents modèles qui incorporent des hypermutabilités contextuelles (M[GTR+ts-CpG], M[GTR+tstv-CpG], M[GTR+ts-tv-CpG] et M[GTR+tstv-CpG+tstv-TpA]). Mais les modèles M[GTR+tstv-CpG] et M[GTR+tstv-CpG+tstv-TpA] semblent avoir les prédictions se rapprochant le plus des RSCU calculés sur les alignements réels. Cependant, les différences sont ténues et impliquent seulement 11% de la variance. Les résultats sont beaucoup plus difficiles à interpréter pour les deux axes suivants (Figure 5 : PCA 3 et 4) ; ils expliquent 4,3 et 2,6% de la variance totale, respectivement. L'axe PCA 4 ne permet pas de discriminer les différents modèles, mais selon l'axe PCA 3 les meilleurs modèles sont le modèle de référence et les modèles M[GTR+tstv-CpG] et M[GTR+tstv-CpG+tstv-TpA], les modèles M[GTR+ts-CpG] et M[GTR+ts-CpG+tv-CpG] semblant faire des prédictions situées trop vers la droite. Cette complexité confirme qu' il nous faut des moyens plus puissants pour ordonner l'habilité de nos modèles à prédire les séquences (et en particulier le RSCU). Nous allons explorer l'utilisation de l'algorithme de forêts aléatoires pour ce faire dans la section suivante.

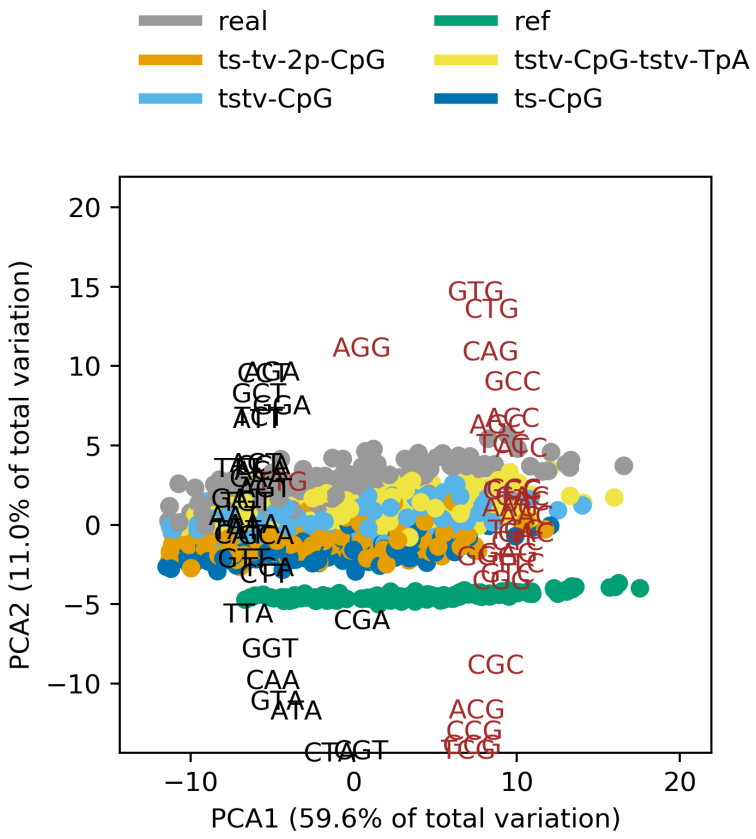

Figure 4.4. Analyse en composantes principales (axes 1 et 2) du RSCU (sans les codons d'arrêt, de la méthionine et du tryptophane) calculé à partir des 137 gènes du jeu Eutheria39 et à partir de la moyenne des RSCU prédits sous les modèles sans et avec hypermutabilités des CpG et TpA. Les codons qui se terminent par G/C sont annotés en rouge alors que les codons qui se terminent par A/T sont annotés en noir. (real, gris) RSCU observée des 137 gènes du jeu Eutheria39. (ref, vert) Prédictions à partir du modèle de référence. (tv-CpG, bleu foncé) Prédictions à partir du modèle M[GTR+ts-CpG]. (tstv-CpG, bleu ciel) Prédictions à partir du modèle M[GTR+tstv-CpG]. (ts-tv-2p-CpG, orange) Prédictions à partir du modèle M[GTR+ts-CpG+tv-CpG]. (tstv-CpG+tstv-TpA, jaune) Prédictions à partir du modèle M[GTR+tstv-CpG+tstv-TpA].

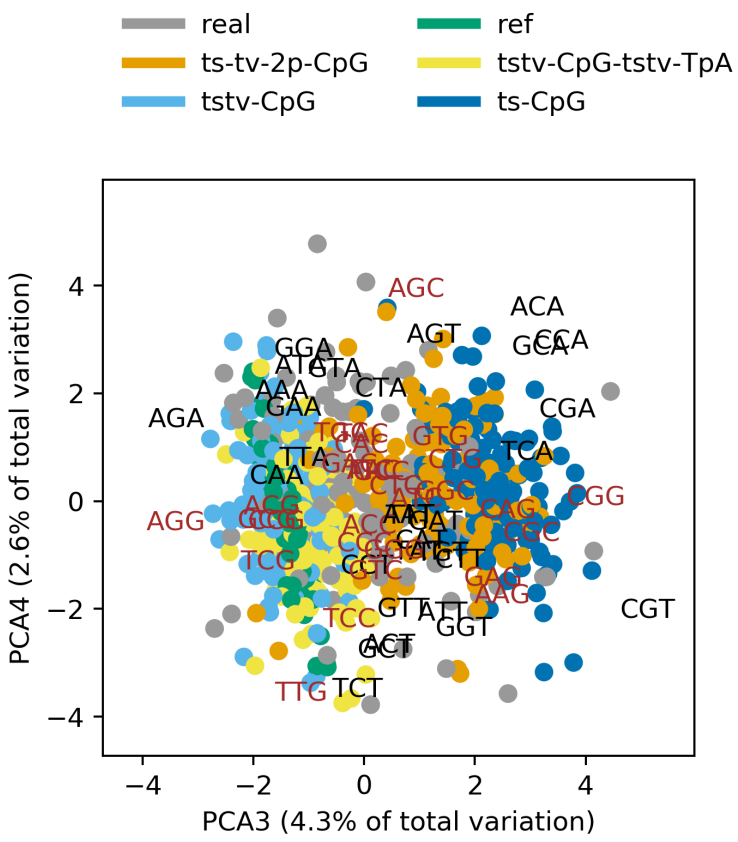

Figure 4.5. Analyse en composantes principales (axes 3 et 4) du RSCU (sans les codons d'arrêt, de méthionine et de tryptophane) calculé à partir des 137 gènes du jeu Eutheria39 et à partir de la moyenne des RSCU prédits sous les modèles sans et avec hypermutabilités des CpG et TpA. Les codons qui se terminent par G/C sont annotés en rouge alors que les codons qui se terminent par A/T sont annotés en noir. (real, gris) RSCU observée des 137 gènes du jeu Eutheria39 ; (ref, vert) Prédictions à partir du modèle de référence. (tv-CpG, bleu foncé) Prédictions à partir du modèle M[GTR+ts-CpG]. (tstv-CpG, bleu ciel) Prédictions à partir du modèle M[GTR+tstv-CpG]. (ts-tv-2p-CpG, orange) Prédictions à partir du modèle M[GTR+ts-CpG+tv-CpG]. (tstv-CpG-tstv-TpA, jaune) Prédictions à partir du modèle M[GTR+tstv-CpG+tstv-TpA].

4.3. Amélioration du CABC en utilisant des algorithmes de forêts aléatoires

Nous venons de voir que l'approche de CABC élaborée dans le chapitre 2 peut facilement s'appliquer à de nouveaux jeux de données tant que nous ne nous intéressons qu'à l'hypermutabilité des transitions en contexte CpG, mais montre un potentiel limité lorsque nous étudions d'autres types d'hypermutabilité. En particulier, à mesure que nous ajoutons des hypermutabilités contextuelles, le choix des statistiques descriptives devient de plus en plus complexe à cause de la combinatoire et de la non-linéarité des relations entre les paramètres des modèles et les statistiques descriptives. Nous allons maintenant explorer l'utilisation de l'algorithme de forêts aléatoires [Breiman, 2001] pour (1) aider au choix des statistiques descriptives, puisque ce dernier est reconnu pour être robuste au fléau de la dimensionnalité ; (2) faire la correction des distributions *a posteriori* produites via l'algorithme kNN-ABC au lieu du modèle de régression linéaire que nous avions utilisé jusqu'à maintenant (LRM). À notre connaissance, personne n'a tenté d'utiliser ce type de correction, bien qu'elle ait été suggérée par Saulnier et al. [2017]. L'algorithme de forêts aléatoires a été récemment utilisé pour inférer non pas la distribution a posteriori dans le contexte de l'ABC [Raynal et al., 2017], mais des statistiques descriptives de cette distribution (e.g., quantiles, espérance et variance). Mais puisque nous tenons à tirer de la distribution a posteriori, cette approche est très limitative pour le moment ; et (3) faire des comparaisons de modèles selon le développement méthodologique proposé par Pudlo et al. [2016].

4.3.1. Choix des statistiques descriptives à l'aide de l'algorithme de forêts aléatoires

Un aspect difficile de la méthodologie CABC, ou plus spécifiquement de l'étape ABC du CABC, est celle de choisir les statistiques descriptives à utiliser pour approximer la distribution a posteriori. La tâche est difficile, car nous voulons garder que les statistiques descriptives les plus informatives de manière à maximiser le taux d'acceptation. Utiliser un trop grand nombre de statistiques descriptives nous expose au fléau de dimensionnalité, et aura pour conséquences potentielles de déformer la distribution a posteriori approximée (revue dans [Prangle, 2018]). Dans le meilleur des cas, chacune des statistiques descriptives est suffisante, ce qui veut dire au sens bayésien que les probabilités a posteriori $p(\theta|S(D),M)$

et $p(\theta|D,M)$ possèdent la même distribution pour toutes distributions *a priori* (revue dans [Prangle, 2018]), où $S(D)$ est une fonction qui permet l'extraction d'une statistique descriptive à partir des données D, et où θ est le paramètre du modèle M.

Dans l'article original de la méthodologie du CABC [Laurin-Lemay et al., 2018c] nous nous sommes servis de notre intuition pour choisir les statistiques descriptives à utiliser dans l'étape ABC du CABC. Ici nous explorons la possibilité d'utiliser l'algorithme de forêts aléatoires sous sa forme de modèle de régression dans le but d'identifier les statistiques descriptives les plus importantes. Nous utilisons l'implémentation de l'algorithme de forêts aléatoires disponible dans la suite scikit-learn [Pedregosa et al., 2011]. La mesure de l'importance des statistiques descriptives nous aide non seulement à sélectionner les variables explicatives les plus utiles, mais aussi potentiellement à mieux comprendre la relation entre les variables explicatives et les variables réponses enfouies dans la boîte noire (black box) du modèle de régression fort complexe que l'algorithme de forêts aléatoires permet de déployer. Cette stratégie permet de proposer des pistes de compréhension aux scientifiques toujours à la recherche d'explications mécanistiques des phénomènes qu'ils étudient (Trends2010)[Rodrigue and Philippe, 2010].

Nous avons utilisé la méthode par permutation pour évaluer l'importance des statistiques descriptives telle que présenté par [Fisher et al., 2018]. L'avantage principal de cette méthode est qu'elle est beaucoup moins gourmande en temps calcul que l'approche où une itération complète de l'algorithme de forêts aléatoires doit être calculée pour évaluer l'importance de chacune des variables explicatives.

Rappelons certains aspects du calcul du R^2 et de son utilisation pour comparer les modèles entre eux. Nous avons l'habitude d'utiliser le coefficient de détermination, R^2, pour évaluer l'habileté des modèles à prédire les variables réponses à partir des variables explicatives. Tout d'abord, un rapport, u/v, permet d'évaluer l'habilité à prédire les données par le modèle alternatif. Le numérateur, u, est l'écart à la prédiction faite par le modèle alternatif et est évalué en faisant la somme des écarts au carré entre les données observées et les prédictions du modèle alternatif. Plus u est petit, plus le modèle prédit bien les données. Le dénominateur, v, la variance, est en fait l'écart à la prédiction faite par le modèle de référence, le modèle le plus simple avec un seul paramètre, soit la moyenne ($Y = w$, voir la sous-section les modèles de régression de l'introduction de la thèse). On s'attend à ce que v soit toujours plus grand

que u, donc u/v entre 0 et 1, puisque v explique moins de variance que u. Puis le R^2, ou le coefficient de détermination est obtenu en calculant $1 - u/v$. Plus le R^2 est près de 1, plus le modèle est habile à prédire les variables réponses. Si le résultat est négatif, c'est que le modèle alternatif u fait moins bien que le modèle de référence,v.

La manière la plus crue de calculer l'importance de chacune des variables explicatives (1) consiste à calculer la différence entre le R^2 calculé par la validation croisée OOB lorsque toutes les variables explicatives sont présentes et le R^2 obtenu lorsqu'une variable explicative est absente du jeu de données [Fisher et al., 2018]. Ainsi, plus la différence sera grande, plus l'importance de la variable sera grande, l'importance étant 1 moins la différence des R^2. En omettant à tour de rôle chacune des variables explicatives, il est possible de les ordonner sur la base de leur importance à la prédiction des variables réponses. Une implémentation alternative, beaucoup moins gourmande en temps calcul, (2) consiste à permuter les valeurs d'une variable explicative de manière à brouiller le signal potentiellement utile à la prédiction de la variable réponse étudiée au moyen d'un modèle de régression ou de classification. De ce fait nous pouvons directement évaluer l'impact de la permutation sur la prédiction du modèle en calculant un R^2 à partir de la validation croisée OOB et cela itérativement pour chacune des variables explicatives, sans avoir à calculer chaque fois un nouvel ensemble d'arbres. Il faut aussi itérer à travers les variables réponses, car l'algorithme de forêts aléatoires prend qu'une seule variable réponse à la fois.

Nous avons d'abord construit une table de référence. Chacune des entrées de la table de référence correspond aux valeurs de paramètres utilisés pour générer les alignements simulées (paramètres d'intérêt et de nuisance fortement corrélés) ainsi que 189 statistiques descriptives calculées à partir des alignements simulés. Chacune des tables de référence comprend $5X10^4$ entrées. Rappelons que les tables de références correspondent à la distribution prédictive *a priori*. Pour chaque modèle, M[GTR+ts-CpG] et M[GTR+ts-CpG+tv-CpG], possède 10 tables de référence sont générées, correspondant aux 10 gènes pour lesquels le modèle de référence, M[GTR]-S[NCatAA*], prédit le plus grand nombre de substitutions (longueur d'arbre $\times$ nombre de sites) sous le modèle mutationnel. Nous avons dans un premier temps évalué la progression de la valeur du R^2 en fonction du nombre d'arbres calculés (entre 100 et 1000) ; nous avons choisi de travailler avec 500 arbres puisque l'amélioration du R^2 par validation croisée OOB atteint un plateau et de ce fait limite le temps calcul.

Tout d'abord, la première chose qui saute aux yeux (figure 4.6 et 4.7) est que certaines des statistiques descriptives étudiées obtiennent une importance plus grande que 1, ce qui est supposé signifier qu'en absence de cette statistique descriptive le modèle obtient un R^2 négatif. D'autre part, nous cherchons toujours à comprendre le sens de certains résultats. Comment expliquer par exemple que pour les échangeabilités, φ, les SS_{dinuc} se retrouve parmi les plus importantes statistiques descriptives pour prédire ces paramètres (figure 4.6). Par exemple, pour ϱ_{AG} les $SS_{pw(A<>G)10}$ et dinucAG obtiennent des importances plus grandes que 3 et 1 respectivement (figure 4.6), cela suggère fortement qu'il y a un problème dans le calcul. Nous avons quelques expériences en tête pour améliorer notre compréhension du fonctionnement de l'algorithme de forêts aléatoires. Tout d'abord, nous pourrions comparer ici les résultats obtenus avec la stratégie du calcul de l'importance par permutation avec celle où une itération complète de l'algorithme de forêts aléatoires est réalisée en absence des statistiques descriptives. Pour éviter de trop prendre de temps calcul, nous pourrions étudier les statistiques descriptives ayant obtenues des valeurs d'importance plus grande que 1. D'autre part, pour faciliter notre compréhension (1) il serait possible de faire le calcul de l'importance sur des tables de références peuplées de simulations faites avec des arbres phylogénétiques plus grands, et (2) avec un plus grand nombre de simulations.

Tout d'abord, les statistiques descriptives identifiées par l'algorithme de forêts aléatoires confirme la majorité des choix que nous avions faits à l'époque du travail [Laurin-Lemay et al., 2018c] (figure 4.8). L'algorithme de forêts aléatoires confirme l'importance de prendre en compte la fréquence relative du dinucléotide en contextes CpG de la position 3-1 des codons (SS_C3pG1) pour identifier le signal en lien avec l'hypermutabilité des transitions du contexte CpG (figure 4.6 et 4.7), mais nous indique aussi préférer le contexte CpA en position 2-3 des codons (SS_C2pA3), aussi produit de la désamination des cytosines méthylées en contexte CpG (tableau 4.11). L'avantage d'utiliser SS_C2pA3 est que les mutations G vers A en position 3 des codons sont synonymes contrairement aux mutations en position 1 des codons. Donc, le signal identifié est libre des contraintes de la sélection.

Nous avons confectionné deux nouveaux ensembles de statistiques descriptives (tableau 4.11 : ssRF15 et ssRF17), pour précisément prendre en compte l'hypermutabilité des transitions (ts-CpG) et des transversions (tv-CpG) en contexte CpG. Sur la base de leur importance dans la prédiction du paramètre λ_{tvCpG}, nous avons inclus la fréquence relative du

dinucléotide CpG de la position 1-2 et 2-3 des codons (SS_{C1pG2} et SS_{C2pG3}) dans l'ensemble ssRF17 (figure 4.7). À noter que les transversions en position 1-2 des codons en contexte CpG permettent des changements entre codons synonymes de l'arginine (CGN vers AGR). Il est possible que l'hypothèse de l'hypermutabilité des transversions en contexte CpG ne soit supportée que parce que l'acide aminé intermédiaire faisant le pont entre les codons de l'arginine CGN et AGR via une hyperméabilité des transitions en contexte CpG est absent des alignements. Cet acide aminé intermédiaire n'apparaît donc pas dans les profils de préférence en acides aminés, la sélection purificatrice étant trop grande pour ce dernier. Nous avons repris les mêmes statistiques descriptives (SS_{A3}, SS_{C3}, SS_{G3} et SS_{T3}) que celles présentes dans l'ensemble ssCABC2018 pour identifier le signal convoité par les propensions nucléotidiques (φ) : le $GC3$ ($SS_{G3} + SS_{C3}$) est un proxy reconnu du processus mutationnel [Sueoka, 1961, Muto and Osawa, 1987, Ermolaeva, 2001, Knight et al., 2001, Chen et al., 2004, Li et al., 2015]. Nous avons calculé les différences entre toutes les paires uniques non ordonnées des séquences et cela pour toutes les combinaisons de nucléotides, comme nous l'avions fait dans l'ensemble ssCABC2018 ($SS_{A<>C}$, $SS_{A<>G}$, $SS_{A<>T}$, $SS_{C<>G}$, $SS_{C<>T}$ et $SS_{G<>T}$), mais n'avons retenu pour calculer les sommes absolues pour chacune des combinaisons de nucléotides que les 10% des paires les plus proches ($SS_{(A<>C)10}$, $SS_{(A<>G)10}$, $SS_{(A<>T)10}$, $SS_{(C<>G)10}$, $SS_{(C<>T)10}$ ey $SS_{(G<>T)10}$). Le calcul de l'importance nous a permis de discriminer entre les sommes absolues calculées pour 10%, 30%, 50% et 100% des paires les plus proches. De cette manière, nous minimisons le biais que pourrait générer la saturation due à de multiples substitutions lorsque deux séquences évolutivement éloignées servent au calcul de la statistique. Les $SS_{(N<>N)10}$ servent à identifier le signal convoité par les échangeabilités, ϱ, et λ_{TBL}. L'algorithme de forêts aléatoires nous a permis d'introduire une nouvelle statistique, soit la distance de Kimura [Kimura, 1980], pour identifier le signal en lien avec l'ajustement de la grandeur de l'arbre phylogénétique, λ_{TBL}. Cette distance modélise l'hétérogénéité du taux de transition au taux de transversion, elle est donc moins sensible à la saturation. La distance est obtenue de la manière suivante $K = -(1/2)ln[(1 - 2P - Q) * racine^2(1 - 2Q)]$, où P et Q correspondent à la fréquence des sites impliquant des transitions et des transversions respectivement. De la même manière que pour les $SS_{(N<>N)10}$, nous avons calculé les distances de Kimura ($SS_{k80nuc10}$) entre toutes les paires uniques non ordonnées de séquences,

mais n'avons retenu que les 10% des paires les plus proches, pour faire la somme des distances. Nous avions dans la liste des 189 statistiques descriptives les sommes des distances de Kimura pour 10%, 30%, 50% et 100% des paires les plus proches. La distance Kimura calculée sur la position 3 des codons et avec 10% des paires les plus proches possèdent la seconde valeur d'importance pour le paramètre λ_{TBL}. Nous avons également calculé le nombre de différences non-synonymes entre toutes les paires uniques non ordonnées de séquences, mais n'avons retenu que les 10% des paires, les plus proches, pour calculer la somme des différences non-synonymes SS_{NS10}.

En résumé, l'algorithme de forêts aléatoires nous a permis d'introduire les SS_{C1pG2} et SS_{C2pG3} pour identifier le signal en lien avec λ_{tvCpG} et d'éliminer les SS_{T3pG1} et SS_{C3pA1}, produits de la désamination du contexte CpG en position 3-1 des codons. Nous avons aussi introduit une nouvelle statistique descriptive $SS_{k80nuc10}$ basée sur la distance de Kimura et modifié la manière dont nous calculons les $SS_{N<>N}$ pour éviter la saturation.

Pour évaluer rapidement l'impact du choix des statistiques descriptives, nous allons comparer les moyennes a posteriori des paramètres d'intérêt λ_{tsCpG} et λ_{tvCpG} des deux modèles, M[GTR+ts-CpG] et M[GTR+ts-CpG+tv-CpG], retrouvées avec l'ensemble originalement utilisé dans [Laurin-Lemay et al., 2018c], ssCABC2018, et les deux nouveaux ensembles choisis à l'aide de l'étude de l'importance (ssRF15 et ssRF17) dans le contexte de l'application de CABC. Les deux nouveaux ensembles, ssRF15 et ssRF17, sont utilisés respectivement avec les modèles M[GTR+ts-CpG] et M[GTR+ts-CpG+tv-CpG]. Les moyennes a posteriori retrouvées avec les trois ensembles sont semblables (figure 4.9 et 4.10). Par contre, l'inférence CABC faite avec les ensembles de statistiques descriptives dont le choix a été guidé par l'étude de l'algorithme de régression de type forêts aléatoires a généré dans quelques cas des résultats aberrants. Pour le paramètre λ_{tsCpG} du modèle M[GTR+ts-CpG] les moyennes sont comparables (figure 4.9). Par contre certains écarts-types des gènes *NR4A2* (24,17), *YTHDF2* (112,07) et *DYNC1I2* (253,68) sont très élevées, ce qui suggère que le LRM éprouve de la difficulté à identifier la relation linéaire entre statistiques descriptives et paramètres λ_{tsCpG}. Toujours dans le contexte CABC+LRM(ssRF15), le paramètre λ_{tsCpG} du modèle M[GTR+ts-CpG+tv-CpG] obtient des moyennes a posteriori aberrantes pour les gènes *NR4A2* et *DYNC1I2* (110,16 et 1637,91 respectivement) alors qu'ils obtiennent, avec *YTHDF2*, des écarts-types élevés (15,98, 1754,5, 36959,38). Les moyennes a posteriori retrouvées pour le paramètre tv-CpG du modèle

M[GTR+ts-CpG+tv-CpG] sont en moyenne moins élevées lorsque CABC+LRM est utilisé non pas avec l'ensemble de statistiques descriptives ssCABC2018, mais avec celles choisies avec l'algorithme de forêts aléatoires (figure 1.11). Par contre, CABC+LRM(ssRF17) génère un écart-type aberrant pour le gène *YTHDF2* (4573410). Les valeurs aberrantes obtenues avec CABC+LRM(ssRF15) et CABC+LRM(ssRF17) suggèrent une mauvaise modélisation des relations entre statistiques descriptives et valeur de paramètres d'intérêt. Faut-il faire une correction de la distribution a posteriori avec le modèle qui a été utilisé pour choisir les statistiques descriptives ? Car, après tout, l'importance a été étudiée avec un modèle de régression capable de prendre en compte des relations non linéaires, ce qui pourrait expliquer que ce choix de statistiques descriptives ne convient pas parfaitement au modèle LRM.

Pour explorer cette piste, nous allons revenir brièvement sur le protocole utilisé pour faire le choix des statistiques descriptives à l'aide de l'algorithme de forêts aléatoires. Rappelons tout d'abord que les statistiques descriptives ont été choisies avec les 10 gènes pour lesquels le modèle de référence prédit le plus de mutations (grandeur de l'arbre), proxy du nombre de substitutions. En d'autres mots, ce sont les gènes pour lesquels nous avons potentiellement le plus de signal évolutifs. Lorsque CABC+LRM est utilisé conjointement avec les statistiques descriptives choisies avec l'algorithme de forêts aléatoires (ssRF15 et ssRF17) versus ssCABC2018, les 10 gènes pour lesquels le modèle de référence prédit le plus de mutation sont ceux qui obtiennent les écarts-types des paramètres d'intérêt inférés les plus petits (0,4 à 1,23 versus 0,5 à 2,42 pour le paramètre λ_{tsCpG} du modèle M[GTR+ts-CpG], 0,41 à 1,07 versus 0,57 à 1,54 pour le paramètre λ_{tsCpG} du modèle M[GTR+ts-CpG+tv-CpG] et 0,36 à 1,36 versus 0,52 à 15,01 pour le paramètre λ_{tvCpG} de M[GTR+ts-CpG+tv-CpG]), ce qui suggère que ces ensembles (ssRF15 et ssRF17) de statistiques descriptives sont optimaux pour les 10 gènes étudiés, mais pas nécessairement pour les autres. Nous pourrions aussi établir l'importance moyenne de chacune des statistiques descriptives à partir de l'ensemble des 137 gènes Eutheria39 et à ce moment, peut être que le choix des statistiques descriptives sur la base de leur importance serait plus adapté en moyenne ?

Table 4.11. Équivalences et changements dans le choix des statistiques descriptives guidés par l'algorithme de forêts aléatoires de trois différents ensembles de statistiques descriptives, dont ssCABC2018.

paramètres	ssCABC2018	ssRF15	ssRF17
λ_{tsCpG}	SS_{C3pG1}	conservée	conservée
λ_{tsCpG}	SS_{T3pG1}	conservée	conservée
λ_{tsCpG}	SS_{C3pA1}	Modifiée pour SS_{C2pA3}	SS_{C2pA3}
λ_{tvCpG}			Ajoutée SS_{C2pG3}
λ_{tvCpG}			Ajoutée SS_{C1pG2}
φ	SS_{N3} (4)	conservées	conservées
ϱ	$SS_{N<>N}$ (6)	Modifiées pour $SS_{(N<>N)10}$	Modifiées pour $SS_{(N<>N)10}$
λ_{omega_*}	SS_{NS}	Modifiée pour $SS_{(NS)10}$	Modifiée pour $SS_{(NS)10}$
λ_{TBL}		$SS_{k80nuc10}$	$SS_{k80nuc10}$

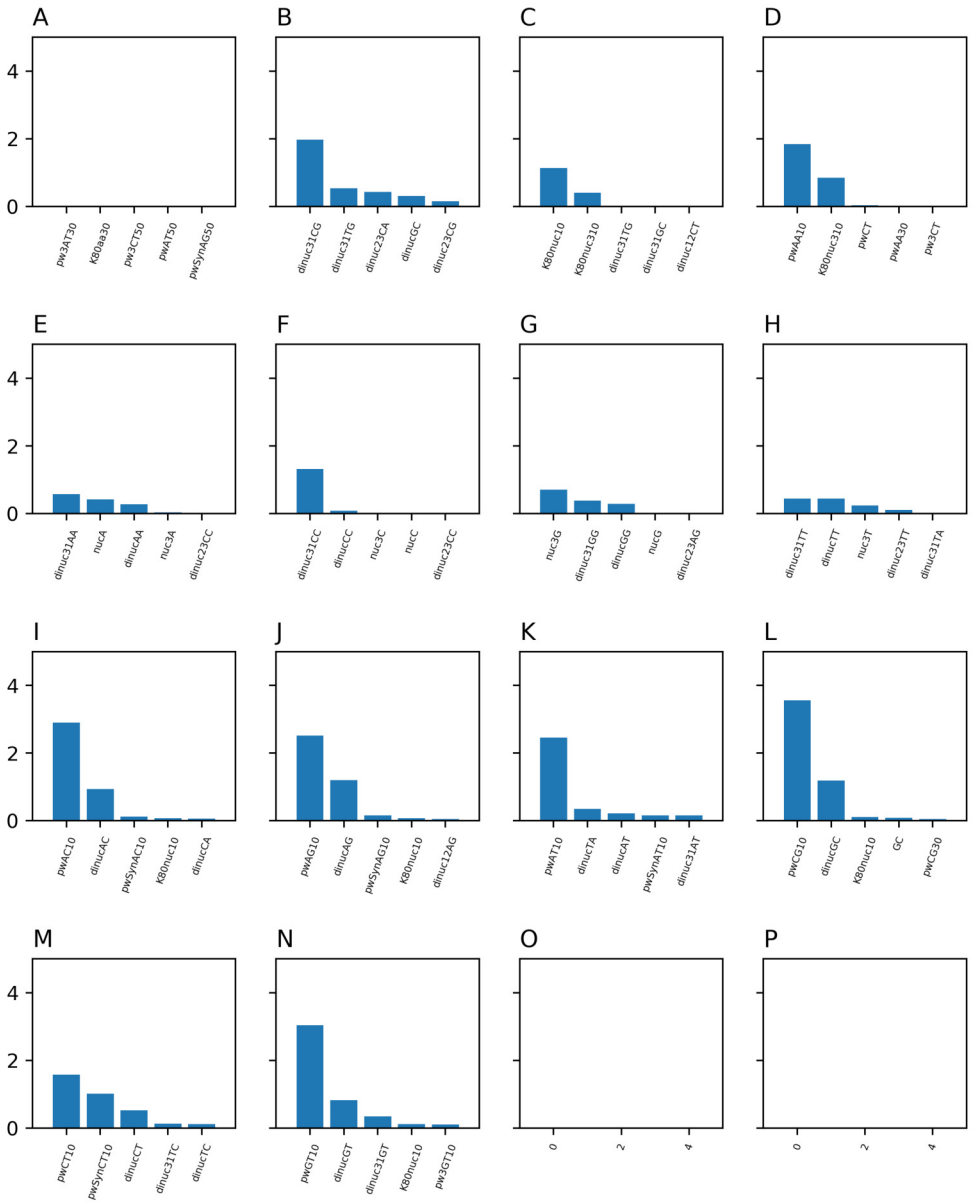

Figure 4.6. Valeur moyenne de l'importance des 5 plus importantes statistiques descriptives, dans l'ordre descendant de gauche à droite, pour chacun des paramètres du M[GTR+tsCpG]-S[NCatAA*], identifiés par les lettres A-N. Se référer aux annexes pour connaître les 189 statistiques descriptives utilisées dans le test. (A) Paramètre root ; (B) Paramètre λ_{tsCpG} ; (C) Paramètre λ_{TBL} ; (D) Paramètre λ_{omega_*} ; (E) Paramètre φ_A ; (F) Paramètre φ_C ; (G) Paramètre φ_G ; (H) Paramètre φ_T (I) Paramètre ϱ_{AC} ; (J) Paramètre ϱ_{AG} ; (K) Paramètre ϱ_{AT} ; (L) Paramètre ϱ_{CG} ; (M) Paramètre ϱ_{CT} ; (N) Paramètre ϱ_{GT}.

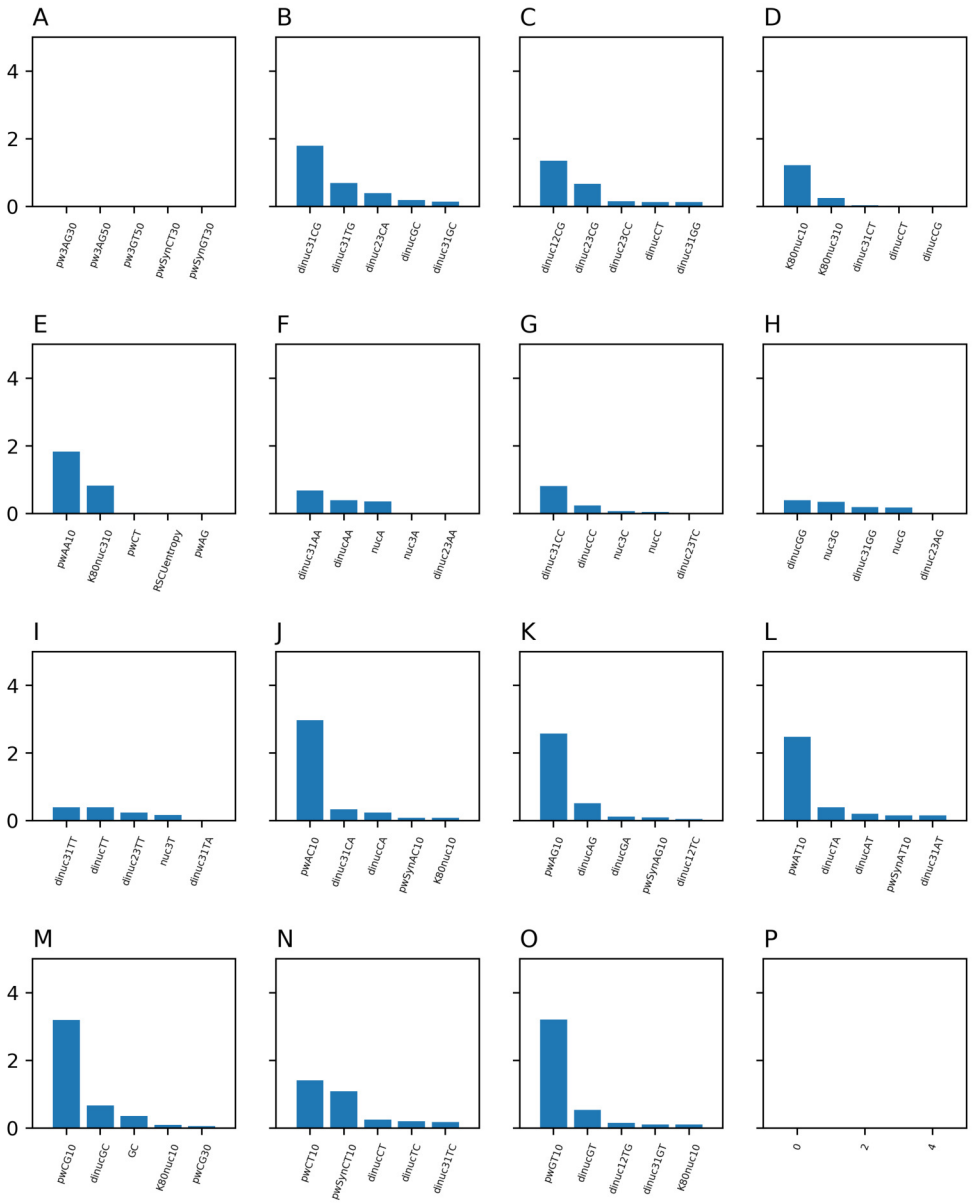

Figure 4.7. Valeur moyenne de l'importance des 5 plus importantes statistiques descriptives, dans l'ordre descendant de gauche à droite, pour chacun des paramètres du M[GTR+ts-CpG+tv-CpG]-S[NCatAA*], identifiés par les lettres A-M. Se référer aux annexes pour connaître les 189 statistiques descriptives utilisées dans le test. (A) Paramètre root ; (B) Paramètre λ_{tsCpG} ; (C) Paramètre λ_{TBL} ; (D) Paramètre λ_{omega_*} ; (E) Paramètre φ_A ; (F) Paramètre φ_C ; (G) Paramètre φ_G ; (H) Paramètre φ_T (I) Paramètre ϱ_{AC} ; (J) Paramètre ϱ_{AG} ; (K) Paramètre ϱ_{AT} ; (L) Paramètre ϱ_{CG} ; (M) Paramètre ϱ_{CT} ; (N) Paramètre ϱ_{GT}.

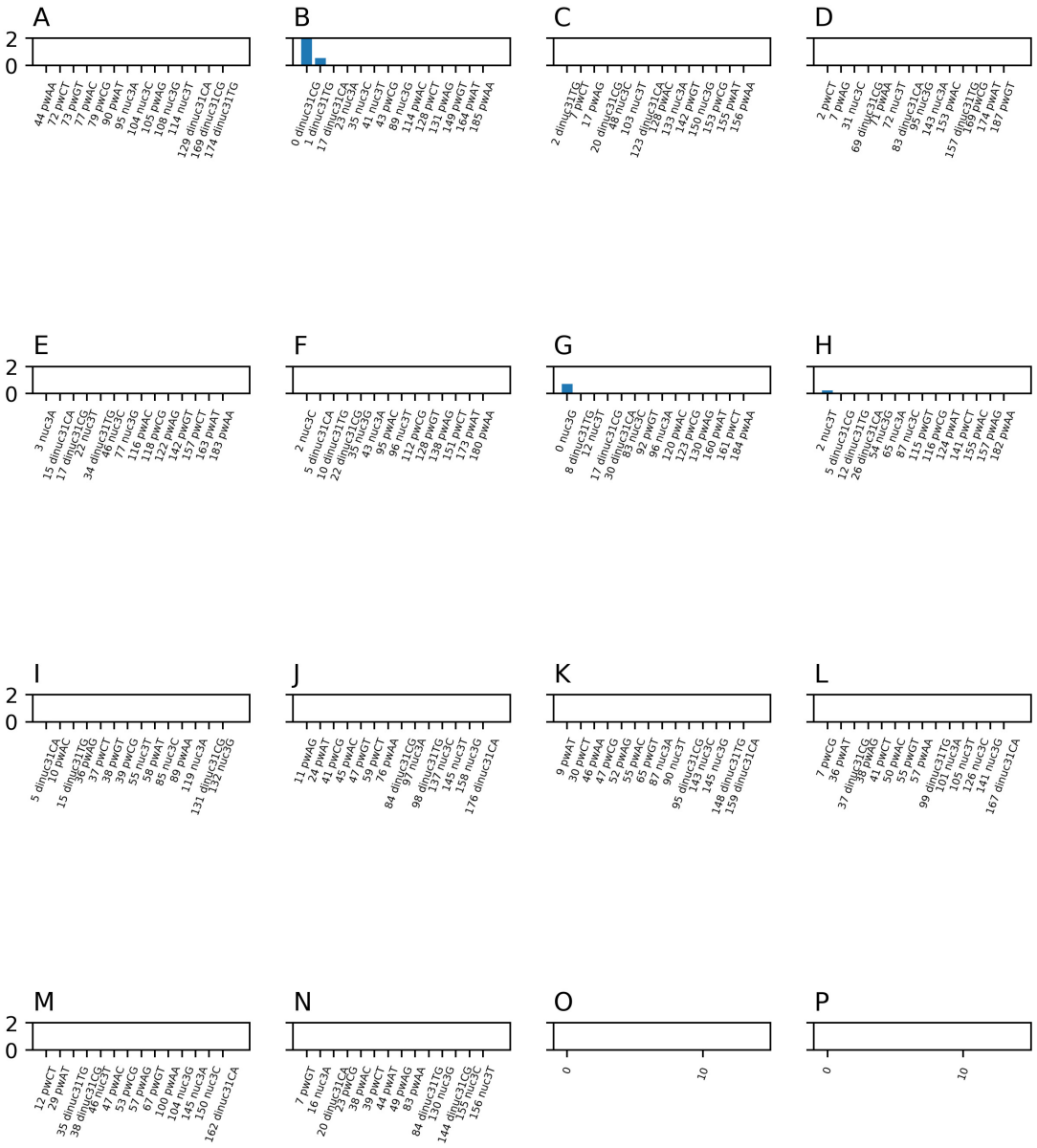

Figure 4.8. Valeur moyenne de l'importance des 14 statistiques decriptives ssCABC2018, dans l'ordre descendant de gauche à droite, pour chacun des paramètres du M[GTR+ts-CpG]-S[NCatAA*], identifiés par les lettres A-N. Le rang relatif des 14 statistiques descriptives de l'ensemble ssCABC2018 est inscrit devant le nom des statistiques descriptives sur l'axe des des ordonnées. Se référer aux annexes pour connaître les 189 statistiques descriptives utilisées dans le test. (A) Paramètre root ; (B) Paramètre λ_{tsCpG} ; (C) Paramètre λ_{TBL} ; (D) Paramètre λ_{omega_*} ; (E) Paramètre φ_A ; (F) Paramètre φ_C ; (G) Paramètre φ_G ; (H) Paramètre φ_T (I) Paramètre ϱ_{AC} ; (J) Paramètre ϱ_{AG} ; (K) Paramètre ϱ_{AT} ; (L) Paramètre ϱ_{CG} ; (M) Paramètre ϱ_{CT} ; (N) Paramètre ϱ_{GT}.

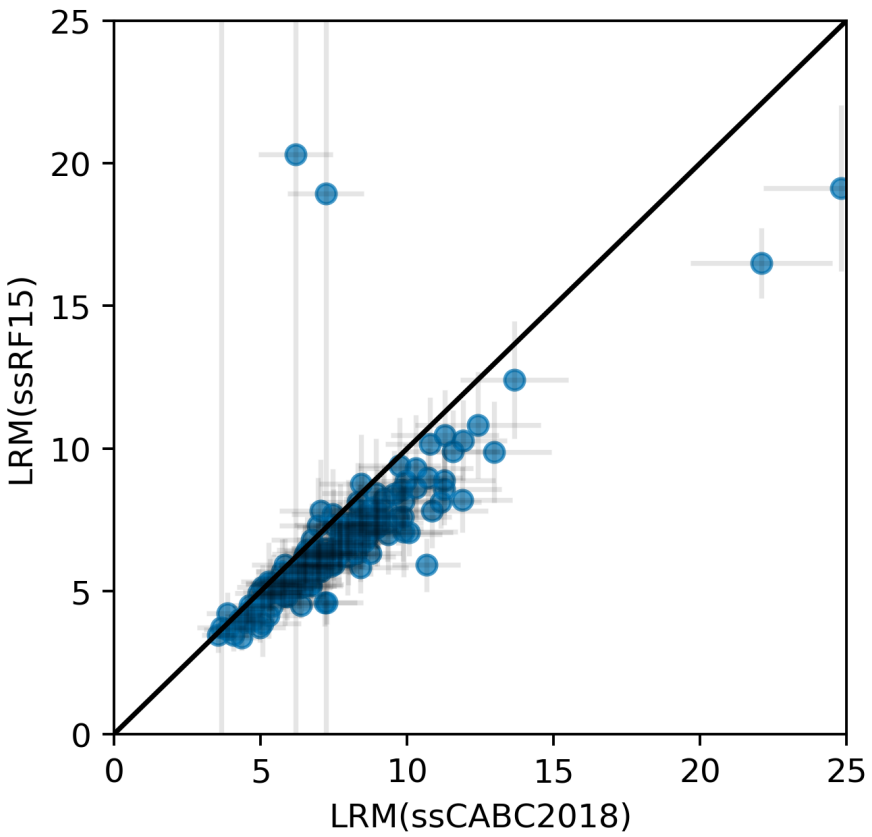

Figure 4.9. Comparaison des moyennes *a posteriori* de λ_{tsCpG} (bleu) du modèle M[GTR+ts-CpG] retrouvées avec CABC+LRM(ssCABC2018), pour l'axe des abscisses, et CABC+LRM(ssRF15) pour l'axe des ordonnées. Les valeurs moyennes de deux gènes ne sont pas présentes : *NR4A2* et *DYNC1I2* obtiennent des valeurs aberrantes pour le paramètre λ_{tsCpG} (y=110,16±1754,51 et y=1637,91±36959,38). Les barres grises correspondent aux barres d'erreur ($\pm\ \sigma$)

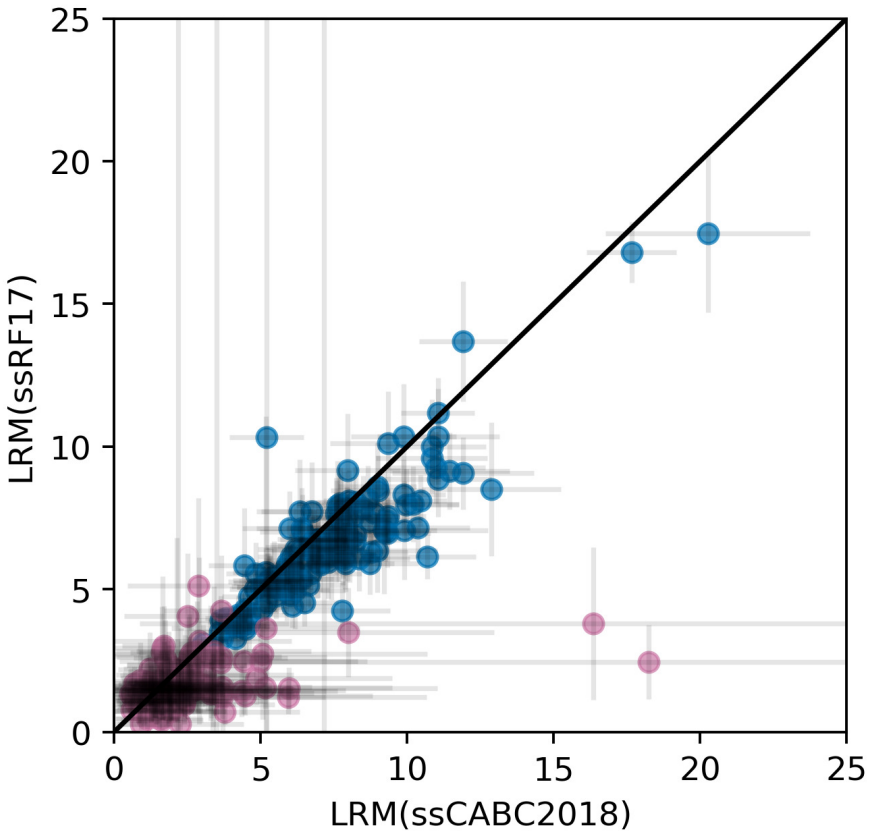

Figure 4.10. Comparaison des moyennes *a posteriori* des paramètres λ_{tsCpG} (bleu) et λ_{tvCpG} (rouge) retrouvés avec CABC(M[GTR+ts-CpG+tv-CpG])+LRM(ssCABC2018), pour l'axe des abscisses, et CABC(M[GTR+ts-CpG+tv-CpG])+LRM(ssRF17) pour l'axe des ordonnées. Les barres grises correspondent aux barres d'erreur ($\pm\ \sigma$).

4.3.2. Correction de la distribution *a posteriori* avec l'algorithme de régression par forêts aléatoires

Jusqu'à présent, nous avons utilisé un modèle de régression linéaire multiple pour faire la correction des distributions a posteriori obtenues avec la méthodologie CABC. Rappelons que nous notons cette procédure de la manière suivante CABC+LRM(ssCABC2018) lorsque l'ensemble de statistiques descriptives de l'article original est utilisé, ssCABC2018. Mais nous avons aussi développé l'utilisation d'un modèle de régression de type forêts aléatoires (RFRM) pour faire la correction des distributions a posteriori obtenues avec l'algorithme CABC, ce qui donne CABC+RFRM(ssCABC2018) par exemple. Pour ce faire nous avons dû

implémenter la correction en utilisant RFRM disponibles dans la suite scikit-learn [Pedregosa et al., 2011].

Algorithm 4 Algorithme CABC+RFRM

Sélectionner les SS et θ avec CABC à partir de la table de référence et des SS_{obs}

Conditionner RFRM à partir de $\theta \sim SS$

Prédire $\theta_{pred-obs}$ avec RFRM à partir de SS_{obs} de l'alignement

Prédire $\theta_{pred-ss}$ avec RFRM à partir de SS

Calculer différence (err) entre θ et $\theta_{pred-ss}$

Calculer la correction $\theta_{pred-obs} + err$

Il faut tout d'abord (1) sélectionner les mille voisins les plus proches sur la base du calcul d'une distance euclidienne entre statistiques descriptives extraites à partir des alignements de gènes, SS_{obs}, et des statistiques descriptives présentes dans la table de référence. La sélection comprend donc les SS retenues de la table de référence, mais aussi les valeurs des paramètres correspondantes, θ. Puis (2) nous voulons conditionner un RFRM avec les valeurs de paramètres, θ, pour variables réponses ($Y = \theta$), et les valeurs de statistiques descriptives, SS, pour variables explicatives ($X = SS$). À partir du RFRM conditionné ($Y \sim X$), (3) il est possible de prédire les valeurs de paramètres, $\theta_{pred-obs}$, en remplaçant SS par les statistiques descriptives de l'alignement de gènes étudié, SS_{obs}. À partir du même RFRM conditionné (4) il est possible de calculer l'erreur sur la prédiction, comme pour calculer un R^2, mais en utilisant le même ensemble de données utilisées pour conditionner RFRM, SS, nous prédisons les valeurs de paramètres de $\theta_{pred-ss}$. Pour ensuite (5) calculer la différence (err) entre les valeurs de paramètres, θ, et celles prédites plus tôt, $\theta_{pred-ss}$. Finalement (6), nous additionnons la différence de prédiction calculée en 5, qui peut varier entre $-inf$ et $+inf$, aux valeurs de paramètres prédites ($\theta_{pred-obs}$) en 3 à partir des statistiques descriptives calculées sur l'alignement de gènes, SS_{obs}. Nous obtenons ainsi une correction qui nous l'espérons nous rapproche de la vraie distribution a posteriori sans trop de biais.

Nous présentons ici quelques résultats préliminaires de l'impact de la correction RFRM lorsque nous analysons le jeu Eutheria39 avec le modèle M[GTR+ts-CpG+tv-CpG]. Dans les quelques résultats préliminaires ici présentés nous comparons les moyennes a posteriori des paramètres λ_{tsCpG} et λ_{tvCpG} retrouvés (1) avec la correction RFRM aux résultats obtenus avec LRM pour l'ensemble de statistiques descriptives de base ssCABC2018 (figure 4.11),

puis nous comparons les moyennes a posteriori retrouvées avec la correction RFRM pour les ensembles ssCABC2018 et ssRF17 (figure 4.12) et finalement pour les ensembles ssCABC2018 et ss73 (figure 4.13). La première chose que nous notons est que les valeurs moyennes inférées restent dans l'espace de la distribution *a priori*, de 0.1 à 10, lorsque l'on utilise RFRM, alors qu'avec LRM, les valeurs sortent de l'espace de la distribution *a priori* (figure 4.11). Effectivement, il n'est pas possible extrapoler avec RFRM, alors LRM permet l'extrapolation. La deuxième chose que nous notons est que les valeurs des écarts-types calculées à partir des distributions a posteriori des paramètres sont beaucoup moins élevées pour RFRM que pour LRM, et cela même pour le paramètre λ_{tvCpG} (figures 4.11-4.13). D'autre part, le choix des statistiques descriptives a un impact sur les valeurs de paramètres λ_{tsCpG} et λ_{tvCpG} inférés. Par exemple, les moyennes a posteriori du paramètre λ_{tsCpG} sont moins grandes lorsque l'ensemble ssCABC2018 est utilisé versus ssRF17 ou encore ss73. Le même phénomène est observé pour le paramètre λ_{tvCpG}, des valeurs d'hypermutabilité des transversions plus grandes sont obtenues avec les ensembles ssRF17 et ss73 (figure 4.13). Il sera intéressant de valider cette approche en étudiant l'erreur sur les inférences et les propriétés de couverture.

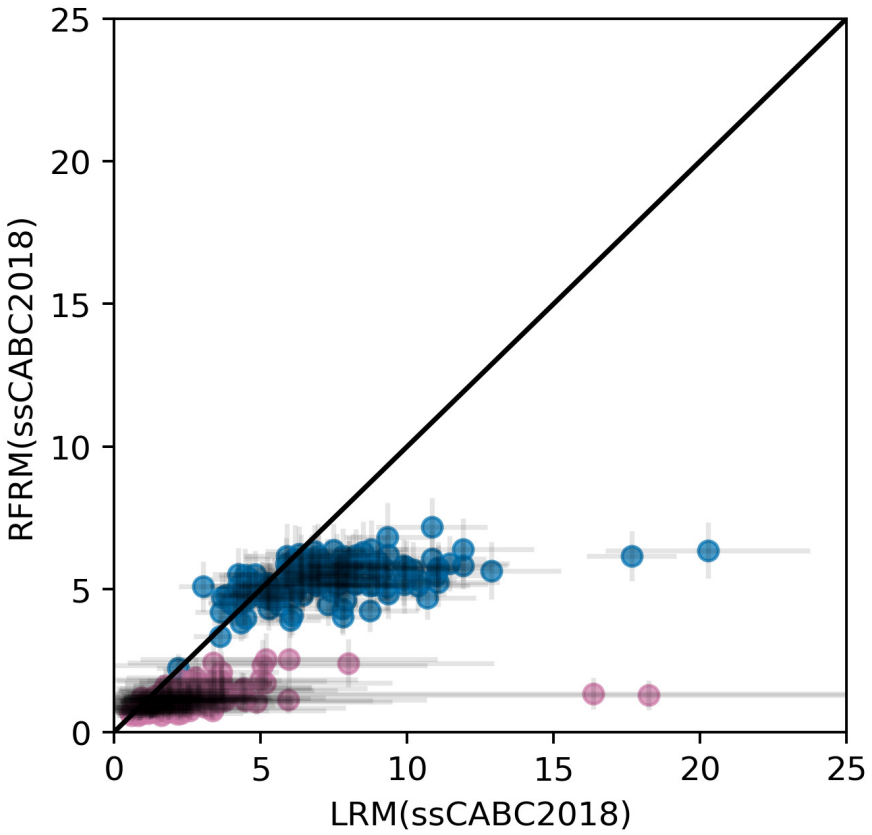

Figure 4.11. Comparaison des moyennes a posteriori des paramètres λ_{tsCpG} (bleu) et λ_{tvCpG} (rouge) retrouvées avec CABC(M[GTR+ts-CpG+tv-CpG])+LRM(ssCABC2018), pour l'axe des abscisses, et CABC(M[GTR+ts-CpG+tv-CpG])+RFRM(ssCABC2018) pour l'axe des ordonnées. Les barres grises correspondent aux barres d'erreur ($\pm\ \sigma$).

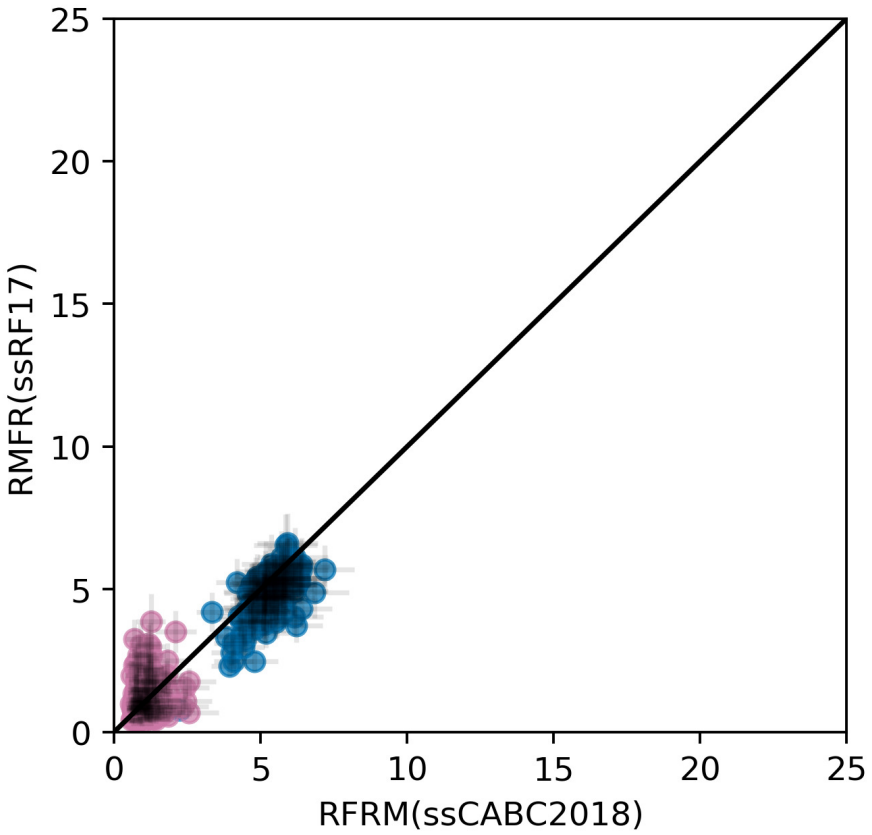

Figure 4.12. Comparaison des moyennes a posteriori des paramètres λ_{tsCpG} (bleu) et λ_{tvCpG} (rouge) retrouvées avec CABC(M[GTR+ts-CpG+tv-CpG])+RFRM(ssCABC2018), pour l'axe des abscisses, et CABC(M[GTR+ts-CpG+tv-CpG])+RFRM(ssRF17) pour l'axe des ordonnées. Les barres grises correspondent aux barres d'erreur ($\pm\ \sigma$).

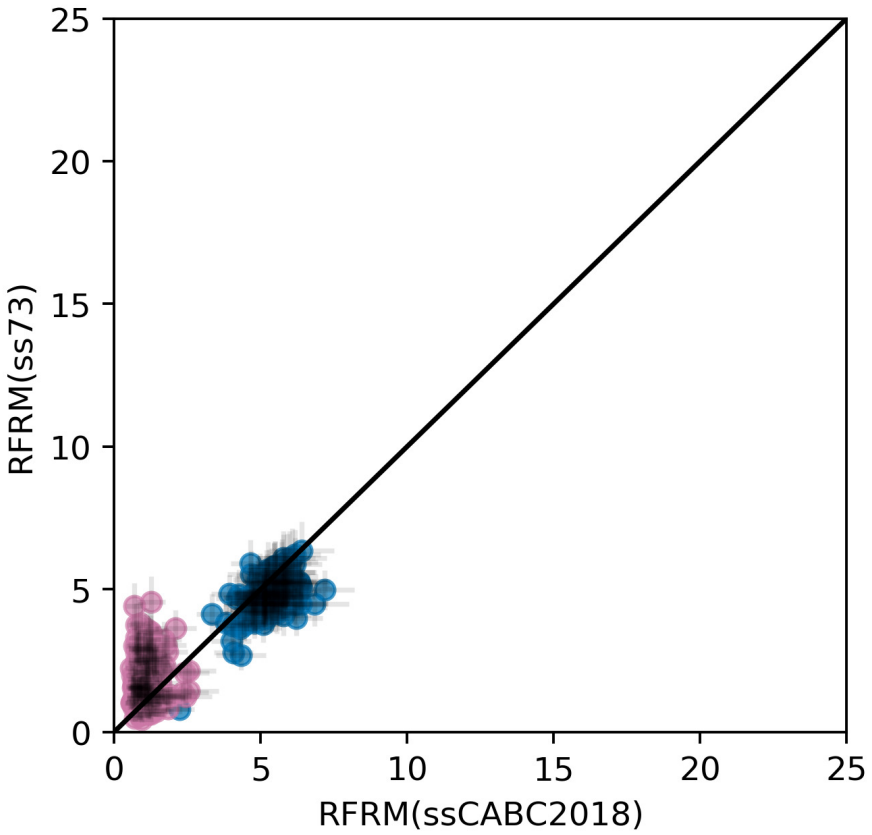

Figure 4.13. Comparaison des moyennes a posteriori des paramètres λ_{tsCpG} (bleu) et λ_{tvCpG} (rouge) retrouvées avec CABC(M[GTR+ts-CpG+tv-CpG])+RFRM(ssCABC2018), pour l'axe des abscisses, et CABC(M[GTR+ts-CpG+tv-CpG])+RFRM(ss73) pour l'axe des ordonnées. Les barres grises correspondent aux barres d'erreur ($\pm\ \sigma$).

4.3.3. Validation

Le système de validation consiste à évaluer la capacité des modèles CABC à correctement inférer les valeurs des paramètres utilisés pour générer les alignements simulés, en pratique les statistiques descriptives utilisées pour les décrire. Plus particulièrement, nous nous intéressons à quantifier (1) l'erreur au carré relative moyenne (Relative Mean Square Error : RMSE) et (2) la couverture à 95%. Rappelons que le RMSE est utilisé dans le contexte de l'ABC par [Beaumont et al., 2002] et dans l'article original de l'approche CABC [Laurin-Lemay et al., 2018c] pour évaluer la performance des modèles. L'étude des propriétés de

couverture consiste à évaluer la fréquence à laquelle la vraie valeur est retrouvée dans l'intervalle de crédibilité à 95% de nos inférences bayésiennes à travers un ensemble de réplicats expérimentaux.

Contrairement au système de validation utilisé dans l'article original du CABC ([Laurin-Lemay et al., 2018c], nous avons réduit considérablement le temps calcul en évitant d'inférer les paramètres faiblement corrélés (longueurs de branches et profils de préférences site-spécifique en acides aminés) aux paramètres d'intérêt. Nous avons montré que ce raccourci méthodologique avait peu d'impact sur l'erreur détectée dans le contexte de l'inférence de l'hypermutabilité des transitions en contexte CpG [Laurin-Lemay et al., 2018c]. La même table de référence peut donc servir à valider autant de simulations que nécessaire. Est-ce que ce raccourci est valable lorsque l'on s'intéresse à hypermutabilité des transversions en contexte CpG ou TpA ? Est-ce que les paramètres pour les transversions sont plus fortement corrélés avec les profils de préférence en acides aminés (sélection), ceux qu'on supposait faiblement corrélés ?

Nous pouvons déjà explorer ce questionnement en calculant le rapport du taux de substitution non-synonyme aux taux de mutation non-synonyme (dN) pour les transitions (TsdN) et les transversions (TvdN) de manière à évaluer l'impact de la sélection sur les processus mutationnels (l'écart au modèle de mutation). Les TvdN obtenus pour chacun des 137 gènes de Mammifères (Eutheria39) avec le modèle de référence, M[GTR]-S[NCatAA*], sont tous plus petits que 1, donc sous l'influence de la sélection négative, alors que les TsdN sont aussi en moyenne plus petite que 1, mais majoritairement plus grande que leur TvdN correspondant (figure 4.14). À noter que le gène TMC1 obtient la plus grande valeur de TsdN, 5,26 (figure 4.14). Ces limitations devront être étudiées ultérieurement, mais il semble que oui, la sélection affecte plus les transversions.

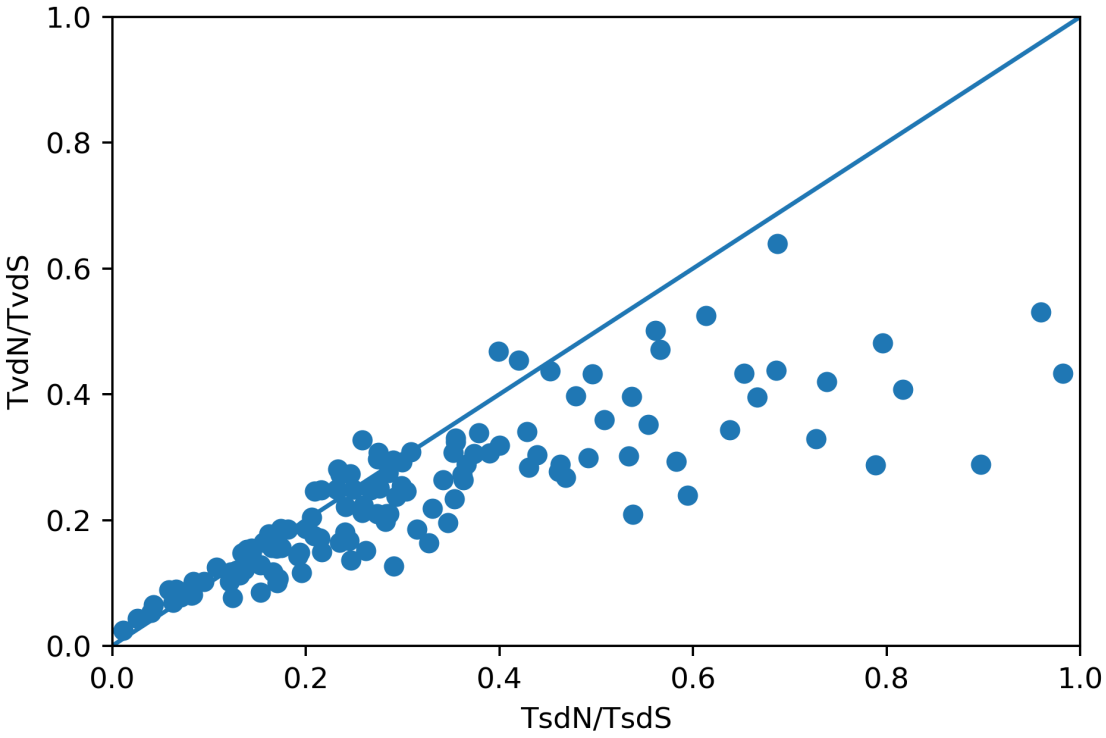

Figure 4.14. Comparaison du taux de substitution non-synonyme aux taux de mutation non-synonyme spécifique aux transitions (axe des abscisses) et spécifique aux transversions (axe des ordonnées) obtenus sous le modèle de référence, M[GTR]-S[NCatAA*], pour les 137 gènes de Mammifères (Eutheria39). TsdS et TvdS valent 1, puisqu'il n'y a pas de sélection sur les mutations synonymes. Une diagonale est tracée pour évaluer l'écart entre TsdN et TvdN. Cinq gènes obtiennent des valeurs plus grandes que 1 pour TsdN/TsdS, ils sont donc absents du graphique présenté.

Pour une seule paramétrisation du CABC avec un seul paramètre d'intérêt (e.g., λ_{tsCpG}), le système de validation compte 25000 analyses. Les 25000 analyses se déclinent par l'utilisation de 250 conditions expérimentales : 50 gènes $\times$ 5 conditions d'hypermutabilité contextuelles (i.e., 0.5, 1, 2, 4 et 8) $\times$ 100 réplicats (simulations). Pour un CABC avec deux paramètres d'intérêt, il faudrait croiser les conditions d'hypermutabilité contextuelles, ce qui fait $50 \times (5 \times 5) \times 100$, donc 125 000 analyses, seulement 25 000 analyses ont étées réalisées jusqu'à présent.

L'erreur retrouvée (RMSE) dans ce travail est très similaire à l'erreur obtenue dans l'article CABC original pour les mêmes conditions [Laurin-Lemay et al., 2018c], ce qui confirme la robustesse de l'approche (tableau 4.12). L'erreur sur le paramètre λ_{tsCpG} diminue à mesure que la quantité de signal évolutif augmente (tableau 4.12), que ce soit pour le paramètre λ_{tsCpG} du modèle M[GTR+ts-CpG] (comme dans [Laurin-Lemay et al., 2018c]) ou du modèle à deux paramètres d'hypermutabilité du contexte CpG, M[GTR+ts-CpG+tv-CpG]. Par contre, l'erreur est toujours plus importante sur le paramètre λ_{tsCpG} appartenant au modèle incorporant un plus grand nombre de paramètres libres (tableau 4.12 : M[GTR+ts-CpG] versus M[GTR+ts-CpG+tv-CpG]), ce qui est attendu étant donné la plus grande combinatoire de l'espace à échantillonner. L'erreur sur le paramètre λ_{tvCpG} augmente à mesure que le niveau d'hypermutabilité des transitions en contexte CpG utilisé pour générer les alignements de la validation augmente. Puisque la taille de l'arbre est fixée, les taux de substitutions ne font que modifier les proportions des différents types de substitutions qui se réalisent. Quand $\lambda_{tsCpG} = 1$, il y a 10% de transitions et 5% de transversions en contexte CpG, alors que quand $\lambda_{tsCpG} = 4$, il y a seulement 2% de transversions et 15% de transitions (tableau 4.13). Il est donc normal que l'erreur sur λ_{tvCpG} augmente avec la valeur du paramètre $\lambda_{tsCpG}1$, car il y a moins d'information pour inférer ce paramètre. De surcroît le choix des statistiques descriptives n'est probablement pas optimal pour identifier le signal en lien avec l'hypermutabilité des transversions en contexte CpG lorsque nous utilisons ssCABC2018. D'autre part, la fréquence moyenne à laquelle la valeur des paramètres d'intérêt est retrouvée pour chacune des conditions d'hypermutabilité contextuelles correspond à l'intervalle de crédibilité avec lequel la couverture est calculée, soit 95% (tableau 4.14), ce qui suggère qu'il n'y a pas de biais dans le système d'inférence utilisé, et cela, que ce soit pour les paramètres λ_{tsCpG} et λ_{tvCpG}. Par contre, quelles sont la sensibilité et la spécificité de notre capacité à détecter les hypermutabilité des transitions et transversion en contexte CpG ?

Maintenant que nous avons vérifié que la validation réalisée sans recalculer les valeurs des paramètres faiblement corrélées donnait des résultats en accord avec [Laurin-Lemay et al., 2018c], nous allons valider l'utilisation de l'algorithme de forêts aléatoires pour guider notre choix de statistiques descriptives pour confectionner les ensembles ssRF15 et ssRF17 (tableau 4.11). En effet, quel est l'impact d'utiliser ces statistiques descriptives sur l'erreur et la couverture des paramètres d'intérêt ?

Globalement lorsque nous utilisons la correction LRM avec ssRF15 ou ssRF17, l'erreur moyenne obtenue, sans l'erreur sur le paramètre root, est plus petite qu'avec les statistiques descriptives que nous avions choisies dans l'article original (ssCABC2018) : 0,429±0,142 versus 0,65±0,125 (tableau 4.15) pour le modèle M[GTR+ts-CpG] et 0,857±0,43 versus 2,43±2,943 (tableau 4.16) pour le modèle M[GTR+ts-CpG+tv-CpG]. Autrement, l'erreur spécifique au paramètre λ_{tsCpG} du modèle M[GTR+ts-CpG] est aussi plus petite pour le nouvel ensemble de statistiques descriptives : soit 0,55±0,56 pour l'ensemble ssRF15 et 0,01±0,066 pour l'ensemble ssCABC2018 (tableau 4.15). L'erreur spécifique au paramètre λ_{tsCpG} est beaucoup plus petite avec l'ensemble ssRF17 dans le cadre du modèle M[GTR+ts-CpG+tv-CpG] : soit 0,57±0,059 versus 0,092±0,77 (tableau 4.16). La différence dans le niveau d'erreur est encore plus grande pour le paramètre λ_{tvCpG} pour les deux ensembles respectivement, ssRF17 versus ssCABC2018 : soit 0,044±0,035 versus 1,691±2,96 (Tableau 18). Par contre, le plus grand gain se fait au niveau du paramètre λ_{TBL} qui obtient maintenant un RMSE de 0,017±0,015 versus 0,285±0,452 et 0,021±0.017 versus 0,33±0,47 pour les modèles M[GTR+ts-CpG] et M[GTR+ts-CpG+tv-CpG] respectivement.

L'algorithme de forêts aléatoires identifie SS_{C1pG2} et SS_{C2pG3} comme des statistiques descriptives importantes pour le paramètre λ_{tvCpG}, d'où probablement la réduction de l'erreur sur ce dernier (tableau 4.16). Par contre, les statistiques descriptives identifiées comme importantes par l'algorithme de forêts aléatoires ne le sont pas nécessairement pour LRM. Une approche similaire, basée sur la différence des R^2, pourrait être employée pour calculer l'importance des statistiques descriptives dans le contexte de la correction LRM et ainsi permettre d'améliorer plus efficacement cette méthode d'ABC+LRM (fisher2018). D'autre part, la couverture à 95% est bonne pour les paramètres d'intérêt lorsque nous utilisons la correction LRM (tableau 4.17 et 4.18).

Lorsque nous utilisons l'ensemble de statistiques descriptives ssCABC2018 l'erreur globale calculée est plus grande avec la correction RFRM qu'avec la correction LRM pour le modèle M[GTR+ts-CpG] : soit 1,148±0,478 versus 0,65±0,125 (tableau 4.15). Cela n'est pas le cas lorsque le paramètre λ_{tvCpG} est introduit avec l'utilisation du modèle M[GTR+ts-CpG+tv-CpG] : l'erreur globale est de 1,955±1,103 avec la correction RFRM versus 2,430±2,943 avec la correction LRM (tableau 4.16). D'autre part, les erreurs globales calculées avec la

correction RFRM sont les plus grandes dans le contexte du modèle M[GTR+ts-CpG], que ce soit pour les ensembles ssCABC2018, ssRF15 ou ss73 (tableau 4.15).

Étonnement, l'erreur diminue pour les paramètres d'intérêt lorsque nous utilisons la correction RFRM avec l'ensemble ssCABC2018 : λ_{tsCpG} passe de 0,061±0,066 à 0,053±0,071 (tableau 4.15) lorsque employé dans le contexte du modèle M[GTR+ts-CpG] et 0,092±0,077 à 0,059±0,078 (tableau 4.16) lorsqu'employé dans le contexte du modèle M[GTR+ts-CpG+tv-CpG]. Par contre, la correction RFRM obtient la plus petite erreur, lorsqu'appliquée sur ss73 pour le paramètre d'intérêt λ_{tsCpG} du modèle M[GTR+ts-CpG], tableau 4.15. De manière surprenante, l'erreur sur λ_{tsCpG} et λ_{tvCpG} est plus grande lorsque les statistiques descriptives choisies avec l'algorithme de forêts aléatoires sont utilisées avec la correction RFRM : 0,096±0,116 (ssRF15) versus 0,053±0,071 (ssCABC2018) et 0,095±0,096 (ssRF17) versus 0,059±0,078 (ssCABC2018) pour le paramètre λ_{tsCpG} des modèles M[GTR+ts-CpG] et M[GTR-ts-CpG+tv-CpG], tableau tableau 4.15 et 4.16 respectivement. De la même manière, le paramètre λ_{tvCpG} du modèle M[GTR-ts-CpG+tv-CpG] obtient une erreur plus grande lorsque la correction se fait sur l'ensemble de statistiques descriptives choisies avec l'algorithme de forêts aléatoires : soit 1,025±2,045 (ssRF17) versus 0,791±0,916 (ssCABC2018) respectivement (tableau 4.15). Par contre, un biais important est généré avec la correction RFRM, au lieu d'avoir une couverture de 95%, les modèles obtiennent une couverture moyenne de 55% pour le paramètre λ_{tsCpG} des modèles M[GTR+ts-CpG] et M[GTR-ts-CpG+tv-CpG] (tableau 4.17 et 4.18).

Nous avons aussi fait une correction de la distribution a posteriori avec l'algorithme ABC+RFRM en utilisant un ensemble comprenant 73 statistiques descriptives (ss73). Cet ensemble de statistiques descriptives comprend les fréquences relatives des dinucléotides en positions 1-2, 2-3 et 3-1 des codons ainsi que les $SS_{N<>N}$, $SS_{(N<>N)10}$, $SS_{k80nuc10}$, $SS_{k80nuc310}$ (calculé pour la position 3 des codons seulement), SS_{k80nuc}, SS_{NS}, SS_{NS10} et les SS_N et SS_{N3}. Utiliser autant de statistiques descriptives nous expose au fléau de dimensionnalité lors de la recherche des plus proches voisins par le calcul de la distance euclidienne : des distorsions potentielles dans l'approximation de la distribution a posteriori [Prangle, 2018]. Néanmoins, ABC+RFRM couplé avec ss73 est l'approche qui génère l'erreur la plus petite sur les paramètres d'intérêt (tableau 4.15 et 4.16 : 0,41±0,061 et 0,048±0,059 pour le paramètre λ_{tsCpG} des modèles M[GTR+ts-CpG] et M[GTR-ts-CpG+tv-CpG] respectivement et

0,365±0,618 pour le paramètre λ_{tvCpG} du modèle M[GTR-ts-CpG+tv-CpG]). Cela confirme que l'approche des forêts aléatoires est robuste au fléau de la dimensionnalité. Par contre, la couverture est toujours aussi mauvaise 55-60% pour le paramètre λ_{tsCpG} des modèles M[GTR+ts-CpG] et M[GTR-ts-CpG+tv-CpG] (tableaux 4.17 et 4.18). À l'inverse, la couverture est relativement bonne pour le paramètre λ_{tvCpG} du modèle M[GTR-ts-CpG+tv-CpG] (tableau 4.18).

L'utilisation de l'algorithme de forêts aléatoires s'est montré utile pour faire le choix des statistiques descriptives à utiliser avec la correction LRM, c'est cette approche du CABC qui génère la plus petite erreur globale : 0,429±0,142 (tableau 4.15) et 0,857±0,43 (tableau 4.16). Par contre la correction RFRM n'est pas efficace à réduire l'erreur, une piste de recherche consisterait à ouvrir, la distribution *a priori*, puisque RFRM n'est pas capable d'extrapolation. Il est aussi possible que le jeu de données, la table des mille plus proches voisins, ne soit pas d'assez grande taille pour les besoins de l'algorithme de forêts aléatoires. Deux pistes de solution s'offrent à nous : (1) une première piste de solution consiste à conditionner le modèle de régression de type forêts aléatoire sur la table de référence (10^5 entrées) et à ensuite utiliser que les mille plus proches voisins pour réaliser les prédictions nécessaires au calcul de la correction et une (2) une deuxième piste serait d'agrandir la table de référence à 10^6 entrées.

Table 4.12. Erreur relative au carré moyenne (RMSE) calculée pour les paramètres d'intérêt, λ_{tsCpG} et λ_{tvCpG}, avec l'ensemble de statistiques descriptives ssCABC2018 et la correction LRM. Le RMSE présenté dans chacune des lignes est la moyenne des RMSE calculés pour chacun des gènes, soit pour 100 réplicats par gène.

λ_{tsCpG}=	λ_{tvCpG}=	modèles	λ_{tsCpG} moy.	λ_{tsCpG} σ	λ_{tvCpG} moy.	λ_{tvCpG} σ
0,5	1	M[GTR+ts-CpG]	0,132	0,125	N/A	N/A
0,5	1	M[GTR+ts-CpG+tv-CpG]	0,219	0,233	0,590	0,746
1	1	M[GTR+ts-CpG]	0,101	0,100	N/A	N/A
1	1	M[GTR+ts-CpG+tv-CpG]	0,151	0,127	0,788	1,380
2	1	M[GTR+ts-CpG]	0,077	0,072	N/A	N/A
2	1	M[GTR+ts-CpG+tv-CpG]	0,111	0,093	1,225	2,653
4	1	M[GTR+ts-CpG]	0,061	0,069	N/A	N/A
4	1	M[GTR+ts-CpG+tv-CpG]	0,086	0,076	2,087	5,731
8	1	M[GTR+ts-CpG]	0,047	0,048	N/A	N/A
8	1	M[GTR+ts-CpG+tv-CpG]	0,065	0,065	4,061	8,166

Table 4.13. Comparaison de la proportion des types de substitutions en contexte CpG (pour un sous-ensemble des 137 gènes du jeu Eutheria39) obtenues en simulant avec $\lambda_{tsCpG}1$ (i.e. le modèle de référence), $\lambda_{tsCpG} = 4$ et $\lambda_{tvCpG} = 4$ (M[GTR+ts-CpG+tv-CpG]) et $\lambda_{tstvCpG} = 4$ (M[GTR+tstv-CpG]) et des valeurs de paramètres du modèle de référence.

	M[GTR]	M[GTR+ts-CpG]	M[GTR+ts-CpG+tv-CpG]	M[GTR+tstv-CpG]
CpG	15±2	18±3	17±3	17±3
CpG (ts)	10±1	15±2	11±2	11±2
CpG (tv)	5±1	2±1	6±1	6±1
N substitutions	5409±1769	6206±2080	6258±2093	6260±2098

Table 4.14. Couverture à 95% calculée pour les paramètres d'intérêt, λ_{tsCpG} et λ_{tvCpG}, avec l'ensemble des statistiques descriptives ssCABC2018 et la correction LRM. La couverture présentée dans chacune des lignes est la moyenne des fréquences à laquelle la vraie valeur est retrouvée dans un ensemble de 100 réplicats des 50 gènes de validation utilisés.

ts-CpG=	tv-CpG=	modèles	λ_{tsCpG} moy.	λ_{tsCpG} σ	λ_{tvCpG} moy.	λ_{tvCpG} σ
0,5	1	M[GTR+ts-CpG]	0,968	0,175	N/A	N/A
0,5	1	M[GTR+ts-CpG+tv-CpG]	0,986	0,117	0,97	0,17
1	1	M[GTR+ts-CpG]	0,961	0,194	N/A	N/A
1	1	M[GTR+ts-CpG+tv-CpG]	0,984	0,127	0,973	0,163
2	1	M[GTR+ts-CpG]	0,96	0,195	N/A	N/A
2	1	M[GTR+ts-CpG+tv-CpG]	0,976	0,152	0,982	0,134
4	1	M[GTR+ts-CpG]	0,947	0,224	N/A	N/A
4	1	M[GTR+ts-CpG+tv-CpG]	0,973	0,162	0,993	0,085
8	1	M[GTR+ts-CpG]	0,934	0,249	N/A	N/A
8	1	M[GTR+ts-CpG+tv-CpG]	0,954	0,21	0,998	0,04

Table 4.15. Comparaison de l'erreur relative au carré moyenne (RMSE) obtenue sous le modèle M[GTR+ts-CpG] avec l'approche CABC pour l'ensemble des paramètres (sans root) et spécifiquement pour le paramètre λ_{tsCpG} lorsque les alignements étudiés sont simulés avec $\lambda_{tsCpG} = 4$ et $\lambda_{tvCpG} = 1$. L'étude ici est faite sur un sous-ensemble de 10 gènes parmi les 50 gènes sélectionnés pour la validation.

			globale		ts-CpG	
modèles	stat, descriptives	cor.,	moy.,	σ	moy.,	σ
M[GTR+ts-CpG]	ssCABC2018	LRM	0,650	0,125	0,061	0,066
M[GTR+ts-CpG]	ssRF15	LRM	0,429	0,142	0,055	0,056
M[GTR+ts-CpG]	ssCABC2018	RFRM	1,148	0,478	0,053	0,071
M[GTR+ts-CpG]	ssRF15	RFRM	0,904	0,328	0,096	0,116
M[GTR+ts-CpG]	ss73	RFRM	1,604	0,971	0,041	0,061

Table 4.16. Comparaison de l'erreur relative au carré moyenne (RMSE) obtenue sous le modèle M[GTR+ts-CpG+tv-CpG] avec l'approche CABC pour l'ensemble des paramètres (sans root) et pour les paramètres λ_{tsCpG} et λ_{tvCpG} lorsque les alignements étudiés sont simulés avec $\lambda_{tsCpG} = 4$ et $\lambda_{tvCpG} = 1$. L'étude ici est faite sur un sous-ensemble de 10 gènes parmi les 50 gènes sélectionnés pour la validation.

			globale		λ_{tsCpG}		λ_{tvCpG}	
modèles	stat. descriptives	cor.	moy.	σ	moy.	σ	moy.	σ
M[GTR+ts-CpG+tv-CpG]	ssCABC2018	LRM	2,430	2,943	0,092	0,077	1,691	2,960
M[GTR+ts-CpG+tv-CpG]	ssRF17	LRM	0,857	0,430	0,057	0,059	0,044	0,035
M[GTR+ts-CpG+tv-CpG]	ssCABC2018	RFRM	1,955	1,103	0,059	0,078	0,791	0,916
M[GTR+ts-CpG+tv-CpG]	ssRF17	RFRM	1,925	2,086	0,095	0,096	1,025	2,045
M[GTR+ts-CpG+tv-CpG]	ss73	RFRM	1,701	7,920	0,048	0,059	0,365	0,618

Table 4.17. Comparaison de la couverture à 95% obtenue sous le modèle M[GTR+ts-CpG] avec l'approche CABC pour le paramètre ts-CpG lorsque les alignements sont simulés avec $\lambda_{tsCpG} = 4$ et $\lambda_{tvCpG} = 1$. L'étude ici est faite sur un sous-ensemble de 10 gènes parmi les 50 gènes sélectionnés pour la validation.

			λ_{tsCpG}	
modèles	stat. descriptives	cor.	moy.	σ
M[GTR+ts-CpG]	ssCABC2018	LRM	0,944	0,230
M[GTR+ts-CpG]	ssRF15	LRM	0,949	0,221
M[GTR+ts-CpG]	ssCABC2018	RFRM	0,554	0,497
M[GTR+ts-CpG]	ssRF15	RFRM	0,520	0,500
M[GTR+ts-CpG]	ss73	RFRM	0,586	0,493

Table 4.18. Comparaison de la couverture à 95% obtenue sous le modèle M[GTR+ts-CpG+tv-CpG] avec l'approche CABC pour les paramètres λ_{tsCpG} et λ_{tvCpG} lorsque les alignements sont simulés avec $\lambda_{tsCpG} = 4$ et $\lambda_{tvCpG} = 1$. L'étude ici est faite sur un sous-ensemble de 10 gènes parmi les 50 gènes sélectionnés pour la validation.

			λ_{tsCpG}		λ_{tvCpG}	
modèles	stat. descriptives	cor.	moy.	σ	moy.	σ
M[GTR+ts-CpG+tv-CpG]	ssCABC2018	LRM	0,970	0,171	0,997	0,055
M[GTR+ts-CpG+tv-CpG]	ssRF17	LRM	0,975	0,156	0,932	0,252
M[GTR+ts-CpG+tv-CpG]	ssCABC2018	RFRM	0,558	0,497	0,896	0,305
M[GTR+ts-CpG+tv-CpG]	ssRF17	RFRM	0,591	0,492	0,814	0,389
M[GTR+ts-CpG+tv-CpG]	ss73	RFRM	0,558	0,497	0,947	0,223

4.3.4. Comparaison de modèles

À l'aide de la méthode développée par [Raynal et al., 2017], nous avons pu observer que le modèle M[GTR] et M[HKY] ne sont pas significativement préférés l'un de l'autre. Par contre, nous notons que la distribution des probabilités a posteriori de préférer M[HKY] est bimodale vers les extrémités, ce qui signifie que certains gènes possèdent une classification préférentielle pour le modèle M[GTR] et d'autres M[HKY]. En fait, la majorité des gènes préfèrent M[GTR] (125/137), mais sans être significatif. En fait, nous obtenons des résultats semblables avec des modèles beaucoup plus simples et une méthodologie plus standard à la phylogénie. Ainsi, en comparant GTR+F+G4 et HKY+F+G4 pour le jeu Eutheria39 (137 alignements) avec [Nguyen et al., 2014], les critères d'information d'Akaike, AIC corrigé et le critère d'information bayésien indiquent que le modèle le plus habile à expliquer les données est GTR+F+G4, dans 137/137, 137/137 et 121/137 respectivement, mais sans jamais être significatif. L'utilisation des bootstraps non paramétriques [Lubke et al., 2017] s'est montrée utile à l'évaluation de ces critères (AIC, AICc et BIC). De la même manière, GTR est préféré par rapport à HKY selon les mêmes critères (136/137 pour AIC, 136/137 pour CAIC et 132/137 pour BIC). Cela suggère donc que la difficulté à préférer un modèle plus qu'un autre de l'approche de Raynald et al 2015 est surtout dû au manque de signal dans les alignements.

Nous avons voulu dans un deuxième temps évaluer la sensibilité (taux de vrais classements) et la spécificité (1-taux de faux classements) du système de comparaison de modèles proposés par Raynald et al 2015 dans nos conditions de modélisation particulières. Nous avons généré 10^4 simulations à partir des deux modèles, M[HKY] et M[GTR], et cela pour les valeurs de paramètres obtenues sous le modèle de référence des 50 gènes utilisés dans la validation précédente, avec 100 réplicats par gène.

Les simulations réalisées sous M[HKY] sont identifiées comme appartenant à M[GTR] avec des probabilités a posteriori qui varient entre 0,46 et 0,97 lorsque les statistiques descriptives ssCACB2018 sont utilisées. Alors que les probabilités a posteriori de préférer le modèle M[GTR], lorsque les simulations sont réalisées sous ce même modèle varient entre 0,77 et 0,97. Il faut augmenter le seuil alpha de significativité à 10% pour avoir une sensibilité plus grande ou égale à 95%, par contre la spécificité (1- taux de faux positifs) diminue à 85% (figure 4.15).

Maintenant que l'on a étudié un cas simple, revenons à l'hypermutabilité en contexte CpG. Environ 70% des gènes du jeu Eutheria39 préfèrent M[GTR+ts-CpG] et M[GTR+tstv-CpG] à M[GTR] pour un seuil alpha de significativité de 5% lorsque l'ensemble ssCABC est utilisé. La préférence pour M[GTR+ts-CpG+tv-CpG] est encore plus grande, environ 86% des gènes possèdent une probabilité a posteriori de préférer ce modèle plutôt que le modèle M[GTR], pour un seuil alpha de significativité de 5%. Cela suggère que cette approche de comparaison de modèles [Raynal et al., 2017] manque de puissance, car la simple analyse de la distribution a posteriori de λ_{tsCpG} (ou de $\lambda_{tstvCpG}$) permet de montrer que le modèle incluant l'hypermutabilité en contexte CpG est toujours significativement préféré.

Finalement, nous avons abordé le cas le plus difficile, mais aussi le plus intéressant biologiquement, la comparaison des modèles M[GTR+ts-CpG], M[GTR+tstv-CpG] et M[GTR+ts-CpG+tv-CpG]. Nous avons voulu évaluer la capacité du système de comparaison à identifier les alignements générés sous le modèle M[GTR+ts-CpG], où le paramètre d'intérêt, λ_{tsCpG}, prend la valeur 4, alors que les choix de modèles possibles sont M[GTR+ts-CpG] versus M[GTR+tstv-CpG] et M[GTR+ts-CpG] versus M[GTR+ts-CpG+tv-CpG]. Cette fois, nous avons un seul ensemble de simulations faites à partir du modèle M[GTR+ts-CpG], soit 5000 simulations où le paramètre λ_{tsCpG} prend la valeur de 4, et où les paramètres de nuisance (M[GTR]) et faiblement corrélés au paramètre d'intérêt (S[NCatAA*]) prennent les valeurs

inférées sous le modèle de référence, et cela pour les 50 gènes utilisés dans la validation précédente, avec 100 réplicats par gène. Nous avons aussi testé deux ensembles de statistiques descriptives ssCABC2018 et ss73. Nos attentes sont que le modèle M[GTR+ts-CpG] devrait être préféré par rapport à M[GTR+tstv-CpG] qui doit faire un compromis entre le niveau d'hypermutabilité des transitions et des transversions en contexte CpG lorsque ces deux hypermutabilités sont très différentes, comme dans les conditions expérimentales que nous avons choisies ici (λ_{tsCpG}=4 et λ_{tvCpG}=1). Lorsque l'ensemble ssCABC2018 de statistiques descriptives est utilisé, la majorité des simulations (4951/5000) sont classées appartenir au modèle M[GTR+tstv-CpG] avec des probabilités a posteriori allant de 0,52 à 0,87, mais donc jamais significatives pour un seuil alpha de 5%. Seulement 49 classifications sont en faveur du modèle M[GTR+ts-CpG], avec une probabilité a posteriori de 0,56 à 0,66. Si l'ensemble de statistiques descriptives ss73 est utilisé, la majorité des simulations (3112/5000) sont classées dans le modèle M[GTR+tstv-CpG] avec des probabilités a posteriori allant de 0,47 à 0,64, donc à la baisse par rapport à l'ensemble ssCABC2018. Certaines probabilités a posteriori suggèrent un mauvais classement de la part de l'algorithme de forêts aléatoires (<0,5). Par contre, avec l'ensemble ss73 de statistiques descriptives, beaucoup plus de simulations (1888/5000) sont classées appartenir au modèle M[GTR+ts-CpG], mais avec des probabilités a posteriori très faibles, de 0,47 à 0,62. Lorsque les deux modèles comparés sont M[GTR+ts-CpG] et M[GTR+ts-CpG+tv-CpG], la majorité des simulations (4382/5000) sont classées appartenir à M[GTR+ts-CpG+tv-CpG] avec ssCABC2018, et une majorité plus forte (4616/5000) pour le même modèle avec ss73. Les probabilités a posteriori sont de 0,48 à 0,82 pour et 0,42 à 0,67 le modèle M[GTR+ts-CpG+tv-CpG] pour les ensembles ssCABC2018 et ss73 respectivement. Alors que les simulations classées appartenir à M[GTR+ts-CpG] obtiennent des probabilités a posteriori de 0,49 à 0,65 et 0,45 à 0,59 pour les ensembles ssCABC2018 et ss73 respectivement.

En résumé, les modèles incorporant l'hypermutabilité des transversions, M[GTR+tstv-CpG] et M[GTR+ts-CpG+tv-CpG], sont préférés au modèle M[GTR+ts-CpG] qui a servi à faire les simulations. Il est possible que (1) les tables de références utilisées soient de trop petites tailles (2) que les distributions *a priori* utilisées soient trop étroites (3) que le nombre d'arbres construits par l'algorithme de forêts aléatoires soit trop petit, ou (4) pas assez profonds ou (5) encore que les ensembles de statistiques descriptives utilisés ne soient

pas adéquats. Il est possible que la comparaison de nos modèles, dans les conditions dans lesquelles nous les utilisons, ne soit pas près d'être efficacement comparée.

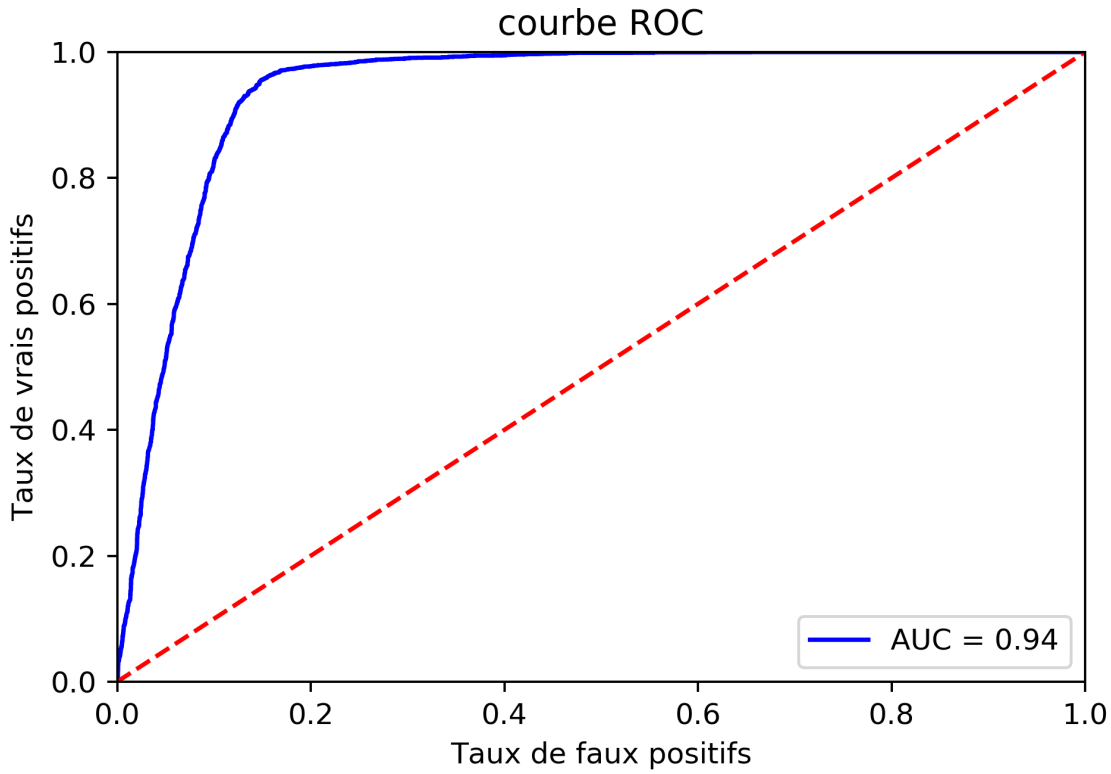

Figure 4.15. Courbe ROC calculée à partir des probabilités a posteriori de préférer le modèle M[HKY] ou M[GTR] pour 10,000 simulations de validations faites équitablement sous ces deux modèles. La préférence est évaluée à l'aide de l'algorithme de forêts aléatoires utilisant un classificateur suivi d'un modèle de régression.

4.4. Perpectives

4.4.1. Modéliser la conversion génique biaisée

La conversion génique biaisée vers GC (gBGC) est un processus mutationnel, qui s'apparente à la sélection en favorisant les allèles G/C par rapport aux allèles A/T, dû à un biais dans la réparation [Duret, 2002b, Duret and Galtier, 2009a, Katzman et al., 2011, Glemin et al., 2015]. La gBGC est un processus hétérogène au cours du temps. La gBGC agit sur des fragments de longueurs variables selon les espèces, l'efficacité de la gBGC serait variable en fonction de la N_e des espèces. Mais, néanmoins, malgré cette hétérogénéité au cours du

temps nous proposons de tester l'importance de la gBGC chez le Eutheria en modélisant la préférence des codons synonymes qui se terminent vers G/C ou vers A/T de manière homogène au cours du temps. Suivant la paramétrisation de [Lartillot, 2013], il est possible de paramétrer la conversion génique biaisée comme de la sélection.

4.4.2. Amélioration du CABC en utilisant des réseaux de neurones convolutifs

Même si l'algorithme de forêts aléatoires paraît prometteur, beaucoup de validations sont encore à faire pour mieux comprendre comment l'algorithme peut aider aux choix des statistiques descriptives ou encore à la correction des distributions a posteriori obtenues avec la méthode CABC. Dans les deux cas, il est nécessaire de définir *a priori* des fonctions qui permettent d'extraire les statistiques descriptives des alignements de gènes et des alignements simulés, alors que certains modèles statistiques (e.g., réseaux de neurones convolutifs : CNN) sont, sous certaines conditions, capables d'extraire le signal convoité via les représentations distribuées pour effectuer leur tâche de classification ou de régression.

Cette capacité des réseaux de neurones convolutifs à traiter des images a permis de développer de nouvelles méthodes d'inférence dans le cadre de la génétique des populations [Flagel et al., 2019]. Mais pour ce faire, les alignements sont tout d'abord transformés en images [Flagel et al., 2019]. De manière surprenante, l'ordre dans lequel les séquences sont présentées dans l'alignement pour produire l'image aura un impact sur la précision du modèle [Flagel et al., 2019]. Pour l'instant, les CNN ont été principalement utilisées pour faire la comparaison de modèles [Flagel et al., 2019], d'une manière très semblable à celle utilisée avec l'algorithme de forêt aléatoire [Pudlo et al., 2016]. Comme nous l'avons fait avec l'algorithme de forêt aléatoire, le CNN pourrait être utilisé pour corriger la distribution a posteriori. L'autre avantage potentiel des CNN est qu'ils peuvent gérer plusieurs variables réponses à la fois. Dans le contexte de la comparaison de modèles, il devrait aussi être possible de calculer une probabilité a posteriori sur l'issue de la classification, comme l'ont fait Pudlo et al 2015, mais cette fois dans le contexte de l'utilisation de CNN. Le transfert conceptuel me semble assez simple, mais l'est-il vraiment ? Les CNN sont des algorithmes complexes à utiliser : il existe plusieurs architectures et de nombreuses paramétrisations sont possibles. Il est donc nécessaire de s'associer avec les experts de ce domaine pour réaliser ce genre de travail efficacement.

4.4.3. Prendre en compte la structure tertiaire des protéines

Dans les années 2000, les modèles structuraux d'évolution moléculaire, un type de mutation-sélection, prennent leur envol (e.g., [Robinson et al., 2003, Kleinman et al., 2010, Rodrigue et al., 2005, 2009]) montrent que prendre en compte la structure tertiaire des protéines ainsi qu'un taux variable de substitutions non-synonymes (ω) entre sites améliore l'habileté des modèles à expliquer les données (sur la base de la comparaison des facteurs de Bayes). Lorsque ces deux hétérogénéités sont prises en compte dans la stratégie de modélisation, bien que les aspects de la structure des protéines soient déterminés *a priori*, l'amélioration de la capacité du modèle à expliquer les données n'est pas que la somme des améliorations prises indépendamment, structure et taux, mais plus. Il y a un effet synergique à utiliser conjointement des contraintes liées à la structure des protéines et aux taux de substitutions variables entre sites.

Nous pensons qu'il serait intéressant de revisiter les modèles de substitution à codons incorporant la prise en compte des contraintes de sélection liées à la structure des protéines pour plusieurs raisons. Tout d'abord parce que les données expérimentales concernant les structures tertiaires des protéines ne cessent d'augmenter, mais aussi parce qu'il y a aujourd'hui de nouvelles capacités de prédictions des structures tertiaires et autres caractéristiques structurales comme l'accessibilité au solvant, et cela par l'utilisation des réseaux de neurones convolutifs (e.g., [Gao et al., 2019]). De plus, grâce aux ressources computationnelles rendues disponibles aujourd'hui, il est possible d'envisager de remplacer les ω site-spécifiques par les profils de préférences en acides aminés.

4.4.4. Utiliser les modèles mutation-sélection pour prédire la pathogénicité des variants de BRCA1

Comme nous l'avons vu, les modèles de type mutation-sélection permettent de dépeindre certains aspects du paysage adaptatif des gènes étudiés. Il est aussi possible de dépeindre le paysage adaptatif des gènes de manière expérimentale (e.g., [Li et al., 2016, Olson et al., 2014, Steinberg and Ostermeier, 2016, Findlay et al., 2018]). Il faut par contre accéder à une grande diversité de variants ainsi qu'à leur phénotype pour ensuite quantifier la fonctionnalité/valeur adaptative des variants/phénotypes. Il est possible d'étudier la diversité naturelle avec un séquençage profond [Lee et al., 2018], ou encore générer des variants expérimentalement. C'est

un sujet de recherche actuellement très en vogue, puisqu'il permet de mieux comprendre quelles sont les contraintes biophysiques, structurales et fonctionnelles des protéines tout autant que les modèles de paysage adaptatifs utilisés soient pertinents.

Récemment des chercheurs ont utilisé la technologie CRISPR/Cas9 pour explorer la valeur adaptative d'un grand nombre de variants du gène BRCA1, plus exactement d'un sous-domaine de ce dernier [Findlay et al., 2018]. Ils ont caractérisé presque l'entièreté des variations simples possibles du sous-domaine, ce qui veut dire que, pour chacune des positions de la séquence codante du sous-domaine du gène, ils ont une expérience fonctionnelle, qui indirectement permet de calculer la valeur adaptative de la mutation. Ils n'ont par contre pas exploré les variations multiples alors que des interactions complexes épistatiques peuvent avoir lieu. Bien que l'approche soit réductrice, ne prenant pas en compte l'effet de l'épistasie, les chercheurs sont néanmoins capables de prédire sur la base de la valeur adaptative retrouvée des pathogénicités déterminées cliniquement [Findlay et al., 2018].

Cette nouvelle capacité de prédiction est très prometteuse, car le problème de la détermination clinique de la pathogénicité réside dans le fait que la probabilité d'étudier les quelques 4000 variants est très faible à cause (1) du faible taux de mutation des mammifères (2) du biais mutationnel qui réduit la probabilité d'observer certains variants et (3) du fait que la caractérisation au niveau de l'ADN est peu fréquente. À noter que nous ne sommes pas des experts du domaine, et il faudrait absolument valider les intuitions avant d'investir plus de ressources dans ce genre de projet.

L'expérience que nous proposons est relativement simple. Il suffit d'appliquer la méthode CABC pour conditionner le modèle M[GTR-ts-CpG]-S[NCatAA*], donc prenant en compte l'hypermutabilité des transitions en contexte CpG ainsi que l'hétérogénéité de préférence site-spécifique en acides aminés, à un alignement du gène BRCA1 à l'échelle des Eutheria ou encore des primates. Puis l'idée serait de comparer la sélection prédite par le modèle M[GTR-ts-CpG]-S[NCatAA*] à la valeur adaptative prédite dans l'expérience [Findlay et al., 2018] et la pathogénicité détectée par les cliniciens. Il serait peut-être nécessaire d'ajuster la sélection sur l'arginine, car le processus mutationnel d'hypermutabilité des transitions en contexte CpG interagit certainement avec les profils de préférence en acides aminés, puisque la pathogénicité semble pouvoir être associée à des mutations de l'arginine (eg., arginine356 CGG [Dunning et al., 1997]).

Bibliographie

AM Aguinaldo, JM Turbeville, LS Linford, MC Rivera, JR Garey, RA Raff, and JA Lake. Evidence for a clade of nematodes, arthropods and other moulting animals. *Nature*, 387 (6632) :489–93, 1997.

H Akaike. NEW LOOK AT STATISTICAL-MODEL IDENTIFICATION. *Ieee Transactions on Automatic Control*, AC19(6) :716–723, 1974. doi : 10.1109/tac.1974.1100705.

H Akashi. Synonymous codon usage in Drosophila melanogaster : natural selection and translational accuracy. *Genetics*, 136(3) :927–35, 1994.

H Akashi. Inferring weak selection from patterns of polymorphism and divergence at silent sites in Drosophila DNA. *Genetics*, 139(2) :1067–1076, 1995.

B Alberts, A Johnson, J Lewis, M Raff, K Roberts, and P Walter. *Molecular Biology of the Cell, Fourth Edition*. Garland Science, 4 edition, 2002.

LB Alexandrov and MR Stratton. Mutational signatures : the patterns of somatic mutations hidden in cancer genomes. *Current Opinion in Genetics & Development*, 24 :52–60, 2014. doi : 10.1016/j.gde.2013.11.014.

LB Alexandrov, S Nik-Zainal, DC Wedge, SA Aparicio, S Behjati, AV Biankin, GR Bignell, N Bolli, A Borg, AL Borresen-Dale, S Boyault, B Burkhardt, AP Butler, C Caldas, HR Davies, C Desmedt, R Eils, JE Eyfjord, JA Foekens, M Greaves, F Hosoda, B Hutter, T Ilicic, S Imbeaud, M Imielinsk, N Jager, DT Jones, D Jones, S Knappskog, M Kool, SR Lakhani, C Lopez-Otin, S Martin, NC Munshi, H Nakamura, PA Northcott, M Pajic, E Papaemmanuil, A Paradiso, JV Pearson, XS Puente, K Raine, M Ramakrishna, AL Richardson, J Richter, P Rosenstiel, M Schlesner, TN Schumacher, PN Span, JW Teague, Y Totoki, AN Tutt, R Valdes-Mas, MM van Buuren, L van 't Veer, A Vincent-Salomon, N Waddell, LR Yates, Initiative Australian Pancreatic Cancer Genome, Icgc Breast Cancer Consortium, Icgc Mmml-Seq Consortium, Icgc PedBrain, J. Zucman-Rossi, PA Futreal,

U McDermott, P Lichter, M Meyerson, SM Grimmond, R Siebert, E Campo, T Shibata, SM Pfister, PJ Campbell, and MR Stratton. Signatures of mutational processes in human cancer. *Nature*, 500(7463) :415–21, 2013a. doi : 10.1038/nature12477.

LB Alexandrov, S Nik-Zainal, DC Wedge, SAJR Aparicio, S Behjati, AV Biankin, GR Bignell, N Bolli, A Borg, A-L Borresen-Dale, S Boyault, B Burkhardt, AP Butler, C Caldas, HR Davies, C Desmedt, R Eils, JE Eyfjord, JA Foekens, M Greaves, F Hosoda, B Hutter, T Ilicic, S Imbeaud, M Imielinsk, N Jaeger, DTW Jones, D Jones, S Knappskog, M Kool, SR Lakhani, C Lopez-Otin, S Martin, NC Munshi, H Nakamura, PA Northcott, M Pajic, E Papaemmanuil, A Paradiso, JV Pearson, XS Puente, K Raine, M Ramakrishna, AL Richardson, J Richter, P Rosenstiel, M Schlesner, TN Schumacher, PN Span, JW Teague, Y Totoki, ANJ Tutt, R Valdes-Mas, MM van Buuren, L van't Veer, A Vincent-Salomon, N Waddell, LR Yates, J Zucman-Rossi, PA Futreal, U McDermott, P Lichter, M Meyerson, SM Grimmond, R Siebert, Elias Co, T Shibata, SM Pfister, PJ Campbell, MR Stratton, Genome Australian Pancreatic Canc, Icgc Breast Canc Consortium, Icgc Mmml-Seq Consortium, and Icgc PedBrain. Signatures of mutational processes in human cancer. *Nature*, 500(7463) :415–+, 2013b. doi : 10.1038/nature12477.

LB Alexandrov, S Nik-Zainal, DC Wedge, PJ Campbell, and MR Stratton. Deciphering Signatures of Mutational Processes Operative in Human Cancer. *Cell Reports*, 3(1) : 246–259, 2013c. doi : 10.1016/j.celrep.2012.12.008.

SF Altschul, W Gish, W Miller, EW Myers, and DJ Lipman. Basic local alignment search tool. *J Mol Biol*, 215(3) :403–10, 1990.

CT Amemiya, J Alfoldi, AP Lee, SH Fan, H Philippe, I MacCallum, I Braasch, T Manousaki, I Schneider, N Rohner, C Organ, D Chalopin, JJ Smith, M Robinson, RA Dorrington, M Gerdol, B Aken, MA Biscotti, M Barucca, D Baurain, AM Berlin, GL Blatch, F Buonocore, T Burmester, MS Campbell, A Canapa, JP Cannon, A Christoffels, G De Moro, AL Edkins, L Fan, AM Fausto, N Feiner, M Forconi, J Gamieldien, S Gnerre, A Gnirke, JV Goldstone, W Haerty, ME Hahn, U Hesse, S Hoffmann, J Johnson, SI Karchner, S Kuraku, M Lara, JZ Levin, GW Litman, E Mauceli, T Miyake, M G Mueller, D R Nelson, A Nitsche, E Olmo, T Ota, A Pallavicini, S Panji, B Picone, CP Ponting, SJ Prohaska, D Przybylski, NR Saha, V Ravi, F J Ribeiro, T Sauka-Spengler, G Scapigliati, SMJ Searle, T Sharpe, O Simakov, PF Stadler, JJ Stegeman, K Sumiyama, D Tabbaa, H Tafer,

J Turner-Maier, P van Heusden, S White, L Williams, M Yandell, H Brinkmann, J N Volff, CJ Tabin, N Shubin, M Schartl, DB Jaffe, JH Postlethwait, B Venkatesh, F Di Palma, ES Lander, A Meyer, and K Lindblad-Toh. The African coelacanth genome provides insights into tetrapod evolution. *Nature*, 496(7445) :311–316, 2013.

W Amos and JI Hoffman. Evidence that two main bottleneck events shaped modern human genetic diversity. *Proceedings of the Royal Society B-Biological Sciences*, 277(1678) :131–137, 2010.

B Arbeithuber, AJ Betancourt, T Ebner, and I Tiemann-Boege. Crossovers are associated with mutation and biased gene conversion at recombination hotspots. *Proceedings of the National Academy of Sciences of the United States of America*, 112(7) :2109–2114, 2015. doi : 10.1073/pnas.1416622112.

PF Arndt and T Hwa. Identification and measurement of neighbor-dependent nucleotide substitution processes. *Bioinformatics*, 21(10) :2322–2328, 2005.

PF Arndt, CB Burge, and T Hwa. DNA sequence evolution with neighbor-dependent mutation. *J Comput Biol*, 10(3-4) :313–322, 2003.

C Auerbach. *Mutation research : Problems, results and perspectives.* Springer-Science+Business Media, B.V., 1976.

OT Avery, CM Macleod, and M McCarty. STUDIES ON THE CHEMICAL NATURE OF THE SUBSTANCE INDUCING TRANSFORMATION OF PNEUMOCOCCAL TYPES : INDUCTION OF TRANSFORMATION BY A DESOXYRIBONUCLEIC ACID FRACTION ISOLATED FROM PNEUMOCOCCUS TYPE III. *The Journal of experimental medicine*, 79(2) :137–58, 1944. doi : 10.1084/jem.79.2.137.

S Barber, Jo Voss, and M Webster. The rate of convergence for approximate Bayesian computation. *Electron J Stat*, 9(1) :80–105, 2015.

W Bateson and G Mendel. *Mendel's principles of heredity, by W. Bateson.* Cambridge [Eng.]University Press,, 1909.

MA Beaumont. Approximate Bayesian Computation. *Annual Review of Statistics and Its Application*, 6(1) :null, 2019. doi : 10.1146/annurev-statistics-030718-105212.

MA Beaumont, WY Zhang, and DJ Balding. Approximate Bayesian computation in population genetics. *Genetics*, 162(4) :2025–2035, 2002.

J Berard and L Gueguen. Accurate estimation of substitution rates with neighbor-dependent models in a phylogenetic context. *Syst Biol*, 61(3) :510–21, 2012.

G Bernardi. Isochores and the evolutionary genomics of vertebrates. *Gene*, 241(1) :3–17, 2000.

E Beutler, T Gelbart, JH Han, JA Koziol, and B Beutler. EVOLUTION OF THE GENOME AND THE GENETIC-CODE - SELECTION AT THE DINUCLEOTIDE LEVEL BY METHYLATION AND POLYRIBONUCLEOTIDE CLEAVAGE. *Proceedings of the National Academy of Sciences of the United States of America*, 86(1) :192–196, 1989. doi : 10.1073/pnas.86.1.192.

N Bierne and A Eyre-Walker. Variation in synonymous codon use and DNA polymorphism within the Drosophila genome. *J Evol Biol*, 19(1) :1–11, 2006.

AP Bird. DNA methylation and the frequency of CpG in animal DNA. *Nucleic Acids Res*, 8(7) :1499–1504, 1980.

S Blanquart and N Lartillot. A site- and time-heterogeneous model of amino acid replacement. *Molecular Biology and Evolution*, 25(5) :842–858, 2008. doi : 10.1093/molbev/msn018.

MGB Blum and O Francois. Non-linear regression models for approximate bayesian computation. *Stat Comput*, 20(1) :63–73, 2010.

L Breiman. Random forests. *Machine Learning*, 45(1) :5–32, 2001. doi : 10.1023/a:1010933404324.

M Bulmer. Coevolution of codon usage and transfer-RNA abundance. *Nature*, 325(6106) :728–730, 1987.

C Burge, AM Campbell, and S Karlin. Over-representation and under-representation of short oligonucleotides in DNA-sequences. *Proc Natl Acad Sci USA*, 89(4) :1358–1362, 1992.

G Cannarrozzi, NN Schraudolph, M Faty, P von Rohr, MT Friberg, AC Roth, P Gonnet, Gn Gonnet, and Y Barral. A Role for Codon Order in Translation Dynamics. *Cell*, 141 (2) :355–367, 2010.

JV Chamary, JL Parmley, and LD Hurst. Hearing silence : non-neutral evolution at synonymous sites in mammals. *Nat Rev Genet*, 7(2) :98–108, 2006.

B Charlesworth. Effective population size and patterns of molecular evolution and variation. *Nat Rev Genet*, 10(3) :195–205, 2009.

N Chatterjee and Graham C Walker. Mechanisms of DNA Damage, Repair, and Mutagenesis. *Environ Mol Mutagen*, 58(5) :235–263, 2017.

SL Chen, W Lee, AK Hottes, L Shapiro, and HH McAdams. Codon usage between genomes is constrained by genome-wide mutational processes. *Proc Natl Acad Sci USA*, 101(10) : 3480–3485, 2004.

OF Christensen. Pseudo-likelihood for non-reversible nucleotide substitution models with neighbour dependent rates. *Stat Appl Genet Mol Biol*, 5, 2006.

OF Christensen, A Hobolth, and JL Jensen. Pseudo-likelihood analysis of codon substitution models with neighbor-dependent rates. *J Comput Biol*, 12(9) :1166–1182, 2005.

JR Coleman, D Papamichail, S Skiena, B Futcher, E Wimmer, and Mueller S. Virus attenuation by genome-scale changes in codon pair bias. *Science*, 320(5884) :1784–1787, 2008.

SR Cook, A Gelman, and DB Rubin. Validation of software for Bayesian models using posterior quantiles. *J Comput Graph Stat*, 15(3) :675–692, 2006.

FH Crick. Codon–anticodon pairing : the wobble hypothesis. *J Mol Biol*, 19(2) :548–55, 1966.

K Csilléry, MGB Blum, OE Gaggiotti, and O Francois. Approximate Bayesian Computation (ABC) in practice. *Trends Ecol Evol*, 25(7) :410–418, 2010.

K Csilléry, O François, and MGB Blum. abc : an R package for approximate Bayesian computation (ABC). *Methods Ecol Evol*, 3(3) :475–479, 2012.

D Darriba, T Flouri, and A Stamatakis. The State of Software for Evolutionary Biology. *Molecular Biology and Evolution*, 35(5) :1037–1046, 2018. doi : 10.1093/molbev/msy014.

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

V Daubin and GJ Szoellosi. Horizontal Gene Transfer and the History of Life. *Cold Spring Harbor Perspectives in Biology*, 8(4), 2016. doi : 10.1101/cshperspect.a018036.

AA Davin, E Tannier, TA Williams, B Boussau, V Daubin, and GJ Szollosi. Gene transfers can date the tree of life. *Nature Ecology & Evolution*, 2(5) :904–909, 2018. doi : 10.1038/s41559-018-0525-3.

N De Maio, C Schloetterer, and C Kosiol. Linking Great Apes Genome Evolution across Time Scales Using Polymorphism-Aware Phylogenetic Models. *Molecular Biology and Evolution*, 30(10) :2249–2262, 2013. doi : 10.1093/molbev/mst131.

F Delsuc, H Brinkmann, and H Philippe. Phylogenomics and the reconstruction of the tree of life. *Nat Rev Genet*, 6(5) :361–75, 2005.

AP Dempster, NM Laird, and DB Rubin. MAXIMUM LIKELIHOOD FROM INCOMPLETE DATA VIA EM ALGORITHM. *Journal of the Royal Statistical Society Series B-Methodological*, 39(1) :1–38, 1977. URL `<GotoISI>://WOS:A1977DM46400001`.

G Diss, I Gagnon-Arsenault, AM Dion-Cote, H Vignaud, DI Ascencio, CM Berger, and CR Landry. Gene duplication can impart fragility, not robustness, in the yeast protein interaction network. *Science*, 355(6325) :630–633, 2017. doi : 10.1126/science.aai7685.

KA Dittmar, JM Goodenbour, and T Pan. Tissue-specific differences in human transfer RNA expression. *PLoS Genet*, 2(12) :e221, 2006. doi : 10.1371/journal.pgen.0020221.

M Doble and SN Gummadi. *Biochemical Engineering*. Prentice-Hall of Indiai, New Delhi, 2007.

A Doherty and JO McInerney. Translational Selection Frequently Overcomes Genetic Drift in Shaping Synonymous Codon Usage Patterns in Vertebrates. *Mol Biol Evol*, 30(10) : 2263–2267, 2013.

A Doron-Faigenboim and T Pupko. A combined empirical and mechanistic codon model. *Molecular Biology and Evolution*, 24(2) :388–397, 2007. doi : 10.1093/molbev/msl175.

M dos Reis and L Wernisch. Estimating Translational Selection in Eukaryotic Genomes. *Mol Biol Evol*, 26(2) :451–461, 2009.

M dos Reis, PCJ Donoghue, and ZH Yang. Bayesian molecular clock dating of species divergences in the genomics era. *Nature Reviews Genetics*, 17(2) :71–80, 2016.

EJP Douzery, C Scornavacca, J Romiguier, K Belkhir, N Galtier, F Delsuc, and V Ranwez. OrthoMaM v8 : A Database of Orthologous Exons and Coding Sequences for Comparative Genomics in Mammals. *Mol Biol Evol*, 31(7) :1923–1928, 2014.

DA Drummond and CO Wilke. Mistranslation-induced protein misfolding as a dominant constraint on coding-sequence evolution. *Cell*, 134(2) :341–352, 2008.

DA Drummond and CO Wilke. The evolutionary consequences of erroneous protein synthesis. *Nature Reviews Genetics*, 10(10) :715–724, 2009. doi : Doi10.1038/Nrg2662.

S Duchene, SYW Ho, and EC Holmes. Declining transition/transversion ratios through time reveal limitations to the accuracy of nucleotide substitution models. *Bmc Evolutionary Biology*, 15, 2015. doi : 10.1186/s12862-015-0312-6.

S Duffy, LA Shackelton, and EC Holmes. Rates of evolutionary change in viruses : patterns and determinants. *Nature Reviews Genetics*, 9(4) :267–276, 2008. doi : Doi10.1038/Nrg2323.

R Dumollard, M Duchen, and J Carroll. The role of mitochondrial function in the oocyte and embryo. *Mitochondrion in the Germline and Early Development*, 77 :21, 2007. doi : 10.1016/s0070-2153(06)77002-8.

KA Dunn, T Kenney, H Gu, and JP Bielawski. Improved inference of site-specific positive selection under a generalized parametric codon model when there are multinucleotide mutations and multiple nonsynonymous rates. *Bmc Evolutionary Biology*, 19, 2019. doi : 10.1186/s12862-018-1326-7.

AM Dunning, M Chiano, NR Smith, J Dearden, M Gore, S Oakes, C Wilson, M Stratton, J Peto, D Easton, D Clayton, and BAJ Ponder. Common BRCA1 variants and susceptibility to breast and ovarian cancer in the general population. *Human Molecular Genetics*, 6(2) :285–289, 1997. doi : 10.1093/hmg/6.2.285.

L Duret. Evolution of synonymous codon usage in metazoans. *Curr Opin Genet Dev*, 12 (6) :640–649, 2002a.

L Duret. Evolution of synonymous codon usage in metazoans. *Curr Opin Genet Dev*, 12 (6) :640–649, 2002b.

L Duret and PF Arndt. The impact of recombination on nucleotide substitutions in the human genome. *PLoS Genet*, 4(5) :e1000071, 2008.

L Duret and N Galtier. The covariation between TpA deficiency, CpG deficiency, and G+C content of human isochores is due to a mathematical artifact. *Mol Biol Evol*, 17(11) : 1620–5., 2000.

L Duret and N Galtier. Biased gene conversion and the evolution of mammalian genomic landscapes. *Annu Rev Genomics Hum Genet*, 10 :285–311, 2009a.

L Duret and N Galtier. Biased Gene Conversion and the Evolution of Mammalian Genomic Landscapes. *Annu Rev Genomics Hum Genet*, 10 :285–311, 2009b.

L Duret and D Mouchiroud. Expression pattern and, surprisingly, gene length shape codon usage in Caenorhabditis, Drosophila, and Arabidopsis. *Proc Natl Acad Sci USA*, 96(8) : 4482–4487, 1999.

Joseph EC. Unifying the derivations for the Akaike and corrected Akaike information criteria. *Statistics & Probability Letters*, 33(2) :201–208, 1997. doi : https://doi.org/10.1016/S0167-7152(96)00128-9.

Julian Echave, Stephanie J. Spielman, and Claus O. Wilke. Causes of evolutionary rate variation among protein sites. *Nat Rev Genet*, 17(2) :109–121, 2016. doi : 10.1038/nrg.2015.18.

B Efron and RJ Tibshirani. *An Introduction to the Bootstrap*. Chapman & Hall/CRC Monographs on Statistics & Applied Probability. Taylor & Francis, 1994.

J Ellson, ER Gansner, E Koutsofios, SC North, and G Woodhull. Graphviz and dynagraph – static and dynamic graph drawing tools. pages 127–148, 2003.

L Eme, SC Sharpe, MW Brown, and AJ Roger. On the Age of Eukaryotes : Evaluating Evidence from Fossils and Molecular Clocks. *Cold Spring Harbor Perspectives in Biology*, 6(8), 2014. doi : 10.1101/cshperspect.a016139.

J Eric, O Travis, P Pearu, et al. SciPy : Open source scientific tools for Python. 2001–.

MD Ermolaeva. Synonymous codon usage in bacteria. *Curr Issues Mol Biol*, 3(4) :91–7, 2001.

Jacob A Esselstyn, CH Oliveros, MT Swanson, and BC Faircloth. Investigating Difficult Nodes in the Placental Mammal Tree with Expanded Taxon Sampling and Thousands of Ultraconserved Elements. *Genome Biology and Evolution*, 9(9) :2308–2321, 2017. doi : 10.1093/gbe/evx168.

JS Farris. Methods for computing Wagner trees. *Sys Zool*, 19(1) :83–&, 1970.

P Fearnhead and D Prangle. Constructing summary statistics for approximate Bayesian computation : semi-automatic approximate Bayesian computation. *J R Stat Soc Series B Stat Methodol*, 74 :419–474, 2012.

J Felsenstein. Maximum-likelihood estimation of evolutionary trees from continuous characters. *Am J Hum Genet*, 25(5) :471–492, 1973.

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

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

T Ferguson. A Bayesian analysis of some nonparametric problems. *Ann Statistics*, 1 :209–230, 1973.

J Filipski, JP Thiery, and G Bernardi. Analysis of bovine genome by cs2so4-ag+ density gradient centrifugation. *J Mol Biol*, 80(1) :177–197, 1973.

GM Findlay, RM Daza, B Martin, MD Zhang, AP Leith, M Gasperini, JD Janizek, X Huang, LM Starita, and J Shendure. Accurate classification of BRCA1 variants with saturation genome editing. *Nature*, 562(7726) :217–+, 2018. doi : 10.1038/s41586-018-0461-z.

Aaron Fisher, Cynthia Rudin, and Francesca Dominici. All models are wrong but many are useful : Variable importance for black-box, proprietary, or misspecified prediction models, using model class reliance. *arXiv preprint arXiv :1801.01489*, 2018.

RA FISHER. THE USE OF MULTIPLE MEASUREMENTS IN TAXONOMIC PROBLEMS. *Annals of Eugenics*, 7(2) :179–188, 1936. doi : 10.1111/j.1469-1809.1936.tb02137.x.

RA Fisher. *The Genetical Theory of Natural Selection*. Oxford University Press, USA, 1 edition, 2000.

WM Fitch. Toward defining course of evolution - minimum change for a specific tree topology. *Sys Zool*, 20(4) :406–&, 1971a.

WM Fitch. Rate of change of concomitantly variable codons. *J Mol Evol*, 1(1) :84–96, 1971b.

WM Fitch and E Markowitz. An improved method for determining codon variability in a gene and its application to the rate of fixation of mutations in evolution. *Biochem Genet*, 4(5) :579–593, 1970.

L Flagel, Y Brandvain, and DR Schrider. The Unreasonable Effectiveness of Convolutional Neural Networks in Population Genetic Inference. *Molecular Biology and Evolution*, 36 (2) :220–238, 2019. doi : 10.1093/molbev/msy224.

NM Foley, MS Springer, and EC Teeling. Mammal madness : is the mammal tree of life not yet resolved ? *Philos Trans R Soc Lond B Biol Sci*, 371(1699), 2016.

PG Foster, LS Jermiin, and DA Hickey. Nucleotide composition bias affects amino acid content in proteins coded by animal mitochondria. *J Mol Evol*, 44(3) :282–288, 1997.

I Fragata, A Blanckaert, Marco A Dias L, DA Liberles, and C Bank. Evolution in the light of fitness landscape theory. *Trends Ecolo Evol*, 34(1) :69–82, 2019. doi : 10.1016/j.tree.2018.10.009.

LC Francioli, PP Polak, A Koren, A Menelaou, S Chun, I Renkens, CM van Duijn, M Swertz, C Wijmenga, G van Ommen, PE Slagboom, D I Boomsma, K Ye, V Guryev, PF Arndt, WP Kloosterman, PIW de Bakker, SR Sunyaev, and Consortium Genome Netherlands. Genome-wide patterns and properties of de novo mutations in humans. *Nat Genet*, 47(7) : 822–+, 2015.

DT Frazier, CP Robert, and J Rousseau. Model Misspecification in ABC : Consequences and Diagnostics. *1708.01974v1 [math.ST]. https ://arxiv.org/pdf/1708.01974*, 2017.

YX Fu and WH Li. Estimating the age of the common ancestor of a sample of dna sequences. *Mol Biol Evol*, 14(2) :195–9, 1997.

N Galtier and M Gouy. Inferring phylogenies from dna sequences of unequal base compositions. *Proc Natl Acad Sci USA*, 92(24) :11317–21, 1995.

N Galtier, C Roux, M Rousselle, J Romiguier, E Figuet, S Glémin, N Bierne, and L Duret. Codon Usage Bias in Animals : Disentangling the Effects of Natural Selection, Effective Population Size, and GC-Biased Gene Conversion. *Mol Biol Evol*, 35(5) :1092–1103, 2018.

S Gangloff, G Achaz, S Francesconi, A Villain, S Miled, C Denis, and B Arcangioli. Quiescence unveils a novel mutational force in fission yeast. *Elife*, 6, 2017.

M Gao, H Zhou, and J Skolnick. DESTINI : A deep-learning approach to contact-driven protein structure prediction. *Scientific Reports*, 9, 2019. doi : 10.1038/s41598-019-40314-1.

A Gelman, JB Carlin, HS Stern, and DB Rubin. *Bayesian Data Analysis*. Chapman and Hall/CRC, Boca Raton, 3nd ed. edition, 2013.

ER Gibney and CM Nolan. Epigenetics and gene expression. *Heredity*, 105(1) :4–13, 2010. doi : 10.1038/hdy.2010.54.

S Glemin, PF Arndt, PW Messer, D Petrov, N Galtier, and L Duret. Quantification of GC-biased gene conversion in the human genome. *Genome Res*, 25(8) :1215–1228, 2015.

T Gojobori, WH Li, and D Graur. PATTERNS OF NUCLEOTIDE SUBSTITUTION IN PSEUDOGENES AND FUNCTIONAL GENES. *Journal of Molecular Evolution*, 18(5) : 360–369, 1982. doi : 10.1007/bf01733904.

N Goldman and Z Yang. A codon-based model of nucleotide substitution for protein-coding DNA sequences. *Mol Biol Evol*, 11(5) :725–36, 1994.

JM Goodenbour and T Pan. Diversity of tRNA genes in eukaryotes. *Nucleic acids research*, 34(21) :6137–46, 2006. doi : 10.1093/nar/gkl725.

R Grantham. AMINO-ACID DIFFERENCE FORMULA TO HELP EXPLAIN PROTEIN EVOLUTION. *Science*, 185(4154) :862–864, 1974. doi : 10.1126/science.185.4154.862.

R Grantham, C Gautier, M Gouy, R Mercier, and A Pave. Codon catalog usage and the genome hypothesis. *Nucleic Acids Res*, 8(1) :r49–r62, 1980.

P Green, B Ewing, W Miller, PJ Thomas, ED Green, and Nisc Comparative Sequencing Progr. Transcription-associated mutational asymmetry in mammalian evolution. *Nature Genetics*, 33(4) :514–517, 2003. doi : 10.1038/ng1103.

L Gueguen and L Duret. Unbiased Estimate of Synonymous and Nonsynonymous Substitution Rates with Nonstationary Base Composition. *Molecular Biology and Evolution*, 35 (3) :734–742, 2018. doi : 10.1093/molbev/msx308.

S Guindon, J-F Dufayard, V Lefort, M Anisimova, W Hordijk, and O Gascuel. New Algorithms and Methods to Estimate Maximum-Likelihood Phylogenies : Assessing the Performance of PhyML 3.0. *Systematic Biology*, 59(3) :307–321, 2010. doi : 10.1093/sysbio/syq010.

YA Guo, MM Chang, WT Huang, WF Ooi, MJ Xing, P Tan, and AJ Skanderup. Mutation hotspots at CTCF binding sites coupled to chromosomal instability in gastrointestinal cancers. *Nat Commun*, 9, 2018.

D Haig. Retroviruses and the placenta. *Current Biology*, 22(15) :R609–R613, 2012. doi : 10.1016/j.cub.2012.06.002.

JBS Haldane. A mathematical theory of natural and artificial selection. V. Selection and mutation. *Mathematical Proceedings of the Cambridge Philosophical Society*, 23 :838 – 844, 1927.

AL Halpern and WJ Bruno. Evolutionary distances for protein-coding sequences : modeling site-specific residue frequencies. *Mol Biol Evol*, 15(7) :910–917, 1998.

Y Harigaya and R Parker. No-go decay : a quality control mechanism for RNA in translation. *Wiley Interdisciplinary Reviews-Rna*, 1(1) :132–141, 2010. doi : 10.1002/wrna.17.

M Hasegawa, H Kishino, and Ta Yano. Dating of the Human Ape Splitting by a Molecular Clock of Mitochondrial-Dna. *J Mol Evol*, 22(2) :160–174, 1985.

WK Hastings. Monte-Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57(1) :97–&, 1970.

T Helleday, S Eshtad, and S Nik-Zainal. Mechanisms underlying mutational signatures in human cancers. *Nature Reviews Genetics*, 15(9) :585–598, 2014. doi : 10.1038/nrg3729.

MD Hendy and D Penny. A framework for the quantitative study of evolutionary trees. *Syst Zool*, 38 :297–309, 1989.

R. Hershberg and D. A. Petrov. Evidence That Mutation Is Universally Biased towards AT in Bacteria. *Plos Genetics*, 6(9), 2010. doi : ARTNe1001115DOI10.1371/journal.pgen.1001115.

R Hershberg and DA Petrov. Selection on Codon Bias. *Annu Rev Genet*, 42 :287–299, 2008.

F Hildebrand, A Meyer, and A Eyre-Walker. Evidence of Selection upon Genomic GC-Content in Bacteria. *Plos Genetics*, 6(9), 2010. doi : ARTNe1001107DOI10.1371/journal.pgen.1001107.

SK Hilton, MB Doud, and JD Bloom. phydms : software for phylogenetic analyses informed by deep mutational scanning. *Peerj*, 5, 2017.

A Hobolth. A Markov chain Monte Carlo expectation maximization algorithm for statistical analysis of DNA sequence evolution with neighbor-dependent substitution rates. *J Comput Graph Stat*, 17(1) :138–162, 2008.

A Hobolth, R Nielsen, Y Wang, FN Wu, and SD Tanksley. CpG plus CpNpG analysis of protein-coding sequences from tomato. *Mol Biol Evol*, 23(6) :1318–1323, 2006.

A Hodgkinson and A Eyre-Walker. Variation in the mutation rate across mammalian genomes. *Nat Rev Genet*, 12(11) :756–766, 2011.

JP Huelsenbeck. Testing a covariotide model of dna substitution. *Mol Biol Evol*, 19(5) : 698–707., 2002.

G Huttley and VB Yap. Robust estimation of natural selection using parametric codon models. In GM Cannarozzi and A Schneider, editors, *Codon Evolution : Mechanisms and Models*, book section 8. OUP Oxford, 2012.

GA Huttley. Modeling the impact of DNA methylation on the evolution of BRCA1 in mammals. *Mol Biol Evol*, 21(9) :1760–8, 2004.

DG Hwang and P Green. Bayesian Markov chain Monte Carlo sequence analysis reveals varying neutral substitution patterns in mammalian evolution. *Proc Natl Acad Sci USA*, 101(39) :13994–14001, 2004.

T Ikemura. Correlation between the abundance of Escherichia coli transfer RNAs and the occurrence of the respective codons in its protein genes : a proposal for a synonymous codon choice that is optimal for the E. coli translational system. *J Mol Biol*, 151(3) : 389–409, 1981.

T Ikemura. Codon usage and transfer-RNA content in unicellular and multicellular organisms. *Mol Biol Evol*, 2(1) :13–34, 1985.

I Irisarri, D Baurain, H Brinkmann, F Delsuc, J-Y Sire, A Kupfer, J Petersen, M Jarek, A Meyer, M Vences, and H Philippe. Phylotranscriptomic consolidation of the jawed vertebrate timetree. *Nature Ecology & Evolution*, 1(9) :1370–1378, 2017. doi : 10.1038/s41559-017-0240-5.

H Jeffreys. Some Tests of Significance, Treated by the Theory of Probability. *Mathematical Proceedings of the Cambridge Philosophical Society*, 31(2) :203–222, 1935. doi : 10.1017/S030500410001330X.

JL Jensen and AMK Pedersen. Probabilistic models of DNA sequence evolution with context dependent rates of substitution. *Adv App Prob*, 32(2) :499–517, 2000.

CT Jones, N Youssef, E Susko, and JP Bielawski. Phenomenological Load on Model Parameters Can Lead to False Biological Conclusions. *Molecular Biology and Evolution*, 35 (6) :1473–1488, 2018. doi : 10.1093/molbey/msy049.

H Jonsson, P Sulem, B Kehr, S Kristmundsdottir, F Zink, E Hjartarson, MT Hardarson, KE Hjorleifsson, HP Eggertsson, SA Gudjonsson, LD Ward, GA Arnadottir, EA Helgason, H Helgason, A Gylfason, A Jonasdottir, T Rafnar, M Frigge, SN Stacey, OT Magnusson, U Thorsteinsdottir, G Masson, A Kong, BV Halldorsson, A Helgason, DF Gudbjartsson, and K Stefansson. Parental influence on human germline de novo mutations in 1,548 trios from Iceland. *Nature*, 549(7673) :519–+, 2017.

TH Jukes and CR Cantor. Evolution of protein molecules. In HN Munro, editor, *Mammalian protein metabolism*, pages 21–132. Academic Press, New York, 1969.

S Katzman, JA Capra, D Haussler, and KS Pollard. Ongoing GC-Biased Evolution Is Widespread in the Human Genome and Enriched Near Recombination Hot Spots. *Genome Biol Evol*, 3 :614–626, 2011.

PD Keightley, U Trivedi, M Thomson, F Oliver, S Kumar, and ML Blaxter. Analysis of the genome sequences of three Drosophila melanogaster spontaneous mutation accumulation

lines. *Genome Research*, 19(7) :1195–1201, 2009. doi : 10.1101/gr.091231.109.

PD Keightley, L Eory, DL Halligan, and M Kirkpatrick. Inference of Mutation Parameters and Selective Constraint in Mammalian Coding Sequences by Approximate Bayesian Computation. *Genetics*, 187(4) :1153–U268, 2011.

I Keller, D Bensasson, and RA Nichols. Transition-transversion bias is not universal : A counter example from grasshopper pseudogenes. *Plos Genetics*, 3(2) :185–191, 2007. doi : 10.1371/journal.pgen.0030022.

MD Kessler and Matthew D Dean. Effective population size does not predict codon usage bias in mammals. *Ecol Evol*, 4(20) :3887–3900, 2014.

M Kimura. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *J Mol Evol*, 16(2) :111–20, 1980.

S Kirkpatrick, CD Gelatt, and MP Vecchi. Optimization by simulated annealing. *Science*, 220 :671–680, 1983.

CL Kleinman, N Rodrigue, C Bonnard, H Philippe, and N Lartillot. A maximum likelihood framework for protein design. *BMC Bioinformatics*, 7(1) :326, 2006.

CL Kleinman, N Rodrigue, N Lartillot, and H Philippe. Statistical potentials for improved structurally constrained evolutionary models. *Mol Biol Evol*, 27(7) :1546–60, 2010.

RD Knight, SJ Freeland, and LF Landweber. A simple model based on mutation and selection explains trends in codon and amino-acid usage and GC composition within and across genomes. *Genome Biol*, 2 :research0010, 2001.

EV Koonin and AS Novozhilov. Origin and Evolution of the Universal Genetic Code. *Annual Review of Genetics*, 51 :45–62, 2017. doi : 10.1146/annurev-genet-120116-024713.

C Kosiol, I Holmes, and N Goldman. An empirical codon model for protein sequence evolution. *Mol Biol Evol*, 24(7) :1464–79, 2007.

A Kousathanas, C Leuenberger, J Helfer, M Quinodoz, M Foll, and D Wegmann. Likelihood-Free Inference in High-Dimensional Models. *Genetics*, 203(2) :893, 2016.

R Krasovec, H Richards, DR Gifford, C Hatcher, KJ Faulkner, RV Belavkin, A Channon, E Aston, AJ McBain, and CG Knight. Spontaneous mutation rate is a plastic trait associated with population density across domains of life. *Plos Biol*, 15(8), 2017.

T Krassowski, AY Coughlan, X-X Shen, X Zhou, J Kominek, DA Opulente, R Riley, IV Grigoriev, N Maheshwari, DC Shields, CP Kurtzman, CT Hittinger, A Rokas, and KH Wolfe.

Evolutionary instability of CUG-Leu in the genetic code of budding yeasts. *Nature Communications*, 9, 2018. doi : 10.1038/s41467-018-04374-7.

S Kumar. Patterns of nucleotide substitution in mitochondrial protein coding genes of vertebrates. *Genetics*, 143(1) :537–548, 1996.

CG Kurland. TRANSLATIONAL ACCURACY AND THE FITNESS OF BACTERIA. *Annual Review of Genetics*, 26 :29–50, 1992. doi : 10.1146/annurev.ge.26.120192.000333.

C Lanave, G Preparata, C Saccone, and G Serio. A new method for calculating evolutionary substitution rates. *J Mol Evol*, 20(1) :86–93, 1984.

N Lartillot. Phylogenetic Patterns of GC-Biased Gene Conversion in Placental Mammals and the Evolutionary Dynamics of Recombination Landscapes. *Mol Biol Evol*, 30(3) : 489–502, 2013.

N Lartillot. Probabilistic models of eukaryotic evolution : time for integration. *Philosophical Transactions of the Royal Society B-Biological Sciences*, 370(1678), 2015. doi : 10.1098/rstb.2014.0338.

N Lartillot and H Philippe. A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. *Mol Biol Evol*, 21(6) :1095–1109, 2004.

N Lartillot and R Poujol. A Phylogenetic Model for Investigating Correlated Evolution of Substitution Rates and Continuous Phenotypic Characters. *Molecular biology and evolution*, 28(1) :729–744, 2011. doi : DOI10.1093/molbev/msq244.

N Lartillot, H Brinkmann, and H Philippe. Suppression of long-branch attraction artefacts in the animal phylogeny using a site-heterogeneous model. *BMC Evol Biol*, 7 Suppl 1 :S4, 2007.

N Lartillot, N Rodrigue, D Stubbs, and J Richer. PhyloBayes MPI : Phylogenetic Reconstruction with Infinite Mixtures of Profiles in a Parallel Environment. *Systematic Biology*, 62(4) :611–615, 2013a. doi : 10.1093/sysbio/syt022.

N Lartillot, N Rodrigue, D Stubbs, and J Richer. PhyloBayes MPI : Phylogenetic Reconstruction with Infinite Mixtures of Profiles in a Parallel Environment. *Syst Biol*, 62(4) : 611–615, 2013b.

T Latrille, L Duret, and N Lartillot. The Red Queen model of recombination hot-spot evolution : a theoretical investigation. *Philosophical Transactions of the Royal Society B-Biological Sciences*, 372(1736), 2017. doi : 10.1098/rstb.2016.0463.

S Laurin-Lemay, H Brinkmann, and H Philippe. Origin of land plants revisited in the light of sequence contamination and missing data. *Current biolog*, 22(15) :R593–4, 2012. doi : 10.1016/j.cub.2012.06.013.

S Laurin-Lemay, H Philippe, and N Rodrigue. Multiple Factors Confounding Phylogenetic Detection of Selection on Codon Usage. *Mol Biol Evol*, 35(6) :1463–1472, 2018a. doi : 10.1093/molbey/msy047.

S Laurin-Lemay, H Philippe, and N Rodrigue. Multiple Factors Confounding Phylogenetic Detection of Selection on Codon Usage. *Mol Biol and Evol*, 35(6) :1463–1472, 2018b.

S Laurin-Lemay, N Rodrigue, N Lartillot, and H Philippe. Conditional Approximate Bayesian Computation : A New Approach for Across-Site Dependency in High-Dimensional Mutation-Selection Models. *Mol Biol Evol*, 35(11) :2819–2834, 2018c. doi : 10.1093/molbev/msy173.

Y Lavner and D Kotlar. Codon bias as a factor in regulating expression via translation rate in the human genome. *Gene*, 345(1) :127–138, 2005.

DS Lawrie, PW Messer, R Hershberg, and D Petrov. Strong Purifying Selection at Synonymous Sites in D. melanogaster. *PLoS Genet*, 9(5), 2013.

HJ Lee, N Rodrigue, and JL Thorne. Relaxing the Molecular Clock to Different Degrees for Different Substitution Types. *Mol Biol Evol*, 32(8) :1948–1961, 2015.

HJ Lee, H Kishino, N Rodrigue, and JL Thorne. Grouping substitution types into different relaxed molecular clocks. *Proc Natl Acad Sci USA*, 371(1699), 2016.

JM Lee, J Huddleston, MB Doud, KA Hooper, NC Wu, T Bedford, and JD Bloom. Deep mutational scanning of hemagglutinin helps predict evolutionary fates of human H3N2 influenza variants. *Proceedings of the National Academy of Sciences of the United States of America*, 115(35) :EB276–EB285, 2018. doi : 10.1073/pnas.1806133115.

C Li, W Qian, CJ Maclean, and J Zhang. The fitness landscape of a tRNA gene. *Science*, 352(6287) :837–840, 2016. doi : 10.1126/science.aae0568.

J Li, J Zhou, Y Wu, SH Yang, and DC Tian. GC-Content of Synonymous Codons Profoundly Influences Amino Acid Usage. *G3 (Bethesda)*, 5(10) :2027–2036, 2015.

M Li, E Kao, X Gao, H Sandig, K Limmer, M Pavon-Eternod, TE Jones, S Landry, T Pan, MD Weitzman, and M David. Codon-usage-based inhibition of HIV protein synthesis by human schlafen 11. *Nature*, 491(7422) :125–128, 2012.

DA Liberles, SA Teichmann, I Bahar, U Bastolla, J Bloom, E Bornberg-Bauer, LJ Colwell, APJ de Koning, NV Dokholyan, J Echave, A Elofsson, DL Gerloff, RA Goldstein, JA Grahnen, MT Holder, C Lakner, N Lartillot, SC Lovell, G Naylor, T Perica, DD Pollock, T Pupko, L Regan, A Roger, N Rubinstein, E Shakhnovich, K Sjölander, S Sunyaev, AI Teufel, JL Thorne, JW Thornton, DM Weinreich, and S Whelan. The interface of protein structure, protein biophysics, and molecular evolution. *Protein Science*, 21(6) : 769–785, 2012.

H Lindsay, VB Yap, H Ying, and GA Huttley. Pitfalls of the most commonly used models of context dependent substitution. *Biol Direct*, 4, 2008.

JE Losey, AR Ives, J Harmon, F Ballantyne, and C Brown. A polymorphism maintained by opposite patterns of parasitism and predation. *Nature*, 388(6639) :269–272, 1997. doi : 10.1038/40849.

Gilles Louppe. Understanding random forests : From theory to practice. *arXiv preprint arXiv :1407.7502*, 2014.

GH Lubke, I Campbell, D McArtor, Miller P, Luningham J, and SM van den Berg. Assessing Model Selection Uncertainty Using a Bootstrap Approach : An Update. *Structural Equation Modeling : A Multidisciplinary Journal*, 24(2) :230–245, 2017. doi : 10.1080/10705511.2016.1252265.

M Lynch. Rate, molecular spectrum, and consequences of human mutation. *Proc Natl Acad Sci USA*, 107(3) :961–8, 2010a. doi : 10.1073/pnas.0912629107.

M Lynch. Evolution of the mutation rate. *Trends in Genetics*, 26(8) :345–352, 2010b. doi : 10.1016/j.tig.2010.05.003.

M Lynch. Mutation and Human Exceptionalism : Our Future Genetic Load. *Genetics*, 202 (3) :869–875, 2016. doi : 10.1534/genetics.115.180471.

M Lynch, W Sung, K Morris, N Coffey, CR Landry, EB Dopman, WJ Dickinson, K Okamoto, S Kulkarni, DL Hartl, and WK Thomas. A genome-wide view of the spectrum of spontaneous mutations in yeast. *Proceedings of the National Academy of Sciences of the United States of America*, 105(27) :9272–9277, 2008. doi : 10.1073/pnas.0803466105.

M Lynch, MS Ackerman, JF Gout, H Long, W Sung, WK Thomas, and PL Foster. Genetic drift, selection and the evolution of the mutation rate. *Nat Rev Genet*, 17(11) :704–714, 2016.

RP Maharjan and T Ferenci. A shifting mutational landscape in 6 nutritional states : Stress-induced mutagenesis as a series of distinct stress input-mutation output relationships. *Plos Biol*, 15(6), 2017.

P Marjoram, J Molitor, V Plagnol, and S Tavare. Markov chain monte carlo without likelihoods. *Proc Natl Acad Sci USA*, 100(26) :15324–15328, 2003.

DM McCandlish and A Stoltzfus. Modeling evolution using the probability of fixation : history and implications. *Q Rev Biol*, 89(3) :225–252, 2014.

JP McCutcheon and NA Moran. Extreme genome reduction in symbiotic bacteria. *Nature Reviews Microbiology*, 10(1) :13–26, 2012. doi : 10.1038/nrmicro2670.

GAT McVean and J Vieira. Inferring Parameters of Mutation, Selection and Demography From Patterns of Synonymous Site Evolution in Drosophila. *Genetics*, 157(1) :245–257, 2001.

DJ Merrell. *The Adaptive Seascape : The Mechanism of Evolution*. University of Minnesota Press, 1994.

J Mersch, MA Jackson, M Park, D Nebgen, SK Peterson, C Singletary, BK Arun, and JK Litton. Cancers Associated With BRCA1 and BRCA2 Mutations Other Than Breast and Ovarian. *Cancer*, 121(2) :269–275, 2015. doi : 10.1002/cncr.29041.

N Metropolis, AW Rosenbluth, MN Rosenbluth, AH Teller, and E Teller. Equation of state calculations by fast computing machines. *J Chem Phys*, 21(6) :1087–1092, 1953.

X Meyer, L Dib, D Silvestro, and N Salamin. Simultaneous Bayesian inference of phylogeny and molecular coevolution. *Proceedings of the National Academy of Sciences of the United States of America*, 116(11) :5027–5036, 2019. doi : 10.1073/pnas.1813836116.

TS Mikkelsen, LW Hillier, EE Eichler, MC Zody, DB Jaffe, SP Yang, W Enard, I Hellmann, K Lindblad-Toh, TK Altheide, N Archidiacono, P Bork, J Butler, JL Chang, Z Cheng, AT Chinwalla, P deJong, KD Delehaunty, CC Fronick, LL Fulton, Y Gilad, G Glusman, S Gnerre, TA Graves, T Hayakawa, KE Hayden, XQ Huang, HK Ji, WJ Kent, MC King, EJ Kulbokas, MK Lee, G Liu, C Lopez-Otin, KD Makova, O Man, ER Mardis, E Mauceli, TL Miner, WE Nash, JO Nelson, S Paabo, NJ Patterson, CS Pohl, KS Pollard, K Prufer, XS Puente, D Reich, M Rocchi, K Rosenbloom, M Ruvolo, DJ Richter, SF Schaffner, AFA Smit, SM Smith, M Suyama, J Taylor, D Torrents, E Tuzun, A Varki, G Velasco, M Ventura, JW Wallis, MC Wendl, RK Wilson, ES Lander, RH Waterston, and Consortium

Chimpanzee Sequencing Analysis. Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature*, 437(7055) :69–87, 2005. doi : 10.1038/nature04072.

B Milholland, X Dong, L Zhang, XX Hao, Y Suh, and J Vijg. Differences between germline and somatic mutation rates in humans and mice. *Nat Commun*, 8, 2017.

K Misawa. A codon substitution model that incorporates the effect of the GC contents, the gene density and the density of CpG islands of human chromosomes. *BMC Genomics*, 12, 2011.

K Misawa and RF Kikuno. Evaluation of the effect of CpG hypermutability on human codon substitution. *Gene*, 431(1-2) :18–22, 2009.

N Mitarai and S Pedersen. Control of ribosome traffic by position-dependent choice of synonymous codons. *Physical Biology*, 10(5), 2013. doi : 10.1088/1478-3975/10/5/056011.

S Miyazawa. Selective Constraints on Amino Acids Estimated by a Mechanistic Codon Substitution Model with Multiple Nucleotide Changes. *Plos One*, 6(3), 2011. doi : 10.1371/journal.pone.0017244.

CF Mugal, PF Arndt, L Holm, and H Ellegren. Evolutionary Consequences of DNA Methylation on the GC Content in Vertebrate Genomes. *G3 (Bethesda)*, 5(3) :441–447, 2015.

HJ Muller. The Measurement of Gene Mutation Rate in Drosophila, Its High Variability, and Its Dependence upon Temperature. *Genetics*, 13(4) :279–357, 1928.

HJ Muller. Our load of mutations. *American Journal of Human Genetics*, 2(2) :111–176, 1950.

K Munch, T Mailund, JY Dutheil, and MH Schierup. A fine-scale recombination map of the human-chimpanzee ancestor reveals faster change in humans than in chimpanzees and a strong impact of GC-biased gene conversion. *Genome Research*, 24(3) :467–474, 2014.

SV Muse and BS Gaut. A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates, with application to the chloroplast genome. *Mol Biol Evol*, 11(5) :715–24, 1994a.

SV Muse and BS Gaut. A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates, with application to the chloroplast genome. *Mol Biol Evol*, 11(5) :715–724, 1994b.

V Mustonen and M Lassig. From fitness landscapes to seascapes : non-equilibrium dynamics of selection and adaptation. *Trends in Genetics*, 25(3) :111–9, 2009. doi : 10.1016/j.tig.

2009.01.002.

A Muto and S Osawa. The guanine and cytosine content of genomic DNA and bacterial evolution. *Proc Natl Acad Sci USA*, 84(1) :166–169, 1987.

JH Myers. Population cycles : generalities, exceptions and remaining mysteries. *Proceedings of the Royal Society B-Biological Sciences*, 285(1875), 2018.

A Nair, T Fountain, S Ikonen, SP Ojanen, and S van Nouhuys. Spatial and temporal genetic structure at the fourth trophic level in a fragmented landscape. *Proceedings of the Royal Society B-Biological Sciences*, 283(1831), 2016.

Y Nakamura, T Gojobori, and T Ikemura. Codon usage tabulated from international DNA sequence databases : status for the year 2000. *Nucleic Acids Res*, 28(1) :292–292, 2000.

PA Nevarez, CM DeBoever, BJ Freeland, MA Quitt, and EC Bush. Context dependent substitution biases vary within the human genome. *Bmc Bioinformatics*, 11, 2010.

L-T Nguyen, HA Schmidt, A von Haeseler, and BQ Minh. IQ-TREE : A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies. *Molecular Biology and Evolution*, 32(1) :268–274, 11 2014. doi : 10.1093/molbev/msu300.

R Nielsen. Mapping mutations on phylogenies. *Syst Biol*, 51(5) :729–39., 2002.

R Nielsen, VL Bauer DuMont, MJ Hubisz, and CF Aquadro. Maximum likelihood estimation of ancestral codon usage bias parameters in Drosophila. *Mol Biol Ecol*, 24(1) :228–35, 2007.

S Nik-Zainal, LB Alexandrov, DC Wedge, P Van Loo, CD Greenman, K Raine, D Jones, J Hinton, J Marshall, LA Stebbings, A Menzies, S Martin, K Leung, L Chen, C Leroy, M Ramakrishna, R Rance, KW Lau, LJ Mudie, I Varela, DJ McBride, GR Bignell, SL Cooke, A Shlien, J Gamble, I Whitmore, M Maddison, PS Tarpey, HR Davies, E Papaemmanuil, PJ Stephens, S McLaren, AP Butler, JW Teague, G Jonsson, JE Garber, D Silver, P Miron, A Fatima, S Boyault, A Langerod, A Tutt, JWM Martens, SAJR Aparicio, A Borg, AV Salomon, G Thomas, AL Borresen-Dale, AL Richardson, MS Neuberger, PA Futreal, PJ Campbell, MR Stratton, and Consortium Int Canc Genome. Mutational Processes Molding the Genomes of 21 Breast Cancers. *Cell*, 149(5) :979–993, 2012. doi : 10.1016/j.cell.2012.04.024.

T Ohta. Slightly deleterious mutant substitutions in evolution. *Nature*, 246(5428) :96–98, 1973.

J Oksanen, FG Blanchet, M Friendly, R Kindt, P Legendre, D McGlinn, PR Minchin, RB O'Hara, GL Simpson, P Solymos, MHH Stevens, E Szoecs, and H Wagner. *vegan : Community Ecology Package*, 2017.

CA Olson, NC Wu, and R Sun. A Comprehensive Biophysical Description of Pairwise Epistasis throughout an Entire Protein Domain. *Current Biology*, 24(22) :2643–2651, 2014. doi : 10.1016/j.cub.2014.09.072.

S Ossowski, K Schneeberger, JI Lucas-Lledo, N Warthmann, RM Clark, RG Shaw, D Weigel, and M Lynch. The Rate and Molecular Spectrum of Spontaneous Mutations in Arabidopsis thaliana. *Science*, 327(5961) :92–94, 2010. doi : 10.1126/science.1180677.

T Pan. Modifications and functional genomics of human transfer RNA. *Cell Research*, 28 (4) :395–404, 2018. doi : 10.1038/s41422-018-0013-y.

M Parisien, XY Wang, and T Pan. Diversity of human tRNA genes from the 1000-genomes project. *Rna Biology*, 10(12) :1853–1867, 2013. doi : 10.4161/rna.27361.

JL Payne and A Wagner. The causes of evolvability and their evolution. *Nature Reviews Genetics*, 20(1) :24–38, 2019. doi : 10.1038/s41576-018-0069-z.

AMK Pedersen, C Wiuf, and FB Christiansen. A codon-based model designed to describe lentiviral evolution. *Mol Biol Evol*, 15(8) :1069–1081, 1998.

F Pedregosa, G Varoquaux, A Gramfort, V Michel, B Thirion, O Grisel, M Blondel, P Prettenhofer, R Weiss, V Dubourg, J Vanderplas, A Passos, D Cournapeau, M Brucher, M Perrot, and E Duchesnay. Scikit-learn : Machine Learning in Python. *Journal of Machine Learning Research*, 12 :2825–2830, 2011.

S Peischl and KJ Gilbert. Evolution of dispersal can rescue populations from expansion load. *bioRxiv*, 2018. doi : 10.1101/483883.

DA Petrov and DL Hartl. Patterns of nucleotide substitution in Drosophila and mammalian genomes. *Proc Natl Acad Sci USA*, 96(4) :1475–1479, 1999. doi : 10.1073/pnas.96.4.1475.

H Philippe and B Roure. Difficult phylogenetic questions : more data, maybe ; better methods, certainly. *BMC biology*, 9 :91, 2011. doi : 10.1186/1741-7007-9-91.

H Philippe, H Brinkmann, DV Lavrov, DTJ Littlewood, M Manuel, G Worheide, and D Baurain. Resolving Difficult Phylogenetic Questions : Why More Sequences Are Not Enough. *Plos Biology*, 9(3), 2011. doi : ARTNe1000602DOI10.1371/journal.pbio.1000602.

ER Pianka. R-SELECTION AND K-SELECTION. *American Naturalist*, 104(940) :592–&, 1970. doi : 10.1086/282697.

D Pitt, N Sevane, EL Nicolazzi, DE MacHugh, SDE Park, L Colli, R Martinez, MW Bruford, and P Orozco-terWengel. Domestication of cattle : Two or three events ? *Evolutionary Applications*, 12(1) :123–136, 2019. doi : 10.1111/eva.12674.

JB Plotkin, H Robins, and AJ Levine. Tissue-specific codon usage and the expression of human genes. *Proceedings of the National Academy of Sciences of the United States of America*, 101(34) :12588–91, 2004. doi : 10.1073/pnas.0404957101.

D Posada. jModelTest : Phylogenetic model averaging. *Mol Biol Evol*, 25(7) :1253–1256, 2008.

F Pouyet, M Bailly-Bechet, D Mouchiroud, and L Gueguen. SENCA : A Multilayered Codon Model to Study the Origins and Dynamics of Codon Usage. *Genome Biol Evol*, 8(8) :2427–2441, 2016.

F. Pouyet, D. Mouchiroud, L. Duret, and M. Semon. Recombination, meiotic expression and human codon usage. *Elife*, 6, 2017.

D Prangle. *Summary Statistics in Approximate Bayesian Computation*. Chapman & Hall/CRC Handbooks of Modern Statistical Methods. CRC Press, Taylor & Francis Group, 2018.

D Prangle, MGB , Blum, G Popovic, and SA Sisson. Diagnostic tools for approximate Bayesian computation using the coverage property. *Aust N Z J Stat*, 56(4) :309–329, 2014a.

D Prangle, P Fearnhead, MP Cox, PJ Biggs, and NP French. Semi-automatic selection of summary statistics for ABC model choice. *Stat Appl Genet Mol Biol*, 13(1) :67–82, 2014b.

JK Pritchard, MT Seielstad, A Perez-Lezaun, and MW Feldman. Population growth of human Y chromosomes : A study of Y chromosome microsatellites. *Molecular Biology and Evolution*, 16(12) :1791–1798, 1999. doi : 10.1093/oxfordjournals.molbev.a026091.

P Pudlo, JM Marin, A Estoup, JM Cornuet, M Gautier, and CP Robert. Reliable ABC model choice via random forests. *Bioinformatics*, 32(6) :859–866, 2016.

Tessa EF Quax, NJ Claassens, D Söll, and J van der Oost. Codon Bias as a Means to Fine-Tune Gene Expression. *Mol Cell*, 59(2) :149–161, 2015.

DC Queller and JE Strassmann. Evolutionary Conflict. *Annual Review of Ecology, Evolution, and Systematics, Vol 49*, 49 :73–93, 2018. doi : 10.1146/annurev-ecolsys-110617-062527.

R Core Team. *R : A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, 2017.

RA Raff, CR Marshall, and JM Turbeville. USING DNA-SEQUENCES TO UNRAVEL THE CAMBRIAN RADIATION OF THE ANIMAL PHYLA. *Annual Review of Ecology and Systematics*, 25 :351–375, 1994. doi : 10.1146/annurev.es.25.110194.002031.

TR Raffel, LB Martin, and J R Rohr. Parasites as predators : unifying natural enemy ecology. *Trends in Ecology & Evolution*, 23(11) :610–618, 2008. doi : 10.1016/j.tree.2008.06.015.

R Rahbari, A Wuster, SJ Lindsay, RJ Hardwick, LB Alexandrov, S Al Turki, A Dominiczak, A Morris, D Porteous, B Smith, MR Stratton, ME Hurles, and UK K Consortium. Timing, rates and spectra of human germline mutation. *Nature Genetics*, 48(2) :126–133, 2016. doi : 10.1038/ng.3469.

L Raynal, JM Marin, P Pudlo, M Ribatet, CP Robert, and A Estoup. ABC random forests for Bayesian parameter inference. *1605.05537v4 [stat.ME]. https ://arxiv.org/pdf/1605.05537*, 2017.

C Rey, L Gueguen, M Semon, and B Boussau. Accurate Detection of Convergent Amino-Acid Evolution with PCOC. *Molecular Biology and Evolution*, 35(9) :2296–2306, 2018. doi : 10.1093/molbev/msy114.

H Robbins and S Monro. STOCHASTIC APPROXIMATION. *Annals of Mathematical Statistics*, 22(2) :316–316, 1951.

DM Robinson, DT Jones, H Kishino, N Goldman, and JL Thorne. Protein evolution with dependence among codons due to tertiary structure. *Mol Biol Evol*, 20(10) :1692–704, 2003.

EP Rocha, A Danchin, and A Viari. Universal replication biases in bacteria. *Mol Microbiol*, 32(1) :11–6, 1999.

N Rodrigue. On the Statistical Interpretation of Site-Specific Variables in Phylogeny-Based Substitution Models. *Genetics*, 193(2) :557–564, 2013.

N Rodrigue and S Aris-Brosou. Fast Bayesian Choice of Phylogenetic Models : Prospecting Data Augmentation-Based Thermodynamic Integration. *Systematic Biology*, 60(6) :881–886, 2011. doi : 10.1093/sysbio/syr065.

N Rodrigue and N Lartillot. Monte Carlo computational approaches in Bayesian codon substitution modeling. In GM Cannarozzi and A Schneider, editors, *Codon Evolution : Mechanisms and Models*, book section 4, pages 45–59. OUP Oxford, 2012.

N Rodrigue and N Lartillot. Site-heterogeneous mutation-selection models within the PhyloBayes-MPI package. *Bioinformatics*, 30(7) :1020–1021, 2014.

N Rodrigue and N Lartillot. Detecting Adaptation in Protein-Coding Genes Using a Bayesian Site-Heterogeneous Mutation-Selection Codon Substitution Model. *Mol Biol Evol*, 34(1) : 204–214, 2017.

N Rodrigue and H Philippe. Mechanistic revisions of phenomenological modeling strategies in molecular evolution. *Trends Genet*, 26(6) :248–252, 2010.

N Rodrigue, N Lartillot, D Bryant, and H Philippe. Site interdependence attributed to tertiary structure in amino acid sequence evolution. *Gene*, 347(2) :207–217, 2005.

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

N Rodrigue, N Lartillot, and H Philippe. Bayesian Comparisons of Codon Substitution Models. *Genetics*, 180(3) :1579–1591, 2008a. doi : 10.1534/genetics.108.092254.

N Rodrigue, N Lartillot, and H Philippe. Bayesian Comparisons of Codon Substitution Models. *Genetics*, 180(3) :1579–1591, 2008b.

N Rodrigue, CL Kleinman, H Philippe, and N Lartillot. Computational Methods for Evaluating Phylogenetic Models of Coding Sequence Evolution with Dependence between Codons. *Mol Biol Evol*, 26(7) :1663–1676, 2009.

N Rodrigue, H Philippe, and N Lartillot. Mutation-selection models of coding sequence evolution with site-heterogeneous amino acid fitness profiles. *Proc Natl Acad Sci U S A*, 107(10) :4629–4634, 2010.

F Rodriguez, JL Oliver, A Marin, and JR Medina. THE GENERAL STOCHASTIC-MODEL OF NUCLEOTIDE SUBSTITUTION. *Journal of Theoretical Biology*, 142(4) :485–501, 1990. doi : 10.1016/s0022-5193(05)80104-3.

AJ Roger, SA Munoz-Gomez, and R Kamikawa. The Origin and Diversification of Mitochondria. *Current Biology*, 27(21) :R1177–R1192, 2017. doi : 10.1016/j.cub.2017.09.015.

F Ronquist and JP Huelsenbeck. MrBayes 3 : Bayesian phylogenetic inference under mixed models. *Bioinformatics*, 19(12) :1572–4., 2003.

MS Rosenberg, S Subramanian, and S Kumar. Patterns of transitional mutation biases within and among mammalian genomes. *Mol Biol Evol*, 20(6) :988–993, 2003. doi : 10.1093/molbev/msg113.

B Roure and H Philippe. Site-specific time heterogeneity of the substitution process and its impact on phylogenetic inference. *Bmc Evolutionary Biology*, 11, 2011. doi : 10.1186/1471-2148-11-17.

B Roure, D Baurain, and H Philippe. Impact of missing data on phylogenies inferred from empirical phylogenomic data sets. *Mol Biol Evol*, 30(1) :197–214, 2013. doi : 10.1093/molbev/mss208.

AM Sackman, LW McGee, AJ Morrison, J Pierce, J Anisman, H Hamilton, S Sanderbeck, C Newman, and DR Rokyta. Mutation-Driven Parallel Evolution during Viral Adaptation. *Mol Biol Evol*, 34(12) :3243–3253, 2017. doi : 10.1093/molbev/msx257.

D Sakkas, M Ramalingam, N Garrido, and CLR Barratt. Sperm selection in natural conception : what can we learn from Mother Nature to improve assisted reproduction outcomes ? *Human Reproduction Update*, 21(6) :711–726, 2015. doi : 10.1093/humupd/dmv042.

A Saniotis and M Henneberg. Medicine could be constructing human bodies in the future. *Medical Hypotheses*, 77(4) :560–564, 2011.

J Sapp. The struggle for authority in the field of heredity, 1900-1932 : new perspectives on the rise of genetics. *J hist biol*, 16(3) :311–42, 1983.

E Saulnier, O Gascuel, and S Alizon. Inferring epidemiological parameters from phylogenies using regression-ABC : A comparative study. *Plos Computational Biology*, 13(3), 2017. doi : 10.1371/journal.pcbi.1005416.

U Schubert, LC Anton, J Gibbs, CC Norbury, JW Yewdell, and JR Bennink. Rapid degradation of a large fraction of newly synthesized proteins by proteasomes. *Nature*, 404 (6779) :770–4, 2000. doi : 10.1038/35008096.

G Schwarz. Estimating dimension of a model. *Annals of Statistics*, 6(2) :461–464, 1978. doi : 10.1214/aos/1176344136.

C Scornavacca, K Belkhir, J Lopez, R Dernat, F Delsuc, EJP Douzery, and V Ranwez. OrthoMaM v10 : Scaling-Up Orthologous Coding Sequence and Exon Alignments with More than One Hundred Mammalian Genomes. *Molecular biology and evolution*, 36(4) : 861–862, 2019. doi : 10.1093/molbev/msz015.

VB Seplyarskiy, MA Andrianova, and GA Bazykin. APOBEC3A/B-induced mutagenesis is responsible for 20% of heritable mutations in the TpCpW context. *Genome Res*, 27(2) : 175–184, 2017.

PM Sharp, M Averof, AT Lloyd, G Matassi, and JF Peden. DNA-SEQUENCE EVOLUTION - THE SOUNDS OF SILENCE. *Philos Trans R Soc Lond B Biol Sci*, 349(1329) :241–247, 1995.

PM Sharp, E Bailes, RJ Grocock, JF Peden, and R E Sockett. Variation in the strength of selected codon usage bias among bacteria. *Nucleic Acids Res*, 33(4) :1141–53, 2005.

PM Sharp, LR Emery, and K Zeng. Forces that influence the evolution of codon bias. *Philos Trans R Soc Lond B Biol Sci*, 365(1544) :1203–1212, 2010.

DC Shields, PM Sharp, DG Higgins, and F. Wright. Silent sites in drosophila genes are not neutral - evidence of selection among synonymous codons. *Mol Biol Evol*, 5(6) :704–716, 1988.

A Siepel and D Haussler. Phylogenetic estimation of context-dependent substitution rates by maximum likelihood. *Mol Biol Evol*, 21(3) :468–88, 2004.

SA Sisson, Y Fan, and Mark M Tanaka. Sequential Monte Carlo without likelihoods. *Proc Natl Acad Sci USA*, 104(6) :1760–1765, 2007.

SA Smith and CW Dunn. Phyutility : a phyloinformatics tool for trees, alignments and molecular data. *Bioinformatics*, 24(5) :715–716, 2008. doi : 10.1093/bioinformatics/btm619.

SA Smith, NG Wilson, FE Goetz, C Feehery, SCS Andrade, GW Rouse, G Giribet, and CW Dunn. Resolving the evolutionary relationships of molluscs with phylogenomic tools. *Nature*, 480(7377) :364–U114, 2011. doi : 10.1038/nature10526.

SA Smith, NG Wilson, FE Goetz, C Feehery, SCS Andrade, GW Rouse, G Giribet, and CW Dunn. Corrigendum : Resolving the evolutionary relationships of molluscs with phylogenomic tools. *Nature*, 493(708), 2012. doi : 10.1038/nature11736.

TCA Smith, PF Arndt, and A Eyre-Walker. Large scale variation in the rate of germ-line de novo mutation, base composition, divergence and diversity in humans. *Plos Genetics*, 14(3), 2018.

A Stamatakis. RAxML version 8 : a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, 30(9) :1312–1313, 2014. doi : 10.1093/bioinformatics/btu033.

B Steinberg and M Ostermeier. Shifting Fitness and Epistatic Landscapes Reflect Trade-offs along an Evolutionary Pathway. *Journal of Molecular Biology*, 428(13) :2730–2743, 2016. doi : 10.1016/j.jmb.2016.04.033.

A Stoltzfus and DM McCandlish. Mutational Biases Influence Parallel Adaptation. *Mol Biol Evol*, 34(9) :2163–2172, 2017. doi : 10.1093/molbev/msx180.

A Stoltzfus and RW Norris. On the Causes of Evolutionary Transition : Transversion Bias. *Mol Biol Evol*, 33(3) :595–602, 2016. doi : 10.1093/molbev/msv274.

N Sueoka. Correlation between base composition of deoxyribonucleic acid and amino acid composition of protein. *Proc Natl Acad Sci USA*, 47(7) :1141–&, 1961.

N Sueoka. On the genetic basis of variation and hetrogeneity of DNA base composition. *Proc Natl Acad Sci USA*, 48 :582–592, 1962.

Y Suzuki, T Gojobori, and S Kumar. Methods for Incorporating the Hypermutability of CpG Dinucleotides in Detecting Natural Selection Operating at the Amino Acid Sequence Level. *Mol Biol Evol*, 26(10) :2275–2284, 2009.

G Talavera and J Castresana. Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. *Syst Biol*, 56(4) :564–77, 2007.

K Tamura and M Nei. Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. *Mol Biol Evol*, 10(3) :512–26, 1993.

Asif U Tamuri, N Goldman, and M dos Reis. A Penalized-Likelihood Method to Estimate the Distribution of Selection Coefficients from Phylogenetic Data. *Genetics*, 197(1) :257–271, 2014.

AU Tamuri, M dos Reis, and RA Goldstein. Estimating the Distribution of Selection Coefficients from Phylogenetic Data Using Sitewise Mutation-Selection Models. *Genetics*, 190 (3) :1101–1115, 2012.

A Tanay and A Regev. Scaling single-cell genomics from phenomenology to mechanism. *Nature*, 541(7637) :331–338, 2017. doi : 10.1038/nature21350.

C Tang, D Garreau, and U von Luxburg. When do random forests fail ? In *Advances in Neural Information Processing Systems 31*, pages 2983–2993. Curran Associates, Inc., 2018.

S Tavare, DJ Balding, RC Griffiths, and P Donnelly. Inferring coalescence times from DNA sequence data. *Genetics*, 145(2) :505–518, 1997.

Gregg WC Thomas, RJ Wang, A Puri, RA Harris, M Raveendran, DST Hughes, SC Murali, LE Williams, H Doddapaneni, DM Muzny, RA Gibbs, CR Abee, MR Galinski, KC Worley, J Rogers, P Radivojac, and MW Hahn. Reproductive Longevity Predicts Mutation Rates in Primates. *Current Biology*, 28(19) :3193, 2018. doi : 10.1016/j.cub.2018.08.050.

A Tubbs and A Nussenzweig. Endogenous DNA Damage as a Source of Genomic Instability in Cancer. *Cell*, 168(4) :644–656, 2017. doi : 10.1016/j.cell.2017.01.002.

T Tuller, A Carmi, K Vestsigian, S Navon, Y Dorfan, J Zaborske, T Pan, O Dahan, I Furman, and Y Pilpel. An evolutionarily conserved mechanism for controlling the efficiency of protein translation. *Cell*, 141(2) :344–54, 2010.

J Van den Eynden and E Larsson. Mutational Signatures Are Critical for Proper Estimation of Purifying Selection Pressures in Cancer Somatic Mutation Data When Using the dN/dS Metric. *Front Genet*, 8 :74, 2017.

DA Vaughan and D Sakkas. Sperm selection methods in the 21st century. *Biology of reproduction*, 2019. doi : 10.1093/biolre/ioz032.

A Venkat, MW Hahn, and Joseph W Thornton. Multinucleotide mutations cause false inferences of lineage-specific positive selection. *Nature Ecology & Evolution*, 2(8) :1280–1288, 2018. doi : 10.1038/s41559-018-0584-5.

GP Wagner and L Altenberg. Perspective : Complex adaptations and the evolution of evolvability. *Evolution*, 50(3) :967–976, 1996. doi : 10.2307/2410639.

J Wakeley. The excess of transitions among nucleotide substitutions : New methods of estimating transition bias underscore its significance. *Trends in Ecology & Evolution*, 11 (4) :158–163, 1996. doi : 10.1016/0169-5347(96)10009-4.

HC Wang, Q Minh, E Susko, and AJ Roger. Modeling Site Heterogeneity with Posterior Mean Site Frequency Profiles Accelerates Accurate Phylogenomic Estimation. *Syst Biol*, 67(2) :216–235, 2018.

KY Wang, CC Chen, SF Tsai, and CKJ Shen. Epigenetic Enhancement of the Post-replicative DNA Mismatch Repair of Mammalian Genomes by a Hemi-(m)CpG-Np95-Dnmt1 Axis. *Scientific Reports*, 6, 2016. doi : 10.1038/srep37490.

JD Watson and FCH Crick. A structure for deoxyribose nucleic acid. *Nature*, 171 :737–738, 1953.

GR Webster, AYH Teh, and JKC Ma. Synthetic gene designThe rationale for codon optimization and implications for molecular pharming in plants. *Biotechnol Bioeng*, 114(3) : 492–502, 2017.

G Weiss and A von Haeseler. Inference of population history using a likelihood approach. *Genetics*, 149(3) :1539–1546, 1998.

WSW Wong, BD Solomon, DL Bodian, P Kothiyal, G Eley, KC Huddleston, R Baker, DC Thach, RK Iyer, JG Vockley, and JE Niederhuber. New observations on maternal age effect on germline de novo mutations. *Nat Commun*, 7, 2016.

S Wright. Evolution in Mendelian Populations. *Genetics*, 16(2) :97–159, 1931.

S Wright. The roles of mutation, inbreeding, crossbreeding and selection in evolution. *Proceedings of the Sixth International Congress of Genetics*, 1 :356–366, 1932.

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

Z Yang. PAML 4 : phylogenetic analysis by maximum likelihood. *Mol Biol Evol*, 24(8) : 1586–91, 2007a.

Z Yang. Paml 4 : phylogenetic analysis by maximum likelihood. *Mol Biol Evol*, 24(8) : 1586–91, 2007b. doi : 10.1093/molbev/msm088.

Z Yang and R Nielsen. Mutation-selection models of codon substitution and their use to estimate selective strengths on codon usage. *Mol Biol Evol*, 25(3) :568–579, 2008.

ZH Yang and B Rannala. Molecular phylogenetics : principles and practice. *Nature Reviews Genetics*, 13(5) :303–314, 2012. doi : 10.1038/nrg3186.

H Ying and G Huttley. Exploiting CpG Hypermutability to Identify Phenotypically Significant Variation Within Human Protein-Coding Genes. *Genome Biol Evol*, 3 :938–949, 2011.

WP You and M Henneberg. Cancer incidence increasing globally : The role of relaxed natural selection. *Evolutionary Applications*, 11(2) :140–152, 2018.

Y Zhou, H Brinkmann, N Rodrigue, N Lartillot, and H Philippe. A dirichlet process covarion mixture model and its assessments using posterior predictive discrepancy tests. *Mol Biol Evol*, 27(2) :371–84, 2010.

S Zoller and A Schneider. A New Semiempirical Codon Substitution Model Based on Principal Component Analysis of Mammalian Sequences. *Molecular biology and evolution*, 29 (1) :271–277, 2012. doi : DOI10.1093/molbev/msr198.

X Zou, M Owusu, R Harris, SP Jackson, JI Loizou, and S Nik-Zainal. Validating the concept of mutational signatures with isogenic cell models. *Nature Communications*, 9, 2018. doi : 10.1038/s41467-018-04052-8.