Modelling the articulation of selective and neutral mechanisms in the evolution of protein-coding DNA sequences

Thibault Latrille

N° d'ordre NNT:
2020LYSE1228

THÈSE de DOCTORAT DE L'UNIVERSITÉ DE LYON

Opérée au sein de :

l'Université Claude Bernard Lyon 1

Ecole Doctorale 341
Écosystèmes Évolution Modélisation Microbiologie

Spécialité de doctorat : Génomique évolutive

Soutenue publiquement le 30/11/2020, par :

Thibault Latrille

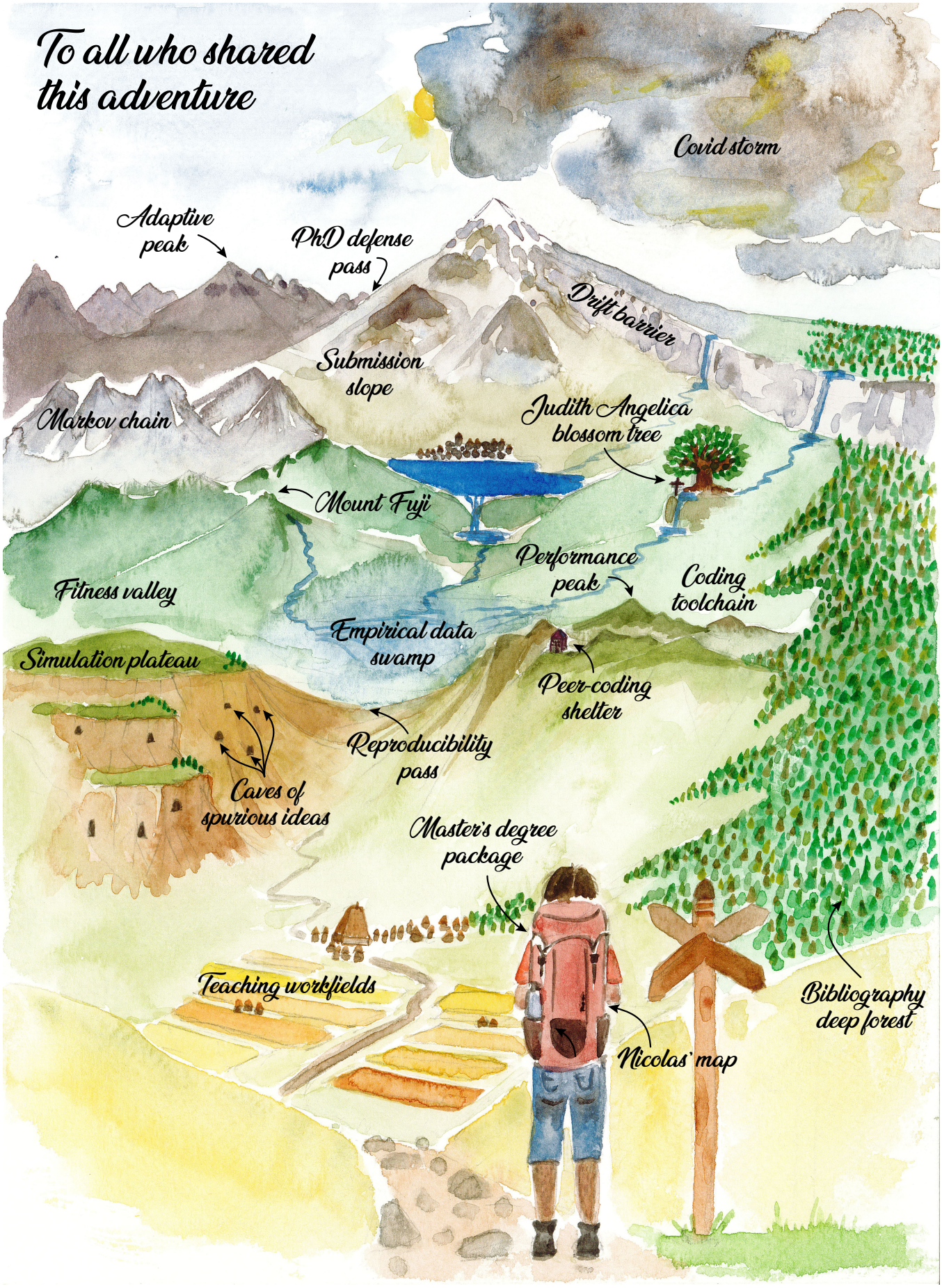
Modélisation de l'articulation des mécanismes sélectifs et neutres dans l'évolution des séquences d'ADN codant pour des protéines.

Devant le jury composé de :

Celine BROCHIER-ARMANET Professeure, Université Claude Bernard Lyon 1	**Présidente**
Julien Yann DUTHEIL Research Group Leader, Max Planck Institute (Allemagne)	**Rapporteur**
Richard GOLDSTEIN Professeur, University College London (Royaume-Uni)	**Rapporteur**
Carina Farah MUGAL Chercheure, Uppsala University (Suède)	**Rapporteure**
Nicolas LARTILLOT Directeur de recherche, CNRS/LBBE	**Directeur de thèse**

L'humanité est constamment aux prises avec deux processus contradictoires dont l'un tend à instaurer l'unification, tandis que l'autre vise à maintenir ou à rétablir la diversification.

Claude Lévi-Strauss.


To all who shared this adventure
Covid storm
Adaptive peak
PhD defense pass
Drift barrier
Submission slope
Markov chain
Judith Angelica blossom tree
Mount Fuji
Performance peak
Coding toolchain
Fitness valley
Empirical data swamp
Simulation plateau
Peer-coding shelter
Reproducibility pass
Caves of spurious ideas
Master's degree package
Teaching workfields
Bibliography deep forest
Nicolas' map

Résumé

L'évolution moléculaire vise à caractériser les mécanismes à l'œuvre dans l'évolution des séquences, régie par un processus stochastique dont les principaux composants sont la mutation, la sélection et la dérive génétique. À long terme, ce processus stochastique se traduit par une histoire d'événements de substitutions le long des arbres d'espèces, induisant des motifs complexes de divergence moléculaire entre les espèces. En analysant ces divergences, les modèles de codons phylogénétiques visent à capturer les paramètres intrinsèques de l'évolution. Dans ce contexte, cette thèse s'est concentrée sur les modèles à codons phylogénétiques et sur la modélisation de l'interaction entre la mutation, la sélection et la dérive génétique dans les séquences d'ADN codant pour des protéines. Parce que la composition de ces séquences ne reflète pas le processus de mutation sous-jacent, mais son filtrage par sélection au niveau des acides aminés, une modélisation minutieuse est nécessaire pour démêler la mutation et la sélection. Ainsi, j'ai développé un modèle d'inférence phylogénétique dans lequel différents taux d'évolution donnent une représentation précise de la manière dont la mutation et la sélection s'opposent à l'équilibre. Deuxièmement, l'équilibre entre mutation et sélection est arbitré par la dérive génétique, qui est médiée par la taille efficace de la population, et ses changements le long d'une phylogénie peuvent être déduits des motifs de substitutions le long des lignées. J'ai ainsi développé un deuxième modèle d'inférence, reconstituant à la fois le paysage de fitness en chaque site, les tendances à long terme de taille efficace de population et les changements de taux de mutation le long de la phylogénie. Ce cadre bayésien a été testé sur des données simulées puis appliqué à des données empiriques. Les estimations de la variation de taille efficace de population correspondent à la direction attendue de la corrélation avec les traits d'histoire de vie ou les variables écologiques, bien que l'ampleur de la variation de la taille efficace de population estimée soit étroite. Afin de comprendre cette variation étroite de la taille efficace de population estimée, j'ai finalement développé un modèle théorique décrivant comment les changements à la fois de taille efficace de population ou du niveau d'expression de la protéine se traduisent par un changement du taux de substitution, sous l'hypothèse que les protéines sont sous sélection directionnelle pour maximiser leur stabilité conformationnelle. Cette réponse est déterminée en fonction des paramètres moléculaires de la biophysique des protéines, et implique une faible réponse du taux de substitution aux changements de niveau d'expression ou de taille efficace de population dans ce contexte. Ce travail démontre que les hypothèses faites sur la structure du paysage de fitness ont une importance critique sur la sensibilité des changements de vitesse d'évolution à des changements de variables écologiques ou moléculaires. Réciproquement, les observations empiriques des motifs de substitutions en réponse à des changements de variables moléculaires ou écologiques nous informent sur la structure sous-jacente du paysage de fitness. En se basant sur l'équilibre mutation-sélection et en intégrant explicitement la taille efficace de population, ce travail présente aussi un cadre conceptuel permettant de relier phylogénie et génétique des populations, dont certaines pistes d'unifications sont envisagées.

Résumé étendu

La théorie neutre de l'évolution a influencé notre compréhension de la génétique des populations et de l'évolution moléculaire. Au-delà des disputes et des controverses entre neutralisme et sélectionnisme, le consensus actuel est de considérer l'évolution des séquences génétiques comme un processus stochastique combinant mutation, sélection et dérive génétique. Les mutations sont source de diversité génétique. La sélection, quant à elle filtre cette diversité. Enfin, l'équilibre entre mutation et sélection est arbitré par la dérive génétique, déterminé par la taille efficace de population (N_e). Sur la longue durée évolutive, mutation, sélection et dérive génétique résultent en une accumulation de substitutions ponctuelles entre les espèces, qui dans les séquences codantes peuvent être soit synonymes, soit non synonymes. S'appuyant ainsi sur ces différences interspécifiques, telles qu'observées dans les alignements multiples de séquences d'ADN codant pour des protéines, l'objectif des modèles à codons phylogénétiques est de mieux caractériser et quantifier les processus mutationnels et sélectifs et de mieux comprendre leur articulation. Les modèles à codons sont toujours un domaine de recherche actif et se scindent en deux philosophies différentes. D'un côté, les modèles phénoménologiques visent à capturer l'effet net de la sélection s'exerçant sur toutes les mutations non synonymes au sein de la protéine, à travers un seul paramètre. De l'autre côté, des approches mécanistes ont pour objectif de capturer l'effet de la sélection sur chaque mutation non synonyme prise individuellement, ce qui requiert de modéliser explicitement le paysage du fitness sous-jacent. En l'état, cependant, de nombreuses questions restent ouvertes et les modèles actuels, qu'ils soient phénoménologiques ou mécanistes, présentent de nombreuses faiblesses. Les approches phénoménologiques n'articulent pas explicitement la relation entre mutation, sélection et dérive génétique, et pourraient encore être améliorées, tout en restant dans l'idée de ne pas modéliser explicitement le paysage sélectif dans ses détails. Quant aux approches mécanistes, dans leurs versions actuelles, elles font des hypothèses très fortes, telles que l'indépendance entre sites, un paysage de fitness fixe au cours du temps, mais aussi une taille efficace de population (N_e) constante le long de la phylogénie. Plus fondamentalement, il existe un certain vide à combler entre ces approches phénoménologiques et mécanistes, et de meilleures connexions conceptuelles et pratiques pourraient être établies entre elles.

Dans ce contexte, mon travail de thèse représente une tentative de démêler les interactions complexes entre mutation, sélection et dérive génétiques en construisant de nouveaux modèles à codons phylogénétiques, selon les deux approches, phénoménologiques et mécanistes. Au cours de ce travail, j'ai lié des idées théoriques à des données empiriques, en utilisant une combinaison d'approches analytiques, d'expériences de simulation de développements statistiques et informatiques utilisant les principes de l'inférence bayésienne par chaines de Markov Monte-Carlo. Les résultats sont divisés en trois manuscrits indépendants, sur le point d'être soumis à des journaux à comité de lecture.

Le premier article revient sur la question de l'équilibre entre biais de mutation et biais de sélection, et de comment cet équilibre doit être correctement formalisé dans le contexte des modèles à codons phénoménologiques. Parce que la composition des séquences d'ADN codant pour les protéines ne reflète pas le processus sous-jacent de mutation, mais son filtrage par sélection au niveau des acides aminés, une modélisation minutieuse est nécessaire pour démêler le processus de mutation et les biais nucléotidiques d'un côté, et la sélection d'un autre côté. Malheureusement, les modèles à codons phénoménologiques actuels, développés à l'origine pour estimer la pression de sélection s'exerçant sur les protéines, ne modélisent pas correctement cet équilibre mutation-sélection. En effet, ils utilisent le biais de composition nucléotidique observé comme proxy pour le biais mutationnel. En conséquence, ils ne fournissent pas une estimation précise du processus de mutation, même s'ils sont capables d'estimer de manière assez fiable la pression de sélection agissant sur les acides aminés. Pour résoudre ce problème, j'ai développé un modèle à codon phylogénétique dans lequel la pression de sélection n'est pas considérée comme un paramètre unique, mais comme un tenseur (95 paramètres libres). Le tenseur capture les faibles différences de pression de sélections dans différentes directions, ce qui donne une représentation précise de la manière dont la mutation et la sélection s'opposent à l'équilibre. Cette paramétrisation représente la forme paramétrique la plus simple, dans un contexte phénoménologique, capable de séparer les effets de la mutation et de la sélection de manière exacte, ou asymptotiquement exacte. Grâce à cela, cette approche de modélisation donne une estimation fiable du processus de mutation, tout en démêlant les pressions de sélection dans différentes directions. Ces développements offrent des outils qui permettront ultimement de mieux comprendre comment le processus mutation-mutation s'articule avec d'autres processus évolutifs impactant la composition nucléotidique, tels que la conversion génique biaisée (gBGC).

Si le premier manuscrit se focalise sur l'articulation entre mutation et sélection, l'équilibre entre ces deux forces est arbitré par la dérive génétique, qui à son tour est modulée par taille efficace de population (N_e). En conséquence, théoriquement, la variation de N_e le long d'une phylogénie peut être déduite de l'histoire des substitutions le long des lignées. Le deuxième manuscrit explore ainsi la question de la prise en compte des variations à long terme de la taille efficace de population (N_e) entre les espèces, dans le contexte d'un modèle à codons mécaniste. Les travaux présentés dans ce second manuscrit représentent la partie la plus intensive du travail de doctorat, en matière de modélisation, d'algorithmes de Monte-Carlo et de développement logiciel. J'ai ainsi développé un modèle à codons mécaniste reconstituant le paysage de fitness en chaque site, les tendances à long terme de la taille efficace de population et du taux de mutation le long de la phylogénie, à partir d'alignements d'ADN de séquences codantes. Simultanément, l'approche estime la corrélation entre les traits d'histoires de vie, le taux de mutation et la taille efficace de population, prenant explicitement en compte l'inertie phylogénétique. Ce modèle a été testé sur des données simulées, puis appliqué à des données empiriques chez les mammifères, les isopodes, les primates et les drosophiles. Les résultats sur données simulées et empiriques suggèrent qu'il existe des signaux per-

sistants dans les séquences d'ADN qui permettent de reconstruire l'histoire évolutive du N_e le long de la phylogénie. Par ailleurs, les variations de taille efficace de population inférées corrèlent avec les traits d'histoire de vie ou les variables écologiques d'une façon qui est attendue d'après les connaissances écologiques disponibles par ailleurs. Cependant, l'ampleur de la variation inférée de N_e à travers la phylogénie est plus étroite que prévu, si l'on compare en particulier aux estimés sur la base du polymorphisme.

Cette dernière observation, qui suggère une violation de certaines hypothèses du modèle, m'a amené à revoir la question de savoir comment la biophysique des protéines, et plus généralement l'épistasie, peut moduler quantitativement la réponse du processus évolutif moléculaire aux changements de la taille efficace de population. Ce dernier travail est présenté comme un troisième manuscrit. En effet, les hypothèses sur la structure sous-jacente du paysage de fitness peuvent avoir une grande influence sur la vitesse d'évolution des protéines, et tout particulièrement sur les changements de cette vitesse d'évolution après un changement de N_e. En plus de N_e, le niveau d'expression des protéines est un autre facteur majeur susceptible de moduler la vitesse d'évolution moléculaire. Les protéines fortement exprimées évoluent généralement moins vite, une corrélation prédite par les modèles biophysiques supposant que les protéines mal repliées sont toxiques et donc soumises à une sélection purificatrice. En conséquence, il convient d'articuler ensemble toutes ces corrélations entre la vitesse d'évolution, la taille efficace de population et le niveau d'expression, en rapport avec la structure du paysage de fitness sous-jacent. Pour ce faire, j'ai dérivé une approximation théorique de la réponse quantitative de vitesse d'évolution à des changements à la fois de N_e et du niveau d'expression, en fonction de la relation génotype-phénotype-fitness sous-jacente. Ce développement est généralement valide pour des traits phénotypiques additifs et une fonction de fitness concave, mais a été appliqué plus spécifiquement à un modèle biophysique dans lequel les protéines sont sous sélection directionnelle pour maximiser leur stabilité conformationnelle. Dans ce cas précis, le modèle prédit une réponse faible du taux d'évolution aux changements de N_e ou de niveau d'expression (qui sont interchangeable), un résultat corroboré par des simulations sous des modèles plus complexes. Sur la base de preuves empiriques, je propose que l'adéquation basée sur la stabilité conformationnelle puisse ne pas fournir un mécanisme suffisant pour expliquer l'amplitude des variations de la vitesse d'évolution observée empiriquement, entre protéines ou entre espèces, induites par les variations de niveau d'expression ou de taille efficace de population. D'autres aspects de la biophysique des protéines pourraient être explorés tels que la sélection pour limiter les interactions non spécifiques entre protéines. Ces aspects pourraient conduire à une réponse plus forte de la vitesse d'évolution aux changements de N_e. Plus généralement, ce travail offre des perspectives pour réduire l'écart entre les prévisions quantitatives des modèles biophysiques et les observations empiriques reliant la réponse de la pression de sélection aux changements de N_e et du niveau d'expression.

Pour conclure, ce travail est une tentative encourageante, quoiqu'encore inaboutie de construire des modèles intégrés d'évolution des séquences d'ADN codant pour les protéines. Ce travail réussit à consolider l'idée que les motifs de substitutions nous informent sur les fluctuations à long terme de la dérive génétique le long des branches et la sélection le long des séquences. Il démontre que les hypothèses faites sur la structure du paysage de fitness ont une importance critique sur la sensibilité des changements vitesse d'évolution à des changements de variables écologiques (N_e) ou de moléculaires (niveau d'expression des protéines). Réciproquement, les observations empiriques des motifs de substitutions en réponse à des changements de variables moléculaires ou écologiques nous informent sur la structure sous-jacente du paysage de fitness. En se basant sur l'équilibre mutation-sélection et en intégrant explicitement la taille efficace de population, ce travail présente aussi un cadre conceptuel permettant de relier phylogénie et génétique des populations, dont certaines pistes d'unifications sont envisagées. Enfin, je pense que cette thèse consolide les modèles théoriques sur lesquels se fonde l'évolution moléculaire et souligne les écueils à éviter, tout en donnant des perspectives pour le développement de méthodes d'inférence permettant d'intégrer différentes données empiriques et niveaux de complexité.

Remerciements

À celles et ceux qui ont parcouru avec moi un bout du chemin lors de cette expédition, vous m'avez fait découvrir des lieux d'émerveillements, autant hors des sentiers battus que sur les lieux pittoresques, vous m'avez sorti des impasses et guidé à travers les ravines et les combes.

Un immense merci à...

I'm immensely thankful to...

Estoy inmensamente agradecido con ...

... bien évidemment, celui qui a dessiné la carte, et sillonné ce périple à mes côtés, Nicolas Lartillot. Tu m'as guidé tout au long de cette traversée, ta vision d'ensemble ainsi que ta précision dans les détails techniques nous ont amenés sur des pics et des vallées que je n'aurai jamais imaginées exister. Tes discussions et échanges autant scientifiques qu'humains ont sillonné d'innombrables sujets, nourri par ton indéniable sens aiguisé de l'observation et tes incroyables qualités d'empathie. J'espère qu'on aura l'occasion de refaire une ou plusieurs randonnées ensemble.

... my jury who accepted reviewing this thesis, Céline Brochier-Armanet, Julien Yann Dutheil, Richard Goldstein, Carina Farah Mugal. These pages are a map of a three years' scientific journey spent in Lyon, I hope you'll be enjoying the views of fitness landscapes, adaptive peaks and Markov chains that Nicolas and I visited along the way.

... vous qui peuplez la vallée du LBBE, autant les habitants endémiques que les individus migrateurs campant le temps d'un passage. Vous avez construit et entretenu une communauté curieuse, incroyablement compétente et bienveillante, merci.

... celles et ceux que j'ai croisés le long de votre aventure et qui se sont envolés vers une autre vallée, Aline Muyle, Wandrille Duchemin, Adrián Arellano Davín, Héloïse Phillippon, Pierre Garcia, Monique Aouad, Anne Oudard, Frédéric Jauffrit, Maud Gautier, Samuel Barreto, Vincent Lanore, François Gindraud et Diego Hartasánchez Frenk.

... vous qui profitez encore du paysage et de la sagesse des résidents du LBBE encore quelque temps, Florian Bénitière, Alexandre Laverré, Alexia Nguyen Trung, Djivan Prentout, Théo Tricou, Marina Brasó Vives, Claire Gayral, Hugo Menet, Louis Duchemin, Antoine Villié, Alice Genestier et Julien Joseph. Je n'oublierai point que nous avons ensemble parcouru les sinuosités des Dombes, des gorges, des canyons, des lacs et même des grottes, autant au sens littéral que figuré.

... celles et ceux qui ont construit le refuge, Anamaria Necsulea, Damien de Vienne, Dominique Mouchiroud, Hélène Badouin, Laurent Duret, et tant d'autres. Ainsi que vous qui avez semé les poireaux et les carottes, Annabelle Haudry, Bastien Boussau, Éric Tannier et Vincent Daubin.

... celles qui m'ont guidé à travers la jungle et les ronces administratives, Nathalie Arbasetti, Odile Mulet-Marquis, Laetitia Catouaria et Aurélie Zerfass, j'en ressors sans trop d'égratignures grâce à vous, alors que nous savons pertinents que j'y aurai laissé des plumes si vous n'étiez pas là pour fournir votre aide si précieuse.

... ceux qui m'ont fourni les outils et le matériel de randonnée et d'escalade, et en

plus m'ont appris à m'en servir en toute sécurité, Bruno Spataro, Stéphane Delmotte, Simon Penel, Adil El Filali, Vincent Miele, Aurélie Siberchicot et Philippe Veber.

... celles et ceux avec qui j'ai participé aux formations des futurs grimpeuses et grimpeurs, Marie Sémon, Carine Rey, Corentin Dechaut, Vincent Lacroix, Arnaud Mary, et tant d'autres déjà cités plus haut. Ainsi qu'aux enseignants de l'Université de Montpellier, Catherine Moulia et toute la cordée, pour m'avoir fait confiance et permettre de venir chaque année en pèlerinage dans les entrailles et tortuosités des controverses.

... à Diego pour tes remarques et commentaires aiguisés, et encore merci pour avoir relu ce manuscrit et corrigé d'innombrables fautes que j'avais laissées jalonner ce périple.

... aux membres du comité de pilotage pour avoir éclairé les passages escarpés, Christophe Douady, Benoit Nabholz, Tristan Lefébur, Laurent Gueguen et Laurent Duret.

... à Nicolas Rodrigue pour les visites et escapades en territoire lyonnais, j'espère pouvoir refaire des excursions avec toi aussi vite que possible, ici ou bien chez toi.

... au Ministère de l'Enseignement Supérieur et de la Recherche, et en réalité à la société pour avoir sponsorisé cette expédition, et financer les victuailles, le matériel, le transport, et le logement.

... à toute la tribu familiale, d'amies et d'amis pour votre confiance et tous ces moments chaleureux que j'ai passés à vos côtés. Autant celles et ceux qui ont déjà soutenu que ceux qui y aspirent, en passant par ceux qui en sont curieux et intéressés, sans oublier ceux qui me demandent pourquoi je fais ça. J'espère vous donner l'envie d'explorer quelques pages de ce paysage, je vous promets qu'au travers de cette marche (aléatoire) il y aura vraiment des arbres (phylogénétiques), des vallées (de fitness), des pics (adaptatifs), des chaînes (de Markov), des barrières (de dérive), des champs (moyens), et tant d'autres curiosités.

... à toute la famille, maman et papa qui m'ont rendu diploïde, et m'ont appris à lire une carte, sans -trop- se perdre et aussi pour m'avoir appris à retrouver un chemin lorsque l'on est perdu. Iris et Myriam, qui ont découvert d'éblouissants et incroyables recoins de paradis, vous avez bien fait de suivre votre instinct et ne pas avoir suivi la route toute tracée, instinct qui ne vous aurez certainement pas amené à élire domicile dans vos oasis si hospitalières et bourgeonnantes. Samuel, qui m'a ravitaillé en patate tout le long de la dernière montée depuis son refuge branché à la fibre.

... Iris pour cette magnifique aquarelle en préface, ce fut un véritable émerveillement de voir se réaliser sous ton pinceau la représentation figurée que je me suis fait de la thèse.

... à ma femme Judith Alexandra, qui a partagé l'ensemble de cette aventure avec moi, sur les plus somptueux pics et les gorges les plus escarpées. Après ce col, je ne sais pas dans quelle vallée nous irons voyager et randonner ensemble, mais avec toi je n'ai pas peur d'y aller même si nous savons pertinemment que Lyon et sa tribu chaleureuse nous manqueront.

... Judith Angelica por su amor. Querida suegra, has regresado a la tierra pero quiero que sepas que tu recuerdo y tus enseñanzas siempre serán parte de nuestro camino.

Thibault Latrille

Modelling the interplay between selective and neutral mechanisms in the evolution of protein-coding DNA sequences

Abstract

Molecular evolution aims to characterize the mechanisms at work in the evolution of genetic sequences. This evolution is governed by a stochastic process whose main components are mutation, selection and genetic drift. In the long term, this stochastic process results in a history of substitution events along species trees, inducing complex patterns of molecular divergence between species. By analysing them, phylogenetic codon models aim at capturing the intrinsic parameters of evolution. In this context, this thesis has been focused on phylogenetic codon models, and on how they can be used to understand the interplay between mutation, selection and drift in shaping protein-coding DNA sequences.

Because the composition of protein-coding DNA sequences does not reflect the underlying mutational process, but its filtering by selection at the level of amino acids, a careful modelling approach is necessary to tease apart mutation and selection. Current codon models are inherently misspecified in this respect and, as a result, do not return accurate estimates of mutation biases. Therefore, I first developed a phylogenetic codon model in which the ratio of the non-synonymous over the synonymous substitution rates is modelled as a tensor, rather than a scalar. This model gives a more accurate representation of how mutation and selection oppose each other at equilibrium and yields accurate estimates of the mutation bias.

Second, the balance between the opposing forces of mutation and selection is arbitrated by genetic drift, which in turn is modulated by effective population size. As a consequence, variation in effective population size along of a phylogeny can theoretically be inferred from the trails of substitutions along the lineages. I thus developed a second model of inference, jointly reconstructing site-specific fitness landscapes and the variation in effective population size and in the mutation rate along the phylogeny. This Bayesian framework was tested against simulated data and then applied to empirical data. Estimated lineage-specific ancestral population sizes show the expected correlation with life-history traits or ecological variables. However, the magnitude of the inferred variation is narrower than expected based on independent estimates.

In order to understand this narrow variation in the estimated effective population sizes, and the possible role of epistasis in this outcome, i finally developed a theoretical model describing how changes in both effective population size or expression level of protein translate into a change in the substitution rate, and how this response depends on the underlying sequence-phenotype-fitness map. I more specifically explored a biophysical model assuming that proteins are under directional selection to maximize their conformational stability. Results of this theoretical and simulation work imply a weak response (or susceptibility) of the substitution rate to changes in expression level

or effective population size (which are interchangeable). Theoretical approximations were also developed, expressing this susceptibility as a function of the parameters of the biophysical model. Finally, these quantitative estimates are discussed in the light of current empirical knowledge.

Altogether, this thesis demonstrates that the assumptions made on the structure of the fitness landscape have a critical importance on the sensitivity of the substitution rate to changes in ecological or molecular variables. Conversely, empirical observations of the patterns of substitutions in response to changes in molecular or ecological variables inform us about the underlying structure of the fitness landscape. Being based on the mutation-selection balance and by explicitly integrating effective population size, my work also presents a conceptual framework relating phylogenetic and population genetics, while proposing conceptual and methodological paths in order to achieve their unification.

Contents

List of Figures

List of Tables

Acronyms

CDS Coding DNA Sequence. 43

DFE Distribution of Fitness Effects. 22, 145

DNA DeoxyriboNucleic Acid. 1, 5, 6, 8, 12, 33–35, 37, 46, 52, 62, 66, 70, 72, 74–76, 82, 85, 90–92, 140, 141, 146, 152

gBGC GC Biased Gene Conversion. 75, 81, 91, 92, 147

GTR Generalized Time Reversible. 38, 58, 88, 97

HB Halpern & Bruno. 48, 49

LHT Life-History Trait. 76, 143

MC Markov Chain. 59, 61

MCMC Markov Chain Monte Carlo. 59, 61, 148

MF Mean-Field. 88–90, 97

MG Muse & Gaut. 48, 87–90

MK McDonald & Kreitman. 145, 146

PCR Polymerase Chain Reaction. 5

RNA RiboNucleic Acid. 34

SFS Site Frequency Spectrum. 21, 22, 145

tRNA transfer RNA. 34, 35

Glossary

allele A variant form of a given gene. 4, 6, 7, 9, 14–27, 56

codon Sequence of three nucleotides coding for a given amino acid. 5, 9, 23, 28, 33–41, 44–56, 61–63, 67, 69–71, 73–76, 80–84, 86–89, 91–97, 140–148, 152

codon usage bias Unequal frequency of the alternative codons that specify the same amino acid. 47, 92

Dirichlet process Family of stochastic processes whose realizations are probability distributions. 46, 58

effective population size The number of individuals in a population who contribute to the next generation. 7–10, 17, 18, 20, 69, 71, 73–77, 141, 148

genetic drift The random fluctuation in allele frequencies due to random sampling of individuals. 7, 8, 17, 19, 27, 80, 152

likelihood Probability of observing the data given the parameters of the statistical model. This is a function of the parameters of the model, the observed data are fixed. 37, 41, 52–57, 59–61, 87, 88, 97, 148

Markov chain Stochastic process with property that the next state of the process depends only on the present state of the process and not on its past. 23–27, 59, 60

Markov chain Monte Carlo Class of algorithms for sampling from a probability distribution based on constructing a Markov chain that has the desired distribution as its equilibrium distribution. 29, 52, 59

nearly-neutral Slightly deleterious or advantageous mutations are effectively neutral when their selection coefficient are lower than one divided by the effective population size. 7–11, 44, 49–51, 67, 69, 73, 77, 140, 143–145

neutral Mutations are nor deleterious neither advantageous, their probability of fixation is one divided by twice the effective population size (for diploids). 6–8, 10, 11, 17, 19, 20, 24, 28, 37, 38, 40, 47, 49, 54, 73, 76, 80–82, 93, 94, 147

non-synonymous Transition that modifies the encoded amino acid. 8–10, 28, 34–41, 44, 47–49, 66, 80, 82, 84, 85, 87–91, 94–97, 144, 145, 149

phenotype The composite of observable traits. 7, 8, 11, 47, 66, 67, 143

polymorphic Which presents several forms. Subject to inter-individual variability. 14, 22

posterior Probability distribution of a parameter of the model conditioned on randomly observed data and the prior distribution. 41, 57, 59, 60

prior Probability distribution that would express one's beliefs about a parameter of the model before the data is taken into account. 54, 56–59, 94

recombination Exchange of DNA sequence information. 21, 23

substitution Point mutation that appeared in only one individual in the population, and subsequently reached fixation in the population. 6–8, 10, 11, 14, 22–26, 28, 35, 37–42, 44–54, 57, 58, 61, 67, 68, 71, 73, 76, 80–82, 84, 85, 87, 88, 92–96, 140–143, 145–149, 152

synonymous substitution that does not modify the encoded amino acid. 9, 10, 34–41, 44, 45, 47, 48, 50, 73, 77, 80–84, 87, 88, 96, 145, 149

Preamble

The diversity of living organisms today is the result of a complex and intricate process, which operates at multiple levels. At the molecular level, the fate of a protein depends on its ability to fold but also on the enzymes it encounters from its creation up to its degradation. Composed of millions of proteins, a cell's own fate depends on its own ability to metabolize substrates and copy its DNA, but also on the fate of surrounding cells and the individual to which it belongs. Moreover, the fate of this individual depends on its own behaviour, but also on its environment and the population to which it belongs. Altogether, scientists have dissected this intricate process into its core components, through molecular biology, enzymology, metabolism, physiology, population genetics, ecology, and so on. Molecular evolution seeks to encompass different levels, relating molecular changes to higher-level evolutionary processes. In this vein, this work is a modest attempt to reconcile several layers of evolution, mechanistically deriving how observable parameters between populations and within populations depends in microscopic molecular and cellular parameters. Altogether, through the framework of population genetics, I seek to draw connections between independent datasets, from molecular parameters of protein biophysics, to diversity and divergence of DNA sequences within and between species, while relating to species' quantitative life-history traits.

This thesis is submitted in partial fulfilment of the requirements for the degree of *Philosophiae Doctor* at the Université de Lyon. The research presented here was conducted at the Laboratoire de Biométrie et Biologie Evolutive (LBBE), under the supervision of research director M. Nicolas Lartillot. This work was conducted from September 2017 onward during a 3-year grant by ENS de Lyon (Contrat Doctoral Spécifique Normalien). The thesis is a collection of three manuscripts preceded by an introduction that relates them and provides background information and motivation for the work.

Part I

Introduction

1

Historical perspective on molecular evolution

Contents

From the discovery of evolution to today's knowledge, the understanding of the mechanisms by which the diversity and the complexity of living forms emerge has seen dramatic changes and has gone through several scientific revolutions. Molecular evolutionary sciences represent one such revolution, a relatively recent scientific development emerging at the crossroads of two scientific fields. On the one hand, evolutionary biology, which has seen tremendous theoretical development in the nineteenth and twentieth centuries. On the other hand, molecular biology, which recruited the advances in biochemistry over the 20th century and has seen many technical revolutions over this time. Being both empirical and theoretical, molecular evolution borrows strength simultaneously from the ever-increasing amount of empirical data available in molecular biology and from the predictive power of theoretical evolutionary biology. From the differences in the observed molecular sequences between individuals of the same population, or between species, biologists can uncover the processes generating this diversity, and unravel the forces governing the underlying evolutionary mechanisms. Can we quantify the relative strength of these forces, shaping both extant populations but also ancient and sometimes extinct lineages? In a nutshell, molecular evolution leverages the patterns of genetic variation

carried by individuals in order to uncover evolutionary mechanisms shaping the evolution of organisms and their ancestral lineages, while at the same time shedding new light on cellular and molecular processes allowing organisms to live and reproduce.

This section will recall the theoretical basis, the assumptions and the limitations on which molecular evolution is based. It is a modest attempt, neither exhaustive nor accurate, probably imprinted with the ideology of our current society on how we perceive and interpret past discoveries. Moreover, this introduction will highlight a few names, while in reality much of the development of molecular evolution also benefited from the contribution of many unmentioned and sometimes forgotten scientists.

1.1 Population-genetics

Molecular evolution is theoretically built upon the framework of population genetics, which in turn historically emerged as a unifying theory between Mendelian inheritance and quantitative genetics, in the early twentieth century. Originally, Johann Gregor Mendel established the statistical laws governing heredity of discrete characters through hybridization experiments on the garden pea plant *Pisum sativum* between 1857 and 1864. This model of inheritance was rediscovered and confirmed in the early twentieth century independently by botanists Hugo de Vries, Carl Correns and Erich von Tschermak (Dunn, 2003).

At first, models of Mendelian inheritance were deemed incompatible with the models of biometricians. The crux of the argument revolved around the evolution of continuous characters[1]. Broadly speaking, supporters of Mendelian genetics held that evolution was driven by mutations transmitted by the discrete segregation of alleles, which biometricians rejected on the basis that this would necessarily imply discontinuous evolutionary leaps (Bowler, 2003). Conversely, biometricians claimed that variation was continuous, which Mendelian geneticists rejected on the basis that the variation measured by biometricians was too small to be impacted by selection (Provine, 2001).

In a series of articles over the 1920s, the statistician Ronald A. Fisher reconciled both theories. First, he proved mathematically that multiple discrete loci could result in a continuous variation (Fisher, 1919). Secondly, Fisher (1930) and Haldane (1932) proved that natural selection could change allele frequencies in a population. Fisher and Haldane hence articulated selection on continuous traits with discrete underlying genetic inheritance, a work that was completed by Wright (1932) for combinations of interacting genes. Wright also proposed the concept of fitness landscape, viewing the evolution of a population as a hill-climbing process. In this context, Wright also explored some of the consequences of random drift, proposing that drift could sometimes allow for a population to cross a valley between multiple fitness peaks. Altogether, Fisher, Haldane and Wright laid the foundations of population genetics, a discipline which basi-

[1]Incompatibility between continuous and discrete evolution can actually be traced back to debates between Jean-Baptiste de Lamarck (1744-1829) defending gradual changes and Georges Cuvier (1869-1932) supporting punctual catastrophic changes, in the late eighteenth century.

cally integrated Mendelian genetics, Darwinism and biometry, easing the debate between continuous and gradual evolution[2].

The emergence of this new scientific field was the first step towards the development of a unified theory of evolution (Huxley, 1942), essentially defined on the basis that natural selection acts on the heritable variation supplied by mutations (Mayr, 1959; Stebbins, 1966; Dobzhansky, 1974).

1.2 Central dogma of molecular biology

During the theoretical development of population genetics, the support of heredity was largely unknown, and the terminology of 'gene', 'alleles' and 'locus' was essentially theoretical and not grounded on directly observable correlates. The first evidence that deoxyribonucleic acid (DNA) is the molecular support of genetic information is in the work of Avery *et al.* (1944), who showed that bacteria treated with a deoxyribonuclease enzyme failed to transform, while otherwise transforming when treated by a protease. The chemical composition of DNA was further elucidated by Chargaff *et al.* (1950), who found that the proportions of adenine (A) and thymine (T) in DNA were roughly the same as the amounts of cytosine (C) and guanine (G), suggesting a relation of complementarity between base pairs (A:T and G:C). On the other hand, the proportion of G+C was found to vary from one species to another, which provided evidence that DNA could encode the genetic information, via a four-letter molecular alphabet.

Ultimately, the double-helix structure of DNA was deciphered by Franklin and Gosling (1953), Watson and Crick (1953) and Wilkins *et al.* (1953). Once the molecular structure of DNA and its role as a support of heredity was elucidated, the work of Crick (1958) on the question of the transfer of information from DNA to proteins resulted in the determination of the genetic code, the translation table from triplets of nucleotides (codons) to amino acids. Ultimately, the establishment of the central dogma of molecular biology detailed the process of protein synthesis. Briefly, the central dogma of molecular biology states that the *"determination of sequence from nucleic acid to nucleic acid, or from nucleic acid to protein may be possible, but transfer from protein to protein, or from protein to nucleic acid is impossible"* (Crick, 1970).

As the support of heredity, DNA gained a central role in evolutionary biology. Moreover, the development of new technologies such as the polymerase chain reaction (PCR) by Kleppe *et al.* (1971), Sanger sequencing (Sanger and Coulson, 1975; Sanger *et al.*, 1977) and more recently the availability of next-generation sequencing techniques, reviewed in Mardis (2008) and Levy and Myers (2016), revolutionized the availability of empirical data on which to test the theoretical predictions of population genetics.

[2]This debate was revived by palaeontologists Gould and Eldredge (1972). As of today it is admitted that both macroevolutive patterns of punctual and gradual changes can be found.

1.3 Neutral theory

Although a unifying theory, population genetics remained rather theoretical for some time, because it deals with the concept of gene frequencies, yet there was no direct way to unambiguously identify the genes with the observable phenotypic traits. For that reason, the connection between theoretical population genetics and empirical and experimental work was only indirect, although quite precisely formalized, through quantitative genetics. Quantitative genetics, or the genetics of complex traits, works by proposing a 'microscopic' model of the genetic architecture of a given observable phenotypic trait. This entails the specification of the number of loci, the effect sizes contributed by each of them, the possible dominance or epistatic interactions between alleles at the same locus or between loci, etc. Population genetics is then used to derive theoretical expectations about the response of the trait to artificial or natural selection, predictions which are then tested against empirical data (Lande, 1976, 1980; Lande and Arnold, 1983). In this framework, however, the detailed genetic basis of the evolutionary process is never accessed directly, but is only indirectly tested.

The situation changed radically during the second half of the 20th century. With the advent of molecular genetics, it became possible to have a direct access to the variability of nucleic and protein sequences within a species, as well as to the differences between closely related species, making it possible to estimate the rate at which allelic genes are substituted. The new observations that were made thanks to these new technological developments turned out to create some surprise.

First, by comparing protein sequences from related species, it was observed that the number of point substitutions between pairs of species was approximately proportional to the time since their last common ancestor (Zuckerkandl and Pauling, 1965; Salser *et al.*, 1976). These observations led to posit the molecular clock hypothesis, which assumes that the rate at which point substitutions accumulate is approximately constant through time. This apparently constant rate of molecular evolution is in sharp contrast with the much more variable rate of morphological evolution observed in the same species, and more generally across the entire fossil record (Simpson, 1944, 1953). Second, electrophoretic methods uncovered surprisingly high levels of genetic variability within natural populations, such that most proteins in diverse organisms were found to be naturally polymorphic (Harris, 1966; Hubby and Lewontin, 1966; Lewontin and Hubby, 1966). In many cases, this molecular polymorphism had no visible phenotypic effects and showed no obvious correlation with any other covariate. Finally, by comparing DNA sequences in related species, it was observed that the overall (genome-wide) rate of DNA substitutions is very high, of least one nucleotide base per genome every two years in a mammalian lineage.

These observations are not easily explained in purely adaptive terms. Instead, they led Kimura (1968), and independently, King and Jukes (1969), to propose the neutral theory of molecular evolution (Kimura *et al.*, 1986; Kimura, 1991). The main tenet of the neutral theory is that most intra- and inter-specific molecular variation is in fact adap-

tively neutral, thus explaining the high protein variability observed in polymorphism datasets, where the diversity is supplied by a high mutational input. Subsequently to origination by mutation, this selectively neutral diversity is reduced by the random extinction of alleles, via the cumulative effect of the random sampling of alleles at each generation. Although the likely fate of a neutral allele just created by mutation is its ultimate extinction, it is also possible that random drift leads to the fixation of this allele in the population. In this context, the frequency of the neutral allele fluctuates through generations, randomly increasing or decreasing over time, because only a relatively small number of gametes are randomly sampled out of the vast number of male and female gametes produced in each generation. As a consequence, the effect of genetic drift at the level of a population results into divergence between lineages, where the majority of the nucleotide substitutions in the course of evolution must have been the result of the random fixation of neutral mutants rather than the result of positive Darwinian selection. Of note, the neutral theory does not say that most mutations are neutral or that adaptation does not take place. A substantial fraction of all mutations are in fact strongly deleterious. However, those mutations are quickly purified away and are generally not visible, neither in the polymorphism within species nor in the divergence between species. The argument of the neutral theory is just that most mutations that are not deleterious are essentially neutral. Adaptive mutations are just rare, relative to neutral mutations, and as a consequence, adaptive arguments do not need to be invoked in order to explain most of the observed intra- and inter-specific variation.

In a second step, Ohta and Kimura (1971) refined the neutral theory, by proposing that mutations can have an effect on the phenotype, and therefore on fitness. However, if their effect on fitness is sufficiently small, they should still behave neutrally and have their fate dictated solely by drift. Ohta (1973) later proposed a mathematical formalization of this argument, incorporating weakly selected mutations to propose the nearly-neutral theory. This theory emphasizes that selective effects lower than the inverse of effective population size are negligible and are expected to behave neutrally. In this regard, effective population size (N_e) is a quantitative measure of genetic drift such that genetic drift decreases with increased effective population size.

The neutral theory sparked a long-standing controversy between neutralist and selectionists. Selectionists maintain that a mutant allele must have some selective advantage to spread through a species, although admitting that a neutral allele may occasionally be carried along by hitchhiking on a closely linked gene that is positively selected. Neutralists, on the other hand, argued that some mutants might spread through a population without having any selective advantage, just by random sampling, such that if a mutant is selectively equivalent to preexisting resident alleles, its fate is thus left to chance. Of note, even if the probability of fixation of any given neutral mutation is low ($p = 1/2N_e$), the high rate of mutation at the gene or genome-wide level and the highly degenerate mapping between genotype and phenotype both leave considerable latitude at the molecular level for random genetic changes that have no effect upon the fitness of the organism (King and Jukes, 1969). As a result, the total flux of neutral substitutions

can in fact be the dominant contribution to intraspecific polymorphism and interspecific differences. This overwhelming combinatorial effect was probably the point that was hard to grasp by many evolutionary biologists at the time, trained in the idea that most mutations should have an effect on the phenotype. Another factor that contributed to the difficulty in accepting the neutral theory is the fact that effective population sizes turn out to be much smaller than true (census) population sizes. This point is important, because, according to the nearly-neutral theory of Ohta (1992), the inverse of effective population size directly determines the proportion of all mutations that are effectively neutral. Once it is recognized that effective population sizes are small, it becomes easier to accept that most mutations with weak effects are effectively neutral.

As of today, it is widely accepted that both genetic drift and natural selection participate in the evolution of genomes. The controversy is no longer strictly dichotomous but rather concerns the quantitative contributions of adaptive and of non-adaptive evolutionary processes, and their articulation with regards to mutation, selection, drift, migration, gene conversion, and other evolutionary processes.

1.4 The legacy of the nearly-neutral theory

The neutral theory, and its nearly-neutral extension, have broad implications in evolutionary biology. Much of its insight has been integrated in modern population genetics, molecular evolutionary sciences, but also phylogenetics and molecular dating. Importantly, because of the marginal role played in this theory by the most unpredictable factor involved in molecular evolution, namely adaptation, the nearly-neutral theory is in a good position to make clear quantitative predictions about the rate and patterns of molecular evolution, or about the structure of genetic diversity within species. As such it gives a well-defined framework to formalize various assumptions about the underlying processes and test them against empirical sequence data, which are becoming increasingly available. Questions within this framework range from the causes of mutational rate variation, to the structure of fitness landscapes, or the impact of changes in effective population size between species. In the following, I summarize several of the most important insights that have been contributed by the neutral and nearly-neutral theory, and how they still play on current research in molecular evolution.

1.4.1 Mostly-purifying selection

First, along with the adoption of the nearly-neutral theory by evolutionary biologists, the common perception about the nature of selection shifted from selection being a driver of changes mediated by adaptive mutations to a mainly purifying force discarding and filtering out strongly deleterious mutations (Lynch and Walsh, 2007). From this perspective, protein sequences are relatively close to their adaptive optimum such that many mutations occurring in their sequence are likely to disrupt their functions. This effect can be observed in underlying DNA sequences, where non-synonymous substitutions oc-

cur less frequently than synonymous substitutions (King and Jukes, 1969), and similarly, radical amino acid replacements are more less than conservative changes (Kimura, 1983). These effects are also observed within populations, non-synonymous single-nucleotide polymorphisms segregate at lower frequencies compared to synonymous polymorphisms, a phenomenon explained by purification of deleterious alleles which cannot reach high frequencies (Akashi, 1999; Cargill *et al.*, 1999; Hughes, 2005). Finally, what determines the rate of non-synonymous evolution of protein-coding genes is primarily the amount of selective constraint acting on them, such that slowly evolving genes are just more constrained than fast-evolving genes Kimura (1983).

1.4.2 The mutation-selection balance

Proteins are relatively close to, but not quite at their optimum. This relates to another important conceptual point contributed by the nearly-neutral theory. From a neutralist perspective, evolution should not be seen as an optimization process, but instead, as a process driving natural protein sequences at their mutation-selection equilibrium. This concept of mutation-selection balance explains important features of natural protein sequences, which cannot be explained only in terms of optimization. Thus, as noted early on by King and Jukes (1969), amino acids that have more codons are more frequently represented in natural protein coding sequences. Similarly, later work by Singer and Hickey (2000) has shown that species with a mutational bias towards AT (respectively GC) tended to have proteomes with a higher frequency of amino acids encoded by AT-rich (respectively GC-rich) codons. Another implication is that proteins are not optimal, either for their enzymatic properties (Cornish-Bowden, 1976; Albery and Knowles, 1976; Hartl *et al.*, 1985) or for their conformational stability (Taverna and Goldstein, 2002). This non-optimality is observed even if proteins are under directional selection for the optimal sequence. All these observations are clear illustrations of the fact that natural sequences are not at their optimum, but instead, are the result of a trade-off between mutation biases and mostly purifying selection. This trade-off between mutation and selection is regulated by the amount of random drift, and thus by effective population size. The concept of mutation-selection balance is not yet fully incorporated in evolutionary thinking. Many evolutionary scientists, and many biologists more generally, still tend to think in terms of optimization. Correctly formalizing this interplay between mutation, selection and drift in the context of phylogenetic codon models is in fact at the core of most of the work presented in the thesis.

1.4.3 The importance of drift

Tempering the effect of selection, drift mediated by effective population size has been repeatedly invoked to explain the relaxation of the selective strength. First, it has been observed that within populations relative diversity of selected site is more reduced for species with smaller effective population size. Indeed, in an intra-specific context, the non-synonymous diversity, relative to the synonymous diversity (i.e. π_N/π_S), is reduced

in species characterized by larger effective population sizes (Piganeau and Eyre-Walker, 2009; Elyashiv *et al.*, 2010; Galtier, 2016; Chen *et al.*, 2017; James *et al.*, 2017). Similarly, in a phylogenetic context, the strength of selection, such as measured by the relative rate of non-synonymous over synonymous substitution, is lower along lineages with small effective population size (Ohta, 1993, 1995; Moran, 1996; Woolfit and Bromham, 2003, 2005; Popadin *et al.*, 2007). It is important to note that, in most cases, the effective population size is not directly measured, but a surrogate measure is used instead, for example synonymous diversity (i.e. π_S) as in (Galtier, 2016), or body size or longevity, expected to be large in lineages with a low N_e (Romiguier *et al.*, 2014). Leveraging the nearly-neutral theory in order to quantitatively measure effective population size in a phylogenetic context is one of the main objectives of this thesis, such as presented in chapter 8. Of note, the quantitative response of the molecular evolutionary process to changes in effective population size appears to strongly depend on the underlying fitness landscapes (Welch *et al.*, 2008), to the point of being entirely absent (Cherry, 1998; Goldstein, 2013). This relationship between the rate of evolution and effective population size is also a main question addressed in this thesis, such as studied in chapter 9.

1.4.4 Unravelling adaptation

The neutralist view of selection as mostly purifying raises an important question: where, and to what extent, does adaptation leave traces in molecular sequences? The fact that the neutral theory has been relatively silent on this question has largely contributed to its rejection by many biologists, and in many respects the question is still open. At first, methods for detecting adaptation have been developed, integrating either the neutral or the nearly-neutral regime as a null model. Departures from one of these null model are then typically interpreted as traces of adaptations. This idea to detect traces of adaptation has been explored in a phylogenetic context, whenever the null model is neutral (Goldman and Yang, 1994; Muse and Gaut, 1994; Yang and Swanson, 2002; Zhang and Nielsen, 2005) or nearly-neutral (Rodrigue and Lartillot, 2016; Bloom, 2017). Similarly, in a population-genetics context, adaptation is detected as a deviation from the null model, considered originally neutral (McDonald and Kreitman, 1991; Charlesworth, 1994; Smith and Eyre-Walker, 2002), and subsequently improved to account for slightly deleterious mutations in a nearly-neutral regime (Eyre-Walker and Keightley, 2009; Galtier, 2016).

These methods have clearly revealed important traces of adaptation (Bustamante *et al.*, 2005; Halligan *et al.*, 2010; Enard *et al.*, 2014), in particular, in genes implicated in host-pathogen interactions (Enard *et al.*, 2016; Grandaubert *et al.*, 2019), or in other specific genes involved in intra-genomic Red-Queen dynamics such as PRDM9 (Thomas *et al.*, 2009; Oliver *et al.*, 2009; Ponting, 2011; Latrille *et al.*, 2017). However, this might represent only the most extreme adaptive events. Much of adaptation might still have been missed at the molecular level. Kimura (1983) proposed a more radical insight about the link between phenotypic adaptation and neutral molecular evolution. By showing an

example of a phenotypic trait under stabilizing selection and controlled by a large number of loci with small effects, phenotype efficiently optimized by selection, but the molecular evolutionary process at each locus essentially indistinguishable from a neutral process. More recent work, using the empirical knowledge acquired by large-scale population-genomics project in humans, draws similar conclusions (Simons *et al.*, 2018). Namely that many traits turn out to be highly polygenic (Pritchard and Cox, 2002), and the frequency changes contributing to their adaptive fine-tuning can be highly stochastic (Sella and Barton, 2019). Analogous to statistical physics, microscopic behaviour of a physical system is dominated by thermal noise, while the macroscopic state looks essentially deterministic and driven by a principle of free-energy minimization.

1.4.5 Molecular evolution is mutation-limited

Originally, the neutral theory was heavily relying on the molecular clock hypothesis of Zuckerkandl and Pauling (1965), which posits that rate of sequence evolution is constant through time and across evolutionary lineages. Although appealing, it became clear that the rate of evolution was not constant (ChungWu and Wen-Hsiung Li, 1985; Li *et al.*, 1987; Bulmer *et al.*, 1991; Gaut *et al.*, 1992). This rejection of the strict clock motivated important methodological developments for modelling the fluctuations of the substitution rate along a phylogeny (Sanderson, 1997; Thorne *et al.*, 1998; Kishino *et al.*, 2001; Aris-Brosou and Yang, 2002; Drummond *et al.*, 2006; Lepage *et al.*, 2007). The primary motivation for these relaxed clock models was to achieve more accurate molecular dating. However, these developments also fostered comparative analyses, trying to explain the causes of the variation of substitution rate between lineages. Methodologically, this motivated the developments of methods able to conduct correlation analyses between molecular evolutionary rates and observable quantitative traits, while correcting for phylogenetic inertia (Lanfear *et al.*, 2010b; Lartillot and Poujol, 2011). Empirically, generation time, but also metabolic rate, or selection for longevity, are potential explanations for the variation in substitution rate (Lartillot and Delsuc, 2012), which can be interpreted in the light of the molecular mechanisms of cell division (Gao *et al.*, 2016).

The exact reasons for the variation in substitution rate between lineages are still debated. However, what is clear is that this variation is mostly reflecting variation in the mutation rate. As such, and in spite of the historically central role played by the molecular clock in the arguments in favour of the neutral theory, the rejection of the molecular clock by empirical data does not contradict the neutral theory. It just confirms that, in a neutral or nearly-neutral regime, the molecular evolutionary process is mutation-limited, or, in other words, that the substitution rate is determined primarily by the mutation rate.

1.4.6 Extending the null hypothesis of molecular evolution

Finally, some patterns have been found inconsistent within the general framework of mutation, selection and drift, thus leading to uncovering new forces such as biased gene conversion which mimics selection but are fundamentally segregation distortion during

recombination (Marais, 2003; Galtier and Duret, 2007; Duret and Galtier, 2009). Such forces are altering the composition of genomes and must be carefully accounted for in models of evolution (Galtier *et al.*, 2009; Ratnakumar *et al.*, 2010; Figuet *et al.*, 2014). However, even though forces such as biased gene conversion are not within the scope of this thesis, some assumptions and design of our models had been taken such as to implement these forces subsequently.

1.4.7 Conclusion

Altogether, evolution of sequences results from the interplay between mutation, selection and drift, where this formalism is developed in chapter 2. Of all these components, selection is the most pervasive, which can be approximated and observed in protein-coding DNA sequences in a phylogenetic context between lineages, presented in chapter 3). Consequently, models are applied to empirical data, and the methodology of Bayesian inference from an alignment of DNA sequences is presented in chapter 4. Finally, selection of protein-coding DNA sequences is related to biochemical and biophysical constraints (chapter 4).

2

The mathematics of molecular evolution

Contents

In molecular evolution, the information contained in empirically observed sequences is leveraged to reconstruct ancestral lineages and to unveil the evolutionary mechanisms having generated this diversity of sequences. In other words, the task is to reconstruct the ancestral path followed by lineages using the knowledge available today, by working backward in time. To do so, however, requires a theoretical model of the generating process forward in time. One can then play this model forward in time and relate the resulting generated sequences to empirically observed patterns.

Working out the long-term molecular evolutionary process first requires to formalize what happens in a short time period within populations. Population genetics, with its assumptions and limitations, provides the theoretical framework for this. The first section thus recalls the basics of mathematical population genetics, and more specifically,

the Wright-Fisher model and its assumptions. This will allow me to relate parameters of evolution such as mutation, selection and drift to observable patterns in molecular sequences such as the probability of fixation of a mutant allele, as well as the expected number of copies of the derived allele we should observe in a population. These relationships between the underlying evolutionary forces and the observable patterns will subsequently be leveraged and recruited in the next section to derive an approximation of the long-term process of sequence substitution, again, parameterized directly in terms of mutation, selection and drift.

Although the mathematical proofs for most of the results presented here are out of the scope of this manuscript, an effort was made to state all definitions and assumptions. Such an effort is meant to clearly define the models, their assumptions and their parameterization from the ground up.

2.1 Population genetics of sequences

2.1.1 The Wright-Fisher model with selection

The Wright-Fisher model describes the change in frequency of a polymorphic gene with two alleles in a diploid population over time. The population is assumed to consist of fixed number of diploid individuals $N \gg 1$. It is also assumed to be panmictic (i.e. non-preferential random mating), with non-overlapping generations. The number of copies of the derived allele B present at the current generation is denoted i and the frequency of this mutant allele B is denoted $p = i/2N$, while the frequency of the resident A alleles is $1-p$.

The ability to survive and produce offspring differs between the three diploid genotypes (AA, AB, BB). Here, selection is assumed to occur between the zygotic and the adult stage, called post-zygotic selection. Quantitatively, selection is captured by a measure called Wrightian fitness (W), which, for a given diploid genotype, is defined as the expected number of offspring produced by an individual having this genotype. Since the population is regulated in size, only the relative fitness matters, which is usually set to 1 for the reference (wild-type) genotype. It is convenient to define the fitness of the other two genotypes, relative to the wild-type, in terms of a selection coefficient. Furthermore, in the following, we will assume additive effects (co-dominance), such that the heterozygote has an intermediate fitness between the two homozygotes. Altogether, fitness of the three diploid genotypes are defined as:

$$\begin{cases} W_{AA} &= 1 \\ W_{AB} &= 1 + s \\ W_{BB} &= 1 + 2s \end{cases} \tag{2.1}$$

More generally than the previous equations, under the assumption that selection is weak $|s| \ll 1$, the selection coefficient can be approximated by the difference in Wrightian

fitness of the mutant and the resident allele as:

$$s = \frac{W_B - W_A}{W_A}, \tag{2.2}$$
$$= \frac{W_B}{W_A} - 1, \tag{2.3}$$
$$\simeq \ln\left(\frac{W_B}{W_A}\right), \tag{2.4}$$
$$\simeq \ln(W_B) - \ln(W_A), \tag{2.5}$$
$$\simeq f_B - f_A, \tag{2.6}$$

where $f = \ln(W)$ is often referred to as the Malthusian fitness, relative fitness or also log-fitness.

2.1.2 Frequency changes across successive generations

Under the Hardy-Weinberg equilibrium of the population, the diploid genotype frequencies in the current generation are distributed as given in table 2.1.

As a result, the mean fitness in the population is a function of the selection coefficient and the frequency of two alleles as:

$$\overline{W} = (1+2s)p^2 + (1+s)2p(1-p) + (1-p)^2 \tag{2.7}$$
$$= 1 + 2ps, \tag{2.8}$$

And the relative fitness of the three different genotypes are also shown in table 2.1.

Genotype	AA	AB	BB
Wrightian fitness (W)	1	$1+s$	$1+2s$
Hardy-Weinberg frequency	$(1-p)^2$	$2p(1-p)$	p^2
Relative Wrightian fitness	$\frac{1}{1+2ps}$	$\frac{1+s}{1+2ps}$	$\frac{1+2s}{1+2ps}$

Table 2.1: *Fitnesses of the different genotypes*

Reproduction proceeds in two steps. In a first step, a very large pool of gametes is produced, in which adults contribute proportionally to the fitness of their genotype. Altogether, the frequency p' of gametes bearing the B allele is a function of p and s, as shown in figure 2.1, and formally derived as:

$$p' = p^2\frac{1+2s}{1+2ps} + p(1-p)\frac{1+s}{1+2ps} \tag{2.9}$$
$$= p\frac{1+s(1+p)}{1+2ps} \tag{2.10}$$

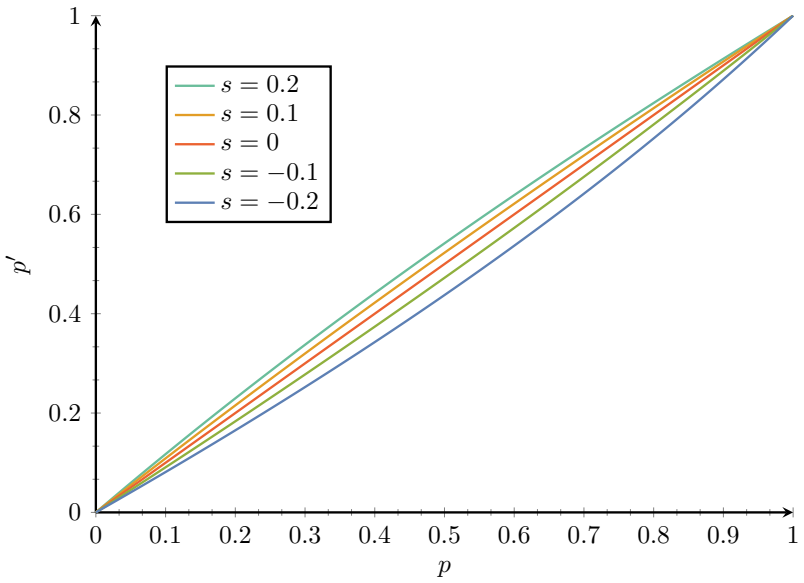


Figure 2.1: *Frequency of derived allele p' after a generation in the vertical axis a function of the frequency in the previous generation p in the horizontal axis, shown for several selection coefficients in coloured solid lines. Positive selection coefficients ($s > 0$) result in increased derived allele frequency at the next generation, which is intuitively expected. The effect is stronger when the derived allele frequency is close to* 0.5*, intuitively because the poll of both alleles must be sufficiently large such that they can be replaced. It is worth noting that even for strong selection coefficients ($s = 0.2$), completely unrealistic in real population, the difference in frequency from one generation to the next is subtle.*

In a second step, the N individuals of the next generation are obtained by randomly sampling from the pool of gametes. As a result, the probability $\mathbb{P}_{ij}$, that there are j copies of the derived allele B present at the next generation, given that there were i copies in the current generation is given by the binomial distribution, with a proportion p' of B alleles in gametes:

$$\mathbb{P}_{ij} = \binom{2N}{j} (p')^j (1-p')^{2N-j} \tag{2.11}$$

$$= \binom{2N}{j} \left(p\frac{1+s(1+p)}{1+2ps}\right)^j \left(1-p\frac{1+s(1+p)}{1+2ps}\right)^{2N-j} \tag{2.12}$$

These transition probabilities define a discrete-space and discrete-time Markov process. It has also been shown to be extremely difficult to explicitly derive formulas for several quantities of evolutionary interest.

Of note, under the assumption that selection is weak $|s| \ll 1$, p' reduces to:

$$p' \simeq p(1+s+ps-2ps) \tag{2.13}$$

$$= p + sp(1-p) \tag{2.14}$$

$$= p + \Delta p, \tag{2.15}$$

where $\Delta p = sp(1-p)$

Intuitively, fluctuations induced by the binomial sampling (equation 2.12) are the underlying cause of random drift. Quantitatively, the expected frequency change from one adult generation to the next adult generation is:

$$\mathbb{E}[\Delta p] = sp(1-p). \tag{2.16}$$

The variance of this binomial distribution is given by:

$$\mathrm{Var}[\Delta p] = \frac{p'(1-p')}{2N} \tag{2.17}$$

Since the change in frequency between two generations is small ($p \simeq p'$), the variance is very close to:

$$\mathrm{Var}[\Delta p] \simeq \frac{p(1-p)}{2N} \tag{2.18}$$

Thus, the variance induced by random drift is inversely proportional to the population size N. Also, if $s \gg 1/2N$, then $\mathbb{E}[\Delta p] \gg \mathrm{Var}[\Delta p]$, or, in other words, the systematic trend imprinted by selection dominates over drift, describing the strong selection regime. In contrast, if $s \ll 1/2N$, drift dominates over selection, describing the effectively neutral regime.

2.1.3 Effective population size

The notion of effective population size, called N_e, only appears when we apply a panmictic model to a population that is not, or to a real population. N_e was originally defined as *"the number of breeding individuals in an idealized population that would show the same amount of dispersion of allele frequencies under random genetic drift or the same amount of inbreeding as the population under consideration"* (Wright, 1931). For most quantities of interest and most real populations, the census population size N of a real population is usually larger than the effective population size N_e. The same population may have multiple effective population sizes for different genetic loci, as for example sex chromosomes do not have the same population sizes as autosomes. For the following development, this idealize population with a single effective population N_e will be assumed.

2.1.4 Probability of fixation

Starting from an initial frequency, the Wright-Fisher process eventually reaches absorption: the derived allele either dies out or invades the population and thus reach fixation. As the effective population size (N_e) approaches infinity (i.e. $N_\mathrm{e} \to \infty$), and assuming that the selection coefficient scaled by effective population size ($N_\mathrm{e}s$) remains constant, the discrete Markov process defined above can be closely approximated by a continuous-time and continuous-space diffusion process. The parameters of this process are summarized in table 2.2 for readability.

Under this diffusive approximation, a partial differential equation known as the Kolmogorov's backward equation can be used to obtain the fixation probability of the derived

Parameter	Symbol	Range
Census population size	N	$[10^2, 10^6]$
Effective population size	N_e	$[10^2, 10^6]$
Absolute Wrightian fitness	W	$\simeq 1$
Relative fitness	$f = \ln(W)$	$\ll 1$
Selection coefficient	s	$\lvert s\rvert \ll 1$
Scaled selection coefficient	$S = 4N_\mathrm{e}s$	Finite (negative or positive)
Mutation rate per generation	u	$[10^{-10}, 10^{-7}]$ per site
Scaled mutation rate	$\theta = 4N_\mathrm{e}u$	$[10^{-8}, 10^{-1}]$ per site

Table 2.2: *Parameters of population genetics*

allele. Formally, for an effective population size N_e, Kimura (1962) derived the probability of fixation ($\mathbb{P}_\mathrm{fix}(s, N_\mathrm{e}, p)$) of a derived allele with selection coefficient s and initial frequency p if the selection coefficient is small ($\lvert s\rvert \ll 1$):

$$\mathbb{P}_\mathrm{fix}(s, N_\mathrm{e}, p) = \frac{1 - \mathrm{e}^{-4N_\mathrm{e}ps}}{1 - \mathrm{e}^{-4N_\mathrm{e}s}}. \tag{2.19}$$

Because s and N_e are confounded parameters, this probability of fixation is denoted $\mathbb{P}_\mathrm{fix}(S, p)$, as a function the scaled selection coefficient $S = 4N_\mathrm{e}s$ and p, as shown in figure 2.2, and formally derived as:

$$\mathbb{P}_\mathrm{fix}(S, p) = \frac{1 - \mathrm{e}^{-pS}}{1 - \mathrm{e}^{-S}}. \tag{2.20}$$

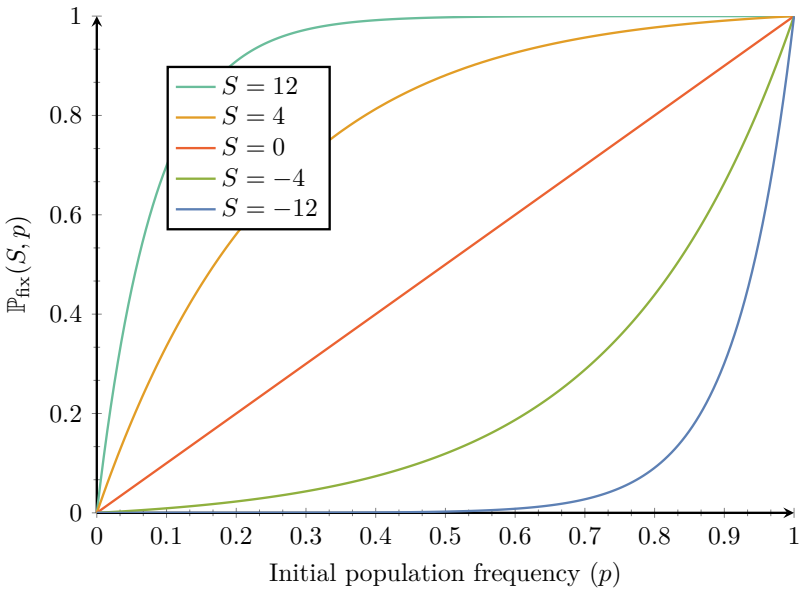


Figure 2.2: *Probability of fixation $\mathbb{P}_{fix}(S, p)$ in the vertical axis as a function of the initial frequency p in the horizontal axis, shown for different scaled effective population size $S = 4N_e s$. In contrast to changes of frequency during a generation, the probability of fixation is sensitive to very weak selection coefficients ($\lvert s\rvert \ll 1$), as long as the scaled selection coefficient is not negligible ($\lvert S\rvert > 1$). Intuitively, selective effects are magnified by population size because the fixation probability is the resultant of the overall trajectory of the allele, integrating small effects throughout its lifespan.*

An interesting special case is obtained for a new mutation appearing in the population. Because it is a single mutant, the initial frequency of the derived allele is $p = 1/2N_e$, and this probability of fixation denoted $\mathbb{P}_{fix}(s, N_e)$ is given by:

$$\mathbb{P}_{fix}(s, N_e) = \frac{1 - e^{-2s}}{1 - e^{-4N_e s}} \tag{2.21}$$

$$\simeq \frac{2s}{1 - e^{-4N_e s}} \tag{2.22}$$

The special case of a neutral allele can be obtained by taking the limit when s goes to 0.

$$\mathbb{P}_{fix}(0, N_e) = \frac{1}{2N_e} \tag{2.23}$$

Altogether, the fixation probability of a selected single mutant relative to the fixation probability of a selectively neutral single mutant is given as:

$$\frac{\mathbb{P}_{fix}(s, N_e)}{\mathbb{P}_{fix}(0, N_e)} \simeq 2N_e \frac{2s}{1 - e^{-S}}, \tag{2.24}$$

$$\simeq \frac{S}{1 - e^{-S}}, \tag{2.25}$$

where this quantity is solely dependent on the scaled selection coefficient S. Such essential result has important consequences, random genetic drift and selection are intrinsically confounded factors. As a an example, increasing population size by a factor of 2 while reducing the selection coefficient by the same amount leads to the exact same equation, such that they are indistinguishable. Moreover, the equation have different limits as a function of the selection coefficient:

$$\begin{cases} \lim\limits_{S \to -\infty} \dfrac{S}{1 - e^{-S}} = -Se^{S} \\ \lim\limits_{S \to 0} \dfrac{S}{1 - e^{-S}} = 1 + \dfrac{S}{2} \\ \lim\limits_{S \to +\infty} \dfrac{S}{1 - e^{-S}} = S. \end{cases} \tag{2.26}$$

More precisely, the scaled fixation probability has different regimes depending on the value of the scaled selection coefficient, as illustrated in figure 2.3. In the regime of a weak selection coefficient, usually defined as $|S| \ll 1$ or $|s| \ll 1/N_e$, known as the drift barrier, the mutant allele is behaving mostly as a neutral allele.

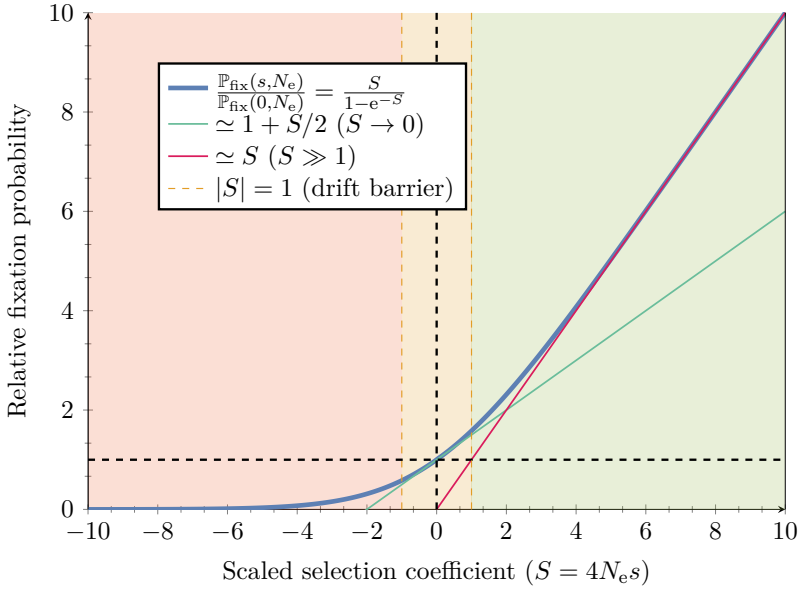


Figure 2.3: *Fixation probability of a selected allele relative to a neutral allele, shown in the vertical axis, as function of the scaled selection coefficient $S = 4N_e s$ in the horizontal axis. For a substantial negative scaled selection coefficient ($s \leq -1/N_e$, red-filled area), the probability of fixation is greatly reduced (by an exponential factor), and the allele will not likely reach fixation. On the other hand, for a positive scaled selection coefficient ($s \geq 1/N_e$, green filled area), the ratio is approximately linear with regard to S. In between, whenever the absolute value of s is close to $1/N_e$ (yellow filled area), the allele behaves approximately neutrally.*

2.1.5 Site frequency spectrum

The probability of fixation of an allele can be empirically observable, and in the context of a Wright-Fisher processes it is related to selection and drift. However, this absorbing fate is not the sole characteristic of the process that relates empirical observable quantities to parameters of the process. Along the whole trajectory of an allele, before fixation or extinction, the probability of this allele to be at a certain frequency can be related to its selection coefficient and to the effective population size. More precisely, $g(x)\mathrm{d}x$ is the expected time for which the population frequency of derived allele is in the range $(x, x + \mathrm{d}x)$ before eventual absorption, as shown in figure 2.4, which is derived using the Kolmogorov forward equation as a function of x and S:

$$g(x, s, N_e) = \frac{\left(1 - e^{-2s}\right)\left(1 - e^{-4N_e s(1-x)}\right)}{s(1 - e^{-4N_e s})x(1 - x)} \tag{2.27}$$

$$\Rightarrow g(x, S) \approx \frac{2\left(1 - e^{-S(1-x)}\right)}{(1 - e^{-S})x(1 - x)} \tag{2.28}$$

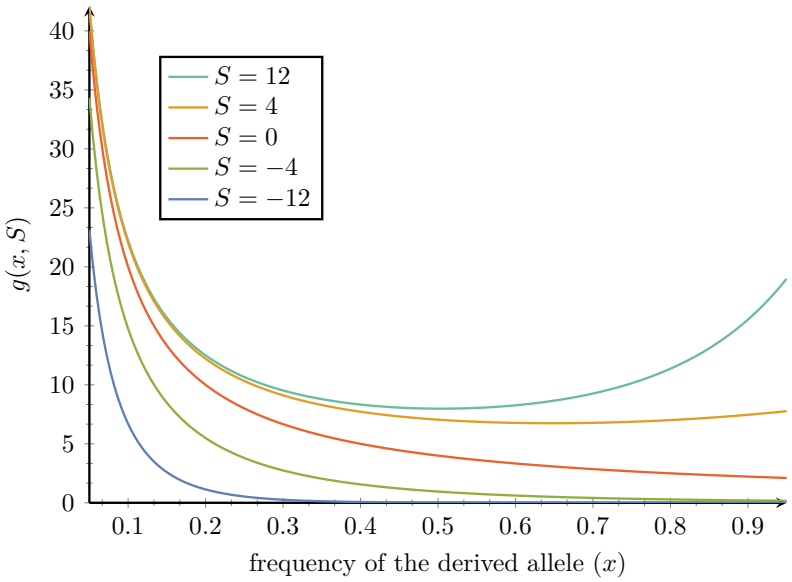


Figure 2.4: *Expected time at a derived frequency $g(x, S)$ in the vertical axis as a function of the frequency x, shown for different scaled selection coefficient. Alleles with a positive selection coefficient can be observed at high frequency, while alleles with negative selection coefficients are unlikely to be observed at high frequency.*

This equation is solely valid for a gene with two alleles, a configuration which is rarely observed in empirical data since more than two variants of a gene are usually present in the population. However, it is frequent to observe sites inside a gene sequence for which only two alleles are segregating. This observation led to the development of a site-specific Wright-Fisher process, assuming that each site follows an independent process (Sawyer and Hartl, 1992). Strictly speaking, this model considers a collection of independently evolving loci, meaning without linkage. It provides a good approximation if there is free recombination between sites. Moreover, the collection is considered infinite whereas the total mutation rate across this infinite collection is considered finite. The assumption of an infinite number of sites is necessary to ensure that each mutation arises at a new site, with a Poisson distribution of total rate u per generation for the whole sequence.

From an empirical perspective, for a sample of n sequences taken in the population, the expected number of sites with i copies of the derived allele (with i ranging from 1 to $n - 1$) is denoted $G(i, n)$. The collection of all $G(i, n)$ generates what is called a site frequency spectrum (SFS), which can intuitively be interpreted as the discrete version of the expected time at a derived frequency (equation 2.28), readily available from a sample of sequences from a population. Given the scaled selection coefficient ($S = 4N_e s$), and the scaled mutation rate per generation for the whole sequence ($\theta =$

$4N_e u$), each entry of the SFS is:

$$G(i,n) = \int_0^1 2N_e u g(x,S) \binom{n}{i} x^i (1-x)^{n-i} \mathrm{d}x \tag{2.29}$$

$$= \theta \int_0^1 \frac{1-\mathrm{e}^{-S(1-x)}}{(1-\mathrm{e}^{-S})x(1-x)} \binom{n}{i} x^i (1-x)^{n-i} \mathrm{d}x \tag{2.30}$$

$$= \frac{\theta}{1-\mathrm{e}^{-S}} \binom{n}{i} \int_0^1 \left(1-\mathrm{e}^{-S(1-x)}\right) x^{i-1} (1-x)^{n-i-1} \mathrm{d}x \tag{2.31}$$

This site frequency spectrum can be confronted to empirical polymorphic data in order to estimate the scaled selection coefficient of new mutations. However, a single selection coefficient for all sites and all mutations is biologically not realistic. Accordingly, a distribution of selection coefficients across sites is assumed, which is usually modelled as a continuous distribution, known as the distribution of fitness effects of mutations (DFE). Mixing over this distribution, the SFS can then be computed as a function of the underlying DFE, and can thus be estimated based on empirical data (Eyre-Walker *et al.*, 2006; Eyre-Walker and Keightley, 2009).

2.2 Mutation-selection process

The previous section recalled the Wright-Fisher process of evolution inside a population, relating selection and drift to the diversity of sequences, which empirically requires gene sequences for at least several individuals. However, modelling sequence evolution between different species along lineages is a different endeavour, in which species are often simplified with a single representative sequence, collapsing the intraspecific diversity. Under this simplification, the interspecific variability and the evolutionary trajectory of sequences are described by the past history of point substitutions along lineages. The rate at which such substitution occurs can nonetheless be decomposed into two mechanisms: their origination through mutation and their final fate of fixation or loss, a modelling approach broadly known as the origin-fixation approximation (McCandlish and Stoltzfus, 2014), illustrated in figure 2.5. Most importantly, this decomposition of substitution events into mutation and fixation events is able to conciliate population genetics and interspecific molecular evolution, where the substitution history is parameterized by mutation, selection and drift. In the field of phylogenetics, the origin-fixation framework is more commonly known as the mutation-selection paradigm, where fixation of an allele encompasses the effect of natural selection and drift (which are confounded factors, see equation 2.25), and origination corresponds to mutation. Since the scope of this manuscript emanates from phylogenetics, I will use the convention mutation-selection terminology hereafter. Of note, a more general mathematical description of the mutation-selection framework recruiting tools from statistical physics can be found in Sella and Hirsh (2005) and Mustonen and Lässig (2009).

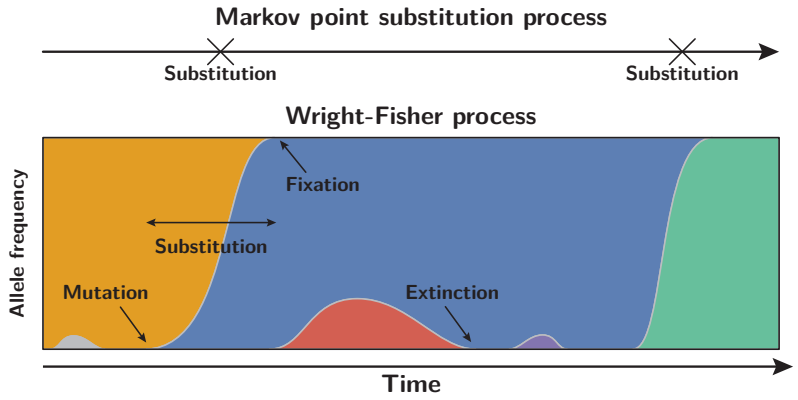


Figure 2.5: *Mutation-selection substitutions models. The trajectory of alleles inside a population is collapsed into a single point substitution process. This approximation is valid under low mutation rates such that a mutation originates uniquely whenever the gene is monomorphic (with a single allele).*

2.2.1 Mutation-limited process

Mutation-selection probabilistic models are usually Markovian with respect to time, such that the next substitution event depends on the current representative sequence but not on earlier sequences visited in the history of a lineage. This continuous-time Markovian process is valid if the mutation rate is sufficiently low, such that the event of a new mutation reaching fixation is completed before the next one occurs. Since the rate of substitution is equal to u (per generation) and that each allele ultimately reaching fixation is segregating for an average of $4N_e$ generations (Kimura and Ohta, 1969), this assumption is broadly applicable whenever the product of population size and mutation rate per generation for the sequence is lower than 1 ($4N_e u \ll 1$). More strictly, the model would require not only that new mutations reaching fixation do so before the next substitution occurs, but before any mutation occurs, even the ones that ultimately become extinct. Since at each generation during the process an average of $2N_e$ mutations are produced, the point substitution is valid under the condition that $8N_e^2 u \ll 1$. In practice, the assumptions that $4N_e u \ll 1$ is a sufficient condition for the process to be well approximated. Throughout this development, it is important to note that u is the mutation rate for the whole sequence under consideration.

For large sequences this approximation is usually not valid, and the sequence is then decomposed into each individual site, forming a collection of independently evolving continuous-time Markov chains. For such a decomposition to be valid, these models have to assume free recombination between sites. The mutation rate u in this condition then refers to the mutation rate for each independent site, rather than the total mutation rate over the collection as a whole. For example, Halpern and Bruno (1998) constructed a model for the evolution of coding sequences where each codon site is modelled as an independent Markov chain.

2.2.2 Substitution rate

The continuous-time Markov chain is defined by the instantaneous rate at which transitions occur between pairs of states. Parameters of this process are summarized in table 2.3 for readability.

Parameter	Symbol	Range
Scaled fitness	$F = 4N_e f$	finite, positive or negative
Mutation rate per time	μ	$[10^{-11}, 10^{-8}]$ per site per year
Substitution rate per time	Q	$[10^{-11}, 10^{-8}]$ per site per year
Equilibrium frequency	π	$[0, 1]$
Equilibrium frequency under mutation	σ	$[0, 1]$
Mean scaled fixation probability	ν	$[0, 1]$ for purifying selection

Table 2.3: *Parameter of mutation-selection processes used in this section (2.2.1)*

Given the current state of allele A, the rate of transition to other states can be derived using the population-genetic equations introduced above. At each generation, the expectation for the number of possible mutants is $2N_e u$, and each of these mutants has a probability $\mathbb{P}_{\text{fix}}(s, N_e)$ to result in a substitution. Altogether, the instantaneous rate of substitution from allele A to B, denoted $Q_{A\to B}$, is equal to the rate of mutation ($\mu_{A\to B}$) multiplied by the probability of fixation of the mutation $\mathbb{P}_{\text{fix}}(s_{A\to B}, N_e)$ and scaled by the number of possible mutants at each generation ($2N_e$):

$$Q_{A\to B} = 2N_e \mu_{A\to B} \mathbb{P}_{\text{fix}}(s_{A\to B}, N_e) \tag{2.32}$$

It is important to note that the substitution rate and the mutation rate are in the same units, such that this equation is valid whether the rate is measured either in units of chronological time or per generation (or in branch length, which will matter later on). As a convention, in what follows, mutation rate is denoted u when measured in units of generation, and denoted μ when measured in units of time. As a consequence, Q is measured in units of time in this section.

In the case of selected mutations, the probability of fixation depends on the difference in log-fitness (f_A and f_B) between the two alleles:

$$\begin{aligned} Q_{A\to B} &= 2N_e \mu_{A\to B} \mathbb{P}_{\text{fix}}(s_{A\to B}, N_e) && (2.33)\\ &= 2N_e \mu_{A\to B} \frac{2(f_B - f_A)}{1 - e^{4N_e(f_A - f_B)}} && (2.34)\\ &= \mu_{A\to B} \frac{F_B - F_A}{1 - e^{F_A - F_B}}, \text{ where } F = 4N_e f && (2.35) \end{aligned}$$

In the case of neutral mutations, the probability of fixation is independent of the original and target sequence, and equals $1/2N_e$. As a consequence, the substitution

rate denoted $Q^0_{A\to B}$ simplifies to:

$$Q^0_{A\to B} = 2N_e\mu_{A\to B}\mathbb{P}_{\text{fix}}(0, N_e) \tag{2.36}$$
$$= 2N_e\mu_{A\to B}\frac{1}{2N_e} \tag{2.37}$$
$$= \mu_{A\to B} \tag{2.38}$$

If the difference of log-fitness tends to 0, the substitution rate is equal to the mutation rate, retrieving equation 2.38:

$$\lim_{|F_B - F_A|\to 0} Q_{A\to B} = \mu_{A\to B} \tag{2.39}$$

Taken together, the transition rates which generate the substitution history and ultimately the interspecific diversity is parameterized solely by mutation, selection and drift. Consequently, from a particular history of substitutions, one can theoretically estimate the parameters of selection, mutation and drift, although it is important to keep in mind that selection and drift are confounded.

2.2.3 Reversibility of the process

The continuous-time Markov chain has so far been defined for 2 alleles but can be generalized to any number of alleles, when the number of alleles is discrete (n) and when transition from any allele to any other allele is possible in one or more substitutions. In this configuration, the transition rates between all possible pairs of alleles is defined by equation 2.35, and equals 0 whenever single step transitions are not possible. Because any state is ultimately connected to any other state, the continuous-time Markov chain is irreducible. Moreover, this substitution process is positive recurrent and aperiodic since any strictly positive transition rate is matched by a strictly positive transition for the reverse substitution. More precisely, the substitution rate between two alleles is null only if the underlying mutation rate is null, in which case the transition rate for the reverse mutation is also null, hence the transition rate for the reverse substitution is also null.

Theoretically, for an irreducible, positive recurrent and aperiodic continuous-time Markov chain, a necessary and sufficient condition to be reversible is given by Kolmogorov's criterion. Kolmogorov's criterion implies that the product of transition rates through any closed loop is the same whenever the traversing is done forward or in reverse. As an example for a Markov chain composed of 3 alleles (i, j and k), as illustrated in figure 2.6, the transition rates must satisfy the equality:

$$Q_{i\to j}Q_{j\to k}Q_{k\to i} = Q_{i\to k}Q_{k\to j}Q_{j\to i} \tag{2.40}$$

Kolmogorov's criterion is satisfied under specific conditions for the substitution pro-

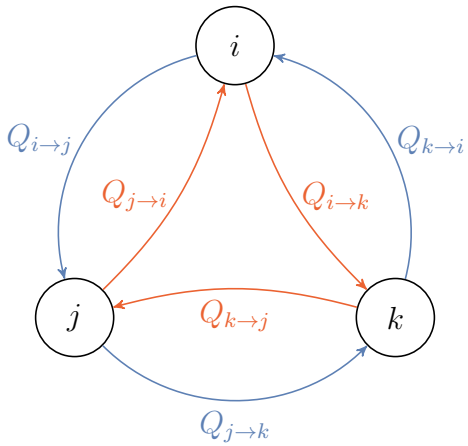


Figure 2.6: *The continuous-time Markov chain is reversible if the process fulfils Kolmogorov's criterion. Namely, the product of the transition rates for a closed loop is equal whether traversed in one sense (blue arrows) or the other (red arrows).*

cess (2.35):

$$1 = \frac{Q_{i\to j}Q_{j\to k}Q_{k\to i}}{Q_{i\to k}Q_{k\to j}Q_{j\to i}} \tag{2.41}$$

$$= \frac{\mu_{i\to j}\mu_{j\to k}\mu_{k\to i}}{\mu_{i\to k}\mu_{k\to j}\mu_{j\to i}} \times \frac{(F_j - F_i)(F_k - F_j)(F_i - F_k)}{(F_k - F_i)(F_j - F_k)(F_i - F_j)} \times \frac{(1 - \mathrm{e}^{F_i - F_k})(1 - \mathrm{e}^{F_k - F_j})(1 - \mathrm{e}^{F_j - F_i})}{(1 - \mathrm{e}^{F_i - F_j})(1 - \mathrm{e}^{F_j - F_k})(1 - \mathrm{e}^{F_k - F_i})}, \tag{2.42}$$

$$= \frac{\mu_{i\to j}\mu_{j\to k}\mu_{k\to i}}{\mu_{i\to k}\mu_{k\to j}\mu_{j\to i}} \times -\frac{\cancel{(F_i - F_j)(F_j - F_k)(F_k - F_i)}}{\cancel{(F_k - F_i)(F_j - F_k)(F_i - F_j)}} \times \frac{(\mathrm{e}^{F_i - F_i} - \mathrm{e}^{F_i - F_k})(\mathrm{e}^{F_k - F_k} - \mathrm{e}^{F_k - F_j})(\mathrm{e}^{F_j - F_j} - \mathrm{e}^{F_j - F_i})}{(\mathrm{e}^{F_i - F_i} - \mathrm{e}^{F_i - F_j})(\mathrm{e}^{F_j - F_j} - \mathrm{e}^{F_j - F_k})(\mathrm{e}^{F_k - F_k} - \mathrm{e}^{F_k - F_i})}, \tag{2.43}$$

$$= -\frac{\mu_{i\to j}\mu_{j\to k}\mu_{k\to i}}{\mu_{i\to k}\mu_{k\to j}\mu_{j\to i}} \times \frac{\cancel{\mathrm{e}^{F_i}}(\mathrm{e}^{-F_i} - \mathrm{e}^{-F_k})\cancel{\mathrm{e}^{F_k}}(\mathrm{e}^{-F_k} - \mathrm{e}^{-F_j})\cancel{\mathrm{e}^{F_j}}(\mathrm{e}^{-F_j} - \mathrm{e}^{-F_i})}{\cancel{\mathrm{e}^{F_i}}(\mathrm{e}^{-F_i} - \mathrm{e}^{-F_j})\cancel{\mathrm{e}^{F_j}}(\mathrm{e}^{-F_j} - \mathrm{e}^{-F_k})\cancel{\mathrm{e}^{F_k}}(\mathrm{e}^{-F_k} - \mathrm{e}^{-F_i})}, \tag{2.44}$$

$$= \frac{\mu_{i\to j}\mu_{j\to k}\mu_{k\to i}}{\mu_{i\to k}\mu_{k\to j}\mu_{j\to i}} \frac{\cancel{(\mathrm{e}^{-F_k} - \mathrm{e}^{-F_i})(\mathrm{e}^{-F_j} - \mathrm{e}^{-F_k})(\mathrm{e}^{-F_i} - \mathrm{e}^{-F_j})}}{\cancel{(\mathrm{e}^{-F_i} - \mathrm{e}^{-F_j})(\mathrm{e}^{-F_j} - \mathrm{e}^{-F_k})(\mathrm{e}^{-F_k} - \mathrm{e}^{-F_i})}}, \tag{2.45}$$

$$= \frac{\mu_{i\to j}\mu_{j\to k}\mu_{k\to i}}{\mu_{i\to k}\mu_{k\to j}\mu_{j\to i}}. \tag{2.46}$$

Namely, Kolmogorov's criterion for the substitution process is satisfied only if the mutation process is also reversible, in which case Kolmogorov's criterion is also fulfilled:

$$\mu_{i\to j}\mu_{j\to k}\mu_{k\to i} = \mu_{i\to k}\mu_{k\to j}\mu_{j\to i}. \tag{2.47}$$

This example can be generalized for any closed loop, such that the reversibility of the substitution process is conditioned on the reversibility of the underlying mutation process, which is often assumed.

2.2.4 Stationary distribution

A realization of the Markov chain for a long period of time results in a given proportion of the time for which the process is fixed for a specific allele, where this proportion

depends of the allele fitness, the mutational process and N_e. Because the continuous-time Markov chain is irreducible, positive recurrent and aperiodic, it has a unique stationary distribution $\boldsymbol{\pi}$, where π_i corresponds to the proportion of time spent in allele i ($1 \leq\leq n$) after the Markov chain has run for an infinite amount of time.

Moreover, under the condition that the Markov chain is time-reversible, the detailed balance for the stationary distribution is satisfied for every pair i and j:

$$\frac{\pi_i}{\pi_j} = \frac{Q_{j\to i}}{Q_{i\to j}} \tag{2.48}$$

$$= \frac{\mu_{j\to i}}{\mu_{i\to j}} \frac{F_i - F_j}{1 - \mathrm{e}^{F_j - F_j}} \frac{1 - \mathrm{e}^{F_i - F_j}}{F_j - F_i} \tag{2.49}$$

$$= -\frac{\mu_{j\to i}}{\mu_{i\to j}} \frac{\mathrm{e}^{F_i}(\mathrm{e}^{-F_i} - \mathrm{e}^{-F_j})}{\mathrm{e}^{F_j}(\mathrm{e}^{-F_j} - \mathrm{e}^{-F_i})} \tag{2.50}$$

$$= \frac{\mu_{j\to i}}{\mu_{i\to j}} \frac{\mathrm{e}^{F_i}}{\mathrm{e}^{F_j}} \tag{2.51}$$

$$\tag{2.52}$$

Under the assumption that the mutational process is also reversible, the detailed balance for the stationary distribution of the mutation process ($\boldsymbol{\sigma}$) is satisfied for every pair i and j:

$$\frac{\mu_{j\to i}}{\mu_{i\to j}} = \frac{\sigma_i}{\sigma_j} \tag{2.53}$$

Altogether, the probability π_i to find the population in allele i is proportional to a function (also called a Boltzmann factor) that depends only on the fitness of allele i, the population size, and details of the mutation process (Sella and Hirsh, 2005; Mustonen and Lässig, 2005):

$$\frac{\pi_i}{\pi_j} = \frac{\sigma_i \mathrm{e}^{F_i}}{\sigma_j \mathrm{e}^{F_j}} \text{ and } \sum_{i=1}^{n} \pi_i = 1, \tag{2.54}$$

$$\iff \pi_i = \frac{\sigma_i \mathrm{e}^{F_i}}{\sum_{j=1}^{n} \sigma_j \mathrm{e}^{F_j}}, \tag{2.55}$$

where the denominator is a normalizing constant such that the sum of probabilities is equal to 1. By analogy with thermodynamic systems, the evolutionary system thus reaches a Boltzmann-like distribution with N_e^{-1} playing the role of evolutionary temperature, and the log-fitness f the role of energy[1].

2.2.5 Mean scaled fixation probability

Occurrence probabilities given by the stationary distribution allows one to calculate all observable quantities of interest, such as the mean fitness, or the mean mutation and

[1] At high mutation rates, the quasi-species theory provides another analogy with statistical mechanics, in which the mutation rate plays the role of temperature instead of genetic drift.

substitution rates, using standard probability theory. One quantity of interest is the ratio of the mean substitution rate over the mean mutation rate, called ν:

$$\nu = \frac{\langle Q \rangle}{\langle \mu \rangle}, \tag{2.56}$$

$$= \frac{\sum_{1 \leq i,j \leq n} \pi_i Q_{i \to j}}{\sum_{1 \leq i,j \leq n} \pi_i \mu_{i \to j}}, \tag{2.57}$$

where the notation $\langle \cdot \rangle$ denotes the statistical average, and the sum is over all possible pairs of codons having a certain property. In other words, ν represents the flow of substitutions at equilibrium, normalized by the mutational flow (or mutational opportunities).

This definition can in principle be applied to any subset of codon pairs. A particularly important case is to sum over all possible pairs of non-synonymous codons (which will be considered in the next chapter). In that case, ν captures the fundamental quantity usually referred to as d_N/d_S. However, the definition is more general.

This ratio can also be interpreted as the mean scaled fixation probability of all mutations that are being proposed at mutation selection equilibrium. Indeed, the scaled fixation probability of a given mutation is the probability of fixation of this mutation, normalized by the fixation probability of neutral mutations:

$$\frac{\mathbb{P}_{\text{fix}}(s_{i \to j}, N_{\text{e}})}{\mathbb{P}_{\text{fix}}(0, N_{\text{e}})} = 2N_{\text{e}}\mathbb{P}_{\text{fix}}(s_{i \to j}, N_{\text{e}}) \tag{2.58}$$

In addition, the probability for a given type of mutation, from i to j, to be proposed at equilibrium, is given by:

$$\mathbb{P}(i \to j) = \frac{\pi_i \mu_{i \to j}}{\mathcal{Z}}, \text{ where } \mathcal{Z} = \sum_{1 \leq i,j \leq n} \pi_i \mu_{i \to j} \tag{2.59}$$

And thus, the statistical average at equilibrium is:

$$\langle 2N_{\text{e}}\mathbb{P}_{\text{fix}} \rangle = \sum_{1 \leq i,j \leq n} \mathbb{P}(i \to j) 2N_{\text{e}}\mathbb{P}_{\text{fix}}(s_{i \to j}, N_{\text{e}}), \tag{2.60}$$

$$= \frac{\sum_{1 \leq i,j \leq n} \pi_i Q_{i \to j}}{\sum_{1 \leq i,j \leq n} \pi_i \mu_{i \to j}}, \text{ from equation 2.32 and 2.59}, \tag{2.61}$$

$$= \nu. \tag{2.62}$$

As a result of this definition, $\nu = 1$ for genes or sites under neutral evolution. Most importantly, departure from 1 would be interpreted as a signature of selection on sequences. First, $\nu > 1$ is interpreted as a signal of adaptive recurrent evolution, since this means that $\mathbb{P}_{\text{fix}} > 1/2N_{\text{e}}$ on average. On the other hand, $\nu < 1$ is a signal of underlying purifying selection such that the sequence is constrained on average. Of note, $\nu > 1$ (or < 1) does not necessarily mean that the selection coefficients are positive (or negative) on average. Finally, a mutation-selection point substitution process at equilibrium under a time-independent fitness landscape results in $\nu \leq 1$, as demonstrated in Spielman and Wilke (2015).

2.3 Mutation-selection analogy in other scientific fields

Presented in the context of phylogenetic evolution of genetic sequences, the mutation-selection process bears many similarities and analogies between other processes present in a variety of scientific fields outside of evolutionary biology, displaying the same underlying mechanism and emerging properties, though with different names and aspirations. This section is an attempt to describe analogous processes and their emerging properties. This effort is made in the aim of giving another view of the mutation-selection process, such as to better appreciate and conceptualize its assumptions, its limits, and the respective role of the different components. Such attempts require to boil down the mutation-selection mechanism into its core components, while at the same time rephrasing the description using lexicography outside of population genetics such as to open new perceiving angles.

2.3.1 Metropolis-Hastings sampling

Obtaining a sequence of random samples from a probability distribution can be difficult, especially when the number of dimensions is high. However, the Metropolis-Hastings procedure based on a Markov chain Monte Carlo can sample from any probability distribution, provided that we know how to compute the probability density, or even less restrictively any function proportional to the density (Hastings, 1970). This stochastic procedure which is based on three steps bears many similarities with the mutation-selection process:

- Generate a stochastic candidate from the current state, analogous to mutation.
- Calculate the acceptance ratio as the ratio of the two densities, analogous to the selection coefficient of the mutated state.
- Stochastic acceptance or rejection based on the acceptance ratio, a process analogous to drift.

Inherently, the Metropolis-Hastings procedure is based on creating and subsequently reducing diversity, which allows to obtain a random sequence of samples from any distribution with a straightforward recipe, and is a critical tool in statistics and statistical physics.

2.3.2 The exploration-exploitation dilemma

Many mathematical, engineering and daily-life problems are not about sampling a state space, but rather about finding the optimal and best state given the criteria or a function to maximize. Naturally, we would prefer deterministic (strictly reproducible) rather than stochastic optimizing strategies to search for an optimal state. Unfortunately, whenever the state space is too large, often due to the curse of dimensionality, a greedy or heuristic search of an optimal state can perform atrociously (Bellman, 1966). In high-dimensional space, stochastic optimization tools have been deemed very valuable, such

as stochastic gradient descent or so-called evolutionary algorithms (Russell and Norvig, 2010; Vikhar, 2017). Inherently, they are based on two processes, one is stochastically creating diversity and exploring the state space, while the other is filtering the explored states and thus reducing the diversity.

In the constrained case of a finite amount of time or attempts to find the best outcome overall, the problem is best described by the multi-armed bandit problem. The name comes from imagining a gambler at a row of slot machines (sometimes known as one-armed bandits), where each slot machine provides a random reward from a probability distribution specific to that machine. The player has to decide which machines to play, how many times to play each machine and in which order to play them, and whether to continue with the current machine or try a different machine, such as to maximize the sum of rewards earned through a sequence of trials. The gambler faces a dilemma at each trial, either reducing his regret by exploiting the best arm, or gaining information through exploration of other arms. The best strategy to solve this dilemma can be mathematically derived in numerous cases, and encompasses mixing strategies with a defined ratio of exploration and exploitation (Auer *et al.*, 2002; Kocsis and Szepesvári, 2006; Fürnkranz *et al.*, 2006). This problem is far from being only theoretical, and has been used to explain a multitude of phenomena, such as the movement of animals in novel landscapes, the most efficient resource allocation for a start-up company, the effects of age on knowledge acquisition in humans, and in the search of the most efficient treatment in clinical trials (Berger-Tal *et al.*, 2014; March, 1991). Another application of the exploration-exploitation dilemma is AlphaGo, the first computational program mastering the board game Go at the professional 9-dan level in 2017, which outcompeted Ke Jie, the world first ranked player at the time (Silver *et al.*, 2017, 2018). AlphaGo has often been publicized and hyped in various media outlets stating that this feat was possible due to machine learning, more specifically due to convolutional neural networks. However, it is more scarcely mentioned that the AlphaGo neural network is combined with an exploration-exploitation algorithm, or more specifically a Monte Carlo tree search. In practice, the convolutional neural network is used as a criterion to measure the advantage of a board configuration[2], but the different moves and paths probed and trimmed are done via an exploration-exploitation procedure.

2.3.3 Interaction between analogies

At the bottom, mutation is a process creating diversity, changing and moving the current viable state to a novel and unknown position, fundamentally allowing exploration of the state space. On the other hand, selection is the criteria on which a new state is deemed a disrupting innovation or a nonviable alteration, and allows to determine which changes to exploit and which to filter out and discard based on its fitness. Fundamentally, mutation creates diversity and selection reduces this diversity by selecting the fittest

[2]Convolutional neural networks also use a stochastic gradient descent to reach convergence, inherently leveraging the stochastic exploration and exploitation procedure to optimize the parameters of the neural network.

mutants. Finally, drift arbitrates between the creation and reduction of the two processes, it dictates how much exploration of novelty is permitted, and conversely how much exploitation of only the fittest states is granted.

Exploration and exploitation, creation and reduction, mutation and selection, are different names (see table 2.4) that ultimately encompass the inherently same process: efficiently sampling and optimizing whenever the state space is too large to be traversed in a finite amount of time.

Mutation	**Selection**	**Drift**
Exploration	Exploitation	Trade-off
Creation	Reduction	Arbitration
Candidate generation	Acceptance	Hastings ratio

Table 2.4: *Mutation, selection and drift lexicographic rephrasing in different fields.*

I argue that evolutionary biologists, studying and leveraging the pervasive process of mutation and selection, can gain knowledge by recruiting insight and developments from other fields, much like there has been many crossovers between economics and evolution in the context of game theory.[3]. From a political standpoint, I also argue that scientific research endeavour is also an exploration-exploitation dilemma, which is arguably externally pressured to pursue exploitation, through funding of impactful research and a publish-or-perish systemic culture in the early career stage.

[3]Game theory was originally developed to model economic actors' behaviour and strategies (Von Neumann and Morgenstern, 1947). It was later adopted within the framework of evolutionary dynamics, helping to explain, for example, the emergence of altruistic behaviour in Darwinian evolution (Smith and Price, 1973; Smith, 1982; Nowak, 2006).

3

Phylogenetic codon models

Contents

Evolutionary trajectories of sequences depend on the forces of mutation, selection and drift, which act conjointly such that each one of them must be well studied and understood. More precisely, models of molecular evolution requires either a given selection coefficient associated to mutation, or that the fitness of each particular sequence is defined. In other words, the relationship between sequence and fitness must be eluci-

dated, which is the focus of the present chapter in the special case of protein-coding DNA sequences. To this aim, this chapter will first present the genetic code and classical phylogenetic codons models, which can quantify the strength of selection acting on proteins through an aggregate parameter (called ω or d_N/d_S). Application of these phylogenetic models to empirical DNA alignments can be extended to model variation of selection across sites of the same protein, or between branches of a phylogenetic tree. Subsequently, mechanistic codon models are presented, assuming that the DNA sequence is at mutation-selection balance under a time-independent fitness landscape over the 20 amino acids. Finally, the relationship between classical and mechanistic models is investigated, and the interpretation of the discrepancy between both models is analysed.

3.1 Protein coding DNA sequences

Proteins have a variety of molecular and cellular roles, they are the enzymes that catalyse chemical bonds, they regulate cell processes and control their rates, they carry signals within the cell and across membranes, they bind and transport small molecules, they form cellular structures, among other functions. This diversity of roles is accomplished by a variety of three-dimensional shapes. A protein's three-dimensional shape is in turn determined by the linear one-dimensional sequence of amino acids of which it is made of, with protein sequences ranging from fewer than 20 to more than 5000 amino acids across the tree of life, with an average of about 350 amino acids. Just as DNA is oriented because of the asymmetry of nucleotides, proteins are oriented due to the asymmetry of amino acids. One end is called the N-terminus, and the other end, the C-terminus, and each amino acid will interact with the other amino acids in its spatial vicinity.

Although each of the 20 different amino acids has unique biochemical properties, they can be classified broadly into four categories determining their solubility and acidity (classification is given in table 3.1). Charged amino acids can be either basic (positively charged) or acidic (negatively charged). However, non-charged amino acids can be polar due to an uneven charge distribution, such that they can form hydrogen bonds with water. Consequently, polar amino acids are called hydrophilic, and are often found on the outer surface of folded proteins. Also, non-charged amino acids can have a uniform charge distribution, and do not form hydrogen bonds with water. Reciprocally, these non-polar amino acids are called hydrophobic and tend to be found in the core of folded proteins.

3.1.1 The genetic code

Because the 20 letter alphabet of proteins is different to the 4 letter alphabet of nucleic acids (DNA and RNA), there is not a one-to-one correspondence between the two alphabets. Instead, amino acids are encoded by codons, a consecutive sequence of 3 nucleotides, yielding $4^3 = 64$ possible permutations, more than sufficient to encode the 20 different amino acids. Moreover, three stop codons (TGA, TAA and TAG) signal the termination of the protein, such that 61 of the 64 codons are used to encode amino acids. Since there

are 61 coding codons and only 20 amino acids, there is a necessary redundancy in the code. Thus, amino acids are encoded by synonymous codons, which are interchangeable in the sense of producing the same amino acid, with the notable exception of methionine and tryptophan, which are only encoded by a singe codon. Altogether, the standard DNA genetic code, which is used by many organisms, translates codon to amino acids as given in table 3.1. To note, there are organisms that use other genetic codes, and in addition many of our genes are mitochondrial, which also use a different genetic code.

	T		C		A		G		
T	TTT	Phenylalanine (Phe/P)	TCT	Serine (Ser/S)	TAT	Tyrosine (Tyr/Y)	TGT	Cysteine (Cys/C)	**T**
	TTC		TCC		TAC		TGC		**C**
	TTA	Leucine (Leu/L)	TCA		TAA	Stop (Ochre)	TGA	Stop (Opal)	**A**
	TTG		TCG		TAG	Stop (Amber)	TGG	Tryptophan (Trp/W)	**G**
C	CTT		CCT	Proline (Pro/P)	CAT	Histidine (His/H)	CGT	Arginine (Arg/R)	**T**
	CTC		CCC		CAC		CGC		**C**
	CTA		CCA		CAA	Glutamine (Gln/Q)	CGA		**A**
	CTG		CCG		CAG		CGG		**G**
A	ATT	Isoleucine (Ile/I)	ACT	Threonine (Thr/T)	AAT	Asparagine (Asn/N)	AGT	Serine (Ser/S)	**T**
	ATC		ACC		AAC		AGC		**C**
	ATA		ACA		AAA	Lysine (Lys/K)	AGA	Arginine (Arg/R)	**A**
	ATG	Methionine (Met/M)	ACG		AAG		AGG		**G**
G	GTT	Valine (Val/V)	GCT	Alanine (Ala/A)	GAT	Aspartic acid (Asp/D)	GGT	Glycine (Gly/G)	**T**
	GTC		GCC		GAC		GGC		**C**
	GTA		GCA		GAA	Glutamic acid (Glu/E)	GGA		**A**
	GTG		GCG		GAG		GGG		**G**

Table 3.1: *The genetic code DNA table translating codons into amino acids. Amino acids are represented into* 4 *categories based on electrochemical properties. Non-polar in yellow ( ), polar in green ( ), basic in blue ( ) and finally acidic in red ( ). Stop codons are represented in gray ( ). The synonymous codons encoding for the same amino acid are usually different in their third codon position, the wooble base.*

Biochemical translation from codon to amino acid mechanistically emanates from transfer RNA (tRNA). More precisely, codons bind to tRNA via an anticodon, three consecutive bases that are complementary and antiparallel to the associated codon. On the other end, a given tRNA binds uniquely with one of the 20 amino acids, where the catalytic reaction is performed by aminoacyl-tRNA synthetase (Rich and RajBhandary, 1976). As a result, tRNA genes along with aminoacyl-tRNA synthetase genes constitute the machinery necessary for translating codons into amino acids . However, there is not a one-to-one correspondence between the 61 codons and tRNA genes. First, the set of unique sequences of anticodon found in tRNAs genes is actually lower than 61, and depends on the species but varies from 41 to 55 (Goodenbour and Pan, 2006). This subset of anticodon sequences necessary to bind all 61 codons is due to non-canonical base pairing[1]. More precisely, the first two positions in the codon bind strongly to the anticodon of the tRNA (second and third positions), while the third base of the codon can be subject to non-standard pairing with the first base of the anticodon. If the anticodon contains a guanine at first position, codons with either U or C at the third position can bind to this anticodon, and this phenomenon explains why there is not any non-synonymous transition from only U to C at the third position, and why synonymous

[1]Canonical base pairing are A-U and G-C, where thymine (T) is replaced by uracil (U) in RNA

codons usually end with T or C. Also, if the anticodon contains an inosine at the first position, codons with either C, U or A at the third position can bind to this anticodon, such that for example leucine encoded by three codons (AUU, AUC, AUA) can be bound by the unique anticodon IAU. Altogether, non-standard pairing explains why the number of unique anticodons is lower than the number of possible codons, and also explains some part of the structure of the genetic code.

Secondly, tRNA genes with the same amino-acid binding site and anticodon, which are called isoacceptor tRNA, may vary in other parts of the tRNA sequence. Effectively, many genes can code for the same isoacceptor tRNA, where each gene can display varying efficiency and errors in translation, adding a layer of regulation to the process of protein synthesis (Lowe and Eddy, 1997; Chan and Lowe, 2008; Jühling *et al.*, 2008; Lin *et al.*, 2019). As a result, in some genes, some codons are more frequently represented than other possible synonymous codons, an effect named codon usage bias. For genes that are expressed at high levels, the codon usage is biased in favour of the codons that have a high tRNA concentration in the cell, ultimately increasing the expression rate and decreasing the rate of mistranslation by reducing the time of occupancy of an open site. Thus, at a fine-grained molecular scope, a synonymous change can influence mRNA stability, splicing process and protein folding during translation (Plotkin and Kudla, 2011; Rak *et al.*, 2018). However in the scope of this manuscript, such selection between synonymous codons will not be considered. Selection for proteins will be framed at the amino-acid level in a first approximation, and mutation, at the nucleotide level.

3.1.2 Amino-acid transitions

Because mutations are at the nucleotide level and affect only one base, any codon can have at most 9 possible transitions to another codon as illustrated in the left panel of figure 3.1 as a graph. Moreover, it is possible that some pairs of amino acids are not accessible through a single non-synonymous transition between the underlying codons. In fact, most pairs of amino acids require at least two non-synonymous transitions (114 pairs), in comparison to pairs of amino acids that are accessible through a single non-synonymous transition (75 pairs). More precisely, the number of possible transitions between the underlying codons for a pair of amino acids is determined by the adjacency matrix shown table 3.2, which is illustrated in the right panel of figure 3.1 as a graph.

3.2 Classical codon models

Under the approximation that selection occurs for proteins, designing substitution models at the amino-acid level has the major shortcoming of not taking into account that the underlying mutation process occurs at the nucleotide level. Conversely, studying evolution of protein-coding DNA sequences only at the nucleotide level, while disregarding the genetic code neglects the consequences that nucleotide variation can have onto protein sequences.

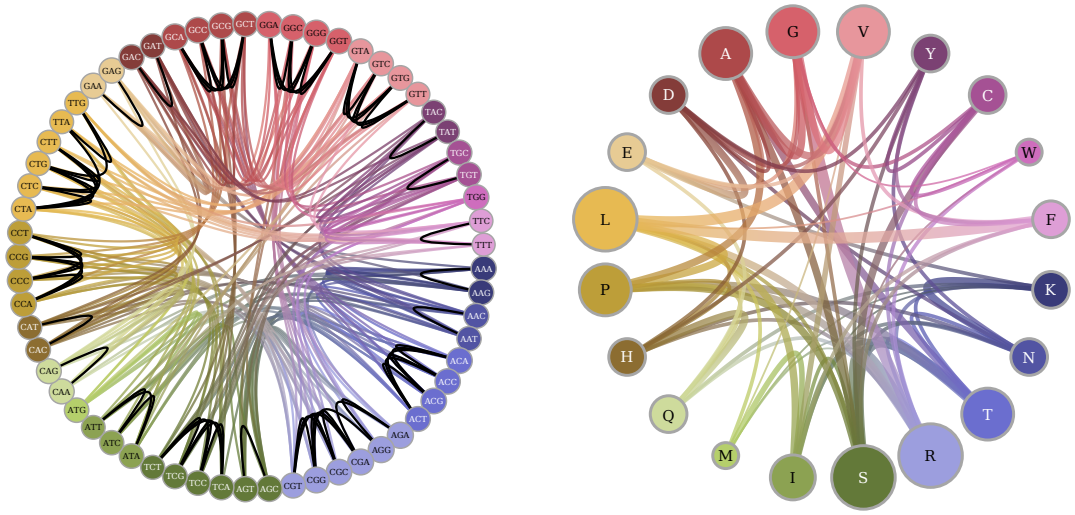


Figure 3.1: *Graphs of possible one nucleotide transitions between codons (left panel) and between amino acids (right panel). Nodes correspond to codons (left panel) and amino acids (right panel), and their colour represents the encoded amino acid. Additionally, for amino acids, the size of nodes represents the number of underlying codons. An edge between two codons depicts a one nucleotide transition such that a codon can have at most* 9 *possible transitions. Similarly, an edge between two amino acids correspond to a one nucleotide non-synonymous transition between the underlying codons, and the width of the edges represents the number of such possible transitions. Non-synonymous transitions are represented in a colour gradient, while synonymous transitions are depicted in black. The graph of the* 61 *codons contains* 263 *transitions,* 67 *of them are synonymous while* 196 *are non-synonymous. Codons encoding for the same amino acid are all fully connected by synonymous changes, except for serine where a transition from the set TCT, TCG, TCC, TCA to the set AGT, AGC requires passing through another amino acid, hence at least two non-synonymous transitions. From the perspective of amino acids, the graph of the* 20 *amino acids contains* 75 *non-synonymous transitions. The graph is not fully connected and does not form a clique. Moreover, the most distant amino acids are at most three transitions away, because a transition from methionine to tyrosine requires at least three non-synonymous transitions. Altogether, for all of the possible* 190 *pairs of amino acids,* 114 *pairs require at least two non-synonymous transitions, and one pair (M-Y) requires at least three non-synonymous transitions.*

	K	N	T	R	S	I	M	Q	H	P	L	E	D	A	G	V	Y	C	W	F
K	-	4	2	2	0	1	1	2	0	0	0	2	0	0	0	0	0	0	0	0
N	-	-	2	0	2	2	0	0	2	0	0	0	2	0	0	0	2	0	0	0
T	-	-	-	2	6	3	1	0	0	4	0	0	0	4	0	0	0	0	0	0
R	-	-	-	-	6	1	1	2	2	4	4	0	0	0	6	0	0	2	2	0
S	-	-	-	-	-	2	0	0	0	4	2	0	0	4	2	0	2	4	1	2
I	-	-	-	-	-	-	3	0	0	0	4	0	0	0	0	3	0	0	0	2
M	-	-	-	-	-	-	-	0	0	0	2	0	0	0	0	1	0	0	0	0
Q	-	-	-	-	-	-	-	-	4	2	2	2	0	0	0	0	0	0	0	0
H	-	-	-	-	-	-	-	-	-	2	2	0	2	0	0	0	2	0	0	0
P	-	-	-	-	-	-	-	-	-	-	4	0	0	4	0	0	0	0	0	0
L	-	-	-	-	-	-	-	-	-	-	-	0	0	0	0	6	0	0	1	6
E	-	-	-	-	-	-	-	-	-	-	-	-	4	2	2	2	0	0	0	0
D	-	-	-	-	-	-	-	-	-	-	-	-	-	2	2	2	2	0	0	0
A	-	-	-	-	-	-	-	-	-	-	-	-	-	-	4	4	0	0	0	0
G	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	4	0	2	1	0
V	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0	0	0	2
Y	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	2	0	2
C	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	2	2
W	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0
F	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-

Table 3.2: *Number of possible one nucleotide non-synonymous transitions between amino acids, integrating over the underlying codons, represented as an adjacency matrix. For all the possible* 190 *pairs of amino acids, only* 75 *pairs contain at least one non-synonymous transition.*

These shortcomings are both addressed by codon models, where the complexity of the genetic code is seen as an asset rather than an encumbrance. Indeed the redundancy in the genetic code can be leveraged to disentangle mutation and selection in protein-coding DNA sequences, under the approximation that selection occurs at the protein level in first approximation, while the mutation process occurs at the DNA level. The genetic code allows to split mutations into synonymous and non-synonymous mutations, where synonymous mutations are deemed neutral, and non-synonymous mutations are considered under selection. Thus, by contrasting the two types of substitutions, non-synonymous against synonymous, one can estimate the impact of selection, effectively factoring out the contribution of the mutation rate and the mutation patterns. This idea was already present in the earliest landmark contributions in molecular evolution (Kimura, 1968; King and Jukes, 1969), using simple statistical approaches. However, the mathematical complexities created by the very irregular nature of the genetic code led to the progressive development of more sophisticated probabilistic models, formalized in a likelihood framework. The first codon models were proposed independently by Muse and Gaut (1994) and Goldman and Yang (1994). The mathematical formalism is now presented in more detail.

3.2.1 The Muse & Gaut formalism

Here, we follow the formalism of codon models pioneered by Muse and Gaut (1994), and further developed by Nielsen and Yang (1998). A 4×4 mutation rate matrix $\boldsymbol{R}$ is first

defined at the nucleotide level. In its most general form consisting of 12 free parameters:

$$\boldsymbol{R} = \begin{array}{c} \\ A \\ C \\ G \\ T \end{array} \begin{array}{c} \begin{array}{cccc} A & C & G & T \end{array} \\ \begin{pmatrix} - & R_{AC} & R_{AG} & R_{AT} \\ R_{CA} & - & R_{CG} & R_{CT} \\ R_{GA} & R_{GC} & - & R_{GT} \\ R_{TA} & R_{TC} & R_{TG} & - \end{pmatrix} \end{array} \tag{3.1}$$

By definition of the instantaneous rate matrix, the sum of the entries in each row of the nucleotide rate matrix $\boldsymbol{R}$ is equal to 0, giving the diagonal entries:

$$R_{aa} = -\sum_{b \neq a} R_{ab}, \, \forall a \in \{A, C, G, T\} \tag{3.2}$$

Most often, this matrix is assumed to be a generalized time-reversible (Tavaré, 1986), or in short GTR, defined by nucleotide equilibrium frequencies ($\boldsymbol{\sigma}$) and by symmetric exchangeability rates ($\boldsymbol{\rho}$) consisting of 9 free parameters:

$$\boldsymbol{R} = \begin{array}{c} \\ A \\ C \\ G \\ T \end{array} \begin{array}{c} \begin{array}{cccc} A & C & G & T \end{array} \\ \begin{pmatrix} - & \rho_{AC}\sigma_C & \rho_{AG}\sigma_G & \rho_{AT}\sigma_T \\ \rho_{AC}\sigma_A & - & \rho_{CG}\sigma_G & \rho_{CT}\sigma_T \\ \rho_{AG}\sigma_A & \rho_{CG}\sigma_C & - & \rho_{GT}\sigma_T \\ \rho_{AT}\sigma_A & \rho_{CT}\sigma_C & \rho_{GT}\sigma_G & - \end{pmatrix} \end{array} \tag{3.3}$$

Then, grouping nucleotides into codons, the mutation rate induced by this nucleotide process from codon i to j depends on the underlying nucleotide change between the two codons. Thus, if codons i and j are only a mutation away, let $\mathcal{M}(i, j)$ denote the nucleotide change between them (e.g. $\mathcal{M}(AAT, AAG) = TG$). With this notation, the mutation rate $\mu_{i,j}$ from codon i to j is:

$$\mu_{i,j} = \begin{cases} R_{\mathcal{M}(i,j)} \text{ if codons } i \text{ and } j \text{ are one mutation away,} \\ 0 \text{ else.} \end{cases} \tag{3.4}$$

In other words, the mutation rate between codons is simply the mutation rate between the underlying nucleotide change.

At the codon level, synonymous mutations are deemed neutral and the rate of synonymous substitutions $Q_{i,j}$ is equal to the mutation rate:

$$Q_{i,j} = \mu_{i,j}, \tag{3.5}$$

$$= R_{\mathcal{M}(i,j)}. \tag{3.6}$$

In contrast, non-synonymous mutations are considered under selection such that the rate of substitution is modulated by a factor ω:

$$Q_{i,j} = \omega\mu_{i,j}, \tag{3.7}$$

$$= \omega R_{\mathcal{M}(i,j)}. \tag{3.8}$$

Altogether, the 61-by-61 codon substitution matrix of Muse and Gaut (1994) is defined entirely by the mutation matrix ($\boldsymbol{R}$), ω and the genetic code:

$$\begin{cases} Q_{i,j} &= 0 \text{ if codons } i \text{ and } j \text{ are more than one mutation away,} \\ Q_{i,j} &= R_{\mathcal{M}(i,j)} \text{ if codons } i \text{ and } j \text{ are synonymous,} \\ Q_{i,j} &= \omega R_{\mathcal{M}(i,j)} \text{ if codons } i \text{ and } j \text{ are non-synonymous.} \end{cases} \tag{3.9}$$

Again, by definition of the instantaneous rate matrix, the sum of the entries in each row of the codon substitution rate matrix $\boldsymbol{Q}$ is equal to 0, giving the diagonal entries:

$$Q_{i,i} = -\sum_{j \neq i, j=1}^{61} Q_{i,j}. \tag{3.10}$$

3.2.2 Interpretation of the model

With the definition given above, ω identifies with the ratio of the rate of non-synonymous substitutions over the rate of synonymous substitutions, hence d_N/d_S. More globally, given how its parameterization carefully distinguishes between synonymous and non-synonymous substitutions, the model can be seen as trying to separate the effects of the mutation rates (captured by $\boldsymbol{R}$) and those of selection at the non-synonymous level (captured by ω).

All non-synonymous mutations are considered equivalent, and ω encompasses the average strength of selection exercised on them. Most importantly, $\omega > 1$ is due to an excess in the rate of non-synonymous substitutions, indicating that the protein is under adaptive evolution. Conversely, a default of non-synonymous substitutions, leading to $\omega < 1$, means the protein is on average under purifying selection. It is worth noting that the protein can be on average under purifying selection ($\omega < 1$), but can have specific regions undergoing positive selection ($\omega > 1$).

3.2.3 Equilibrium properties

Under the Muse & Gaut formalism, the codon equilibrium frequencies ($\boldsymbol{\pi}$) depend only on the equilibrium nucleotide frequencies ($\boldsymbol{\sigma}$), but not on ω:

$$\pi_i = \frac{\left[\prod_{k \in \{1,2,3\}} \sigma_{i[k]}\right]}{\sum_{j=1}^{61} \sigma_{j[1]}\sigma_{j[2]}\sigma_{j[3]}} \tag{3.11}$$

$$= \frac{\left[\prod_{k \in \{1,2,3\}} \sigma_{i[k]}\right]}{(1 - \sigma_T\sigma_A\sigma_A - \sigma_T\sigma_A\sigma_G - \sigma_T\sigma_G\sigma_A)}, \tag{3.12}$$

where $i[k]$ denotes the nucleotide at position $k \in \{1, 2, 3\}$ of codon i, and the sum in the denominator can be obtained by simply correcting for the stop codons (TAA, TAG and TGA).

As a result of equation 3.12, the Muse & Gaut formalism predicts that the nucleotide composition is the same for all 3 positions of the codons. However it has empirically been observed that the nucleotide compositions are in fact not identical (Singer and Hickey, 2000). These modulations across the three coding positions have been accommodated using the so-called 3x4 formalism (Muse and Gaut, 1994; Goldman and Yang, 1994), allowing for different nucleotide rate matrices at the three positions. However, this is problematic, since this modelling has the consequence that synonymous substitutions occur at different rates at the first and third positions. For instance, mutations from codon CTC to CTT or from CTA to TTA are both synonymous (leucine) and from C to T, but the 3x4 formalism would give them different rates. Yet, in reality, the mutation process is blind to the coding structure, and should be homogeneous across coding positions, and if neutral, all mutations from C to T should have the same rate. In any case, this suggests that the mutation matrices estimated by codon models are not correctly reflecting the mutation rates between nucleotides.

3.2.4 The Goldman & Yang formalism

In the alternative Goldman and Yang (1994) formalism, the mutation rate between two codons does not depend only on the exchangeability between the underlying nucleotide change ($\rho_{\mathcal{M}(i,j)}$), but also on the frequency of the target codon (π_j):

$$\mu_{i,j} = \rho_{\mathcal{M}(i,j)}\pi_j. \tag{3.13}$$

Careful examination of this model reveals a number of peculiar properties, which seem undesirable. For example, under a mutational bias toward T, a synonymous mutation from codon AAC to AAT (asparagine) would have a lower instantaneous rate than a substitution from codon TTC to TTT (phenylalaline), both being synonymous and from C to T at third position. In this formalism, the mutation involving a specific codon position depends on the nucleotide states at the other two positions, even if the mutation is synonymous (neutral). Moreover, it has been shown that this alternative formalism induces different estimations of the strength of selection ω (Kosakovsky Pond and Muse, 2005b; Yap *et al.*, 2010; Spielman and Wilke, 2015). Altogether, such alternative formalisms are theoretically problematic, and the original Muse & Gaut formalism remains the mechanistically justified framework (Rodrigue *et al.*, 2008a).

As a result, throughout this manuscript the symbol ω will be used specifically for the multiplicative factor appearing in the Muse and Gaut (1994) formalism (see section 3.2.1), whereas d_N/d_S will be used to refer generically to the ratio of non-synonymous over synonymous substitution rates, regardless of the specific formalism. Hence, whenever d_N/d_S is used in this manuscript instead of ω, the underlying specific formalism is not considered necessary to the point raised. Contrarily, whenever ω is used, it refers to the specific Muse & Gaut formalism of section 3.2.1. A notable exception for this conventions is in the third article (chapter 9 and supplementary materials in chapter 12), where ω will be used for readability while having a slightly different meaning (mean scaled

fixation probability of non-synonymous mutations) but still identifies with the ratio of non-synonymous over synonymous substitution rates (see section 3.4.1).

3.2.5 Complexification of classical codon models

Classical models of codon substitutions have been extensively applied to protein-coding sequence alignments, to estimate the ratio of non-synonymous over synonymous substitution rates, d_N/d_S. Such models capture the average effect of selection on non-synonymous mutations, without seeking to discriminate between different types of mutations. To circumvent such limitation, Yang *et al.* (1998) introduced a codon model in which d_N/d_S depends on the distance between amino acids, measured in terms of the Grantham (1974) distance. Additionally, models introduced several d_N/d_S to account for amino-acid chemical properties (polarity, volume, charge, and so on) in classical codon models (Dutheil, 2008).

One particularly important application of classical codon models has been to characterize genes under positive selection (i.e. with a $d_N/d_S > 1$), or sites within genes or specific lineage under accelerated evolution. As a result, variants of codon models have been developed that can provide estimates of d_N/d_S for each site within a gene, or for each branch within a phylogenetic tree. Moreover, these codon models have also proved to be valuable to quantify and assess the modulation of the selective constraints more generally imposed on protein-coding sequences (see section 5.2).

3.2.6 Variation across sites

The strength of selection is not typically homogeneous along the protein sequence, and it has been rapidly recognized that it could be useful to estimate the d_N/d_S for each site individually, as opposed to globally over the entire sequence. This turns out to be particularly important for detecting recurrent diversifying selection. Indeed, recurrent positive selection might often be concentrated in a small region of the protein (e.g. domain or site of the protein that is more directly interacting with a pathogen), the rest of the protein being under a regime of purifying selection. Estimating d_N/d_S at the site level will make it possible to detect such regions under positive selection. In contrast, the gene-level d_N/d_S will generally be below 1.

However, the statistical information available along the tree for a specific site is sparse such that sites sharing similar patterns are merged together to gather enough signal. Practically, in a popular approach of so-called random-site phylogenetic codon models, d_N/d_S is allowed to vary across sites, via a finite mixture model (Nielsen and Yang, 1998; Yang *et al.*, 2000, 2005; Huelsenbeck *et al.*, 2006). Generally, for detecting positive selection a category of sites is constrained to be under $d_N/d_S > 1$. Both proportions of sites and values of the different d_N/d_S categories are then estimated by maximum likelihood or Bayesian inference (see chapter 4). Sites under adaptive evolution are then detected based on their empirical Bayes posterior probability $d_N/d_S > 1$ (Huelsenbeck and Dyer, 2004; Yang *et al.*, 2005). To note, in this context of site-specific finite mixture

models, methods have also proposed to estimate both d_N and d_S separately (Kosakovsky Pond and Muse, 2005b; Spielman *et al.*, 2016).

A long series of site models has been proposed, most of which have been implemented in `PAML` (Yang, 1997, 2007), but also in MrBayes for the infinite mixture version (Huelsenbeck and Ronquist, 2001; Ronquist *et al.*, 2012). Specific applications at the level of the entire exome have uncovered sites of the sequence under positive recurrent selection (Kosiol *et al.*, 2008). Other analyses have revealed the importance of host-pathogen or host-virus interactions in contributing to strong signals of ongoing adaptation in protein-coding sequences (Enard *et al.*, 2016).

Finally, independently of the question of detecting positive selection, site models also turn out to be very valuable models, in the aim of uncovering selective pressures acting on specific sites. This can be used, for instance, to investigate the biophysical correlates of the strength of purifying selection at the site level (see section 5.2.2).

3.2.7 Variation across branches

Beside variation across sites, the strength of selection is not typically homogeneous along the phylogenetic tree, and it has also been recognized that it could be useful to model this variation. A first approach allows for a different d_N/d_S only on a given branch, or on a subset of the phylogeny, chosen a priori based on biological assumptions (Yang and Nielsen, 1998). For example, such models can detect an adaptive process ongoing during the divergence of one lineage, which can allow for the detection of the proteins responsible for speciation (Yang and Nielsen, 1998; Zhang and Nielsen, 2005). The most extreme version of this model simply assumes that each branch has its own d_N/d_S, without any constraints (Popadin *et al.*, 2007). To avoid overfitting, branches can be clustered based on their substitution rates, using a sequential testing approach (Dutheil *et al.*, 2012).

Alternatively, d_N/d_S can be modelled as a continuous trait, varying continuously along the phylogeny, and susceptible to show phylogenetic inertia. To account for this, d_N/d_S is not mathematically formalized as a parameter anymore, but instead, it is modelled as a stochastic process, and more specifically, a log-Brownian process, splitting at each node of the tree into independent processes. This modelling approach was previously used in the context of the comparative method, to model the evolution of quantitative traits observable at the tips (Felsenstein, 1985; Huelsenbeck and Rannala, 2003). It was then recruited to model the variation in the total rate of substitution, in the context of the so-called auto-correlated relaxed clock models, used to estimate divergence times (Thorne *et al.*, 1998). Finally, it was used to model the variation, independently, of d_S and d_N (Seo *et al.*, 2004), or of d_S and d_N/d_S (Lartillot and Poujol, 2011).

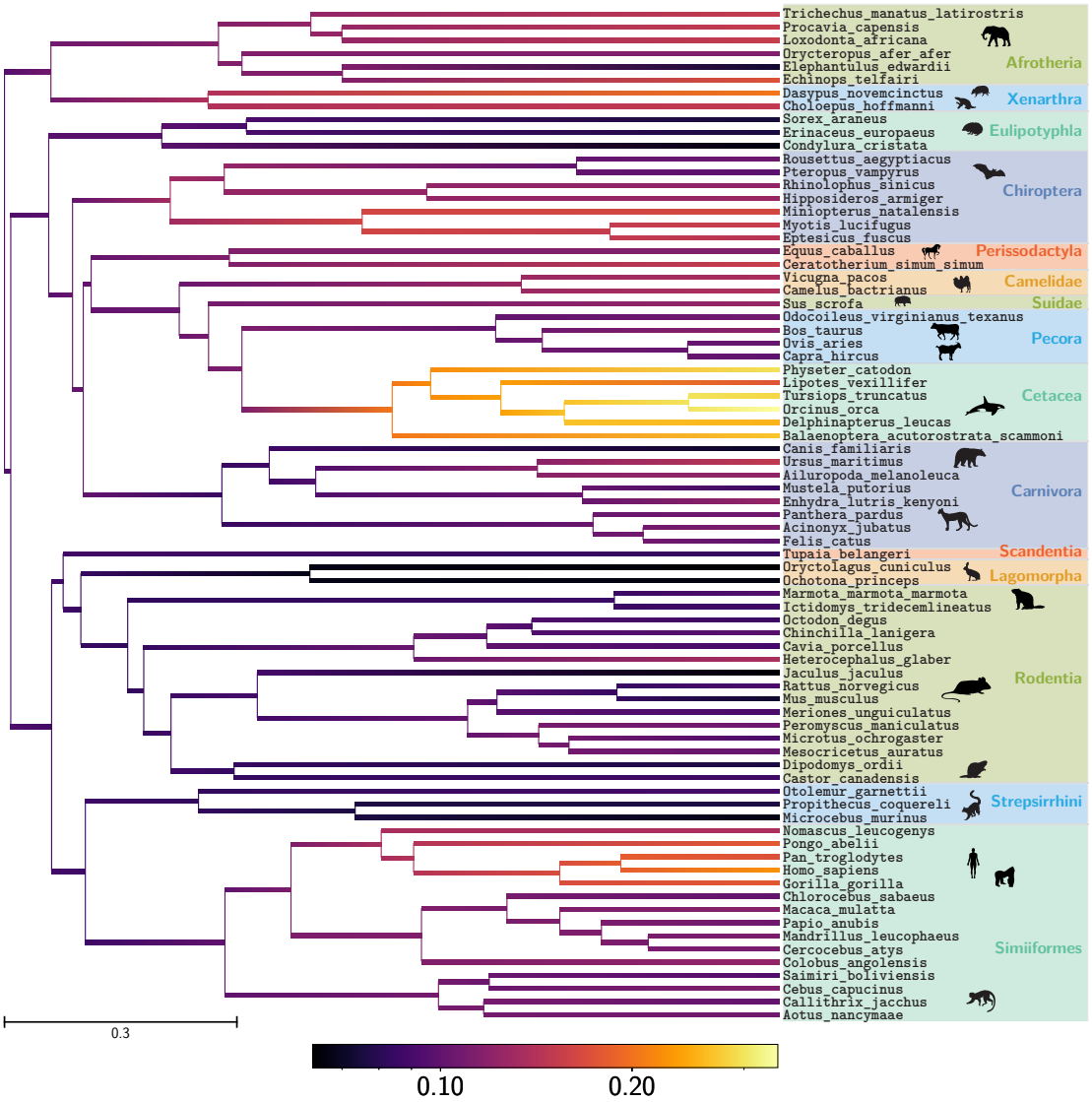


Figure 3.2: *d_N/d_S variations across branches in mammals. The Brownian process (i.e. logarithm of d_N/d_S) starts at the root of the dated tree, runs along branches and splits at each node of the tree into two independent children processes until reaching the extant species. Along each branch, the value of d_N/d_S used in the substitution matrix is taken as the average of the trajectory between the two nodes at the tips of the branch (i.e. child and parent). However, for the representation, a gradient between the child and parent node highlight the change of d_N/d_S along this specific branch. The dataset consist of 77 extant taxa on a randomly chosen set of* 18 *coding sequences (CDS) from OrthoMam database (Ranwez* et al., *2007; Scornavacca* et al., *2019). This analysis was performed under the Muse & Gaut formalism and conducted on the software* `BayesCode` *(see chapter 4). Variations in d_N/d_S along the tree can also be related to ecological variables, or life-history traits.*

The external factors determining the variation d_N/d_S across lineages have subsequently been investigated, primarily focused on environmental variables and life-history traits that can vary between species. This has been done using either sequential approaches, first estimating the variation in d_N/d_S using some of the methods mentioned

above, and then using the classical comparative method to correlate the estimated variation with independently observed quantitative or life-history traits (Popadin *et al.*, 2007; Lanfear *et al.*, 2010a; Romiguier *et al.*, 2014).

Thereafter, integrative inference methods combining both molecular sequences and quantitative traits have been developed, jointly modelling the variation of all of these variables using a single multivariate Brownian process (Lartillot and Poujol, 2011). Each entry of the process describes the evolution of one of the variables of interest: d_S, d_N/d_S, quantitative traits, etc. The model can then be fitted on an empirical data set consisting of a multiple sequence alignment of coding sequences and a matrix of quantitative traits observed in extant species. This leads to a joint estimation of the stochastic process and the covariance matrix, thus giving estimates of the covariance between d_N/d_S and traits, corrected for phylogenetic inertia.

Applications of this integrative approach also found that d_N/d_S correlates positively with traits such as longevity and body mass (Lartillot and Poujol, 2011; Figuet *et al.*, 2017). Since lineages with a large body size and extended longevity typically correspond to low N_e (Romiguier *et al.*, 2014), these empirical correlations suggest a negative correlation between d_N/d_S and N_e, thus confirming the theoretical prediction of the nearly-neutral theory of evolution. Similarly, and more directly, d_N/d_S was found to correlate negatively with the synonymous diversity ($\pi_S = 4N_e u$), which is a molecular proxy of effective population size (Brevet and Lartillot, 2019). These important results confirm one of the key predictions of the nearly-neutral theory. However, the universality and robustness of the correlation between d_N/d_S and life-history traits is still debated, and further investigations are required (Nabholz *et al.*, 2013; Lanfear *et al.*, 2014; Figuet *et al.*, 2016; Bolívar *et al.*, 2019).

3.2.8 Variation across sites and branches

Naturally, both space (site-specific) and time (branch-specific) refinements mentioned above led to the development of the so-called branch-site models (Yang and Nielsen, 2002; Zhang and Nielsen, 2005; Kosakovsky Pond *et al.*, 2011; Murrell *et al.*, 2012, 2013). The fine-grained tuning of site-branch models increased statistical power by seeking short and strong episodes of adaptive selection on a background of purifying selection. However, in the case of Red-Queen processes ongoing on the protein, the episodes detected by branch-site models would merely be a small fraction of the underlying adaptation. Indeed the overall tree is under adaptive process and one cannot contrast a branch to the rest of the tree.

3.3 Mechanistic codon models

Classical codon models presented above capture the average effect of selection on non-synonymous mutations, without seeking to discriminate between different types of mutations. In contrast, mechanistic codon models seek to predict individually all substi-

tution rates, for each position and between each pair of codons, in an explicit model of the adaptive landscape.

3.3.1 The Halpern & Bruno formalism

The Halpern and Bruno (1998) formalism assumes that the protein-coding sequence is at mutation-selection balance under a time-independent fitness landscape, with a fitness that is multiplicative across sites (i.e. without epistasis). As a result, the fitness landscape is characterized by a fitness vector over the 20 amino acids at each site. Furthermore, the substitution process at each position is independent of the current state at all other positions, and it will generally be different at each site (Rodrigue *et al.*, 2010; Tamuri and Goldstein, 2012).

In the following equations, I omit the dependence on sites, such that the fact that this process is site-specific is implicit. Consider a given site, the probability of fixation depends on the difference in fitness between the amino acid encoded by the mutated codon ($f_{\mathcal{A}(j)}$) and the amino acid encoded by the original codon ($f_{\mathcal{A}(i)}$), where $\mathcal{A}(i)$ denotes the amino acid encoded by codon i. The rate of substitution from codon i to j is derived from equation 2.35:

$$Q_{i,j} = \mu_{i,j} \frac{4N_\mathrm{e}\left(f_{\mathcal{A}(j)} - f_{\mathcal{A}(i)}\right)}{1 - \mathrm{e}^{4N_\mathrm{e}\left(f_{\mathcal{A}(i)} - f_{\mathcal{A}(j)}\right)}}, \tag{3.14}$$

$$= \mu_{i,j} \frac{F_{\mathcal{A}(j)} - F_{\mathcal{A}(i)}}{1 - \mathrm{e}^{F_{\mathcal{A}(i)} - F_{\mathcal{A}(j)}}}. \tag{3.15}$$

Altogether, the 61-by-61 codon substitution matrix of mechanistic codon models $\boldsymbol{Q}$ is defined entirely by the mutation matrix ($\boldsymbol{R}$), the vector of 20 amino-acid relative fitness ($\boldsymbol{f}$) and the genetic code:

$$\begin{cases} Q_{i,j} &= 0 \text{ if codons } i \text{ and } j \text{ are more than one mutation away,} \\ Q_{i,j} &= \mu_{i,j} \text{ if codons } i \text{ and } j \text{ are synonymous,} \\ Q_{i,j} &= \mu_{i,j} \dfrac{F_{\mathcal{A}(j)} - F_{\mathcal{A}(i)}}{1 - \mathrm{e}^{F_{\mathcal{A}(i)} - F_{\mathcal{A}(j)}}} \text{ if codons } i \text{ and } j \text{ are non-synonymous.} \end{cases} \tag{3.16}$$

Because the process is time-reversible (see chapter 2), from equation 2.55, the stationary distribution equals to:

$$\pi_i = \frac{\left[\prod\limits_{k \in \{1,2,3\}} \sigma_{i[k]}\right] \mathrm{e}^{F_{\mathcal{A}(i)}}}{\sum\limits_{j=1}^{61} \sigma_{j[1]}\sigma_{j[2]}\sigma_{j[3]} F_{\mathcal{A}(j)}}. \tag{3.17}$$

The stationary frequency of a codon is ultimately the product of the nucleotide frequencies ($\boldsymbol{\sigma}$) at its three positions and the scaled Wrightian fitness of the amino-acid ($\mathrm{e}^{F_{\mathcal{A}(i)}}$).

3.3.2 Empirical calibration of the model

Fitting the mutation-selection model on a sequence alignment, via equation (3.16), results in an estimation of the nucleotide mutation rate matrix as well as the amino-acid fitness

landscapes at each site of the sequence. Several approaches have been used to do this. In the original approach, Halpern and Bruno (1998) leveraged the detailed balance:

$$\frac{\pi_i}{\pi_j} = \frac{Q_{j,i}}{Q_{i,j}} \tag{3.18}$$

$$= \frac{\mu_{j,i}\left(F_{\mathcal{A}(i)} - F_{\mathcal{A}(j)}\right)\left(1 - \mathrm{e}^{F_{\mathcal{A}(i)} - F_{\mathcal{A}(j)}}\right)}{\mu_{i,j}\left(F_{\mathcal{A}(j)} - F_{\mathcal{A}(i)}\right)\left(1 - \mathrm{e}^{F_{\mathcal{A}(j)} - F_{\mathcal{A}(i)}}\right)} \tag{3.19}$$

$$= \frac{\mu_{j,i}\left(\mathrm{e}^{F_{\mathcal{A}(i)} - F_{\mathcal{A}(j)}} - 1\right)}{\mu_{i,j}\left(1 - \mathrm{e}^{F_{\mathcal{A}(j)} - F_{\mathcal{A}(i)}}\right)} \tag{3.20}$$

$$= \frac{\mu_{j,i}\mathrm{e}^{F_{\mathcal{A}(i)}}\left(\mathrm{e}^{-F_{\mathcal{A}(j)}} - \mathrm{e}^{-F_{\mathcal{A}(i)}}\right)}{\mu_{i,j}\mathrm{e}^{F_{\mathcal{A}(j)}}\left(\mathrm{e}^{-F_{\mathcal{A}(j)}} - \mathrm{e}^{-F_{\mathcal{A}(i)}}\right)} \tag{3.21}$$

$$= \mathrm{e}^{F_{\mathcal{A}(i)} - F_{\mathcal{A}(j)}}\frac{\mu_{j,i}}{\mu_{i,j}} \tag{3.22}$$

Such that the scaled selection coefficients are related to the stationary codon frequencies:

$$F_{\mathcal{A}(i)} - F_{\mathcal{A}(j)} = \ln\left(\frac{\pi_i\mu_{i,j}}{\pi_j\mu_{j,i}}\right) \tag{3.23}$$

And finally the substitution rate between codon i and j is:

$$Q_{i,j} = \mu_{i,j}\frac{F_{\mathcal{A}(j)} - F_{\mathcal{A}(i)}}{1 - \mathrm{e}^{F_{\mathcal{A}(i)} - F_{\mathcal{A}(j)}}} \tag{3.24}$$

$$= \mu_{i,j}\frac{\ln\left(\frac{\pi_j\mu_{j,i}}{\pi_i\mu_{i,j}}\right)}{1 - \frac{\pi_i\mu_{i,j}}{\pi_j\mu_{j,i}}} \tag{3.25}$$

As a result, the substitution rate from codon i to j can be approximated based on a plugin estimator for both the mutational process and the amino-acid frequencies, independently estimated. Alternatively, site-specific amino-acid preferences have been estimated either by penalized maximum likelihood (Tamuri and Goldstein, 2012; Tamuri *et al.*, 2014), or in a Bayesian context using an infinite mixture based on a Dirichlet process prior (Rodrigue *et al.*, 2010; Rodrigue and Lartillot, 2014). Comparison of both inference approaches yields similar results in terms of estimated profiles and their induced selective constraint on protein-coding DNA sequences (Spielman and Wilke, 2016). Finally, instead of estimating the fitness landscape directly on the multiple sequence alignment, deep mutational scanning approaches can be used to estimate fitness profiles experimentally (Bloom, 2014b,a), as presented in chapter 5.

3.3.3 Modulating the fitness landscape across branches

Thus far, in the mutation-selection formalism, fitness landscape has been considered static. In practice, fitness landscapes are dynamic and changing with time (Naumenko *et al.*, 2012; Bazykin, 2015). In particular, selective pressures may change following one (or more) transitions to a new environment (e.g.: a new host). Changes in selective pressures induced by environmental changes can be modelled in a mutation-selection

framework by introducing different fitness profiles in different parts of the tree (Tamuri *et al.*, 2009). Similarly, phenotypic convergent evolution has been investigated in relation to underlying molecular convergence at the level of codons. In this context, if a specific codon site is responsible for the phenotypic convergence, the species sharing the convergent phenotype should also share convergence in amino-acid profiles at this specific site (Parto and Lartillot, 2017, 2018)

3.3.4 Mutation-selection and codon usage

Another example of a mutation-selection mechanistic codon model is one in which codon usage bias is modelled, in particular, a model in which each synonymous codon of the same amino acids have different fitness (i.e. F_i for all 61 codons) as in Yang and Nielsen (2008). It is important to note that contrarily to the Halpern & Bruno formalism, codon preferences are not site-specific but instead are estimated gene-wide. In this model, substitution rates are defined as:

$$\begin{cases} Q_{i,j} & = 0 \text{ if codons } i \text{ and } j \text{ are more than one mutation away,} \\ Q_{i,j} & = R_{\mathcal{M}(i,j)} \dfrac{F_j - F_i}{1 - \mathrm{e}^{F_i - F_j}} \text{ if codons } i \text{ and } j \text{ are synonymous,} \\ Q_{i,j} & = \omega R_{\mathcal{M}(i,j)} \dfrac{F_j - F_i}{1 - \mathrm{e}^{F_i - F_j}} \text{ if codons } i \text{ and } j \text{ are non-synonymous.} \end{cases} \tag{3.26}$$

With such a definition, this model is hybrid between the classical model (due to ω) and the mechanistic mutation-selection codon model (due to the selection coefficients for codons F_i). Such hybrid models have the interest of measuring the average effect of selection on non-synonymous mutations through d_N/d_S without making the assumption that synonymous mutations are neutral.

3.4 Relationship between mechanistic and classical codon models

Even though classical codon models have fewer parameters than mechanistic codon models, it is important to realize they are not nested. Indeed, it is impossible to find a given set of parameters for which the two models are equivalent, except by assuming all sites to have a uniform fitness distribution over amino acids in the Halpern & Bruno mutation-selection model, and setting $\omega = 1$ in the Muse & Gaut model, but this is really a trivial case. They are inherently different and proceed from a different philosophy. On one hand, mechanistic models rely on an explicit fitness landscape, while, on the other hand, classical models capture the average effect of selection through a single ω parameter.

The difference can be highlighted by considering the case of reverse mutations. In a mechanistic model (section 3.3), a negative selection coefficient associated with a given non-synonymous mutation is always matched by a positive selection coefficient for the reverse mutation. As a result, the rate of substitution will be lower than the mutation rate in one direction, but higher in the other direction. In contrast, in classical codon

models (section 3.2), if $\omega < 1$ (respectively, $\omega > 1$), the rate of substitution is lower (respectively, higher) than the synonymous substitution rate in the two directions.

Nevertheless, it is possible to make conceptual and quantitative connections between these two modelling paradigms. This point was explored in detail by Spielman and Wilke (2015), Dos Reis (2015), Jones *et al.* (2016) and Rodrigue and Lartillot (2016), summarized in table 3.3.

Symbol	Interpretation
d_N	Non-synonymous substitution rate.
d_S	Synonymous substitution rate.
d_N/d_S	Ratio of non-synonymous over synonymous substitution rate.
ν	Mean scaled fixation probability of non-synonymous mutations.
ω	Scaling factor for all non-synonymous substitutions in the Muse and Gaut (1994) formalism.
ω_0	Induced ν_{HB} in the Halpern and Bruno (1998) mechanistic formalism.
ω_*	Scaling factor for all non-synonymous substitutions in the Halpern and Bruno (1998) formalism.

Table 3.3: *Relationship between classical and mechanistic codon models*

3.4.1 The Halpern & Bruno mechanistic codon model as a nearly-neutral model

Once fitted to the data, the classical Muse & Gaut (MG) formalism returns estimates of mutation rates and ω (see subsection 3.2.1). From there, one can compute the substitution and mutation rates of each codon substitution. Using equation 2.56 on the subset of non-synonymous mutations thus gives ν_{MG} at stationarity:

$$\nu_{\text{MG}} = \frac{\sum_{i=1}^{61} \pi_i \sum_{j \in \mathcal{N}_i} Q_{i,j}}{\sum_{i=1}^{61} \pi_i \sum_{j \in \mathcal{N}_i} \mu_{i,j}} \tag{3.27}$$

$$= \frac{\sum_{i=1}^{61} \pi_i \sum_{j \in \mathcal{N}_i} \omega \mu_{i,j}}{\sum_{i=1}^{61} \pi_i \sum_{j \in \mathcal{N}_i} \mu_{i,j}} \tag{3.28}$$

$$= \omega, \tag{3.29}$$

where $\mathcal{N}_i$ is the set of non-synonymous codons neighbours to codon i. Such equation is also true for any classical codon model formalism, where this identity between ν and d_N/d_S bears much importance.

This rate of non-synonymous substitutions over mutations (ν) can be interpreted as the mean scaled fixation probability of non-synonymous mutations (see section 2.2.5), such that even if classical codon models are not mechanistic in essence, the parameter d_N/d_S can be interpreted a posteriori as the mean scaled fixation probability of non-synonymous mutations.

On the other hand, the mechanistic codon models in the Halpern & Bruno (HB) formalism return estimates of mutation rates and fitness profiles of amino acids (see sub-

section 3.3.1). From there, one can also compute the fixation probability individually for each codon substitution. Likewise, using equation 2.56 on the subset of non-synonymous mutations gives (ν_{HB}) at stationarity:

$$\nu_{\text{HB}} = \frac{\sum_{i=1}^{61} \pi_i \sum_{j \in \mathcal{N}_i} Q_{i,j}}{\sum_{i=1}^{61} \pi_i \sum_{j \in \mathcal{N}_i} \mu_{i,j}} \tag{3.30}$$

$$= \frac{\sum_{i=1}^{61} \pi_i \sum_{j \in \mathcal{N}_i} \frac{F_{\mathcal{A}(j)} - F_{\mathcal{A}(i)}}{1 - \mathrm{e}^{F_{\mathcal{A}(i)} - F_{\mathcal{A}(j)}}}}{\sum_{i=1}^{61} \pi_i \sum_{j \in \mathcal{N}_i} \mu_{i,j}}. \tag{3.31}$$

Hence, for the mutation-selection mechanistic model, ν_{HB} can be interpreted as the resulting d_N/d_S induced by the model (Spielman and Wilke, 2015; Dos Reis, 2015). Indeed, simulation experiments conducted by Spielman and Wilke (2015) under a mutation-selection model then analysed using a classical codon model indeed showed agreement between the induced and estimated d_N/d_S. To note, inference under the Muse & Gaut formalism showed the best agreement compared to other formalisms of classical codon models.

Moreover, Spielman and Wilke (2015) showed mathematically that, if the underlying process is at equilibrium under a time-independent fitness landscape (nearly-neutral regime), then the mean scaled fixation probability ν_{HB} induced by the model will always be lower than 1. In other words, they showed that mechanistic mutation-selection codon models display the important feature of genuinely accounting for purifying selection. From a dynamic perspective, a non-synonymous mutation from a codon with high fitness to another codon will have a low probability of fixation, since the mutated codon will have a lower fitness. At equilibrium, this low probability of fixation of the other codon results in a high frequency of the codon with higher fitness. Essentially, at equilibrium the codon frequencies only fluctuate at the mutation-selection balance, and all the mutations are neutral on average, but slightly deleterious or advantageous, hence the name nearly-neutral models (Ohta, 1973, 1992; Rodrigue and Lartillot, 2016). This justifies the interpretation of the Halpern & Bruno mechanistic codon models as an implementation of the nearly-neutral regime.

Altogether, classical codon substitution models will interpret a mechanistic mutation-selection model as purifying selection ($\omega < 1$). Accordingly, the mean scaled probability of fixation ν_{HB} has also been denoted ω_0 (Rodrigue and Lartillot, 2016).

3.4.2 The Halpern & Bruno mechanistic codon model as a nearly-neutral null model

As seen above, under the assumption that the protein is under a nearly-neutral regime, the predicted ω_0 (mutation-selection model) and the estimated ω (classical model) should be the same (Spielman and Wilke, 2015). But assumptions of the models can be bro-

ken, resulting in discrepancy between the ω_0 induced (or predicted) by the Halpern & Bruno mechanistic model, once fitted on the data, and ω directly estimated by classical codon models.

This deviation can be captured as a gene-wide multiplying factor ω_* (Rodrigue and Lartillot, 2016):

$$\begin{cases} Q_{i,j} & = 0 \text{ if codons } i \text{ and } j \text{ are more than one mutation away,} \\ Q_{i,j} & = \mu_{i,j} \text{ if codons } i \text{ and } j \text{ are synonymous,} \\ Q_{i,j} & = \omega_* \mu_{i,j} \dfrac{F_{\mathcal{A}(j)} - F_{\mathcal{A}(i)}}{1 - \mathrm{e}^{F_{\mathcal{A}(i)} - F_{\mathcal{A}(j)}}} \text{ if codons } i \text{ and } j \text{ are non-synonymous.} \end{cases} \tag{3.32}$$

Since fitness profiles are capturing ω_0, the resulting ω which is a function of the model parameters, can be interpreted as:

$$\omega = \omega_* \times \omega_0 \tag{3.33}$$

This modelling approach is hybrid between mechanistic and phenomenological model, since the parameter ω_* cannot be interpreted mechanistically. Moreover, the deviation of ω_* can bend upward or downward, where different interpretations can be given of both cases.

3.4.3 Adaptive evolution

The Halpern & Bruno formalism assumes that fitness landscapes are not dependent on time. Alternatively, time-dependent fitness landscapes are known as seascape (Mustonen and Lässig, 2009). Because of the external movement of the fitness landscape, similarly to Red-Queen dynamics, the current sequence is more likely to slide into a fitness valley rather than on top of a peak when the landscape is moving. In other words, because the current sequence is at mutation-selection-drift balance and the movement of the landscape is external, the fitness of the sequence is not likely to increase in the new fitness landscape. As a result, external changes of the landscape results in lower fitness of the current sequence on average. The resulting dynamics is that selection pushes the sequence to climb up the time-dependent fitness landscape constantly, and the protein sequence is tracking a constantly moving fitness optimum.

Since the protein sequence is always lagging behind the moving target defined by the amino acid preferences, and since substitutions are accepted preferentially if they are in the direction of this target, substitutions are on average adaptive. In other words, the sequence would become increasingly maladaptive in the absence of such positively selected substitutions. Thus, breaking the assumption of time independence of amino-acid preferences leads to the estimation of an induced ω_0 lower than the realized ω:

$$\omega \geq \omega_0 \iff \omega_* \geq 1 \tag{3.34}$$

3.4.4 Epistasis and entrenchment

The nearly-neutral assumption of the Halpern & Bruno formalism can also be broken if there is no independence between sites, known as epistasis between sites. Unfortunately,

one consequence of epistatic interactions is that even if a mutation is nearly-neutral upon fixation, subsequently fixed mutations on other sites make the original substitution more and more deleterious to revert over time (Gong and Bloom, 2014; Lunzer *et al.*, 2010; Mccandlish *et al.*, 2013). This effect called entrenchment results in the current amino acids reinforcing their relative fitness with time, in opposition to constantly lagging behind a moving target (Pollock *et al.*, 2012). In other words, at the moment of a substitution, the target amino acid has a nearly equal relative fitness, which on average then increases with time (Goldstein and Pollock, 2016, 2017). Contradictory to what happens during adaptation, breaking the assumption of independence between sites leads to entrenchment and the realized ω being lower than the induced ω_0 (Rodrigue and Lartillot, 2016):

$$\omega \leq \omega_0 \iff \omega_* \leq 1 \tag{3.35}$$

Altogether, a departure from near-neutrality with a $\omega \geq \omega_0$ is a signature of an ongoing Red-Queen process and that the protein is under ever-changing adaptation. On the other hand, a $\omega \leq \omega_0$ is a signature of epistatic interaction between amino acids. However, one shortcoming of nearly-neutral codon substitution models is that if one does not get a statistical departure from near-neutrality ($\omega = \omega_0$), it could be due to a mixture of both Red-Queen and epistatic processes that cannot be disentangled.

4

Probabilistic inference and parameter estimation

Contents

The previous chapter treated how substitution rates are defined and parameterized in phylogenetic codon models, either classical or mechanistic, but not how these parameters are inferred and estimated. In contrast, the goal of this chapter is to present the methodology for estimating the parameters from a set of observed protein-coding DNA sequences in different species or lineages. To do so, I will first introduce the concept of likelihood and how the likelihood is computed in the context of phylogenetic models (section 4.1). Then, I will briefly introduce the maximum likelihood method of inference (section 4.1.4) and, finally, the principles of Bayesian inference using Markov chain Monte Carlo (section 4.2).

4.1 Likelihood of the data

To define the likelihood, it is important to realize that codon models presented previously can also be used in forward mode, so as to generate a simulated alignment of protein-

coding sequences. Given specific parameter values for the model (generically noted θ), the probability of simulating a replicate of the sequence data exactly identical to the empirical dataset D (noted $\mathbb{P}(D \mid \theta)$) can then be taken as a measure of how well this alignment is explained by the model, under the specific parameter values θ. This defines the likelihood, which is thus a function of the parameter θ:

$$L(\theta) = \mathbb{P}(D \mid \theta) \tag{4.1}$$

Unfortunately, even with an astronomical number of simulations, it is very unlikely to generate precisely our observed alignment, and even more difficult to precisely pin down the probability that our observed alignment has been generated by the model under a given set of parameters. Deriving this probability analytically is thus the first theoretical question to answer, which is the focus of this section. The challenge is that only the data for extant species are observed whereas sequence at the root of the tree and subsequent evolutionary events of speciation are not directly observed. In other words, all possible trajectories leading to the observed alignment must be integrated and weighted by their respective probabilities.

Throughout this development, the tree topology (τ) is considered known and fixed. This restriction emanates from the fact that the scope of this work is not to infer the topology, but rather the parameters of the molecular evolutionary process. Moreover, the development conducted below does not delve into the details of how multiple sequence alignments are obtained in practice, and assumes in particular that they are correct. However, it has been shown that outputs of different sequence alignment methods tend to produce different results that are not always mutually consistent. The main determining factor of alignment accuracy is evolutionary divergence, such that if alignments are restricted to orthologs from closely related taxa, or to slowly evolving genes, alignment errors become rare and may not cause significant problems.

Importantly, the models of sequence evolution considered in this thesis all assume site independence, such that changes at one sequence position have no impact on whether and how other positions will change. This assumption of independence between sites allows the probability of an observed alignment to be expressed as the product over alignment columns of the probability of observing each of them. This independence assumption is a simplification. However it greatly facilitates likelihood-based inference.

This development of likelihood computation is divided into three sections, first integrating over all trajectory along a single branch of the tree (section 4.1.1), and subsequently over the entire tree (section 4.1.2), while finally efficiently computing the probability of the data given the parameters (section 4.1.3).

4.1.1 Finite-time transition probabilities over a branch at a given site

The point substitution process implied by the codon model defines the instantaneous rates of change between the different codons through the substitution matrix $\boldsymbol{Q}$. Given a

starting (ancestral) codon state and a given amount of time over which the substitution process runs, the first task is to derive the probability of the descendant sequence presenting each of the 61 possible codon states. In practice, the substitution rate matrix must be normalized, such that time is measured in units of branch length, expressing the expected number of neutral changes that have occurred since the ancestor. For example, a branch length of 2 implies that 2 changes are expected to be seen on average along the branch under the condition that substitutions are neutral. At a given site (z) of the sequence, and along a given branch with branch length l, the codon probability matrix $\boldsymbol{P}^{(\mathrm{z})}(l)$ is related to the transition matrix ($\boldsymbol{Q}^{(\mathrm{z})}$ at site z) through the first-order differential equation:

$$\frac{\mathrm{d}\boldsymbol{P}^{(\mathrm{z})}(l)}{\mathrm{d}l} = \boldsymbol{P}^{(\mathrm{z})}(l)\boldsymbol{Q}^{(\mathrm{z})}, \tag{4.2}$$

which has solution:

$$\boldsymbol{P}^{(\mathrm{z})}(l) = \mathrm{e}^{l\boldsymbol{Q}^{(\mathrm{z})}}. \tag{4.3}$$

This integration of the substitution rate matrix over the branch takes into account all possible histories of substitution events compatible with the states at the two ends, leading to a compact probability matrix computed as an exponential of the rate matrix. In practice, exponentiating the rate matrix is usually performed using decomposition in eigenvalues and eigenvectors.

4.1.2 Integrating over ancestral states

The challenge for generalizing this argument from a single branch to a complete tree is that only the data at the tips of the tree are observed whereas the states at the internal nodes are not. If they were known, the likelihood would be readily calculated, by taking the product of the transition probabilities over all branches. As an example, and for better readability, a simple illustrative tree given in figure 4.1 will be used prior to giving to general formulas.

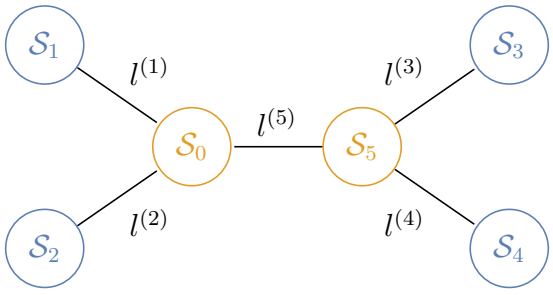


Figure 4.1: *Illustrative phylogenetic tree. Internal states of nodes ($\mathcal{S}_0$ and $\mathcal{S}_5$) are represented in yellow, while states of extant nodes ($\mathcal{S}_1$ to $\mathcal{S}_4$) for which the state is known from the observed data is represented in blue.*

In the example of the illustrated tree, from the observed data for the extant nodes $\boldsymbol{D}^{(\mathrm{z})} = \{\mathcal{S}_1, \mathcal{S}_2, \mathcal{S}_3, \mathcal{S}_4\}$, at site z, given that the states of the internal nodes are known,

the likelihood is computed as:

$$\mathbb{P}\left(\boldsymbol{D}^{(\mathrm{z})} \mid \mathcal{S}_0, \mathcal{S}_5, \boldsymbol{Q}^{(\mathrm{z})}, l^{(\mathrm{b})}\right) = \boldsymbol{P}^{(\mathrm{z})}_{\mathcal{S}_0,\mathcal{S}_1}(l^{(1)})\boldsymbol{P}^{(\mathrm{z})}_{\mathcal{S}_0,\mathcal{S}_2}(l^{(2)}), \times \boldsymbol{P}^{(\mathrm{z})}_{\mathcal{S}_5,\mathcal{S}_3}(l^{(3)})\boldsymbol{P}^{(\mathrm{z})}_{\mathcal{S}_5,\mathcal{S}_4}(l^{(4)})\boldsymbol{P}^{(\mathrm{z})}_{\mathcal{S}_0,\mathcal{S}_5}(l^{(5)}). \tag{4.4}$$

In the general case of an arbitrary topology with B branches, $\mathcal{S}_{\mathrm{b}\uparrow}$ and $\mathcal{S}_{\mathrm{b}\downarrow}$ are used to denote the parent and descendant nodes of a branch, the likelihood conditional on internal states is given as:

$$\mathbb{P}\left(\boldsymbol{D}^{(\mathrm{z})} \mid \mathcal{S}_I, \boldsymbol{Q}^{(\mathrm{z})}, l^{(\mathrm{b})}\right) = \prod_{\mathrm{b}=1}^{B} \boldsymbol{P}^{(\mathrm{z})}_{\mathcal{S}_{\mathrm{b}\downarrow},\mathcal{S}_{\mathrm{b}\uparrow}}(l^{(\mathrm{b})}), \tag{4.5}$$

where I runs over all the internal nodes, and there are B branches.

Because the states of the internal nodes are actually unknown, the likelihood must be summed over all possible configurations for them, including at the root. At the root, the states are produced according to equilibrium frequencies of the process $(\pi^{(\mathrm{z})})$.

In the case of the illustrative example, the total probability is given as:

$$\mathbb{P}\left(\boldsymbol{D}^{(\mathrm{z})} \mid \boldsymbol{Q}^{(\mathrm{z})}, l^{(\mathrm{b})}\right) = \sum_{\mathcal{S}_0=1}^{61} \pi_{\mathcal{S}_0} \sum_{\mathcal{S}_5=1}^{61} \pi_{\mathcal{S}_5} \mathbb{P}\left(\boldsymbol{D}^{(\mathrm{z})} \mid \mathcal{S}_0, \mathcal{S}_5, \boldsymbol{Q}^{(\mathrm{z})}, l^{(\mathrm{b})}\right) \tag{4.6}$$

$$= \sum_{\mathcal{S}_0=1}^{61} \pi_{\mathcal{S}_0} \sum_{\mathcal{S}_5=1}^{61} \pi_{\mathcal{S}_5} \boldsymbol{P}^{(\mathrm{z})}_{\mathcal{S}_0,\mathcal{S}_1}(l^{(1)})\boldsymbol{P}^{(\mathrm{z})}_{\mathcal{S}_0,\mathcal{S}_2}(l^{(2)}), \times \boldsymbol{P}^{(\mathrm{z})}_{\mathcal{S}_5,\mathcal{S}_3}(l^{(3)})\boldsymbol{P}^{(\mathrm{z})}_{\mathcal{S}_5,\mathcal{S}_4}(l^{(4)})\boldsymbol{P}^{(\mathrm{z})}_{\mathcal{S}_0,\mathcal{S}_5}(l^{(5)}). \tag{4.7}$$

And because the process is reversible, the codon equilibrium frequencies satisfy the equations:

$$\mathbf{0} = \boldsymbol{\pi}^{(\mathrm{b,z})}\boldsymbol{Q}^{(\mathrm{b,z})}, \tag{4.8}$$

$$\iff \boldsymbol{\pi}^{(\mathrm{b,z})} = \boldsymbol{\pi}^{(\mathrm{b,z})}\boldsymbol{P}^{(\mathrm{b,z})}, \tag{4.9}$$

$$\iff \frac{\pi^{(\mathrm{z})}_i}{\pi^{(\mathrm{z})}_j} = \frac{Q^{(\mathrm{z})}_{j,i}}{Q^{(\mathrm{z})}_{i,j}} \text{ for all pairs } i, j. \tag{4.10}$$

In the general topology with B branches, the likelihood is thus given as:

$$\mathbb{P}\left(\boldsymbol{D}^{(\mathrm{z})} \mid \boldsymbol{Q}^{(\mathrm{z})}, l^{(\mathrm{b})}\right) = \sum_{\mathcal{S}_0=1}^{61} \pi_{\mathcal{S}_0} \cdots \sum_{\mathcal{S}_k=1}^{61} \pi_{\mathcal{S}_k} \mathbb{P}\left(\boldsymbol{D}^{(\mathrm{z})} \mid \mathcal{S}_I, \boldsymbol{Q}^{(\mathrm{z})}, l^{(\mathrm{b})}\right), \tag{4.11}$$

$$= \sum_{\mathcal{S}_0=1}^{61} \pi_{\mathcal{S}_0} \cdots \sum_{\mathcal{S}_k=1}^{61} \pi_{\mathcal{S}_k} \prod_{b=1}^{B} \boldsymbol{P}^{(\mathrm{z})}_{\mathcal{S}_{\mathrm{b}+},\mathcal{S}_{\mathrm{b}-}}(l^{(\mathrm{b})}). \tag{4.12}$$

And finally, the assumption of independence between sites allows the probability of an observed set of aligned sequences at the tips of an evolutionary tree to be expressed as the product over alignment columns (Z sites) of the observed nucleotides or amino acids in those columns:

$$\mathbb{P}\left(\boldsymbol{D} \mid \boldsymbol{Q}^{(\mathrm{z})}, l^{(\mathrm{b})}\right) = \prod_{\mathrm{z}=1}^{\mathrm{Z}} \mathbb{P}\left(\boldsymbol{D}^{(\mathrm{z})} \mid \boldsymbol{Q}^{(\mathrm{z})}, l^{(\mathrm{b})}\right) \tag{4.13}$$

4.1.3 Pruning algorithm

The likelihood at a specific column of a multiple sequence alignment given by equation 4.13 requires extensive computation, but can, however, be computed in linear time (as a function of the number of branches) using the pruning algorithm of Felsenstein (1981).

$$\mathbb{P}\left(\boldsymbol{D}^{(\mathrm{z})} \mid \boldsymbol{Q}^{(\mathrm{z})}, l^{(\mathrm{b})}\right) = \sum_{i=1}^{61} \pi_i^{(\mathrm{z})} \psi_0^{(\mathrm{z})}(i)\,, \tag{4.14}$$

where $\psi_{\mathrm{n}}^{(\mathrm{z})}(i)$ is computed recursively from the 2 descendant children n_1 and n_2 of an internal node n, as:

$$\psi_{\mathrm{n}}^{(\mathrm{z})}(i) = \left[\sum_{j=1}^{61} \boldsymbol{P}_{i,j}^{(\mathrm{z})}(l^{(\mathrm{n}\to\mathrm{n}_1)})\psi_{\mathrm{n}_1}^{(\mathrm{z})}(j)\right] \cdot \left[\sum_{j=1}^{61} \boldsymbol{P}_{i,j}^{(\mathrm{z})}(l^{(\mathrm{n}\to\mathrm{n}_2)})\psi_{\mathrm{n}_2}^{(\mathrm{z})}(j)\right]. \tag{4.15}$$

And if the node n is a node with no descendant, meaning an extant taxa:

$$\psi_{\mathrm{n}}^{(\mathrm{z})}(i) = \begin{cases} 1, & \text{if } \mathcal{S}_{\mathrm{n}} = i \\ 0, & \text{otherwise.} \end{cases} \tag{4.16}$$

4.1.4 Maximum likelihood

The previous sections introduced the computational procedure to compute the likelihood. Combining this procedure with numerical optimization methods allows one to find the parameter values $\widehat{\theta}$ maximizing the likelihood. In other words, our point estimate for the parameters is taken such as to maximize the probability for the model to reproduce the empirical alignment. This approach, which enjoys many desirable theoretical properties, such as asymptotic consistency and efficiency, was introduced in phylogenetics by Cavalli-Sforza and Edwards (1967) with reconstruction based on allele frequencies, and then by Felsenstein (1981) for phylogenies based on nucleotide sequences. It has also been extensively used for estimating the parameters of codon models, and in particular, classical d_N/d_S based codon models (Yang, 1997; Kosakovsky Pond and Muse, 2005a; Dutheil *et al.*, 2006; Yang, 2007; Guéguen *et al.*, 2013; Kosakovsky Pond *et al.*, 2020).

4.2 Bayesian inference

An alternative to maximum likelihood is Bayesian inference. This inference methodology, which dates back from Laplace and Bayes (1763), was introduced in phylogenetics by Yang and Rannala (1997), Mau *et al.* (1999), Larget and Simon (1999), Li *et al.* (2000) and Huelsenbeck and Ronquist (2001). Broadly speaking, the Bayesian paradigm can be seen as a way to model uncertainty in a probabilistic way. More specifically, the parameters of the model (collectively denoted θ) are considered as random variables, from a prior distribution describing our uncertainty about their value before having seen the data. The probability of those parameters are going to be modified after acquisition of

information supplied by observed data. Formally, this update of our knowledge is captured by the computation of the posterior distribution, which is obtained by conditioning the random variable theta on the observed value for the data:

$$\mathbb{P}(\theta \mid \boldsymbol{D}) = \frac{\mathbb{P}(\boldsymbol{D} \mid \theta)\mathbb{P}(\theta)}{\mathbb{P}(\boldsymbol{D})}. \tag{4.17}$$

Here the denominator is the marginal likelihood, meaning the likelihood integrated over the prior:

$$\mathbb{P}(\boldsymbol{D}) = \int \mathbb{P}(D \mid \theta)\mathbb{P}(\theta)\mathrm{d}\theta. \tag{4.18}$$

In an inference context, because we are only interested in the relative posterior probabilities of alternative values of θ, the marginal likelihood is a constant. For that reason, Bayes theorem can also be presented as:

$$\mathbb{P}(\theta \mid \boldsymbol{D}) \propto \mathbb{P}(\boldsymbol{D} \mid \theta)\mathbb{P}(\theta). \tag{4.19}$$

Simply stating that posterior is proportional to likelihood multiplied by prior. In other words, updating our knowledge, such as initially represented by our prior, is done multiplicatively, using the likelihood, and renormalizing to obtain a proper probability distribution (the posterior).

4.2.1 Bayesian statistics and model complexity

Bayesian statistics and maximum likelihood are often opposed to each other and sometimes fiercely defended by their respective proponents. There are indeed fundamental philosophical differences. In particular, Bayesian inference is potentially sensitive to the prior, although, practically, prior sensitivity can be investigated. In addition posterior and prior can be presented next to each other, such that differences between the two can be interpreted as the amount of signal extracted from the data, and potential issues with the choice of the prior can be pointed out.

However, the recent success of Bayesian inference relates more fundamentally to the way it deals with model complexity (Huelsenbeck *et al.*, 2000; Lartillot, 2020).

First, by sampling from the posterior distribution, Bayesian inference offers a method for integrating over the uncertainty about the parameters. This leads to more robust inference (Huelsenbeck *et al.*, 2000). A corollary is that over parametrization is not such a drastic issue as in maximum likelihood inference. In the worst possible case of over-parametrization, namely that of confounded parameters, such that the model is exactly the same for different set of parameters, confounded parameters can be identified afterward through parameters correlation in their joint posterior distribution. However, over-parameterized models are still a misappropriate use of computing resources, which results in a greater environmental cost.

Second, and most importantly, Bayesian inference gives a natural language to combine multiple levels of random variables, in the form of hierarchical models. Thus, for instance, in Bayesian molecular dating, the substitution process depends on divergence times and

substitution rate variations across the tree. In turn, divergence times, such as specified by the phylogeny, are the result of a birth-death process, while variation in the substitution rate across branches is naturally expressed by modelling the rate itself as a log-Brownian process. Finally, the parameters of the birth-death and log-Brownian processes are endowed with a prior. The model is thus hierarchical, with four levels. This expressiveness in model structure, combined with generic Monte Carlo approaches for dealing with complex random effects and multi-level evolutionary processes, has played a fundamental role in the recent success and popularity of Bayesian inference in evolutionary genetics.

4.2.2 Hierarchical model

The relationship between the random variables defining a hierarchical model can be formalized as a Bayesian network, which is a probabilistic graphical representation of the set of variables and their conditional dependencies via a directed acyclic graph (DAG). In the example of the prior distribution for the rate substitution matrix in the case of the mutation-selection model, the prior is defined as the joint distribution of the prior over the selection coefficient over amino acids and the mutation rate matrix, which itself is a deterministic function of the equilibrium frequencies of nucleotides and the exchangeability rates for the general-time-reversible (GTR) mutation matrix (see figure 4.2). Seeing the DAG the other way around (following the arrows), simple prior distributions are combined together to form more complex joint prior distribution which ultimately defines the prior distribution over the model parameter vector (θ). This hierarchy can naturally be extended across sites, across branches or across genes, and include the data, which are themselves a random variable produced by the substitution process.

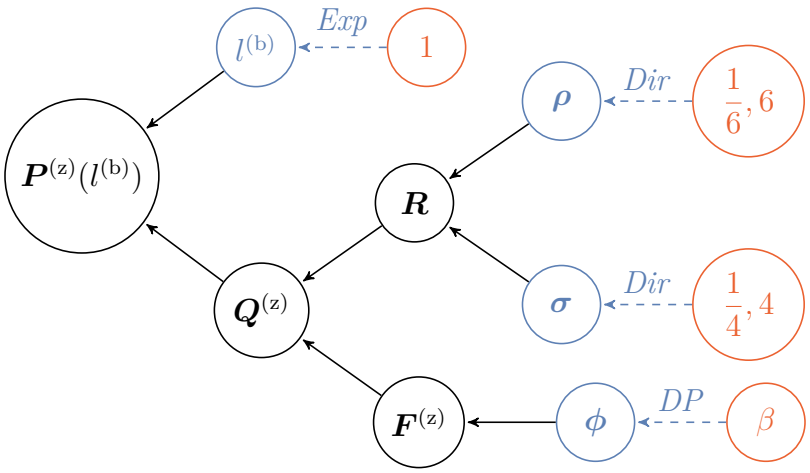


Figure 4.2: *Directed acyclic graph of dependencies between variables. Nodes of the directed acyclic graph are the variables, and edges are the functions. Hyper-parameters are depicted in red circle, random variables in blue circles, and transformed variables in black. blue dashed line denotes a drawing from a random distribution, and black solid lines denote a function. Exp denotes an exponential distribution, Dir denotes a Dirichlet distribution, and finally DP denotes a Dirichlet process.*

4.2.3 Markov chain Monte Carlo (MCMC)

Once realizing that the prior distribution can be boiled down to a set of simpler distributions over the components of the parameter vector, the difficulty in computing the posterior distribution arises from the high dimensionality of the parameter space, known as the curse of dimensionality. More precisely, the number of states increases exponentially with the number of dimensions of the space, such that the explicit evaluation for both the prior and the likelihood for a sufficiently fine-grained set of parameter values is unrealistic. In addition, the posterior distribution takes negligibly small values over most of the parameter space. Reduction in the exploration of the state space, and focusing of most of the computational effort in the relevant region, is obtained by employing Monte Carlo (MC) methods, which effectively approximate the target posterior distribution by sampling from it.

Historically, the first MC algorithm is associated with the army laboratory in Los Alamos under the direction of Metropolis in early 1952 [1]. Published by Metropolis *et al.* (1953), the primary focus of MC algorithm is on computing the mean energy of random configurations for a system of many particles. This energy is not available analytically and requires integration across all realizations of the random configurations of the particle system. Because dimensionality is high (proportional to the number of particles), numerical integration is impossible using a deterministic algorithm. Moreover, because the probability of a given configuration can be very small, even Monte Carlo integration by sampling randomly the prior (uniform) distribution over configurations fails to correctly approximate this integral.

This problem can, however, be formalized in terms of a Markov chain, where each state of the process is a particular configuration of particles. The transition probabilities between states must generate a stationary distribution equal to the target distribution of particle configurations. Given this requirement, and given one can also sample from the transition probabilities, the Markov chain Monte Carlo starts from an arbitrary state and can be updated by random sampling from the transition probabilities. After a period of burn-in, the Markov chain reaches the dynamic equilibrium, and the energy of each configuration can be computed. Finally, the average of this energy is an approximate solution for the integral of energy over the thermal equilibrium distribution of atomic configurations.

4.2.4 Metropolis-Hastings sampling

One specific algorithm designed such that the MCMC stationary distribution match the specified target distribution is the Metropolis algorithm, presented in the original paper (Metropolis *et al.*, 1953). This algorithm is composed of an acceptance/rejection rule such that the algorithm proceeds as follows at each step of the Markov chain. Start-

[1] Both a physicist and a mathematician, Nicolas Metropolis was one of the first scientists to work on the Manhattan Project that led to the production of the atomic bomb. Almost as early, he became obsessed with the hydrogen bomb, which he eventually contributed to make.

ing from a state X_t at step t:

- Generate a random candidate state X' according to $g(X' \mid X_t)$.
- Calculate the acceptance ratio $r = \min\left(1, \frac{\mathbb{P}(X')}{\mathbb{P}(X_t)} \frac{g(X_t \mid X')}{g(X' \mid X_t)}\right)$.
- Generate a uniform random number $u \in [0, 1]$. If $u \leq r$, then accept the new state and set $X_{t+1} = X'$. Otherwise reject the new state and set $X_{t+1} = X$

The algorithm requires the ability to calculate the acceptance ratio r for all possible jump, and to draw a jump from any state. In addition, the last step above requires the generation of a uniform random number. The Metropolis procedure has been initially developed in the context of a symmetric distribution $g(X' \mid X) = g(X \mid X')$, and was later generalized to incorporate any proposal distribution, in which case an additional factor named the Hastings ratio $(g(X' \mid X)/g(X \mid X'))$ as to be accounted for.

4.2.5 Gibbs sampling

Whenever a joint distribution of variables is not known explicitly or is difficult to sample from directly, but the conditional distribution of each variable is easier to sample from, a specific algorithm known as Gibbs sampling is applicable. The original implementation of the Gibbs sampler by Geman and Geman (1984) was applied to a discrete image processing problem, a problem somewhat remote from statistical inference in the classical sense[2].

The individual random variables are sampled one at a time, with each variable being conditioned on the most recent values for all other variables. It can be shown that the sequence of samples constitutes a Markov chain, and the stationary distribution of that Markov chain is just the joint distribution. Gibbs sampling is particularly well adapted to sampling the posterior distribution of a Bayesian network, since they are composed of a set of individual random variables in which each variable is conditioned on only a small number of other variables.

Gibbs sampling, or more generally conditional Metropolis-Hastings can be considered a general framework for sampling from a large set of variables by sampling each variable (or in some cases, each group of variables) in turn. Various algorithms can be used to sample these individual variables, depending on the exact form of the multivariate distribution, it can incorporate the Metropolis–Hastings algorithm, or more sophisticated methods such as slice sampling, adaptive rejection sampling or adaptive rejection Metropolis.

4.2.6 Sufficient statistics & data augmentation

MCMC samplers target the distribution over the model parameters by repeatedly invoking the pruning algorithm to recalculate the pruning-based likelihood. This is most

[2]This paper is also responsible for the name Gibbs sampling, because it implemented this method for the Bayesian study of Gibbs random fields, which in turn, derive their name from the physicist Josiah Willard Gibbs (1839-1903).

often the limiting step of the MCMC. An alternative is to augment the observed sequence data with a realization of the random process resulting in a detailed substitution history over the tree (Nielsen, 2002; Rodrigue *et al.*, 2008b). Conditionally on the detailed substitution history $\mathcal{H}$, compatible with the data $\boldsymbol{D}$, the MC can be performed over the augmented configuration $(\mathcal{H}, \theta \mid \boldsymbol{D})$. The key idea that makes this strategy efficient is that the mapping-based likelihood depends on compact summary statistics of $\mathcal{H}$, leading to very fast evaluation of the likelihood (Lartillot, 2006; De Koning *et al.*, 2010; Romiguier *et al.*, 2012; Irvahn and Minin, 2014; Davydov *et al.*, 2016; Guéguen and Duret, 2018). On the other hand, this requires to implement more complex MC procedures that have to alternate between:

1. sampling $\mathcal{H}$ conditionally on the data and the current parameter configuration;
2. re-sampling the parameters conditionally on $\mathcal{H}$.

This strategy plays an essential role in the case of the complex phylogenetic codon model introduced in chapter 8.

4.2.7 Implementation

The software implementation of Bayesian phylogenetic models is generally a difficult endeavour. They must be flexible to adapt to different models of variations, while at the same time be reliable, reproducible, maintainable and fast. This is even more true for models integrating variation across sites, across branches or across genes. All these constraints led to the (still ongoing) development of a new Bayesian phylogenetic software platform called `BayesCode`, conducted by multiple maintainers with different goals and different models of evolution in mind. `BayesCode` adopts a modular design, using the graphical model formalism (see section 4.2.2) at a coarse-grained level, resulting in a flexible approach for model design by combining building blocks, corresponding to the fundamental distributions, the stochastic processes, and the likelihood computation routines that form the basis of a large family of phylogenetic models. Historically, the development of this software platform was initiated concurrently to the beginning of this thesis, and chapter 8 which model variation of selection across sites and drift across branches has been implemented under this framework. This software written in modern `C++` (version 14) is available at `https://github.com/bayesiancook/bayescode`.

5 Protein thermodynamics

Contents

The previous chapters introduced codon models and methodology for estimating parameters of mutation, selection and drift from empirical data, but remained elusive on the nature of the fitness landscape underlying proteins and did not question the causal determining factor for the strength of selection. This chapter will seek to clarify the relationship between phylogenetic codon models and biophysics of protein, such as to uncover the underlying properties of the selective pressures shaping protein-coding DNA sequences. Consequently, this chapter will present work at the interface between phylogenetic codon models and protein biophysics, where both fields are corroborated and consolidated by the other. Within this interface, many questions arise regarding the compatibility and feedback between these fields. Are the predictions of biophysical models of protein evolution compatible and confirmed by the application of phylogenetic codon models on empirical data? Or the other way around, can phylogenetic codon models

be informed by the underlying biophysics of proteins? To answer such questions, the first section of this chapter will present the theoretical foundations of protein biophysics, focusing on globular protein stability imposed by structural constraints. Subsequently, the second section will present how these models can explain in part the observed variation of selective constraints across genes, across sites and across branches observed with classical codon models. Thirdly, moving from classical to mechanistic codon model, the next section will discuss how fitness landscapes estimated by mechanistic codon models can also be related to the underlying protein biophysics. Finally, phylogenetic models augmented and incorporating the underlying biophysics are presented and the implications of such models is discussed.

Several authors have adequately reviewed the interface between both fields from a broad perspective (Tokuriki and Tawfik, 2009b; Liberles *et al.*, 2012; Serohijos and Shakhnovich, 2014; Sikosek and Chan, 2014; Arenas, 2015; Echave and Wilke, 2017; Bastolla *et al.*, 2017). This chapter, being aimed at evolutionary biologists familiar with phylogenetic codon models (already presented in chapter 3), addresses more specifically how such models fit within the prediction of protein biophysics of globular proteins.

5.1 The link between protein biophysics and molecular evolution

The ability of a protein to perform its function depends on the stability of its 3-dimensional folding structure, but also on its ability to bind ligands and/or interact with other proteins, both in terms of kinetics and stability. Theoretically, thermodynamics and kinetics of proteins are expected to be related to their function, and hence to selective constraints (Tokuriki and Tawfik, 2009a; Bastolla *et al.*, 2017).

5.1.1 Conformational stability of proteins

In thermodynamics, the stability of a protein is determined by the Gibbs free energy of its folded conformation, in comparison to the free energy of its possible unfolded conformations. Similarly to the mutation-selection Markov process defined in chapter 2, it is possible to derive the equilibrium distribution of conformations, where fitness is analogous to the opposite of free energy (less energetic conformations are more stable) and population size is analogous to inverse temperature. As a result, the probability of observing a protein in its folded conformation, given by the Boltzmann equation, is proportional to the exponential of the free energy of its folded conformation (G_{F}):

$$\mathbb{P}_{\mathrm{F}} = \frac{\mathrm{e}^{-G_{\mathrm{F}}/kT}}{\mathcal{Z}}, \tag{5.1}$$

where k is the Boltzmann constant and T is temperature in Kelvin, $\mathcal{Z}$ is a normalizing constant summed over all possible conformations, also called the conformational partition function. The conformational partition function is related to the free energy of the folded

state and of all the possible unfolded states:

$$\mathcal{Z} = \mathrm{e}^{-G_\mathrm{F}/kT} + \sum_{\text{unfolded}} \mathrm{e}^{-G_\text{unfolded}/kT} \tag{5.2}$$

$$= \mathrm{e}^{-G_\mathrm{F}/kT} + \mathrm{e}^{-G_\mathrm{U}/kT}, \tag{5.3}$$

where G_U encompasses the free energy of all the possible unfolded conformations. Altogether, the probability of observing the protein in folded conformation can then be re-expressed as:

$$\mathbb{P}_\mathrm{F} = \frac{\mathrm{e}^{-G_\mathrm{F}/kT}}{\mathrm{e}^{-G_\mathrm{F}/kT} + \mathrm{e}^{-G_\mathrm{U}/kT}}, \tag{5.4}$$

$$= \frac{\mathrm{e}^{-\Delta G/kT}}{1 + \mathrm{e}^{-\Delta G/kT}}, \tag{5.5}$$

$$= \frac{1}{1 + \mathrm{e}^{\Delta G/kT}}. \tag{5.6}$$

where $\Delta G = G_\mathrm{F} - G_\mathrm{U}$. Thus, in order for a protein to fold into its native state with high probability, the free energy gap (ΔG) has to be both negative and large in absolute value, as depicted in figure 5.1.

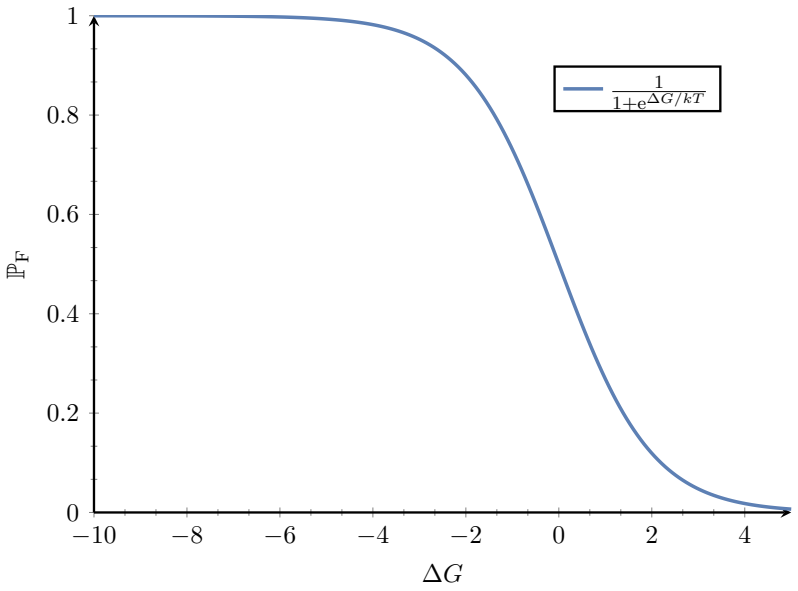


Figure 5.1: *Probability of folding ($\mathbb{P}_F$) as a function of the free energy gap ($\Delta G = G_F - G_U$). ΔG is in kcal/mol and $1/kT = 1.686$ mol/kcal at $25°C$ (or $298.2K$). The free energy gap has to be both negative and large in absolute value for the protein to be folded.*

In this context, mutations stabilize the protein only if they decrease the free energy of the folded conformations more than they decrease the free energy of unfolded conformations. For example, a transition to an amino acid that decreases by the same amount the free energy of both folded and unfolded conformations will have no impact on the stability of the protein. As a result, protein stability can be increased by stabilizing the folded conformation (positive design) or destabilizing the competing unfolded conformations (negative design). We can thus characterize the destabilizing effect of a mutation

by its effect on ΔG, denoted $\Delta\Delta G$:

$$\Delta\Delta G = \Delta G\,(\text{Mutant}) - \Delta G\,(\text{Wild type})\,, \tag{5.7}$$

where by definition $\Delta\Delta G < 0$ for stabilizing mutations, and conversely $\Delta\Delta G > 0$ for destabilizing mutations.

Free energy gaps ΔG can be experimentally measured, and fall within the range of -25 to -5 kcal/mol (Kumar *et al.*, 2006; Gromiha *et al.*, 2016). Moreover, empirical measurements of folding stability changes due to single point mutations can also be obtained experimentally. This process is costly and has to be done for each single mutation (Rocklin *et al.*, 2017).

Alternatively, the free energy gap of a protein can be computed with a biophysical model of the protein, by modelling the atomic structure and the potential energy of contact between residues at the atomic level in a 3-dimensional structure. Computing the free energy gap for a given protein sequence is challenging for any given conformation of the backbone, since it also depends on the conformation of the side chains as well as the solvent. To allow for faster computation, coarse-grained approximations have been proposed, in the form of statistical potentials, which approximate free energy (G) as a sum of free energy terms over all pairwise contacts between residues across the protein (Miyazawa and Jernigan, 1985).

For a given folded conformation of the protein, the statistical potential gives G_{F}. However, in order to get ΔG, one still needs to sum over all unfolded conformations to compute $\mathcal{Z}$, or G_{U}. Models typically approximate the distribution of unfolded Gibbs free energy using representative decoy conformations for which energy is computed, assuming a quasi-chemical or normal approximation (Goldstein, 2011).

Alternatively, in order to explicitly sum over all possible conformations, some models approximate the structure and dynamics of proteins by 2-dimensional lattice models with regular pavement (Taverna and Goldstein, 2002; Noivirt-Brik *et al.*, 2009). Lattice models are designed to sum over all possible conformations, and are useful as a theoretical construct to gain new insights about biophysics and protein evolution. However, lattice models are empirically less directly usable. In between these two extremes, many models can approximate with various degrees of freedom and parametrization the stability of a protein from its sequence.

5.1.2 From stability to fitness

Empirically, a large body of evidence indicates that the stability, or, in other words, the ability to fold in globular conformation, is a target of natural selection (Sikosek and Chan, 2014). Even though the association between protein sequence and protein stability is within reach and can be obtained with various degrees of approximations, the association between protein stability and fitness is more elusive and difficult to apprehend. It is known that protein stability is related to fitness, as demonstrated by a study of beta-lactamase TEM-1 mutants (Jacquier *et al.*, 2013), or illustrated by the use of functional assays to identify stabilizing mutations (Araya *et al.*, 2012). However, it is not clear

whether protein stability increases fitness by being more efficient, or whether it is the deleterious cytotoxic effect of unfolded proteins that results in purifying selection for destabilizing mutations. Additionally, the ability to bind other proteins may interfere with stability against misfolding, and large functional movements may imply a stability cost.

The relationship between stability and fitness raises the more general question of why globular proteins are marginally stable. Indeed, the optimal stability is never achieved, and two types of explanation have been proposed. Firstly, that it could be the consequence of the stability-activity trade-off such that proteins are selected for an intermediate stability. Secondly, and more fundamentally, that it is an expectation of the mutation-selection equilibrium even under directional selection for stability (Taverna and Goldstein, 2002). These two explanations are not mutually incompatible, and can both explain the observed marginal stability of proteins.

Furthermore, translation errors act like point mutations, with a fairly high translation error rate. They have measurable destabilizing effects in terms of $\Delta\Delta G$, just like non-synonymous mutations. The fitness associated with a sequence variant at the DNA level thus integrates the average effect of these destabilizing mutations induced at the translation level. For this reason, at mutation-selection equilibrium, the protein encoded at the DNA level tends to have a more negative ΔG than without error, as if to anticipate these additional destabilizing effects (Wilke and Drummond, 2006).

5.1.3 Conformational stability and epistasis

Computing the free energy gap ΔG requires knowledge of interacting energy contact between amino acids in close proximity. It is important to remember that proximity can exist even between amino acids far apart in the folded structure, inasmuch as they may be in contact in unfolded structures. As a result, the ΔG impact of a mutation at a specific position of the protein depends on the context and the amino acids at other positions. Specifically, amino-acid changes can be stabilizing or destabilizing depending on the amino acids present at other positions. Moreover, even if ΔG would be an additive trait, in the sense that each position contributes independently to ΔG without pairwise interaction terms, the selective effect of a mutation would still depend on the amino acids at other positions. The reason is that even if ΔG is an additive trait, the log-fitness is still not a linear function of ΔG. The former case of site interdependence due to interacting terms is called specific epistasis, while the latter case of non-linearity of the fitness function is called by contrast non-specific epistasis.

Formally, the relation between sequence ($\mathbb{S}$) and log-fitness (f) is complex, and can be abstracted by an intermediate phenotype. In the specific case of conformational stability, the phenotype is the free energy gap, and the ternary relationship develops as:

$$\mathbb{S} \rightarrow \Delta G(\mathbb{S}) \rightarrow f\left(\Delta G\right). \tag{5.8}$$

Withing this ternary relationship, the fitness effect of a mutation is site-specific in only one specific case, namely that the phenotype is additive and that the log-fitness is linear

with the phenotype. Whenever one of these two assumptions is not valid, the fitness effect of a mutation at a specific site depends on the overall sequence. This site interdependence represents a challenge for phylogenetic codons, generally not modelled explicitly with some exceptions (see section 5.3).

Site interdependence has important consequences on molecular evolution of protein sequences, and results in entrenchment and Stokes shifts (Pollock *et al.*, 2012; Shah *et al.*, 2015) Briefly speaking, even if a mutation is nearly-neutral upon fixation, subsequently fixed mutations on other sites make the original substitution more and more deleterious to revert over time (Lunzer *et al.*, 2010; Naumenko *et al.*, 2012; Mccandlish *et al.*, 2013).

5.1.4 Aggregation avoidance

So far, proteins have been seen as independent machinery of cells, however, within the crowded intracellular space, proteins are not independent entities but are interacting with other proteins and engaged in non-specific interactions (Yang *et al.*, 2012; Zhang *et al.*, 2013). In non-specific interactions at the protein surface, stabilizing amino acids are hydrophilic and destabilizing amino acids are hydrophobic, sticking to hydrophobic residues in other proteins (Dixit and Maslov, 2013; Manhart and Morozov, 2015a). The misinteraction avoidance hypothesis predicts that, compared with lowly expressed proteins, highly expressed proteins disfavour residues that promote misinteraction, exhibit a lower misinteraction probability per molecule and have higher conservation for misinteraction avoiding residues.

5.2 Confronting classical codon models with protein biophysics

Application of phylogenetic codon models to empirical data has made it possible to infer the variation in the overall strength of the selective constraints across genes, sites, and branches. These results have been interpreted in the light of the underlying biophysics.

5.2.1 Variation across genes

Phylogenetic codon models can readily be applied to independent single-gene multiple-sequence alignments. The d_N/d_S estimated for each gene can then be related to the selective constraints acting on the gene. As a result, increased availability of genomic data together with the advancement of computing resources and algorithms prompted an extensive search for the major determining factor of a gene's d_N/d_S. Surprisingly, the functional importance of a protein, widely thought to approximate the level of functional constraint, has only a minor role, whereas protein expression level (mRNA concentration) is found to be a major determinant (Zhang and Yang, 2015). Most importantly, this relationship is negative such that genes with a high expression level are under stronger purifying selection, and have a lower d_N/d_S at the level of the gene (Duret and Mouch-

iroud, 2000; Drummond *et al.*, 2005; Zhang and Yang, 2015). In unicellular organisms, the mRNA concentration of a gene varies across cell cycle stages and environments, but most studies used data collected from the mid-log phase of growth under rich media, which presumably reflects average concentrations across cell cycle stages. In multicellular organisms, mRNA concentration data used are typically from the whole organism or are averaged from several examined tissues. Because of the strong correlation between mRNA and protein concentrations, the negative correlation between protein concentration and evolutionary rate is also strong.

Theoretical models based on protein stability presented previously have been invoked to explain the negative correlation between d_N/d_S and expression level (Wilke and Drummond, 2006; Drummond and Wilke, 2008). The rationale is that for the same fraction of misfolded proteins, a strongly expressed protein will produce more macromolecules toxic to the cell than a weakly expressed protein. As a result, selection against protein misfolding induces abundant proteins to evolve to greater stability, where the protein is more constrained and evolves more slowly (Serohijos *et al.*, 2012).

However, even for those proteins of comparable expression levels, their d_N/d_S still spans several orders of magnitude (Drummond and Wilke, 2008). This observation suggest that protein abundance, although a major determinant of d_N/d_S, is not its only causal variable. As an example, some topologies are more robust (depending on the density of contacts, in particular), and therefore evolve faster, which may be one of the contributing factors of the residual variance (Echave and Wilke, 2017).

5.2.2 Variation across sites

Similarly to the search for determining factors of d_N/d_S at the gene level, an extensive search had been conducted at the site level, within a protein. The major determinant of site-specific d_N/d_S proved to be relative solvent accessibility (RSA), where sites with higher RSA display a higher d_N/d_S (Ramsey *et al.*, 2011). It was later shown that the number of native inter-residue contacts formed by a protein site, which is negatively correlated with the RSA, is a stronger predictor of site-specific d_N/d_S (Yeh *et al.*, 2013).

The observations that surface residues of globular proteins undergo substitution more rapidly than those in the core is generally attributed to the fact that natural selection imposes stronger constraints on buried sites. In fact, selection for protein stability induces stronger constraints on amino-acid residues located inside a protein structure (that is, core residues), which indeed have more central roles than surface residues in the protein's Gibbs free energy of folding.

Altogether, d_N/d_S changes dramatically between exposed and buried sites in such a way that buried sites tend to evolve more slowly than exposed sites, compatible with models of selection for protein stability (Echave *et al.*, 2016).

5.2.3 Variation across branches

As already mentioned in chapter 1, the nearly-neutral theory predicts a lower d_N/d_S in species with a higher N_e, due to a better purification of weakly deleterious mutants. Biophysical knowledge can be useful here to get more insight about the magnitude of the response of d_N/d_S to changes in N_e. Surprisingly, under simple biophysically inspired models assuming that proteins are under selection for their thermodynamic stability, with fitness being proportional to the folded fraction, computational experiments have led to the observation that d_N/d_S is essentially independent of N_e (Goldstein, 2013). This observation has been explained by the equimutability of the free energy of folding, namely, that the distribution of changes in free energy of folding ($\Delta\Delta G$) due to mutations is approximately independent of the current free energy (ΔG), a necessary and sufficient condition (under the condition that fitness is log-concave) to obtain independence between d_N/d_S and N_e (Cherry, 1998). In reality, however, the distribution of $\Delta\Delta G$ is expected to at least weakly depend on ΔG due to combinatorial considerations. For example, if a protein sequence is already maximally stable, only destabilizing (or neutral) mutations can occur, which has been empirically observed (Serohijos *et al.*, 2012).

5.2.4 Integrating several levels

Ultimately, studies presented in this section focus on the scaling of d_N/d_S to either protein abundance, or to effective population size, and also to relative solvent accessibility. However, these various factors susceptible to modulate the d_N/d_S have been rarely investigated simultaneously. Why is d_N/d_S supposedly independent of N_e but depend on protein abundance? For example, is the relationship between d_N/d_S and either protein abundance or population size expected to be different? I argue the integration and unification between these levels are scarcely made. For example, under which models for the relation between biophysics and fitness are the relations of d_N/d_S to protein abundance and population size expected to be the same or different? What do empirical data have to say quantitatively about this question? Can we derive quantitative estimates of the magnitude of these responses, and compare them with empirical estimates? I argue that some work is still needed toward a better integration and unification between these multiple aspects of the role of biophysics in molecular evolution.

5.3 Informing mutation-selection codon models using protein biophysics and experimental data

In section 5.2, I reviewed how the selective patterns inferred using classical models could then be confronted with insights from biophysics. In the case of mutation-selection codon models, on the other hand, knowledge obtained from biophysics can be more directly introduced into the model.

5.3.1 Experimentally informed site-specific codon models

In an experimental context, it is possible to mutate the DNA of an organism and establish an experiment where the mutant competes with the resident in a specific medium, and the difference in growth of the two variants allows to determine the fitness impact of the mutation. In the case of free-living unicellular organisms, such process can be automated to estimate selection coefficients of a wide variety of mutants, an experiment called deep mutational scanning. Technically, for each site of the protein, the fitness of the 20 amino acids can be experimentally determined and the resulting fitness landscape (also named preferences or fitness profile) can be estimated, as shown in figure 5.2. Such experimentally determined fitness landscapes are directly comparable to statistical estimates by phylogenetic codon models, under the assumption that the site-specific fitness landscape is kept constant along the phylogeny. Bloom (2014b,a) found that site-specific evolutionary models informed by experimentally determined profiles greatly outperformed non site-specific alternatives in fitting phylogenies of proteins, from humans, swine, equine, and avian influenza (Doud *et al.*, 2015). Moreover, Bloom (2017) recruited experimentally determined fitness profiles to determine which site of the protein are sufficiently different from their phylogenetic counterpart to be considered under adaptation.

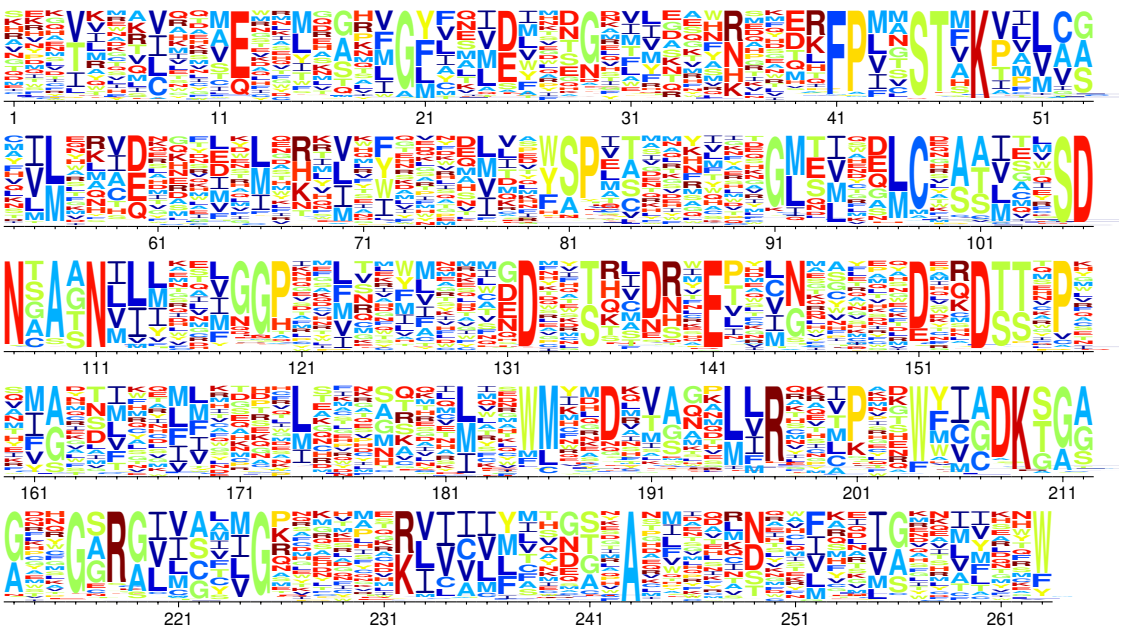


Figure 5.2: *Site-specific deep mutational scanning data on β-lactamase (bacteria) of Stiffler* et al. *(2015). For each site, preferences of the 20 amino acids are represented by the height of the corresponding amino-acid letter. The analysis is performed using phydms (Hilton* et al.*, 2017).*

5.3.2 Structurally constrained site-interdependent codon models

It has long been realized that inter-residue interactions within proteins leads to amino-acid fixation probabilities that are dependent upon amino acids present at other sites. More generally, site-specific fixation probabilities may change along an evolutionary trajectory because the selection coefficient of a given mutation may depend on the specific sequence background in which it occurs (Goldstein and Pollock, 2016). However, both classical codon models and mechanistic codon models rely on the assumption of site independence, where each site of the protein is modelled as an independent Markov process. Accordingly, each site is considered separately, and defines an independent Markov substitution process along the branches of a tree.

From a modelling and inference perspective, accounting for epistasis is challenging both in terms of parametrization and computational complexity (Manhart and Morozov, 2015b). Means of relaxing this assumption have been pursued, usually with dependence introduced between a limited number of sites (Felsenstein and Churchill, 1996). In particular, models explicitly treating protein structure and site interdependencies have been developed, recruiting a coarse-grained protein structure conjointly to a statistical potential scoring the compatibility between sequence and structure, in order to evaluate the probability of fixation of a given mutation (Robinson *et al.*, 2003; Rodrigue *et al.*, 2005).

Subsequently, methods to assess the statistical fit of such computationally complex models had been developed (Rodrigue *et al.*, 2009), as well as refinement of statistical potentials (Kleinman *et al.*, 2010). These structurally constrained models have been shown to fit data better than the corresponding models that ignore protein structure. However, some of the available site-specific phylogenetic codon models still better fit the data than structurally constrained models, possibly indicating that alternative models should be explored in order to better incorporate structural constraints and protein biophysics.

Alternatively, the assumption of site independence can be understood as considering that substitution processes at the level of sites are averaged over time, where the dependencies to other sites are integrated over the course of the process. As a result, statistical methods relying on site-specific processes while accounting for epistasis consist in obtaining the marginal process for a specific site, derived analytically from the joint process integrated over the other sites. Projecting a joint process of several sites into a single site process leverages mean-field theory developed in statistical physics, and has been used to develop phylogenetic models accounting for protein structure (Chi *et al.*, 2018) and protein stability (Arenas *et al.*, 2015, 2017). Unfortunately, these methods are not parameterized directly in terms of parameters of evolution, namely mutation and effective population size, and the estimated fitness parameters cannot be related to empirically determined parameters.

5.4 General conclusions

Finally, models of protein biophysics are appealing to evolutionary biologists since they are based on theoretical grounds and can also be confronted to empirical data. However, integration of protein biophysics models into the framework of phylogenetic inference is difficult, and inference models have to balance the trade-off between complexity and simplicity. Moreover, I argue that phylogenetic models should be mechanistic in principle, or, in other words, they should be defined in terms of parameters that can be accessed by independent experimental means, such as to confront estimates. As an example, analytical models of protein biophysics relating probability of fixation to molecular and thermodynamic parameters can be fitted to protein-coding DNA sequences. Parameter estimates can be compared to their empirically determined counterparts, such as to verify and solidify the soundness of both phylogenetic inference and protein biophysics.

6 Thesis objectives

Contents

The neutral theory, and its nearly-neutral extension, such as historically reviewed in chapter 1, have deeply influenced our understanding of population genetics and molecular evolution. Beyond the disputes and the controversies between neutralism and selectionism, the current consensus is to view the evolution of genetic sequences as a stochastic process. One component of this process is creating diversity through mutation; an antagonistic component is filtering out this diversity through selection; and finally, the balance between these components is tuned by the effective population size, which determines the amount of random drift, formally presented in chapter 2. The long-term outcome of this evolutionary process is an accumulation of point substitutions (both synonymous and non-synonymous) between species. Relying on this primary source of information contained in multiple sequence alignments of protein-coding genes obtained from contemporaneous species, the aim of phylogenetic codon models, as discussed in chapter 3, is to better characterize and quantify the interplay between mutation, selection and random drift. Codon models are still an active area of research, and proceed from two different philosophies: on one side, phenomenological models, aiming to capture the net effect of selection through $\omega = d_N/d_S$; on the other side, mechanistic approaches, with the more ambitious aim of modelling the fine-grained fitness landscape. As it stands, however, many questions are still open, and current models, whether phenomenological or mechanistic, present many weaknesses. Phenomenological approaches could still be improved, while staying in the idea of not explicitly modelling the detailed fitness landscape. As for mechanistic approaches, in their current version, are making very strong assumptions, such as site independence, a time-independent fitness landscape, but also constant effective population size across the whole phylogeny. More fundamentally, there

is a certain gap to be filled between these two alternative approaches, and better connections could be made between them.

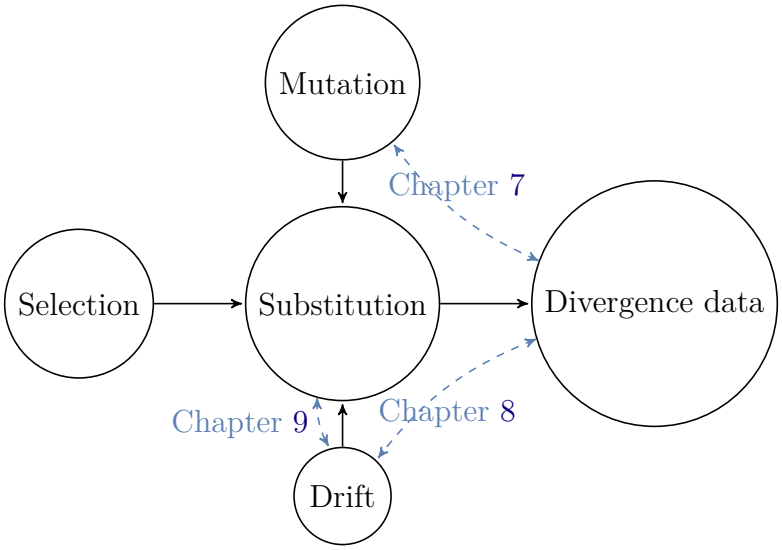


Figure 6.1: *In this thesis, several aspects of the mutation, selection and drift equilibrium are studied and related to empirical data, in the context of protein-coding DNA sequences. Firstly, because the composition of protein-coding DNA sequences does not reflect the underlying mutational process but its filtering by selection at the level of amino acids, a careful phenomenological modelling is necessary to uncover mutational process and nucleotide fixation bias, a study presented in chapter 7. Secondly, the balance between mutation and selection is arbitrated by drift, which is mediated by effective population size and its changes along a phylogeny can be estimated by mechanistic codon models, a study presented in chapter 8. Finally, selection for protein stability implies an analytical relationship between the rate of evolution and effective population size and protein expression level, a study presented in chapter 9.*

In this context, my thesis work represents an attempt at revisiting the question of how to correctly disentangle the complex interactions between mutation, selection and random drift using phylogenetic codon models, under both approaches, either phenomenological or mechanistic. During this work, I have confronted theoretical insights with empirical data, using a combination of analytical developments, simulation experiments and Bayesian inference. The results are divided in three chapters, each written in the form of an independent manuscript, that shall be submitted to peer-reviewed journals. The first article (chapter 7) revisits the question of the balance between mutation bias and selection, and how this balance should be properly formalized in the context of classical (phenomenological) codon models. The second manuscript (chapter 8, with supplementary materials in chapter 11), explores the question of accounting for the variation in long-term effective population size (N_e) between species, in the context of a mechanistic mutation-selection model. The work presented in this manuscript represents the most intensive part of the PhD work, in terms of modelling, Monte Carlo algorithmics (see chapter 4) and software development. Finally, some of the observations made during this second part of my work, in particular the relatively narrow dynamic range of varia-

tion in N_e uncovered using this fully mechanistic approach, prompted me to revisit the question of how protein biophysics (see chapter 5), and more generally epistasis, can quantitatively modulate the response of the molecular evolutionary process to changes in effective population size. This last work is presented as a third manuscript (chapter 9, with supplementary materials in chapter 12).

6.1 Robustness of codon models to mutational bias

Nucleotide composition in protein-coding sequences is the result of the equilibrium between mutation and selection. Because of selection, the nucleotide composition of protein-coding sequences is different from what would be expected under a pure mutational process. In particular, it differs between the three coding positions, with the third position showing more extreme composition than the first and the second positions. This empirical observation is well known. Yet, classical codon models (see chapter 3) do not correctly capture this phenomenon. Instead, in their classical parameterization, in terms of a 4x4 nucleotide rate matrix and a single ω parameter, phenomenological codon models predict that the nucleotide composition should be the same for all 3 positions of the codons, and should be equal to the equilibrium frequencies of the underlying 4x4 nucleotide process. Alternatively, to accommodate this variation across coding positions, some models allow for different nucleotide rate matrices at the three positions. However, this approach is problematic since the mutation process should in principle be blind to the coding structure, and should be homogeneous across coding positions. Although this misconception has probably minor impact on the detection of positive selection, it is a clear symptom of a more fundamental issue with teasing apart mutation rates and fixation biases in the context of phenomenological codon models. Practically, this could have important consequences, in particular, given the current interest in modelling the impact of GC-biased gene conversion (gBGC) on the evolution of protein-coding sequences, a factor which requires mutation and fixation biases to be carefully disentangled. Conceptually, the problem comes from the fact that, at the mutation-selection equilibrium, there is a net selection differential, or net fixation bias, acting against the mutational pressure. In other words, at equilibrium, ω is not the same in different mutational directions. Because they capture selection through a single parameter ω, classical codon models cannot correctly capture this net fixation bias. To address this problem, chapter 7 presents an alternative modelling approach, where ω is not seen as a scalar anymore, but as an array of ω values unfolding along multiple directions. This model is tested against empirical and simulated protein coding DNA alignments.

6.2 Inferring long-term population size

Presented in section 3.2, mechanistic phylogenetic codon models are grounded on population genetics first principles. Being explicitly parameterized in terms of mutation rates and population-scaled fitness coefficients, these models represent a principled approach for investigating the intricate interplay between mutation, selection and drift. In their current form, mutation-selection models assume a fixed and site-specific fitness landscape, without epistasis. As a result, they are entirely characterized by the collection of site-specific amino-acid fitness profiles. However, thus far, they have relied on the assumption of a constant effective population size across the phylogeny, clearly an unreasonable hypothesis. Selection and drift are confounded parameters, but they can nevertheless be disentangled by assuming that fitness is fixed along the phylogeny but changing along the sequence, and orthogonally, by assuming that effective population size is constant across sites, but variable across the phylogeny. In addition to effective population size (N_e), the mutation rate (μ) is also susceptible to vary between lineages. Furthermore, both N_e and μ are expected to co-vary with life-history traits (LHTs). This suggests that the model should more globally account for the joint evolutionary process followed by all of these lineage-specific variables (N_e, μ, and LHTs). In this direction, chapter 8 introduces an extended mutation-selection model jointly reconstructing the fitness landscape across sites and long-term trends in effective population size, mutation rate and LHTs along the phylogeny, from an alignment of DNA coding sequences and a matrix of observed LHTs in extant species. The model was implemented in a Bayesian Monte Carlo framework (see chapter 4.2). Together, the model estimates correlation between reconstructed life-history traits, mutation rate and effective population size, intrinsically including phylogenetic inertia. It was tested against simulated data, and finally applied to empirical data in mammals, isopods, primates and Drosophila. The reconstructed history of N_e in these groups appears to correlate with LHTs or ecological variables in a way that suggests that the reconstruction is reasonable, at least in its global trends. On the other hand, the range of variation in N_e inferred across species is surprisingly narrow. This last point suggests that some of the assumptions of the model, in particular concerning the structure of the assumed fitness landscape, are potentially problematic.

6.3 Substitution rate response to changes in effective population size and expression level

The surprisingly narrow range of variation in N_e inferred across large phylogenies by the mechanistic mutation-selection model such as mentioned above (section 6.2), prompted me to conduct a more detailed theoretical investigation of the quantitative impact of changes in N_e on the molecular evolutionary process followed by protein-coding sequences. A particularly important variable to investigate in this direction is the substitution rate of selected mutations relative to the neutral substitution rate $\omega = d_N/d_S$.

Under the nearly-neutral theory of evolution, lineages with large effective population size (N_e) are expected to undergo stronger purifying selection, and consequently a decrease in ω. Empirical correlation patterns between ω and either life-history traits or synonymous diversity (which is a proxy of N_e), have tended to confirm this prediction. However, simulations using computational models based on the biophysics of protein conformational stability (presented in section 5.1) have suggested that ω can in fact be virtually independent of N_e. The discrepancy between these conclusions suggests that a more detailed quantitative investigation of what determines the quantitative response of ω to changes in N_e, depending on the exact model of the mapping from sequences to fitness, would be useful. Another related question is how ω varies between proteins, depending on their expression level. Empirically, there is a robust negative correlation between ω and expression level across genes. Theoretically, many biophysically inspired models suggest that the response of ω to changes in expression levels should be the same as, or similar to, its response to changes in N_e. This suggests that the two questions, the impact of changes in N_e and in expression levels, would benefit from a simultaneous theoretical investigation. To address these questions, chapter 9 derives a theoretical approximation for the quantitative response of ω to changes in N_e and in expression level, under an explicit genotype-phenotype-fitness map. The method presented is generally valid for an additive trait and log-concave fitness functions, but more specifically applied to proteins undergoing selection for their conformational stability. The analytical results, obtained under simplified models, are corroborated by simulations under more complex models. Finally, analytical predictions of the response of ω to changes in N_e and expression level are confronted with empirical data, while other aspects of protein biophysics such as protein-protein interactions are also discussed.

Part II

Studies

7

Robustness of codon models to mutational bias

Thibault Latrille[1, 2], Nicolas Lartillot[1]

[1]Université de Lyon, Université Lyon 1, UMR CNRS 5558 Laboratoire de Biométrie et Biologie Évolutive, 69622 Villeurbanne, France
[2]École Normale Supérieure de Lyon, Universite de Lyon, 69007 Lyon, France

Contents

7.1 Introduction

Phylogenetic codon models are now routinely used in many domains of bioinformatics and molecular evolutionary studies. One of their main applications has been to characterize the genes, sites (Nielsen and Yang, 1998), or lineages (Zhang and Nielsen, 2005) having experienced positive selection. More generally, these models highlight the respective contributions of mutation, selection, genetic drift and biased gene conversion (Kosiol and Anisimova, 2019), and the causes of their variation between genes (Zhang and Yang, 2015) or across species (Lartillot and Poujol, 2011).

Conceptually, codon models take advantage of the fact that synonymous and non-synonymous substitutions are differentially impacted by selection. Assuming synonymous mutations are neutral, the synonymous substitution rate is equal to the underlying mutation rate (Kimura, 1983). Non-synonymous substitutions, on the other hand, reflect the combined effect of mutation and selection (Ohta, 1995). Classical codon models formalize this idea by invoking a single parameter ω, acting multiplicatively on non-synonymous substitutions rates (Muse and Gaut, 1994; Goldman and Yang, 1994). Using a parametric model automatically corrects for the multiplicity issues created by the complex structure of the genetic code and by uneven mutation rates between nucleotides. As a result, ω captures the net, or aggregate, effect of selection on non-synonymous mutations.

Classical codon models, so defined, are phenomenological, in the sense that they capture a complex mixture of selective effects through a single parameter (Rodrigue and Philippe, 2010). In reality, the selective effects associated with non-synonymous mutations depends on the context (site-specificity) and the amino acids involved in the transition (Kosiol *et al.*, 2007). Attempts at an explicit modelling of these complex selective landscapes have also been done, leading to mechanistic codon models, based on the mutation-selection formalism (Halpern and Bruno, 1998). These models, further developed in multiple inference frameworks (Rodrigue *et al.*, 2010; Tamuri and Goldstein, 2012), sometimes using empirically informed fitness landscapes (Bloom, 2014b), could have many interesting applications, such as inferring the distribution of fitness effects (Tamuri and Goldstein, 2012) or detecting genes under adaptation (Rodrigue and Lartillot, 2016), or even phylogenetic inference. However, they are computationally complex and potentially sensitive to the violation of their assumptions about the fitness landscape (such as site independence). For this reason, classical codon models remain an attractive, potentially more robust, although still perfectible approach.

The parametric design of current classical models, relying on a single aggregate parameter ω, raises the question whether they reliably estimate the underlying mutational process. Several observations suggest that this may not be the case. For instance, in their simplest form (Muse and Gaut, 1994; Goldman and Yang, 1994), classical codon models predict that the nucleotide composition should be the same for all three positions of the codons, and should be equal to the nucleotide equilibrium frequencies implied by the underlying nucleotide substitution rate matrix. In reality, the nucleotide composition differs: the third position shows more extreme GC composition, reflecting the underlying

mutation bias, compared to the first and second positions, which are typically closer to 50% GC (Singer and Hickey, 2000).

These modulations across the three coding positions have been accommodated using the so-called 3x4 formalism (Goldman and Yang, 1994; Kosakovsky Pond and Muse, 2005b), allowing for different nucleotide rate matrices at the three coding positions. However, this is also problematic, since this modelling approach has the consequence that synonymous substitutions, say, from A to C, occur at different rates at the first and third positions. Yet, in reality, the mutation process is blind to the coding structure, and should be homogeneous across coding positions, and if neutral, all mutations from A to C should thus have the same rate.

These observations suggest that the mutation matrix (1x4) or matrices (3x4) estimated by codon models are not correctly reflecting the mutation rates between nucleotides (Rodrigue *et al.*, 2008a). Instead, what these matrices are capturing is the result of the compromise between mutation and selection at the level of the realized nucleotide frequencies. For detecting selection, this problem is probably minor, although it still bears consequences on the estimation of ω (Spielman and Wilke, 2015). Conceptually, however, it is a clear symptom of a more fundamental problem: mutation rates and fixation probabilities are not correctly teased apart by current codon models.

Practically, this misconception could have important consequences in contexts other than tests of positive selection. In particular, there is a current interest in investigating the variation between species in GC content, and its effect on the evolution of protein-coding sequences (Bolívar *et al.*, 2019). An important factor here is biased gene conversion toward GC (called gBGC), which can confound the tests for detecting positive selection and, more generally, the estimation of d_N/d_S (Galtier *et al.*, 2009; Ratnakumar *et al.*, 2010; Figuet *et al.*, 2014). Even in the absence of gBGC, however, uneven mutation rates varying across species can have an important impact on the estimation of the strength of selection. All this suggests that, even before introducing gBGC in codon models, correctly formalizing the interplay between mutation and selection in current codon models would be an important first step.

In this direction, the key point that needs to be correctly formalized is the following. If the nucleotide's realized frequencies are the result of a compromise between mutation and selection, then this implies that the strength of selection is not the same between all nucleotide or amino-acid pairs. For instance, if the mutation process is AT-biased, then, because of selection, the realized nucleotide frequencies at equilibrium will be less AT-biased than expected under the pure mutation process. However, this implies that, at equilibrium, there will be a net mutation pressure toward AT, which has to be compensated for by a net selection differential toward GC. In other words, at equilibrium, mutations toward AT will be more deleterious on average than those toward GC, or equivalently, the d_N/d_S toward AT will be lower than the d_N/d_S toward GC.

All this suggests that, in order for a codon model to correctly formalize this subtle interplay between mutation and selection, the component of the parameter vector responsible for absorbing the net effect of selection (i.e. ω) should not be a scalar, as is

currently the case. Instead, it should be a tensor, that is, an array of ω values unfolding along multiple directions. In the present work, we address the question of whether we can derive a parametric structure being able correctly tease apart mutation rates and selection, and this, without having to explicitly model the underlying fitness landscape. In order to derive a codon model along those lines, our strategy is to first assume a true site-specific evolutionary process, following the mutation-selection formalism. Then, we derive the mean substitution process implied across all sites by this mechanistic model and identify the mean fixation probabilities appearing in this mean-field process with the ω tensor to be estimated. Inferring parameters on simulated alignments, we show that the model correctly estimates the mutation rates, as well as the mean effect of selection.

7.2 Results

To illustrate the problem, we first conduct simulation experiments under a simple mutation-selection substitution model assuming site-specific amino-acid preferences. We use these simulation experiments to explore through summary statistics the intricate interplay between mutation and selection. Then, we explore how codon models with different parameterizations are able to infer the mutation rates and the strength of selection on these simulated alignments. Finally, these alternative models are applied to empirical data.

7.2.1 Simulations experiments

Simulations of protein-coding DNA sequences were conducted under an origination-fixation substitution process (McCandlish and Stoltzfus, 2014) at the level of codons (see section 7.4.1). We assume a simple mutation process with a single parameter controlling the mutational bias toward AT, denoted $\lambda = (\sigma_A+\sigma_T)/(\sigma_C+\sigma_G)$, where σ_x is the equilibrium frequency of nucleotide x. This mutational process is shared by all sites of the sequence. With regards to selection, synonymous mutations are considered neutral, such that the synonymous substitution rate equal to the underlying mutation rate. At the non-synonymous level, selection is modelled by introducing site-specific amino-acid fitness profiles (i.e. a vector of 20 fitnesses for each coding site), which are drawn from a uniform dirichlet distribution of concentration parameter α. A low α induces site-specific profiles having a large variance, with some amino acids with a high fitness while all other have a low fitness. Conversely, a large value for α induces more even amino-acid fitness profiles at each site. Thus, ultimately, the stringency of selection increase with decreasing α. Altogether, the two parameters of the model tune the mutation bias (λ) and the stringency of selection (α), respectively, as depicted in figure 7.1. All simulations are obtained using the same underlying topology of 180 species and 498 codon sites.

Simulation of this origination-fixation process along a species tree result in a multiple sequence alignment of coding sequences for the extant species, from which summary statistics can then be computed. One such straightforward summary statistic is the frequency of the different nucleotides, and the resulting nucleotide bias AT/GC observed

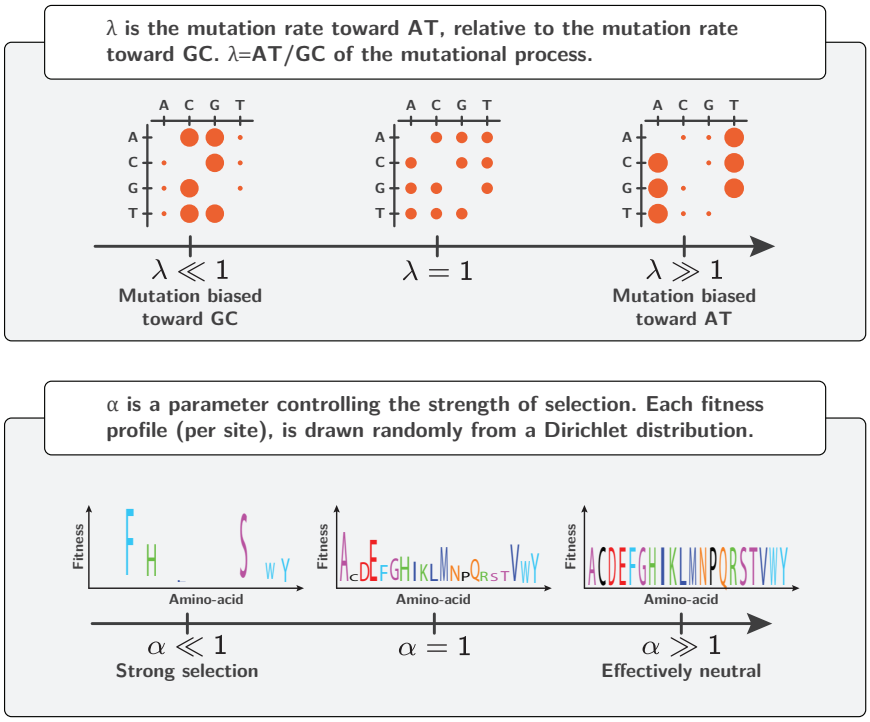


Figure 7.1: *Parameters of the mutation-selection model. Mutational bias (toward A and T) is shared by all sites of the sequence, and tuned by the parameter λ. Conversely, each codon site of the sequence is defined by a unique fitness profile, drawn from a Dirichlet distribution with concentration parameter α. Stringency of selection increase with decreasing α.*

in the alignment. This observed nucleotide bias can be computed separately for each coding position (first, second and third) and compared to the underlying true mutational bias λ. As can be seen from figure 7.2, the third position of codons reflects the underlying mutational bias quite faithfully, while the first and second positions are impacted by the strength of selection and display nucleotide biases that are less extreme than the one implied by the mutational process. This differential effect across the three coding positions is explained by nucleotide mutations at the third codon position being more often synonymous, while mutations at the first and second positions are more often changing the amino-acid and are thus more often under purifying selection.

Apart from the observed nucleotide bias in the alignment, the diversity of amino acids is an important indicator of the selective constraints that the sequence experiences. This diversity can be quantified by the frequencies of amino acids observed across all taxa in the alignment, and then summarized through a single statistic, namely the Shannon entropy of amino-acid frequencies (Goldstein and Pollock, 2017), as described in section 7.4.4. Diversity can be quantified for a given site of the sequence, and this site-specific diversity can subsequently be averaged over all sites of the sequence (yielding what is hereafter called the site-specific diversity). Alternatively, the diversity can be quantified directly on the amino-acid frequencies observed across the whole sequence alignment (which we refer to as the sequence diversity).

These two variants of the amino-acid diversity are computed for alignments simu-

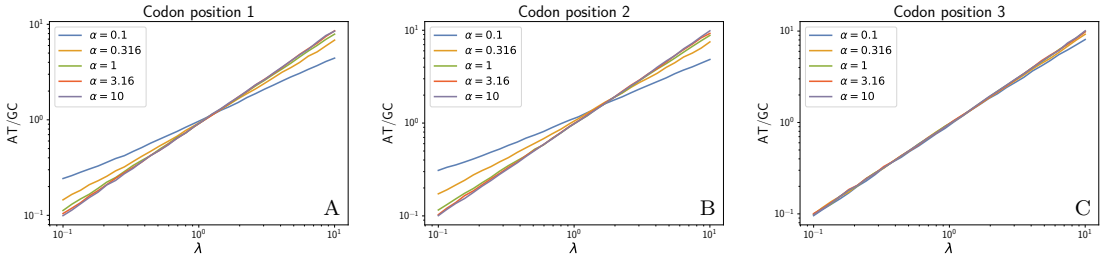


Figure 7.2: *Observed AT/GC composition of the alignment, represented at the different positions of codons (first, second and third), summed over all sites. The horizontal axis represents the underlying mutational bias (λ) of the nucleotide matrix, and the vertical axis represent the observed AT/GC of the codon position across the alignment. Stringency of selection is represented by 5 coloured solid lines with decreasing α. AT/GC at the third codon position (panel C) matches the mutational bias, whereas in contrast first and second positions (panel A and B) are less extreme than the underlying bias. With increase stringency of selection (i.e. with decreasing α), the observed bias is less strongly reflecting the underlying mutational bias, because selection is opposing the mutational bias.*

lated under different values of α and λ (figure 7.3). Under stringent selection, only a small number of amino acids are typically permissible any given site, resulting in a low site-specific diversity. Yet all amino acids occur at comparable frequencies in the alignment, resulting in a high sequence diversity. These observations highlight the distinction between averaged site-specific diversity and global sequence diversity. Of note, varying α has a strong impact on the site-specific diversity (figure 7.3), directly reflecting the fact that more stringent selection amounts to reducing the number of acceptable amino-acids at each site. On the other hand, it has a minor impact on sequence diversity, which merely reflects the fact that the strength of selection does not impact the global composition of the sequence.

Imposing a stronger mutational bias (either toward AT or toward GC) greatly reduces both site-specific and sequence diversity. This shows that the composition in amino acids is highly dependent on the underlying mutational bias, but also, that a more extreme mutational bias results in a more constrained substitution process: in effect, under a strong mutational bias, only those amino-acids that have both a high fitness and codons enriched in the nucleotides favored by the mutational process are eventually observed. This effect is less visible whenever selection is more stringent (i.e. with decreasing α), but can still be observed even for stringent selection.

The observed diversity is the result of a mix between mutation and selection. An alternative statistic, more directly relevant for measuring the intrinsic effect of selection, is the mean scaled fixation probability of non-synonymous mutations (ν). This summary statistic ν can be quantified from the substitutions recorded along the simulation trajectory (see section 7.4.5). For very long trajectories, it identifies with the ratio of non-synonymous over synonymous substitution rates (or d_N/d_S) induced by the underlying mutation-selection model (Spielman and Wilke, 2015; Dos Reis, 2015; Jones *et al.*, 2016). As expected, ν is always lower than 1 for simulations at equilibrium, under a time-independent fitness landscape (Spielman and Wilke, 2015). Quite expectedly (figure 7.4, panel A) ν depends strongly on the stringency of selection (α), which it is meant

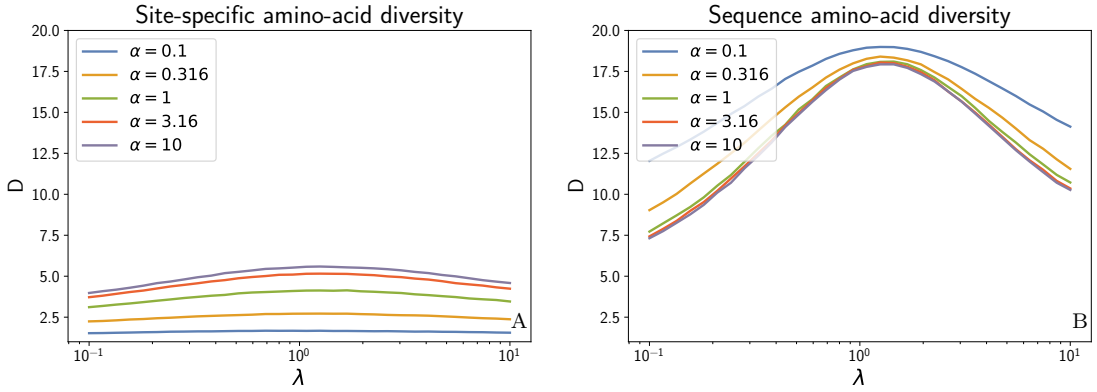


Figure 7.3: *Diversity of the amino-acid frequencies is quantified as the exponential of Shannon's entropy in the vertical axis, either as site-specific diversity in the panel A or as sequence diversity in the panel B. Sequence diversity is higher than site-specific diversity, because at any given site only a small number of amino acids are actually permissible. From a selective perspective, site-specific diversity decreases with stringency of selection (decreasing α represented by 5 different solid lines) because at given site only a few amino acids are permitted. Conversely, because site-specific fitness profiles are randomly drawn, each site has different permitted amino acid, increasing the sequence diversity as the stringency of selection increases. From a mutational perspective, diversity decreases with increased mutational bias toward either toward AT or GC (λ in horizontal axis). This effect is explained by the high frequency of amino acids containing nucleotides favoured by the underlying mutational bias. Finally, under stringent selection, diversity is less sensitive to the underlying mutational bias.*

to measure. On the other hand, ν depends weakly on the mutational bias (λ). This is in stark contrast with the amino-acid diversity, which is dependent on λ (figure 7.3).

The proxy of selection represented by ν concerns all non-synonymous mutations, but we can also consider the mean scaled fixation probability only for the subset of non-synonymous mutations from weak nucleotides (A or T) to strong nucleotides (G or C), called $\nu_{\mathrm{AT}\to\mathrm{GC}}$. Interestingly, $\nu_{\mathrm{AT}\to\mathrm{GC}}$ increases with the strength of the mutational bias toward AT (i.e. with increasing λ, figure 7.4, panel B). This distortion of the selective effects toward GC is stronger under an increased stringency of selection (i.e. under a lower α). Likewise, the non-synonymous mutations could also be restricted from strong (GC) to weak nucleotides (AT). This ratio decreases with the strength of the mutational bias toward AT (not shown). As a result, the ratio ratio between $\nu_{\mathrm{AT}\to\mathrm{GC}}$ and $\nu_{\mathrm{AT}\to\mathrm{GC}}$ is higher than 1 under an mutational bias toward AT (and lower than 1 respectively for a bias toward GC). It is monotonously increasing with the mutational bias toward AT (i.e. with increasing λ, figure 7.4, panel C). Altogether, fixation probabilities are opposed to mutational bias, and the realized equilibrium frequencies are thus at an equilibrium point between these two opposing forces.

7.2.2 Parameter inference on simulated data

From an alignment of protein-coding DNA sequences, without knowing the specific history of substitutions, can one estimate the mutational bias (λ) and the mean scaled fixation probability (ν)? In other words, can we tease apart mutation and selection?

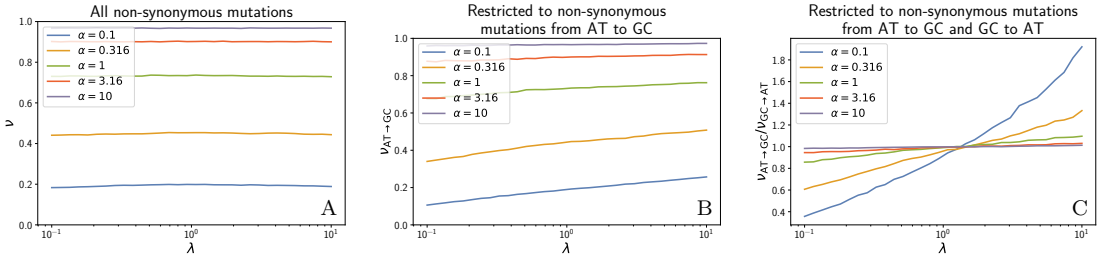


Figure 7.4: *Mean scaled fixation probability (ν) in vertical axis as a function of mutational bias (λ) in the horizontal axis, for different stringency of selection (α) in coloured solid lines. In panel A, expectedly, ν decrease with increased strength of selection (i.e. with decreasing α). However, ν is relatively unaffected by the mutational bias (λ). In panel B, ν is restricted to mutations from weak nucleotides (AT) to strong nucleotides (GC), called $\nu_{AT \to GC}$, represented in the vertical axis. A mutational process biased towards AT leads to an increased fixation probability toward GG, in the opposite direction. In panel C, $\nu_{AT \to GC}$ is divided by the fixation probabilities in the opposing direction $\nu_{GC \to AT}$, represented in the vertical axis and increasing monotonously with λ. Altogether, mutational bias is balanced by selection in the opposite direction, where this effect increases with the stringency of selection.*

To address this question, here we consider two codon models for inference, differing only by their parametrization of the codon matrix $\boldsymbol{Q}$. Both are homogeneous along the sequence (i.e. not site-specific). The first is based on Muse and Gaut (1994) formalism and uses a scalar ω parameter, while the second is based on a tensor representation of ω.

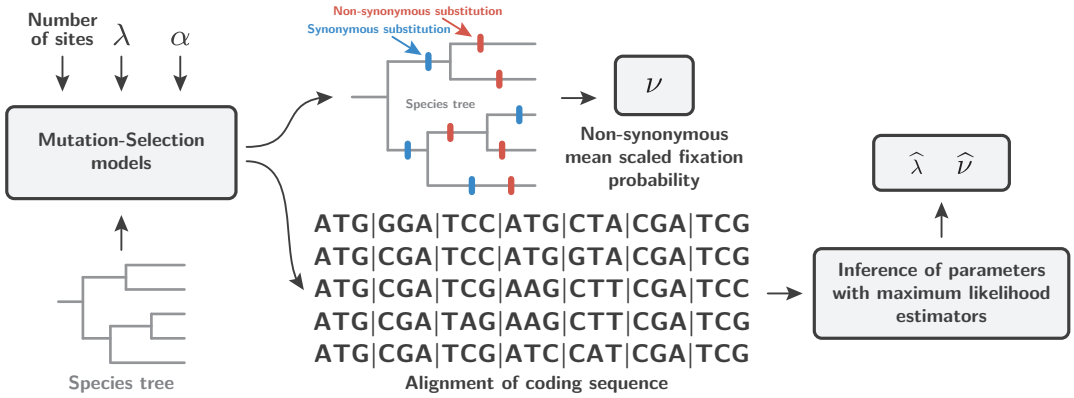


Figure 7.5: *Inferred value ($\widehat{\lambda}$, $\widehat{\nu}$) compared to underlying value (λ, ν) of the simulation. The different parameterization of the inference model can result in different estimates of mutational bias ($\widehat{\lambda}$) and mean scaled fixation probability ($\widehat{\nu}$). The main goal is to derive a model of inference that can reliably estimate these parameters. Two models of inference are proposed, the first is based on Muse & Gaut formalism, and the second based on a tensor of mean scaled fixation probabilities.*

ω as a scalar: the Muse & Gaut formalism

This model is defined in terms of a generalized time-reversible nucleotide rate matrix $\boldsymbol{R}$ and a scalar parameter ω. The matrix $\boldsymbol{R}$ is a function of the nucleotide frequencies $\boldsymbol{\sigma}$

and the symmetric exchangeability rates $\boldsymbol{\rho}$ (Tavaré, 1986):

$$R_{a,b} = \rho_{a,b}\sigma_b \tag{7.1}$$

At the level of codons, the substitution rate between the source (i) and target codon (j) depends on the underlying nucleotide change between ($\mathcal{M}(i,j)$) and whether or not the change is non-synonymous (e.g. $\mathcal{M}(AAT, AAG) = TG$). Altogether, the substitution rates between codons $Q_{i,j}$, formalized by Muse and Gaut (1994) are a function of the mutation matrix $\boldsymbol{R}(\boldsymbol{\sigma}, \boldsymbol{\rho})$, a single parameter of selective strength ω, and the genetic code as:

$$\begin{cases} Q_{i,j} & = 0 \text{ if codons } i \text{ and } j \text{ are more than one mutation away,} \\ Q_{i,j} & = R_{\mathcal{M}(i,j)} \text{ if codons } i \text{ and } j \text{ are synonymous,} \\ Q_{i,j} & = \omega R_{\mathcal{M}(i,j)} \text{ if codons } i \text{ and } j \text{ are non-synonymous.} \end{cases} \tag{7.2}$$

The model can be fitted by maximum likelihood. Then, from the estimate of $\widehat{\boldsymbol{R}}$, one can derive a nucleotide bias toward AT as:

$$\widehat{\lambda}_{\text{MG}} = (\widehat{\sigma_A} + \widehat{\sigma_T})/(\widehat{\sigma_G} + \widehat{\sigma_C}). \tag{7.3}$$

As for the mean strength of selection $\widehat{\nu}_{\text{MG}}$, a direct estimate is given by $\widehat{\omega}$.

As shown in the left panel of figure 7.6, estimate of the mutational bias is halfway between the nucleotide bias observed in the alignment and the true mutational bias used during the simulation. Thus, the MG model cannot reliably infer the mutational bias. On the other hand, $\widehat{\omega}$ is close to the underlying mean scaled fixation probability ν computed during the simulation, with a precision of 98.2% (not shown). Thus, the failure to correctly estimate the mutation process does not seem to have a strong impact on the inference of selection, at least in the present case.

ω as a tensor: mean-field derivation

We would like to derive a codon model that would be more accurate than the Muse & Gaut model, but that would still be site-homogeneous. However, the true process is site-specific. The link between the two can be formalized by projecting the site-specific processes onto a gene-wise process, using what can be seen as a mean-field approximation. The gene-wise process obtained by this procedure is expressed in terms of mutation rates and mean scaled fixation probabilities. Finally, the mean scaled fixation probabilities can be identified with the ω-tensor.

Specifically, at each site z, the true codon process is:

$$\begin{cases} Q^{(\text{z})}_{i,j} & = 0 \text{ if codons } i \text{ and } j \text{ are more than one mutation away,} \\ Q^{(\text{z})}_{i,j} & = R_{\mathcal{M}(i,j)} \text{ if codons } i \text{ and } j \text{ are synonymous,} \\ Q^{(\text{z})}_{i,j} & = R_{\mathcal{M}(i,j)} 2N_{\text{e}}\mathbb{P}^{(\text{z})}_{\text{fix}}(i,j) \text{ if codons } i \text{ and } j \text{ are non-synonymous.} \end{cases} \tag{7.4}$$

Where $2N_{\text{e}}\mathbb{P}^{(\text{z})}_{\text{fix}}(i,j)$ is the scaled fixation probability of codon j against codon i, at site z.

At equilibrium of the process, averaging over sites gives the mean-field gene-level process:

$$\begin{cases} \langle Q_{i,j} \rangle & = 0 \text{ if codons } i \text{ and } j \text{ are more than one mutation away,} \\ \langle Q_{i,j} \rangle & = R_{\mathcal{M}(i,j)} \text{ if codons } i \text{ and } j \text{ are synonymous,} \\ \langle Q_{i,j} \rangle & = R_{\mathcal{M}(i,j)} \langle 2N_e \mathbb{P}_{\text{fix}}(i,j) \rangle \text{ if codons } i \text{ and } j \text{ are non-synonymous.} \end{cases} \tag{7.5}$$

However, because selection between codons reduces to selection between pairs of amino-acids, $\langle 2N_e \mathbb{P}_{\text{fix}}(i,j) \rangle$ only depends on the amino-acids encoded by i and j (section 7.4.6 in methods). Thus, by identification, the inference model should be parameterized by a set of ω values for all pairs of amino acids, denoted $\omega_{x,y}$. For 20 amino acids, the total number of pairs of amino acids is 190, hence 380 parameters by counting in both directions. However, because of the structure of the genetic code, there are 75 pairs that are one nucleotide away, since some amino acids are not directly accessible through a single non-synonymous mutation. As a result, the number of parameters necessary to determine all non-zero entries of the tenser ($\omega_{x,y}$) in both directions is 150. Finally, under the assumption of a reversible process, the number of parameters can be reduced to 75 symmetric exchangeabilities ($\beta_{x,y}$) and 20 stationary effects (ϵ_x):

$$\omega_{x,y} = \epsilon_y \beta_{x,y}, \text{ where } \beta_{x,y} = \beta_{y,x}. \tag{7.6}$$

Altogether, the substitution rates between codons $Q_{i,j}$ are defined as:

$$\begin{cases} Q_{i,j} & = 0 \text{ if codons } i \text{ and } j \text{ are non neighbors,} \\ Q_{i,j} & = R_{\mathcal{M}(i,j)} \text{ if codons } i \text{ and } j \text{ are synonymous,,} \\ Q_{i,j} & = R_{\mathcal{M}(i,j)} \omega_{\mathcal{A}(i),\mathcal{A}(j)} \text{ if codons } i \text{ and } j \text{ are non-synonymous,} \end{cases} \tag{7.7}$$

where $\mathcal{A}(i)$ is the amino acid encoded by codon i and $\omega_{x,y}$ is given by equation 7.6.

This mean-field (MF) model is fitted by maximum likelihood, giving an estimate for its parameters, $\widehat{\boldsymbol{R}}$, $\widehat{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{\epsilon}}$. Then, from the estimate of the GTR nucleotide matrix ($\widehat{\boldsymbol{R}}$), a mutation bias $\widehat{\lambda}_{\text{MF}}$ can be estimated as previously (equation 7.3 above).

As shown in the right panel of figure 7.6, $\widehat{\lambda}_{\text{MF}}$ under the MF model provides an accurate estimate of the true mutational. In other words, the MF model can tease out the observed AT/GC bias of the alignment and the underlying mutational bias.

The mean scaled fixation probability of non-synonymous mutations $\widehat{\nu}_{\text{MF}}$ can also be computed. It is now a compound parameter, expressed as a function of $\widehat{\boldsymbol{R}}$, $\widehat{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{\epsilon}}$ (see section 7.4.7). Under this model, $\widehat{\nu}_{\text{MF}}$ is close to the true mean scaled fixation probability ν computed during the simulation, with a precision of 97.0% (not shown).

7.2.3 Estimation of empirical sequence data

The two alternative models of inference just considered, namely the classical Muse & Gaut (MG) and the mean-field (MF) codon models, were then applied to empirical protein-coding sequence alignments. Two examples were analysed: the nucleoprotein in *Influenza Virus* assembled in Bloom (2017), and the β-lactamase in *bacteria* gathered in Bloom (2014b), as shown in table 7.1.

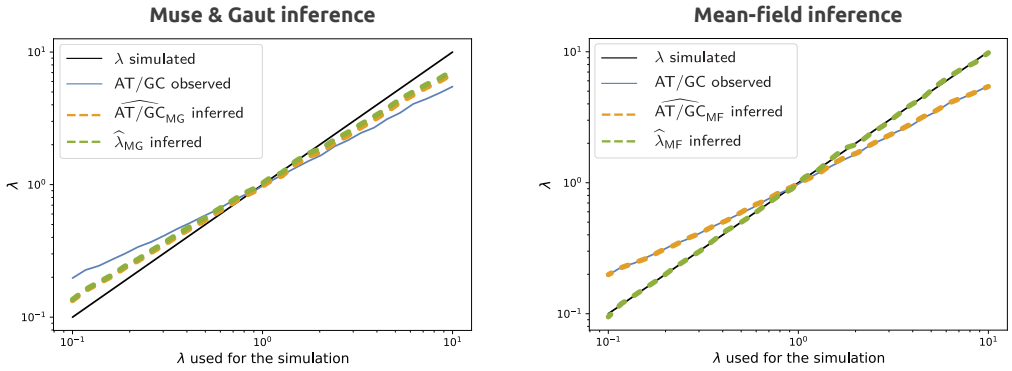


Figure 7.6: *Different estimates of the mutational bias in the vertical axis are represented as a function of the underlying true mutational bias (λ) of the simulation in the horizontal axis. Mutational bias can be estimated directly from the observed nucleotide frequencies in the alignment (AT/GC in blue solid line), similarly to figure 7.2, which is skewed by selection and always less extreme than the underlying mutational bias. The robustness of mutational bias estimation of two different inference models are shown. Selection is modelled as a single ω parameter in the Muse & Gaut formalism (MG) in the left panel, while selection is modelled as a tensor of ω parameters in different directions using a mean-field (MF) approximation in the right panel. In both panels, the true value of the mutational bias is represented in black solid line. The estimated mutational bias $\widehat{\lambda}_{MG}$ (in yellow dotted line) in the MG formalism is between the true value and the observed AT/GC. Conversely, estimated mutational bias $\widehat{\lambda}_{MG}$ (in yellow dotted line) in the MF approximation equal to the underlying value. Moreover, the expected $\widehat{AT/GC}$ from the parameter of the model fits the observed value in the MF approximation, while being skewed in the MG formalism. Altogether, the MF approximation can tease apart mutation and selection, while the MG formalism has to reach a compromise between observed AT/GC and underlying mutational bias.*

The nucleoprotein alignment of 498 amino acids is available for 180 virus strains (as human host), and is globally biased toward AT. Similarly to what was observed in the simulation experiments presented above, the mutational bias estimates under the two codon models are greater than the observed nucleotide bias (i.e. $1 < \text{AT/GC} < \widehat{\lambda}$). This effect is, as previously, probably due to selection at the level of amino acids, partially opposing the mutational bias. More importantly, the mutational bias estimated by the MF model is more extreme than the MG estimate (i.e. $1 < \widehat{\lambda}_{\text{MG}} < \widehat{\lambda}_{\text{MF}}$). This example behaves identically to the observations made with simulated alignments, where, compared to MG, the MF model estimates a stronger mutational bias, which was also closer to the real value. Thus, a reasonable interpretation is that MG is also underestimating the underlying mutational bias in the present case, and that the estimate of the MF model is more accurate.

Concerning selection, the estimated mean scaled fixation probability of non-synonymous mutations, denoted $\widehat{\nu}$, is similarly estimated in the MG and MF models ($\widehat{\nu}_{\text{MG}} \simeq \widehat{\nu}_{\text{MF}}$). Additionally, in the MF model, $\widehat{\nu}_{\text{MF}}$ can be restricted to mutations from weak nucleotides (AT) to strong (GC), or vice versa (see section 7.4.7). We observe that under a mutational bias favouring AT (i.e. $\lambda > 1$), the mean fixation probability of non-synonymous mutations is higher toward GC than toward AT (i.e. $\widehat{\nu}_{MF,\text{AT}\to\text{GC}} > \widehat{\nu}_{MF,\text{GC}\to\text{AT}}$), as

expected if the mutational bias is toward AT.

Reciprocally, for the β-lactamase, the alignment of 263 amino acids available for 85 species is globally biased toward GC. Expectedly, the mean-field estimate is even more strongly biased toward GC (i.e. $\widehat{\lambda}_{\mathrm{MF}} < \mathrm{AT/GC} < 1$). Curiously, the MG model estimates a weaker underlying mutational bias than the observed bias (i.e. $\mathrm{AT/GC} < \widehat{\lambda}_{\mathrm{MG}} < 1$). Concerning selection, we observe that the fixation probability of non-synonymous mutations is higher on average toward AT than toward GC (i.e. $\widehat{\nu}_{MF,\mathrm{GC}\to\mathrm{AT}} > \widehat{\nu}_{MF,\mathrm{AT}\to\mathrm{GC}}$), again, as expected under a GC-biased mutation process.

Altogether, the results obtained on empirical data are globally in agreement with the observations gathered from the simulation experiments, namely that the presence of a mutational bias results in a selection differential, taking the form of a slightly higher mean fixation probability of non-synonymous mutations opposing the mutational bias. Our MF model detects this effect and simultaneously estimates more extreme (and probably more accurate) mutational biases compared to the MG model.

	Nucleoprotein	**Lactamase**
Number of sites	498	263
Number of taxa	180	85
AT/GC of the alignment	1.15	0.79
AT/GC at first coding position	1.06	0.58
AT/GC at second coding position	1.22	1.18
AT/GC at third coding position	1.19	0.71
Muse & Gaut mutational bias ($\widehat{\lambda}_{\mathbf{MG}}$)	1.39	0.85
Mean-field mutational bias ($\widehat{\lambda}_{\mathbf{MF}}$)	1.64	0.68
Site diversity	1.10	1.37
Muse & Gaut scaled fixation probability ($\widehat{\nu}_{\mathbf{MG}}$)	0.085	0.29
Mean-field scaled fixation probability ($\widehat{\nu}_{\mathbf{MF}}$)	0.086	0.30
Mean-field scaled fixation probability from AT to GC ($\widehat{\nu}_{\mathbf{MF},\mathbf{AT}\to\mathbf{GC}}$)	0.14	0.31
Mean-field scaled fixation probability from GC to AT ($\widehat{\nu}_{\mathbf{MF},\mathbf{GC}\to\mathbf{AT}}$)	0.10	0.44
$\widehat{\nu}_{\mathbf{MF},\mathbf{AT}\to\mathbf{GC}}/\widehat{\nu}_{\mathbf{MF},\mathbf{GC}\to\mathbf{AT}}$	1.36	0.71

Table 7.1: *Estimated parameters of mutational bias ($\widehat{\lambda}$) from two models of inference, namely classical Muse & Gaut (MG) and mean-field (MF). These models are applied to two distinct datasets of protein-coding DNA alignment, nucleoprotein in the left column and β-lactamase in the right column. By taking into account selection in multiple direction, MG models estimates a stronger mutational bias than the MG model. For the MG model the mean scaled fixation probability of non-synonymous mutations ($\widehat{\nu}_{MF}$) can be obtained either from weak (AT) to strong nucleotides (GC), or vice versa. The fixation probability of non-synonymous mutations is opposed to the underlying mutational bias, such that a skewed mutational process results in a skewed selection, justifying that they must be articulated together.*

7.3 Discussion

In protein-coding DNA sequences, the nucleic composition results from a subtle interplay between mutation at the nucleic level and selection at the protein level. As a result, the observed nucleotide bias in the alignment is different from the underlying mutational bias.

However, current parametric codon models are inherently misspecified and, for that reason, are unable to tease apart these opposing effects of mutation and selection correctly. As a result, they don't estimate the mutational process reliably.

In this work we sought to find the simplest parametric codon model able to correctly tease apart mutation rates on one hand, and net mean fixation probabilities on the other hand, and this, without having to explicitly model the underlying fitness landscape. In order to derive a codon model along those lines, our strategy is to first assume an underlying microscopic model of sequence evolution (here, a mutation-selection model based on a site-specific, time-independent fitness landscape). Then, we derive the gene-wise mean fixation probabilities between all pairs of codons, implied by the underlying microscopic process. Finally, we observe that this mean-field process should in fact invoke as many distinct ω parameters as there are pairs of amino acids that are nearest neighbours in the genetic code. There are reversibility conditions, reducing the dimensionality and allowing for a GTR-like parameterization of this tensor (95 parameters for selection).

Inferring parameters on simulated alignments, we show that the model derived using this mean-field argument correctly estimates the underlying mutational bias and selective pressure. Applied to empirical alignments, we also observe that there is a selection differential opposing the mutational bias.

This work first points to a fundamental property of natural genetic sequences, namely that they are not optimized but are the result of an equilibrium between forces. In the specific case highlighted in this work, mutational bias at the nucleotide-level results in suboptimal amino-acid being overrepresented in the sequence. This was pointed out previously (Singer and Hickey, 2000), although never directly formalized in the context of a phylogenetic codon model.

One important consequence of this tradeoff between mutation and selection at equilibrium is that the observed higher mean fixation probability toward GC is mimicking the effect of biased gene conversion toward GC (gBGC), although unlike gBGC, the phenomenon described here corresponds to a genuine selective effect. Although we did not explore the consequences of this at the level of intra-specific polymorphism, the selection differential uncovered here also implies that the distribution of fitness effects is not the same in the two directions, either toward AT or toward GC. Specifically, in the presence of an AT-biased mutation process, the non-synonymous GC polymorphisms are expected to segregate at higher frequencies, compared to non-synonymous AT polymorphisms.

These observations have some practical implications: for instance, experiments observing a fixation (or segregation) bias toward GC at the non-synonymous level must also rule out that this fixation bias is not a simple consequence of the mutation-selection balance. More generally, our observations and modelling principles offer a useful pre-

liminary basis to better understand how mutation and selection will work together with biased gene conversion (gBGC), and therefore will help better understand how gBGC will impact both nucleotide composition and d_N/d_S. It is worth mentioning that in our result, we focused on the fixation probability from AT to GC ($\nu_{\mathrm{AT}\to\mathrm{GC}}$) because of the relationship to gBGC. However, in practice, the same analysis and methods can be applied to any subset of nucleotides or codons.

Our mean-field parametric model uses gene-level parameters (in the form of a tensor) that is meant to capture the mean scaled fixation probabilities. This derivation, and its validation on simulated data, shows that, even though the underlying selective landscape is site-specific, a gene-level approximation can nonetheless accurately disentangles mutation and selection. As a result, this study demonstrates that phenomenological models derived out of mechanistic models are more compact (i.e. not site-specific), and in certain cases are sufficient to extract the relevant parameters.

The methodology proposed here for deriving inference models consists in proceeding in two steps, first assuming an underlying mechanistic model of sequence evolution, parameterized by variables that are derived from first principles (fitness landscape, mutations rates, ...). Subsequently, the phenomenological inference model is obtained by matching its parameters (here, the entries of the ω tensor) with the aggregate parameters derived from the application of the mean-field procedure to the mechanistic model. Altogether, we believe that the approach used here could be applied more generally: inference models can be phenomenological in practice, but should nonetheless be derived from an underlying mechanistic model, so as to correctly formalize the interplay between mutation, selection, drift and other evolutionary forces.

Finally, this work is still preliminary since the mean-field model should be tested against a more diverse range of empirical data, in terms of phylogenetic depth, strength of selection, and codon usage bias to assert the validity of our empirical results. Also, the empirical fit to the data between the models (e.g. using AIC) should be more carefully examined. Indeed, by setting $\boldsymbol{\epsilon} = \mathbf{1}$ and $\boldsymbol{\beta} = \omega \times \mathbf{1}$ in our mean-field model, we retrieve the nested Muse & Gaut model, hence, both models are directly comparable. In addition, several other codon models (Rodrigue *et al.*, 2008a; Kosakovsky Pond *et al.*, 2020) should be included in a broader comparison of the accuracy of the estimation of the underlying mutational bias and strength of selection on protein-coding DNA sequences.

7.4 Materials & Methods

7.4.1 Simulation model

We seek to simulate the evolution of protein-coding sequences along a specie tree. Starting with one sequence at the root of the tree, the sequences evolve independently along the different branches of the tree by point substitutions, until they reach the leaves. At the end of the simulation, we get one sequence for each leaf of the tree, meaning one sequence per species. The substitution is modelled using the origination-fixation approximation,

i.e. substitution rates are the product of the mutation rate at the nucleotide level, and fixation probabilities, based on selection at the amino-acid level.

The mutation process is assumed homogeneous across sites. On the other hand, selection is assumed to be varying along the sequence. During the simulation, given the current sequence, the substitution rates toward all possible mutants (one nucleotide change) are computed and the next substitution event is drawn randomly based on Gillespie's algorithm (Gillespie, 1977).

7.4.2 Mutational bias at the nucleotide level

The mutation rate between nucleotides is always proportional to μ. Moreover, mutations from any nucleotide to another weak nucleotide is increased by the factor λ compared with mutations to another strong nucleotide. The mutation rate matrix is thus:

$$\boldsymbol{R} = \begin{array}{c} \\ A \\ C \\ G \\ T \end{array} \begin{array}{c} \begin{array}{cccc} A & C & G & T \end{array} \\ \begin{pmatrix} -\mu(2+\lambda) & \mu & \mu & \mu\lambda \\ \mu\lambda & -\mu(1+2\lambda) & \mu & \mu\lambda \\ \mu\lambda & \mu & -\mu(1+2\lambda) & \mu\lambda \\ \mu\lambda & \mu & \mu & -\mu(2+\lambda) \end{pmatrix} \end{array} \tag{7.8}$$

Which has the following stationary distribution:

$$\boldsymbol{\sigma R} = \mathbf{1}, \tag{7.9}$$

$$\iff \boldsymbol{\sigma} = \left(\frac{\lambda}{2+2\lambda}, \frac{1}{2+2\lambda}, \frac{1}{2+2\lambda}, \frac{\lambda}{2+2\lambda} \right). \tag{7.10}$$

The process is reversible and fulfills the detailed balance condition, i.e. for any pair of distinct nucleotides:

$$\sigma_a R_{a,b} = \sigma_b R_{b,a}. \tag{7.11}$$

As a result, the ratio of weak over strong nucleotide frequencies at stationarity is equal to λ:

$$\frac{\sigma_A + \sigma_T}{\sigma_C + \sigma_G} = \frac{\lambda(2+2\lambda)^{-1} + \lambda(2+2\lambda)^{-1}}{(2+2\lambda)^{-1} + (2+2\lambda)^{-1}}, \text{ from eq. 7.10,} \tag{7.12}$$

$$= \lambda. \tag{7.13}$$

7.4.3 Selection at the amino-acid level

The substitution rate is considered null between any two codons differing by more than one nucleotide. Otherwise, the mutation rate between a pair of codons is given by the mutation rate of the underlying single nucleotide change. Selection is modelled at the amino-acid level, i.e. we assume that all codons encoding for one particular amino acid are selectively neutral.

To take into account the heterogeneity of selection between different sites of the protein, we assume that each site z of the sequence is independently evolving under

a site-specific fitness landscape, characterized by a 20-dimensional frequency vector of scaled (Wrightian) fitness parameters $\boldsymbol{\phi}^{(\mathrm{z})} = \{\phi_{\mathrm{a}}^{(\mathrm{z})},\ 1 \leq \mathrm{a} \leq 20\}$. The (Wrigthian) fitness vectors across sites are drawn IID from a uniform Dirichlet distribution prior to the simulation over the tree:

$$\boldsymbol{\phi}^{(\mathrm{z})} \sim \mathrm{Dirichlet}(\alpha, \ldots, \alpha),\ \mathrm{z} \in \{1, \ldots, \mathrm{Z}\}, \tag{7.14}$$

The malthusian fitness (or log-fitness) of amino acid a, denoted $F_{\mathrm{a}}^{(\mathrm{z})}$, is accordingly:

$$F_{\mathrm{a}}^{(\mathrm{z})} = \ln\left(\phi_{\mathrm{a}}^{(\mathrm{z})}\right),\ \mathrm{z} \in \{1, \ldots, \mathrm{Z}\},\ \mathrm{a} \in \{1, \ldots, 20\} \tag{7.15}$$

At site z, the substitution rate between non-synonymous codons i and j is given by the product of the mutation rate and the probability of fixation:

$$Q_{i,j}^{(\mathrm{z})} = R_{\mathcal{M}(i,j)} \frac{F_{\mathcal{A}(j)}^{(\mathrm{z})} - F_{\mathcal{A}(i)}^{(\mathrm{z})}}{1 - \mathrm{e}^{F_{\mathcal{A}(i)}^{(\mathrm{z})} - F_{\mathcal{A}(j)}^{(\mathrm{z})}}} \tag{7.16}$$

where $\mathcal{A}(i)$ denotes the amino-acid encoded by codon i. At the root of the tree, for each site z, the sequence is drawn from the stationary distribution of the process specified by $\boldsymbol{\pi}^{(\mathrm{z})}$, which is given by:

$$\pi_i^{(\mathrm{z})} = \mathcal{Z}^{(\mathrm{z})} \left[\prod_{k \in \{1,2,3\}} \sigma_{i[k]}\right] \mathrm{e}^{F_{\mathcal{A}(i)}^{(\mathrm{z})}}, \tag{7.17}$$

where $i[k]$ denotes the nucleotide at position $k \in \{1, 2, 3\}$ of codon i, and $\mathcal{Z}^{(\mathrm{z})}$ is the normalizing constant at site z:

$$\mathcal{Z}^{(\mathrm{z})} = \left(\sum_{j=1}^{61} \left[\prod_{k \in \{1,2,3\}} \sigma_{j[k]}\right] \mathrm{e}^{F_{\mathcal{A}(j)}^{(\mathrm{z})}}\right)^{-1} \tag{7.18}$$

The substitution process is reversible and fulfils detailed balance conditions at each site z and between each pair of codons $(i,\ j)$:

$$\pi_i^{(\mathrm{z})} Q_{i,j}^{(\mathrm{z})} = \pi_j^{(\mathrm{z})} Q_{j,i}^{(\mathrm{z})} \tag{7.19}$$

Of note, by modelling fitness at the amino-acid level, we assume that all codons encoding for one particular amino acid are selectively neutral. In addition, in this modelling framework, the genetic code is of particular importance since the number of codons encoding for a particular amino acid varies greatly. As an example, tryptophan is encoded by one codon, while leucine is encoded by 6 codons. Intuitively, this variation makes the mutation bias more pronounced among codons encoding for the same amino acid, since there are more mutations possible that are selectively neutral (i.e. synonymous). On the other hand, the mutation bias is more constrained if the amino acid is encoded by few codons.

7.4.4 Site and sequence diversity of amino-acids

For a site z, the diversity $D^{(\mathrm{z})}$ (or effective number of amino acids) is computed as the exponential of Shannon's entropy from the frequencies of the different amino-acids $\boldsymbol{\psi}^{(\mathrm{z})} = \{\psi_{\mathrm{a}}^{(\mathrm{z})},\ \mathrm{a} \in \{1,\ldots,20\},$ as:

$$D\left(\boldsymbol{\psi}^{(\mathrm{z})}\right) = \exp\left[-\sum_{\mathrm{a}=1}^{20} \psi_{\mathrm{a}}^{(\mathrm{z})} \ln\left(\psi_{\mathrm{a}}^{(\mathrm{z})}\right)\right] \tag{7.20}$$

The diversity is a measure of the sparsity of the amino-acid frequencies, with a value of 1 corresponding to the minimum diversity (i.e. with only one amino acid permissible), and a value of 20 corresponding to maximum diversity, where each amino acid has the same frequency. The diversity can be first computed for each site and then averaged over all sites as:

$$\langle D\left(\boldsymbol{\psi}\right)\rangle = \frac{1}{\mathrm{Z}}\sum_{\mathrm{z}=1}^{\mathrm{Z}} D\left(\boldsymbol{\psi}^{(\mathrm{z})}\right) \tag{7.21}$$

Alternatively, average frequencies of the different amino acids can first be computed over the alignment and then used to compute the global sequence diversity:

$$\langle\boldsymbol{\psi}\rangle = \frac{1}{\mathrm{Z}}\sum_{\mathrm{z}=1}^{\mathrm{Z}} \boldsymbol{\psi}^{(\mathrm{z})} \tag{7.22}$$

Then the sequence diversity is simply:

$$D\left(\langle\boldsymbol{\psi}\rangle\right) = \exp\left[-\sum_{\mathrm{a}=1}^{20} \langle\psi\rangle_{\mathrm{a}} \ln\left(\langle\psi\rangle_{\mathrm{a}}\right)\right] \tag{7.23}$$

7.4.5 Mean scaled fixation probability

The sequence at time t is denoted $\mathbb{S}(t)$ and the codon present at site z is denoted $\mathbb{S}_{\mathrm{z}}(t)$. For a given sequence, the mean scaled fixation probability over mutations away from $\mathbb{S}(t)$ (weighted by their probability of occurrence) is given by the ratio:

$$\nu(t) = \frac{\sum_{\mathrm{z}=1}^{\mathrm{Z}} \sum_{j\in\mathcal{N}(\mathbb{S}_{\mathrm{z}}(t))} Q_{\mathbb{S}_{\mathrm{z}}(t)\to j}}{\sum_{\mathrm{z}=1}^{\mathrm{Z}} \sum_{(j\in\mathbb{S}_{\mathrm{z}}(t))} \mu_{\mathbb{S}_{\mathrm{z}}(t)\to j}}, \tag{7.24}$$

where $\mathcal{N}(i)$ is the set of non-synonymous codons neighbours of codon i and $Q_{i,j}^{(\mathrm{z})}$ are defined as in equation 7.16. Averaged over all branches of the tree, the mean scaled fixation probability is :

$$\nu = \langle\nu(t)\rangle, \tag{7.25}$$

$$= \int_t \nu(t)\mathrm{d}t, \tag{7.26}$$

where the integral is taken over all branches of the tree, while the integrand $\nu(t)$ is a piecewise function changing after every point substitution event. The mean scaled fixation

probability from weak (AT) to strong (GC) nucleotides, denoted $\nu_{\mathrm{AT}\to\mathrm{GC}}$, is obtained similarly by restricting the sums (in the numerator and the denominator) from weak to strong mutations. A similar computation can be done from strong to weak.

7.4.6 Derivation of mean-field model

The mean-field codon model $\langle \boldsymbol{Q} \rangle$ is defined such that $\langle Q_{i,j} \rangle$ is the average rate of substitution to codon j, conditional on currently being on codon i, the average being taken across sites. Importantly, sites differ in their probability of being currently in state i. The average should therefore be weighted by this probability.

Assuming an underlying site-specific mutation-selection process at equilibrium, given we know that a mutation is from codon i, the probability that this mutation is occuring at site z is:

$$\mathbb{P}(\mathrm{z} \mid i) = \frac{\pi_i^{(\mathrm{z})}}{\sum\limits_{\mathrm{z}=1}^{\mathrm{Z}} \pi_i^{(\mathrm{z})}} \tag{7.27}$$

The site-averaged (mean-field) substitution rate from codon i to j is as result given as:

$$\langle Q_{i,j} \rangle = \sum_{\mathrm{z}=1}^{\mathrm{Z}} \mathbb{P}(\mathrm{z} \mid i) Q_{i,j} \tag{7.28}$$

If codon i and codon j are synonymous, this equation simplifies to the underlying mutation rate $R_{\mathcal{M}(i,j)}$. Otherwise, if codon i and codon j are non-synonymous, the mean-field substitution rate is:

$$\begin{aligned}
\langle Q_{i,j} \rangle &= \left\langle R_{\mathcal{M}(i,j)} 2N_{\mathrm{e}} \mathbb{P}_{\mathrm{fix}}(i,j) \right\rangle, && (7.29)\\
&= R_{\mathcal{M}(i,j)} \left\langle 2N_{\mathrm{e}} \mathbb{P}_{\mathrm{fix}}(i,j) \right\rangle, && (7.30)\\
&= R_{\mathcal{M}(i,j)} \frac{\sum\limits_{\mathrm{z}=1}^{\mathrm{Z}} \pi_i^{(\mathrm{z})} \dfrac{F^{(\mathrm{z})}_{\mathcal{A}(j)} - F^{(\mathrm{z})}_{\mathcal{A}(i)}}{1 - \mathrm{e}^{F^{(\mathrm{z})}_{\mathcal{A}(i)} - F^{(\mathrm{z})}_{\mathcal{A}(j)}}}}{\sum\limits_{\mathrm{z}=1}^{\mathrm{Z}} \pi_i^{(\mathrm{z})}}, && (7.31)\\
&= R_{\mathcal{M}(i,j)} \frac{\sum\limits_{\mathrm{z}=1}^{\mathrm{Z}} \mathcal{Z}^{(\mathrm{z})} \dfrac{F^{(\mathrm{z})}_{\mathcal{A}(j)} - F^{(\mathrm{z})}_{\mathcal{A}(i)}}{\mathrm{e}^{-F^{(\mathrm{z})}_{\mathcal{A}(i)}} - \mathrm{e}^{-F^{(\mathrm{z})}_{\mathcal{A}(j)}}}}{\sum\limits_{\mathrm{z}=1}^{\mathrm{Z}} \mathcal{Z}^{(\mathrm{z})} \mathrm{e}^{F^{(\mathrm{z})}_{\mathcal{A}(i)}}} && (7.32)
\end{aligned}$$

As a result, $\langle 2N_{\mathrm{e}} \mathbb{P}_{\mathrm{fix}}(i,j) \rangle$ is dependent on the source and target codon solely through the source amino acid ($\mathcal{A}(i)$) and target amino acid ($\mathcal{A}(j)$), hence the parameter $\omega_{\mathcal{A}(i),\mathcal{A}(j)}$ identifies with the average fixation probability:

$$\omega_{\mathcal{A}(i),\mathcal{A}(j)} = \frac{\sum\limits_{\mathrm{z}=1}^{\mathrm{Z}} \mathcal{Z}^{(\mathrm{z})} \dfrac{F^{(\mathrm{z})}_{\mathcal{A}(j)} - F^{(\mathrm{z})}_{\mathcal{A}(i)}}{\mathrm{e}^{-F^{(\mathrm{z})}_{\mathcal{A}(i)}} - \mathrm{e}^{-F^{(\mathrm{z})}_{\mathcal{A}(j)}}}}{\sum\limits_{\mathrm{z}=1}^{\mathrm{Z}} \mathcal{Z}^{(\mathrm{z})}}. \tag{7.33}$$

7.4.7 Mean scaled fixation probability (ν) under the mean-field model

The mean-field model is parameterized by a GTR mutation matrix $\boldsymbol{R}(\boldsymbol{\sigma}, \boldsymbol{\rho})$ and the selection coefficient $\boldsymbol{\omega}(\boldsymbol{\beta}, \boldsymbol{\epsilon})$. As a result, the mean scaled fixation probability of non-synonymous mutations is:

$$\nu_{\mathrm{MF}} = \frac{\sum_{i=1}^{61} \pi_i \sum_{j \in \mathcal{N}(i)} Q_{i,j}}{\sum_{i=1}^{61} \pi_i \sum_{j \in \mathcal{N}(i)} \mu_{i,j}}, \tag{7.34}$$

$$= \frac{\sum_{i=1}^{61} \left[\prod_{k \in \{1,2,3\}} \sigma_{i[k]} \right] \epsilon_{\mathcal{A}(i)} \sum_{j \in \mathcal{N}(i)} R_{\mathcal{M}(i,j)} \epsilon_{\mathcal{A}(j)} \beta_{\mathcal{A}(i),\mathcal{A}(j)}}{\sum_{i=1}^{61} \left[\prod_{k \in \{1,2,3\}} \sigma_{i[k]} \right] \epsilon_{\mathcal{A}(i)} \sum_{j \in \mathcal{N}(i)} R_{\mathcal{M}(i,j)}}, \tag{7.35}$$

where $i[k]$ denotes the nucleotide at position $k \in \{1, 2, 3\}$ of codon i.

Similarly, the mean scaled fixation probability from weak (AT) to strong (GC) nucleotides denoted $\nu_{MF,\mathrm{AT}\to\mathrm{GC}}$ is obtained similarly by restricting the sums (in the numerator and the denominator) to one nucleotide mutations only from weak to strong. Conversely, by restricting the sum from strong (GC) to weak (AT), we obtain $\nu_{MF,\mathrm{GC}\to\mathrm{AT}}$.

7.4.8 Inference method with `Hyphy`

Maximum likelihood estimation has been performed with the software `Hyphy` (Kosakovsky Pond and Muse, 2005a). The `Python` scripts generating the `Hyphy` batch files (for both Muse & Gaut and mean-field), as well as scripts for the post-analysis of the experiments, are available at `https://github.com/ThibaultLatrille/NucleotideBias` under MIT license.

7.5 Author contributions

TL developed the simulator `SimuEvol` and conducted all analyses, in the context of a PhD work (Ecole Normale Superieure de Lyon). TL and NL both contributed to the writing of the manuscript.

7.6 Acknowledgements

We gratefully acknowledge the help of Laurent Gueguen, Benoit Nahbolz and Laurent Duret for their input and comments on this work.

8

Inferring long-term effective population size with Mutation-Selection models

Thibault Latrille[1, 2], Vincent Lanore[1], Nicolas Lartillot[1]

[1]Université de Lyon, Université Lyon 1, UMR CNRS 5558 Laboratoire de Biométrie et Biologie Évolutive, 69622 Villeurbanne, France
[2]École Normale Supérieure de Lyon, Universite de Lyon, 69007 Lyon, France

Contents

8.1 Introduction

Since the realization, by Zuckerkandl and Pauling (1965) that genetic sequences are informative about the evolutionary history of the species, molecular phylogenetics has developed into a mature and very active field. A broad array of models and inference methods have been developed, using DNA sequences for reconstructing the phylogenetic relationships among species (Felsenstein, 1981), for estimating divergence times (Thorne and Kishino, 2002), or for reconstructing the genetic sequences of remote ancestors (Liberles, 2007). However, genetic sequences might contain information about other aspects of the evolutionary history and, in particular, about past population-genetic regimes.

Interspecific divergence is the long-term outcome of population-genetic processes, in which point mutations at the level of individuals are then subjected to selection and genetic drift, leading to substitutions at the level of the population. As a result, the substitution patterns that can be reconstructed along phylogenies are modulated by the underlying population-genetic parameters (mutation biases, selective landscapes, effective population size), suggesting the possibility to infer the past variation of these parameters over the phylogeny. Independently, ecological properties such as phenotypic characters or life-history traits can be observed in extinct or in present-day species. Using the comparative method (Felsenstein, 1985), these traits can be reconstructed for the unobserved ancestral species. Combined together, genetic and phenotypic ancestral reconstructions can then be used to unravel the interplay between evolutionary and ecological mechanisms.

Practically, in order to disentangle mutation, selection and genetic drift, we need to classify individual substitutions into different categories, differing in the strength of mutation, selection or genetic drift. In protein-coding DNA sequences, the mutational process occurs at the nucleotide level. Assuming that synonymous mutations are selectively neutral and that selection mostly acts at the protein level, synonymous substitutions can be used to infer the patterns of mutation, without any interference contributed by selection. Then, by comparing the non-synonymous substitution rate relative to the synonymous substitution rate (the ratio d_N/d_S), one can estimate the global strength of selection acting on proteins. This idea was formalized using phylogenetic codon models (Muse and Gaut, 1994; Goldman and Yang, 1994). This led to a broad range of applications, either to detect proteins under adaptive selection (Kosiol *et al.*, 2008), or to measure the modulations of the strength of purifying selection between sites (Echave *et al.*, 2016), genes (Zhang and Yang, 2015), or lineages (Lartillot and Poujol, 2011).

Concerning variation in d_N/d_S between lineages, and in a context mostly characterized by purifying selection, the nearly-neutral theory predicts that changes in the global strength of selection (measured as d_N/d_S) is related to changes in the relative strength of genetic drift, which is in turn mediated by changes in effective population size (N_e) (Ohta, 1992). Mechanistically, populations with high N_e are characterized by more efficient purifying selection against mildly deleterious mutations, resulting in lower d_N/d_S (Kimura, 1979; Welch *et al.*, 2008).

Codon models have been used to empirically measure such changes in the efficacy

of purifying selection along phylogenies, either by allowing for different d_N/d_S values in different parts of the tree (Dutheil *et al.*, 2012), or by estimating d_N/d_S independently for every branch of the tree (Popadin *et al.*, 2007). Alternatively, d_N/d_S can be modelled as a continuous trait, varying along the phylogeny as a stochastic process, splitting at each node of the tree into independent processes (Seo *et al.*, 2004). Once empirical estimates of the variation in d_N/d_S between lineages or groups has been obtained, these can be compared to changes in N_e across lineages, so as to test the validity of the predictions of the nearly-neutral theory. Independent empirical estimation of N_e is usually done vie proxies, such as the neutral diversity within species (Galtier, 2016), or life-history traits. For instance, animal species characterized by a large body size or an extended longevity are typically expected to also have a low N_e (Romiguier *et al.*, 2014). Alternatively, a Bayesian integrative framework has been proposed (Lartillot and Poujol, 2011), extending the approach of Seo *et al.* (2004), in which the joint variation in d_S, d_N/d_S and in life-history traits or other proxies of N_e is modelled as a multivariate Brownian process, with a variance-covariance matrix capturing the signal of their correlated evolution.

Analyses using these approaches and these proxies of N_e have suggested a negative correlation between d_N/d_S and N_e (Popadin *et al.*, 2007; Lanfear *et al.*, 2010a; Lartillot and Poujol, 2011; Lartillot and Delsuc, 2012; Romiguier *et al.*, 2014; Figuet *et al.*, 2017), thus confirming the theoretical prediction of the nearly-neutral theory. However, the universality and robustness of the correlation between d_N/d_S and N_e is still debated (Nabholz *et al.*, 2013; Lanfear *et al.*, 2014; Figuet *et al.*, 2016; Bolívar *et al.*, 2019), and further investigation might be required. Moreover, these analyses do not explicitly formalize the quantitative relationship between N_e and d_N/d_S. This relation is in principle dependent on the underlying fitness landscape (Welch *et al.*, 2008; Cherry, 1998; Goldstein, 2011), and can show complicated behavior due to non-equilibrium properties (Jones *et al.*, 2016). These questions could be addressed in the context of a mechanistic modelling approach.

As an alternative to classical d_N/d_S-based codon models, mechanistic codon models explicitly introduce population genetic equations into the codon substitution process (Halpern and Bruno, 1998). Specifically, these so-called mutation-selection codon models explicitly assign a fitness parameter to each amino acid. As a result, the substitution rate between each pair of codons can be predicted, as the product of the mutation rate and the fixation probability of the new codon, which is in turn dependent on the fitness of the initial and the final codons. Since the strength of selection is typically not homogeneous along the protein sequence, and depends on the local physicochemical requirements (Echave *et al.*, 2016; Goldstein and Pollock, 2016, 2017), local changes in selective strength are usually taken into account by allowing for site-specific amino-acid fitness profiles. Site-specific amino-acid preferences are typically estimated either by penalized maximum likelihood (Tamuri and Goldstein, 2012; Tamuri *et al.*, 2014), or in a Bayesian context, using an infinite mixture based on a Dirichlet process prior (Rodrigue *et al.*, 2010; Rodrigue and Lartillot, 2014). This second approach is further considered below.

Although not directly expressed in terms of this variable, the mutation-selection formalism induces an equilibrium d_N/d_S, which is theoretically lower than 1, thus explicitly

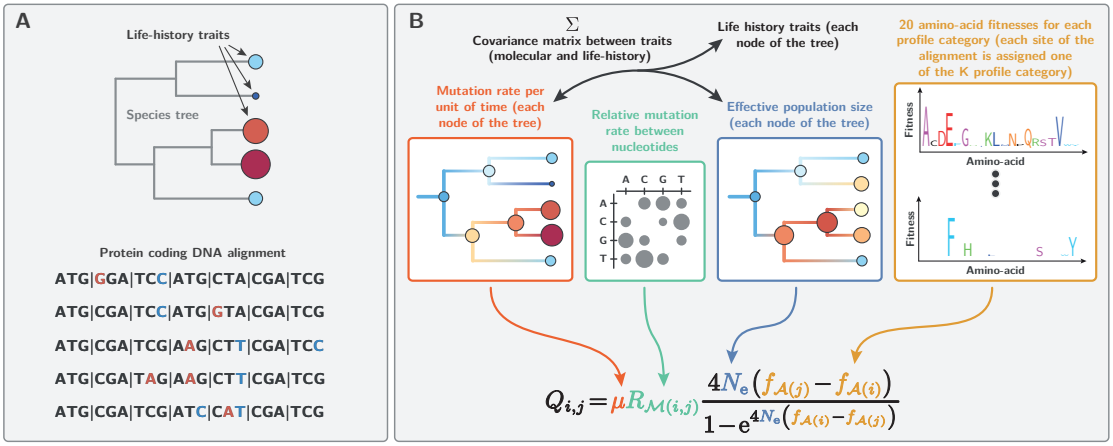


Figure 8.1: *Model summary. Panel A. Our method requires a (given) rooted tree topology, an alignment of protein-coding DNA and (optionally) quantitative life-history trait for the extant species. Panel B. Relying on a codon model based on the mutation-selection formalism, assuming an auto-correlated log-Brownian process for the variation through time in effective population size (N_e), mutation rate (μ) and life-history traits, our Bayesian inference method estimates amino-acid fitness profiles across sites, variation in mutation rate and effective population size along the tree, as well as the node ages and the nucleotide mutation rates.*

modelling purifying selection (Spielman and Wilke, 2015; Dos Reis, 2015). As a result, the mutation-selection codon framework proved to be a valuable null (nearly-neutral) model, against which to compare the observed d_N/d_S by classical codon models, so as to test for the presence of adaptation (Rodrigue and Lartillot, 2016; Bloom, 2017).

However, these mutation-selection methods have so far assumed the strength of genetic drift, or equivalently N_e, to be constant across the phylogeny. This assumption is clearly not realistic, as attested by the empirically measured variation in d_N/d_S between lineages using classical codon models or, more directly, by the broad range of synonymous neutral diversity observed across species (Galtier, 2016). The impact of this assumption on the estimation of the fitness landscape across sites (Tamuri *et al.*, 2014; Rodrigue and Lartillot, 2014), or on the tests for the presence of adaptation (Rodrigue and Lartillot, 2016; Bloom, 2017) is totally unknown. Relaxing this assumption of a constant N_e is thus necessary.

Conversely, since the mutation-selection formalism explicitly incorporates N_e as a parameter of the model, extending the model so as to let N_e vary across lineages is relatively straightforward, at least conceptually. Doing this would then provide an occasion to address several important questions: do we have enough signal in empirical sequence alignments, to estimate the evolutionary history of N_e along a phylogeny? Can we more generally revisit the question of the empirical correlations between N_e and ecological life-history traits (longevity, maturity, weight, size, ...), previously explored using classical d_N/d_S based models, but now in the context of this mechanistic framework?

8.2 New approaches

To address these questions, here we introduce a variant of the mutation-selection codon model, in which selection is modulated along the sequence (using site-specific amino-acid profiles), while the mutation rate (μ), the effective population size (N_e) and life-history traits are allowed to vary along the phylogeny (figure 8.1). Methodologically, our model is fundamentally an integration between the Bayesian non-parametric version of the Halpern and Bruno (1998) mutation-selection model (Rodrigue and Lartillot, 2014), and the molecular comparative framework modelling the joint evolution of life-history and molecular traits (Lartillot and Poujol, 2011).

Formally, the substitution rate (per unit of time) from codon i to j, denoted $Q_{i,j}$, is equal to the total rate of mutation (per unit of time) at the level of the population ($2N_e\mu_{i,j}$) multiplied by the probability of fixation of the mutation $\mathbb{P}_{\text{fix}}(i,j)$:

$$Q_{i,j} = 2N_e\mu_{i,j}\mathbb{P}_{\text{fix}}(i,j) \tag{8.1}$$

In the case of synonymous mutations, which we assumed are neutral, the probability of fixation is independent of the original and target codon, and equals $1/2N_e$, such that $Q_{i,j}$ simplifies to:

$$Q_{i,j} = \mu_{i,j} \tag{8.2}$$

In the case of non-synonymous mutations, the probability of fixation depends on the difference in fitness between the amino acid encoded by the initial and final codons:

$$Q_{i,j} = \mu_{i,j}\frac{4N_e\left(f_{\mathcal{A}(j)} - f_{\mathcal{A}(i)}\right)}{1 - e^{4N_e\left(f_{\mathcal{A}(i)} - f_{\mathcal{A}(j)}\right)}} \tag{8.3}$$

where $\boldsymbol{f}$ is a 20-dimensional vector specifying the log-fitness for each amino acid, and $\mathcal{A}(i)$ is the amino acid encoded by codon i.

In the model introduced here, N_e and μ are allowed to vary between species (across branches) as a multivariate log-Brownian process, but are assumed constant along the DNA sequence. Conversely, amino-acid fitness profiles $\boldsymbol{f}$ are considered constant along the tree but are assumed to vary across sites, being modelled as independent and identically distributed random-effects from an unknown distribution estimated using a Dirichlet process prior.

This model was implemented in a Markov chain Monte Carlo framework, allowing for joint inference of site-specific selection profiles and reconstruction of life-history traits and population-genetic regimes along the phylogeny. After validating our model and our inference framework against simulated data, we apply it to several cases of interest across metazoans (placental mammals, primates and isopods), for which some proxies of N_e are available.

8.3 Results

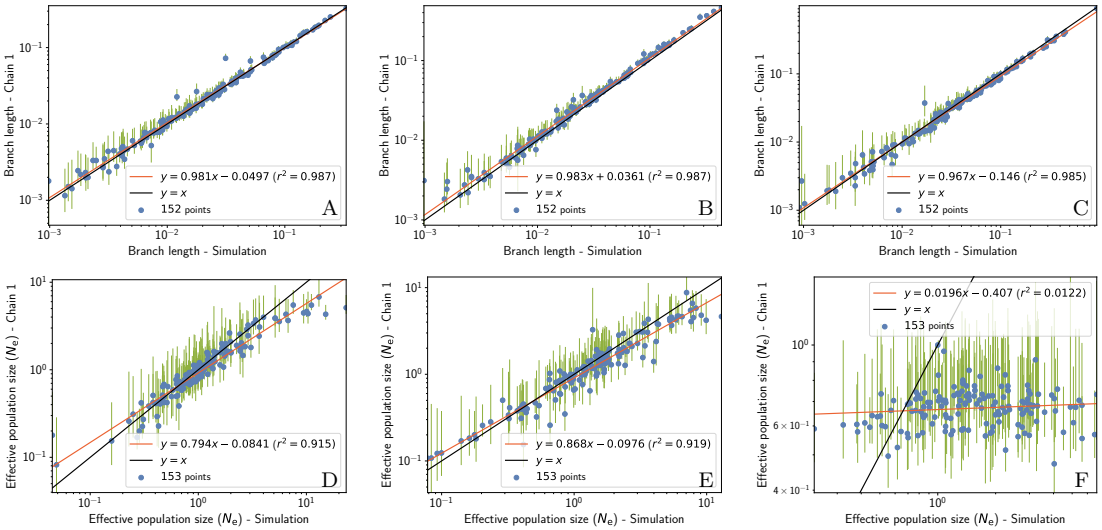


Figure 8.2: *A-C: branch lengths in expected number of substitutions per site. D-F: N_e values across nodes (including the leaves) relative to N_e at the root. From left to right: simulation under the mutation-selection approximation (A,D), under a Wright-Fisher model accounting for small population size effects (5000 individuals at the root), site linkage and short term fluctuation of N_e (B,E) and accounting for site epistasis in the context of selection for protein stability. The tree root is* 150 *million years old, where the initial population start with a mutation rate of* $1e^{-8}$ *per site per generation, and generation time of* 10 *years. These experiments confirm that signal in the placental mammalian tree can allow to reliably infer the direction of change in N_e, even if linkage disequilibrium, short term fluctuation of N_e and finite population size effects are not accounted for in the inference framework. However, the presence of epistasis between sites is a serious threat to the inference of N_e.*

8.3.1 Validation using simulations

The inference framework was first tested on independently simulated multiple sequence alignments (see methods). With the aim of applying the inference method to empirical datasets, the simulation parameters were chosen so as to match an empirically relevant empirical regime. Thus, the tree topology and the branch lengths were chosen based on a tree estimated on the mammalian dataset further considered below. The other aspects of the simulation model (fitness landscape, variation in N_e) were then varied along a gradient of increasing complexity, so as to test the inference framework under increasingly challenging conditions.

A first series of simulations was meant to test the soundness of our inference framework, by simulating essentially under the model used for inference, although with an independently developed software. Thus, the mutation-selection approximation was assumed to be valid, and sites were simulated under different fitness profiles empirically determined (Bloom, 2017), and finally, N_e was assumed to undergo discrete shifts at the tree nodes but otherwise to remain constant along each branch. In this context, branch lengths and branch-specific values of N_e were accurately estimated by our inference method (figure 8.2, panel A & D). Concerning N_e, the slope of the linear regression between true and estimated branch-specific N_e is 0.794 ($r^2 = 0.915$).

However, the assumptions made for this first round of simulations are almost certainly violated in practice. First, N_e is expected to undergo continuous changes along the lineages of the phylogeny. Second, the diffusion approximation for the probability of fixation (equation 8.3) may not hold in small finite populations. Third, assuming a separate substitution process for each site is equivalent to assuming no linkage between sites (free recombination). In practice, however, there is limited recombination, at least within exons, and this could induce deviations from the mutation-selection approximation, due to Hill-Robertson effects.

The finite population was now modelled explicitly, using a Wright-Fisher simulator, tracking the frequency of each allele at the gene level and at each generation along the phylogeny. No recombination was implemented within genes. These more complex simulation settings account for small population size effects, for hitchhiking of weakly deleterious mutations during selective sweep and for background selection due to linkage disequilibrium. In addition, the effective population size N_e and the mutation rate were allowed to fluctuate continuously along the branches of the tree (changing by a small amount after each generation of the underlying Wright-Fisher process). Finally, short-term fluctuations of N_e, of the order of 20% per generation, were accounted for by adding a random noise to the Brownian process describing the long-term evolution of N_e. In spite of these deviations between the simulation and the inference models, branch lengths and branch-specific effective population sizes could again be robustly recovered by the inference framework (slope of 0.868, $r^2 = 0.919$, figure 8.2, panel B & E).

These results are encouraging. However, they still rely on the assumption of a site-independent fitness landscape, which is equivalent to assuming no epistasis. Yet this assumption is almost certainly violated in practice (Pollock and Goldstein, 2014; Shah *et al.*, 2015). Accordingly, we implemented a more complex, site-dependent fitness landscape accounting for the selective interactions between sites induced by the 3-dimensional structure of protein. In this model, the conformational stability of the protein determines its probability of being in the folded state, which is in turn taken as a proxy for fitness (Williams *et al.*, 2006; Goldstein, 2011; Pollock *et al.*, 2012). Under this evolutionary model, and at any given time, the fitness landscape at a particular codon site is dependent on the amino acids that are currently present at those sites that are in the vicinity of the focal site in 3D space (see supplementary). When applied to data simulated using this model, our inference framework could accurately recover the simulated branch lengths (figure 8.2, panel D). On the other hand, the distribution of N_e across the tree could not be accurately recovered (slope of 0.0196, $r^2 = 0.0122$, figure 8.2, panel F). In fact, no meaningful variation in N_e is detected, and the little variation in N_e that is inferred shows no correlation with the true branch-specific mean N_e values. This effect can be explained by the predicted independence of d_N/d_S, and more generally of the scaled selection coefficients associated with non-synonymous mutations, to changes in N_e in this specific model of protein stability, as shown theoretically by Goldstein (2013).

As an alternative model of epistasis between sites, a Fisher geometric model was also considered for the simulations (see supplementary). The results under this model are in-

termediate between simulations without epistasis and simulations under the biophysically-inspired model considered above. More specifically, under data simulated using Fisher's geometric model, the true and estimated branch-specific N_e are strongly correlated with each other ($r^2 = 0.73$). On the other hand, the slope of the correlation is substantially less than 1 (0.571). In other words, the trends in N_e across the tree are correctly recovered, but the range of the variation in effective population size over the tree is substantially under-estimated. As for the branch lengths, they are again correctly estimated. In summary, our simulation experiments show that our inference framework is reliable in the absence of model mis-specification and is robust to violations concerning short-versus long-term variation in N_e or to the presence of empirically reasonable levels of Hill-Robertson interference. On the other hand, and very importantly, epistasis, which is ignored by the inference model, appears to lead to a general underestimation of the true variation in N_e, to an extent that depends on the exact epistatic model but can go as far as completely obliterating any signal about the true variation in N_e across the tree in the most extreme situations.

8.3.2 Empirical experiments

We next applied our inference framework to a series of 4 empirical datasets spanning different taxnonomic groups within metazoans. As a first empirical case, we considered a dataset of 77 placental mammals, for which complete genome sequences and information about life-history traits is available. Placental mammals offer an interesting example, for which effective population size is likely to show substantial variation across lineages. This variation in N_e is expected to covary with life-history traits (LHTs), such that large-bodied species are expected to have smaller effective population sizes, compared to small-bodied species.

For computational reasons, we restricted our analyses to small concatenates made of 18 randomly sampled alignments of orthologous genes. Since the mutation-selection model considered here assumes a mostly nearly-neutral regime, genes for which positive selection was detected using a site codon model were excluded. To assess the reproducibility of our inference and check that the signal about variation in N_e is not driven by particular genes, we analysed 4 concatenated random samples of 18 genes. The different concatenate showed similar trends in the change of μ ($r^2 = [0.92, 0.95]$) and N_e ($r^2 = [0.51, 0.68]$) between pairs of experiments (see supplementary).

The reconstructed long-term changes in effective population size (N_e) is displayed in figure 8.3. We visually observe a global trend of increasing N_e throughout the tree around 90 and 60 My. We also observe N_e to be lower in some clades, such as Cetacea and Camelidae, while being higher in other clades, such as Rodentia and Pecora. In some cases, a decrease in N_e can be observed along an isolated branch of the tree, for example on the branches leading to the Alpaca (*Vicugna pacos*) or the cheetah (*Acinonyx jubatus*).

The estimated covariance matrix (table 8.1) gives a global synthetic picture about the patterns of covariation between the mutation rate per unit of time μ, the effective

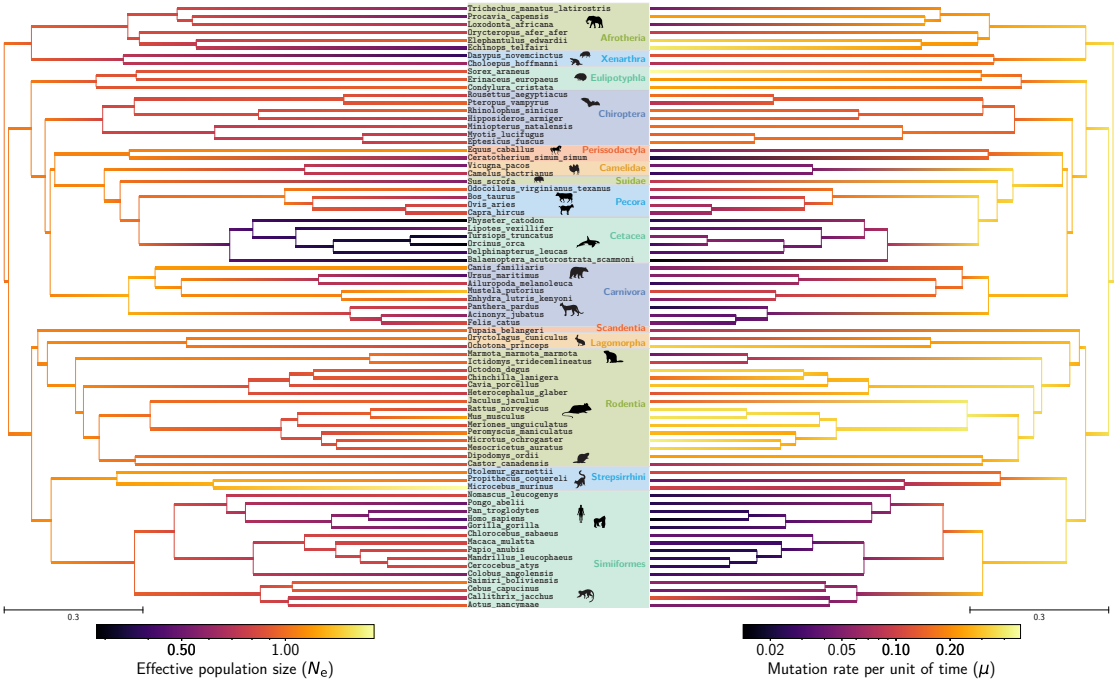


Figure 8.3: *Inferred phylogenetic history of N_e (left) and μ (right) across placental mammals. Inference was conducted on a randomly chosen set of 18 out of 226 highly conserved CDS (< 1% of gaps). Only highly conserved CDS were retained such that the assumption of constant fitness landscape is not incautiously broken by protein with changing function and/or adaptive selection. N_e values are relative to the root, which is arbitrarily set to one. Mean values of MCMC (after burn-in) are obtained at each node of the tree, hence a gradient can be extrapolated along each branch. μ spanned almost 2 order of magnitude, and if we assume the root to be 105My old (Kumar et al., 2017), the rescaled mutation rate per site per year in extant species is between $1.1e^{-10}$ and $7.8e^{-9}$. N_e at the root of the tree is arbitrarily set to 1, and all values are relative to the root, which spans at most an order of magnitude.*

population size N_e and the three LHTs. First, the variation in μ across species is negatively correlated with variation in body mass, age at sexual maturity and longevity ($\rho = [-0.84, -0.83]$, table 8.1). These correlations, which were previously reported (Lartillot and Delsuc, 2012; Nabholz *et al.*, 2013) probably reflect generation time effects (Lanfear *et al.*, 2010a; Gao *et al.*, 2016). Similarly, and more interestingly in the present context, the variation in N_e between species is also negatively correlated with LHTs ($\rho = [-0.54, -0.47]$, table 8.1). This is consistent with the expectation that small-sized and short-lived species tend to be characterized by larger effective population sizes (Romiguier *et al.*, 2014). Of note, these results mirror previous findings, based on classical codon models, showing that d_N/d_S tends to be positively correlated with LHTs (Lartillot and Delsuc, 2012; Nabholz *et al.*, 2013; Figuet *et al.*, 2017). Result which was also recovered on the present dataset, using a classical d_N/d_S based codon model (supplementary materials). Interestingly, the correlation of d_N/d_S with LHTs is weaker than that of our inferred N_e with LHTs, as expected if the variation in d_N/d_S indirectly (and imperfectly) reflects the underlying variation in Ne. Finally, N_e and μ are positively correlated in their variation ($\rho = 0.44$), which might simply reflect the fact that both negatively correlate with LHTs. The partial-correlation coefficients (see sup-

Correlation (ρ)	N_e	μ	Maximum longevity	Adult weight	Female maturity
N_e	-	0.439^{**}	-0.523^{**}	-0.544^{**}	-0.47^{**}
μ	-	-	-0.832^{**}	-0.835^{**}	-0.833^{**}
Maximum longevity	-	-	-	0.827^{**}	0.845^{**}
Adult weight	-	-	-	-	0.809^{**}
Female maturity	-	-	-	-	-

Table 8.1: *Correlation coefficient between effective population size (N_e), mutation rate per site per unit of time (μ), and life-history traits (Maximum longevity, adult weight and female maturity). Asterisks indicate strength of support of the posterior probability to be different than 0 (pp) as $^{*}pp > 0.95$ and $^{**}pp > 0.975$. Observed correlations are compatible with the interpretation that large populations are composed of small, short-lived individuals. Moreover if the mutation rate per generation is considered constant in first approximation, the mutation rate per unit of time is positively correlated to generation rate, hence to population size.*

plementary) between N_e and LHTs are not significantly different from 0. However, this might simply be due to the very strong correlation between the three LHTs considered here ($\rho = [0.81, 0.85]$), such that controlling for any one of them removes most of the signal contributed by the available empirical variation between species.

Thus, altogether, the inferred trends in N_e across species appear to be as expected, based on considerations about life-history evolution. On the other hand, the total range of the inferred variation in N_e across the entire extant taxa is surprisingly narrow, with one order of magnitude (9.2) at most between high and low N_e (see supplementary). This almost certainly represents an underestimate of the true range of variation across placental mammals.

As another case study, we analysed a group of isopod species that have made multiple independent transitions to subterranean environments. The transition from a terrestrial to a subterranean lifestyle is typically associated with a global life-history and ecological syndrome characterized by a loss of vision, longer generation times and, most interestingly, smaller population sizes, due to a lower carrying capacity of the subterranean environment (Capderrey *et al.*, 2013). Protein coding DNA sequence alignments and qualitative life-history traits such as habitat (surface or underground), pigmentation (depigmented, partially depigmented or pigmented) and ocular structure (anophthalmia, microphthalmia, or ocular) are available for these species (Eme *et al.*, 2013; Saclier *et al.*, 2018). The assumption of a Brownian auto-correlated process for describing the changes in N_e along the tree may not be so well adapted to the present case, since the changes in N_e associated with the transition to a subterranean environment are likely to correspond to relatively sudden shifts, rather than continuous variation, and the ecological correlate (subterranean versus terrestrial) is not a quantitative trait. However, the dataset considered here contains independent transitions to a subterranean lifestyle, thus offering an opportunity to test for a potential correlation between inferred N_e variation and terrestrial versus subterranean lifestyles over the terminal branches. In our analysis across 4 concatenated random samples of 12 genes, we observe a reproducible (see supplementary) and statistically significant reduction in N_e for underground or depigmented species, or for species with visual impairment (see figure 8.4). Of note, the species that dit not un-

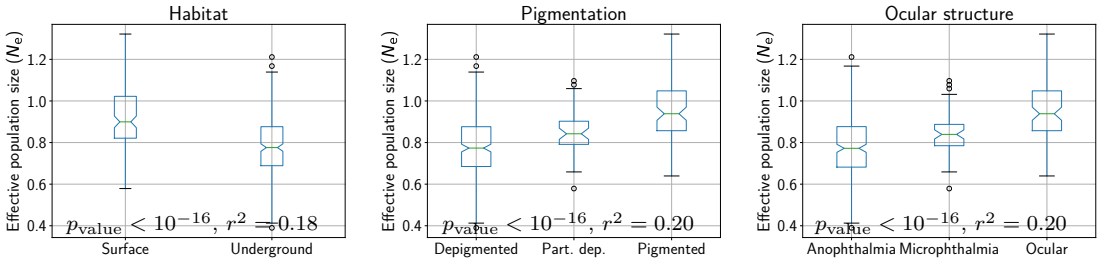


Figure 8.4: *N_e estimation for extant isopods species, sorted according to their habitat (left), pigmentation (middle), and ocular structure (right). All three qualitative trait statistically correlate with changes in N_e. Underground, or depigmented species, or species with visual impairment are characteristic of low N_e species.*

dergo a transition to subterranean environments feature a relative N_e close to 1, meaning that N_e has not changed much along the lineages (since the root of the tree). Again, the total range of the inferred variation in N_e across the entire extant taxa is surprisingly narrow, with ratio of 3.3 at most between high and low N_e (see supplementary).

Next, our empirical framework was also applied on a set of genes sampled across primates, taken from Perelman *et al.* (2011) and reanalysed in Brevet and Lartillot (2019). In addition to LHTs (mass, female maturity, generation time and longevity), information about nuclear synonymous diversity (π_S) and non-synonymous over synonymous diversity (π_N/π_S), are available for 10 species across the dataset and are expected to correlate with N_e according to population genetics (Eyre-walker and Keightley, 2007; Galtier, 2016). However, the correlation coefficient between our inferred N_e and π_S or π_N/π_S and LHTs are not statistically significant, nor with LHTs (see supplementary). Again, the total range of the inferred variation in N_e across the entire tree is narrow, with ratio of 6.4 at most between high and low N_e. This results contrast with the finding of Brevet and Lartillot (2019) on the same dataset based on d_N/d_S-based codon models, where the estimated N_e was found to span several orders of magnitude, and correlated positively with π_S.

8.4 Discussion

Mechanistic phylogenetic codon models express the substitution rates between codons as a function of the mutation rates at the nucleotide level, selection over amino-acid sequences and effective population size. Thus far, the development of mutation-selection models of the HB family (Rodrigue *et al.*, 2010; Tamuri and Goldstein, 2012) has mostly focused on the question of fully accounting for the fine-scale modulations of selection between amino-acids and across sites (Rodrigue *et al.*, 2010; Tamuri and Goldstein, 2012). However, the issue of the variation in the global population-genetic regime between species has received much less attention. In particular, effective population size (N_e) is expected to vary substantially over the species of a given clade, yet current mutation-selection models all invariably assume N_e to be constant across the phylogeny.

Here, we have introduced an extension of the mutation-selection model that accounts for this variation. When applied to an alignment of protein coding sequences, this mech-

anistic model returns an estimate of the modulations of amino-acid preferences across sites. Simultaneously, it reconstructs the joint evolution of life-history traits and molecular and population-genetic parameters (mutation rate μ and effective population size N_e) along the phylogeny, while estimating the correlation matrix between these variables, intrinsically accounting for phylogenetic inertia.

8.4.1 Reliability of the inference of the phylogenetic history of N_e

The reconstructions obtained on several empirical datasets, in particular in mammals and in isopods, suggest that the method is able to correctly infer the directional trends of the changes in N_e across species. In particular, in mammals, the inferred variation in N_e correlates negatively with body size and, more generally, with life-history traits, as expected under the reasonable assumption that large-bodied mammals would tend to have smaller effective population sizes Popadin *et al.* (2007); Lartillot and Delsuc (2012); Nabholz *et al.* (2013); Figuet *et al.* (2017). Similarly, in isopods, smaller effective population sizes are inferred in subterranean species, again, as expected (Capderrey *et al.*, 2013).

However, if the trends are in right direction, the magnitude of the changes inferred across the phylogeny is surprisingly narrow and does not match independent empirical estimates of the variation in those clades. In particular, in mammals, synonymous diversity varies by a factor at least 10 between species (Galtier, 2016). In animals, the synonymous diversity roughly spans two orders of magnitude, whereas N_e varies considerably more across species, by a factor of 10^3 (Galtier and Rousselle, 2020). For instance, effective population sizes estimated based on population genomic data are of the order of 10 000 in humans (Li and Durbin, 2011), and 100 000 in mice (Geraldes *et al.*, 2008). Thus, clearly, our approach underestimates the true variation. Different mechanisms not accounted for by the model could explain this result.

First, genetic hitchhiking, Hill-Robertson interference, and short-term fluctuations of N_e could generate this effect. However, inference conducted on alignments simulated under a Wright-Fisher model accounting for linkage and for short-term variation in N_e suggests that empirically reasonable levels of Hill-Robertson interferences are not strong enough to explain this observation, at least in the regimes explored. Second, μ and N_e could also be fluctuating along the genome (Gossmann *et al.*, 2011; Ellegren *et al.*, 2003; Eyre-Walker and Eyre-Walker, 2014). This assumption needs to be tested, though we expect that relaxing this assumption would not change drastically the magnitude of inferred N_e since some of this fluctuation should be absorbed by the inferred site-specific fitness profiles. Third, the DNA sequences could also be misaligned at some sites. However we observe the same magnitude of inferred N_e for different sets of genes indicating this might not be the primary reason. Fourth, the genes selected in our alignments could be under adaptive evolution, or their function could have changed. However, at least in mammals, the impact of this potential problem was minimized by the use of genes for which no positive selection was detected using standard phylogenetic codon site models.

Finally, one key assumption of the mutation-selection model that is likely to be violated in practice is the assumption of site-independence. In reality, epistasis might be prevalent in protein coding sequence evolution (Pollock and Goldstein, 2014; Shah *et al.*, 2015). Our simulations under an epistatic landscape point to epistasis being a major factor to be investigated. Indeed, N_e could not be appropriately estimated under these simulation settings, although the outcome more specifically depends on the exact model for the fitness landscape. An extreme case is obtained using a biophysically-inspired model, assuming purifying selection for conformational stability. This model was previously explored using simulations and theoretical developments Goldstein (2013), and it was shown that, under this model, d_N/d_S and more generally the substitution process is virtually insensitive to N_e. This is confirmed by our experiments, showing that the mutation-selection approach explored here cannot infer the true variation in N_e under this model.

A less extreme outcome is obtained under an alternative model also implementing epistatic interactions between sites via Fisher's geometric model (Tenaillon, 2014; Blanquart and Bataillon, 2016). Interestingly, under this model, our inference framework is able to infer the correct trends of N_e, although with a substantially underestimated range of inferred variation, thus mirroring the results obtained on placental mammals. Of note, these results do not necessarily imply that models based on biophysics are empirically less relevant than Fisher's geometric model. Instead, they might just betray that the response of the substitution process to changes in N_e may be sensitive to the exact quantitative details of the underlying fitness landscape. More work is probably needed here to characterize these exact conditions. Nevertheless, our simulation experiments suggest a global pattern: epistatic interactions induce a buffering of the response of the substitution process to changes in N_e. The meaningful correlation patterns observed with LHTs in the case of placental mammals suggest that this buffering is not complete. Nevertheless, ignoring epistatic interactions at the inference level appears to result in a substantial underestimation of the range over which N_e varies across species.

Interestingly, the magnitude of the inferred range of N_e variation is similar for the placental and the primate datasets (with a 9-fold and 6-fold variation in mammals and primates, respectively), whereas one would have expected a much larger range of variation over the broader phylogenetic scale of placental mammals, compared to primates. An explanation could be that the effects of epistasis are more apparent at longer time-scales. Indeed, the total number of substitutions from root to leaves is greater, and as a result, the local environment, and therefore the fitness landscape at the level of each site, has been less stable across the phylogeny.

Although modelling epistasis in an inference framework is a complex biological, mathematical and computational problem, our work points to a potential signal of epistasis that could be retrieved in a phylogenetic context. More specifically, since the slope of the response of the substitution process to changes in N_e appears to be informative about the epistatic regime, then, conversely, by relying on independent estimates of N_e (e.g. using polymorphism), this effect could be used to leverage a quantitative estimate of the statistical distribution of epistatic effects.

Other methods have recently been developed to reconstruct phylogenetic changes in N_e. For example, a method recently developed uses polymorphism and generation time for some present-day species to reconstruct N_e along the phylogeny, based on a classical (d_N/d_S-based) codon model (Brevet and Lartillot, 2019). This method implicitly relies on a nearly-neutral model, assuming a fixed and gamma-shaped distribution of fitness effects across non-synonymous mutations. The approach is calibrated using fossils, and as a result, returns estimates of the absolute value of N_e and of its phylogenetic variation. Here, in contrast, our method requires neither generation times nor polymorphism data, and the fitness effects are not constrained to a specific distribution. On the other hand, the inferred effective population sizes are only relative. In addition, the empirical fitting of the model requires more computing resources.

8.4.2 Potential applications and future developments

Apart from reconstructing the phylogenetic history of N_e and investigating its causes and covariates, another potentially interesting application of our approach is in detecting adaptation. In this direction, mutation-selection models represent a useful null nearly-neutral model, explicitly modelling the background of purifying selection acting over protein coding genes. Adaptation can then be detected by measuring the deviation from this null model (Rodrigue and Lartillot, 2016; Bloom, 2017).

However, by assuming a constant N_e along a phylogeny, the statistical power of this approach to detect sites under adaptive evolution may not be optimal. In particular, the site-specific fitness profiles inferred by the model are averaged along the phylogeny and are seemingly more diffuse than those estimated profiles under our present framework (see supplementary materials). Thus, our method should provide a better null model of purifying selection against which to test for the presence of adaptive evolution.

This approach can be further extended in other directions. First, currently, our model also assumes no selection on codon usage. In the case of primates or placental mammals, this assumption is probably reasonable (Yang and Nielsen, 2008), although it is more questionable for other groups, in particular Drosophila (Duret and Mouchiroud, 1999; Plotkin and Kudla, 2011). In principle, this assumption can be relaxed by implementing selective codon preferences that are shared across all sites. Such an implementation would provide the advantage of estimating codon usage biases, while simultaneously accounting for its confounding effect when estimating selection on amino-acids and inter-specific variation in N_e.

Second, the Bayesian analysis conducted here was based on relatively small alignments (20 000 sites at most), and with strong limits on the parametrization of the underlying mixture model (allowing for at most 50 distinct profile categories). Profiling of the program (not shown) shows that the number of components of the profile mixture is the limiting step of the computation. Yet, a larger number of components might be required, in order to achieve more accurate inference of the site-specific profiles. One possible development, leading to statistically more stable genome-wide estimates of N_e,

would be to develop a multi-gene parallelized version of the model, in which each coding sequence would have its own mixture model, and would run on a separate thread, while the history of N_e would be shared by all computing processes.

Finally, estimating N_e in a mutation-selection phylogenetic model relies on the relation between N_e and the relative strength of drift, in a context where, ultimately, the signal about the intensity of drift comes from the relative rate of non-synonymous substitutions. However, this purely phylogenetic approach does not leverage a second aspect of N_e at the population level, namely, the fact that N_e also determines the levels of neutral genetic diversity that can be maintained ($\pi = 4N_e u$, where u is the mutation rate per generation). Hence, neutral diversity yields an independent empirical estimate of N_e. In principle, our mechanistic model could be extended so as to incorporate polymorphism data within species at the tips of the phylogeny. A similar method has been previously pioneered in the case of 3 species and using a distribution of fitness effect (Wilson *et al.*, 2011). More generally, the nearly-neutral theory of evolution defines a long-term N_e, which might be different from the short-term definition of N_e (Platt *et al.*, 2018). Thus we could ask if empirical independent estimations of N_e from within species (based on genetic diversity) and between species (based on the substitution process) are congruent, and if not, what are the mechanisms responsible for this discrepancy.

Notwithstanding theoretical considerations on the nearly-neutral theory of evolution, empirical clues about the long-term trends in the modulations of the intensity of genetic drift opens up a large diversity of ecological and evolutionary questions. Spatial and temporal changes of genetic drift along ecological niches and events can now be investigated, so as to disentangle the underlying evolutionary and ecological pressures.

8.5 Materials and Methods

In the model presented here, N_e and μ and quantitative traits are allowed to vary between species across branches as a multivariate log-Brownian process, but assumed constant along the DNA sequence. Conversely, amino-acid fitness profiles are assumed to vary across sites, but are considered constant along the tree. The model makes several assumptions about the evolutionary process generating the observed alignment. First, the species tree topology is supposed to be known, and each gene should match the species tree, meaning genes are strict orthologs (no paralogs and no horizontal transfers). Second, there is no epistasis (interaction between sites), such that any position of the sequence has its own independent evolutionary process and a substitution at one position does not affect the substitution process at other positions. Third, from a population genetics perspective, we assumed sites of the protein to be unlinked, or equivalently the mutation rate is low enough such that there is no Hill-Robertson interference nor genetic hitchhiking. Fourth, polymorphism is ignored in extant species.

The parameterization of the models is described as a Bayesian hierarchical model, including the prior distributions and the parameters of the model. This hierarchical model is formally represented as directed acyclic graph, depicted in figure 8.5.

8.5.1 Nucleotide mutation rates

The generalized time-reversible nucleotide mutation rate matrix $\boldsymbol{R}$ is a function of the nucleotide frequencies $\boldsymbol{\sigma}$ and the symmetric exchangeability rates $\boldsymbol{\rho}$ (Tavaré, 1986). $\boldsymbol{\sigma} = (\sigma_A, \sigma_C, \sigma_G, \sigma_T)$ is the equilibrium base frequency vector, giving the frequency at which each base occurs at each site. $\boldsymbol{\rho} = (\rho_{AC}, \rho_{AG}, \rho_{AT}, \rho_{CG}, \rho_{CT}, \rho_{GT})$ is the vector of exchangeabilities between nucleotides. Altogether, the rate matrix is:

$$\boldsymbol{R} = \begin{array}{c} \\ A \\ C \\ G \\ T \end{array} \begin{array}{c} \begin{array}{cccc} A & C & G & T \end{array} \\ \begin{pmatrix} - & \rho_{AC}\sigma_C & \rho_{AG}\sigma_G & \rho_{AT}\sigma_T \\ \rho_{AC}\sigma_A & - & \rho_{CG}\sigma_G & \rho_{CT}\sigma_T \\ \rho_{AG}\sigma_A & \rho_{CG}\sigma_C & - & \rho_{GT}\sigma_T \\ \rho_{AT}\sigma_A & \rho_{CT}\sigma_C & \rho_{GT}\sigma_G & - \end{pmatrix} \end{array} \tag{8.4}$$

By definition, the sum of the entries in each row of the nucleotide rate matrix $\boldsymbol{R}$ is equal to 0, giving the diagonal entries:

$$R_{a,a} = - \sum_{b \neq a, b \in \{A,C,G,T\}} R_{a,b} \tag{8.5}$$

The prior on the exchangeabilities $\boldsymbol{\rho}$ is a uniform Dirichlet distribution of dimension 6:

$$\boldsymbol{\rho} \sim \mathrm{Dir}\left(\frac{1}{6}, 6\right). \tag{8.6}$$

The prior on the equilibrium base frequencies $\boldsymbol{\sigma}$ is a uniform Dirichlet distribution of dimension 4:

$$\boldsymbol{\sigma} \sim \mathrm{Dir}\left(\frac{1}{4}, 4\right) \tag{8.7}$$

The general time-reversible nucleotide matrix is normalized such that the total flow equals to 1:

$$\sum_{a \in \{A,C,G,T\}} -\sigma_a R_{a,a} = 1. \tag{8.8}$$

8.5.2 Site-dependent selection

Site-specific amino-acid fitness profiles are assumed i.i.d. from a mixture model, itself endowed with a truncated Dirichlet process prior. Specifically, the mixture has K components (K = 50 by default). The prior on component weights ($\boldsymbol{\theta}$) is modeled using a stick-breaking process, truncated at K and of parameter β:

$$\begin{aligned} &\boldsymbol{\theta} \sim \mathrm{StickBreaking}\left(\mathrm{K}, \beta\right) \\ \iff &\theta_{\mathrm{k}} = \psi_{\mathrm{k}} \cdot \prod_{a=1}^{\mathrm{k}-1} \left(1 - \psi_a\right), \ \mathrm{k} \in \{1, \ldots, \mathrm{K}\}, \end{aligned} \tag{8.9}$$

where ψ_{k} are i.i.d. from a beta distribution

$$\psi_{\mathrm{k}} \sim \mathrm{Beta}\left(1, \beta\right), \ \mathrm{k} \in \{1, \ldots, \mathrm{K}\}. \tag{8.10}$$

Of note, the weights decrease geometrically in expectation, at rate β, such that lower values of β induce more heterogeneous distributions of weights.

Each component of the mixture defines a 20-dimensional fitness profile $\boldsymbol{\phi}^{(\mathrm{k})}$ (summing to 1), for $\mathrm{k} \in \{1, \ldots, \mathrm{K}\}$. These fitness profiles are i.i.d. from a Dirichlet of center $\boldsymbol{\gamma}$ and concentration α:

$$\boldsymbol{\phi}^{(\mathrm{k})} \sim \mathrm{Dir}\left(\boldsymbol{\gamma},\ \alpha\right),\ \mathrm{k} \in \{1, \ldots, \mathrm{K}\}. \tag{8.11}$$

Site allocations to the mixture components $\kappa\left(\mathrm{z}\right) \in \{1, \ldots, \mathrm{K}\}$, for $\mathrm{z} \in \{1, \ldots, \mathrm{Z}\}$ running over the Z sites of the alignment, are i.i.d. multinomial of parameter $\boldsymbol{\theta}$:

$$\boldsymbol{m} \sim \mathrm{Multinomial}\left(\boldsymbol{\theta}\right), \tag{8.12}$$

$$\text{where } m_{\mathrm{k}} = \sum_{\mathrm{z} \in \{1, \ldots, \mathrm{Z}\}} \mathbb{1}_{\kappa(\mathrm{z})=\mathrm{k}} \tag{8.13}$$

For a given parameter configuration for the mixture, the Malthusian fitness selection coefficients $\boldsymbol{f}^{(\mathrm{z})}$ at site z, are obtained by taking the logarithm of the fitness profile assigned to this site:

$$\boldsymbol{f}^{(\mathrm{z})} = \ln\left(\boldsymbol{\phi}^{(\kappa(\mathrm{z}))}\right),\ \mathrm{z} \in \{1, \ldots, \mathrm{Z}\}. \tag{8.14}$$

8.5.3 Dated tree

The topology of the rooted phylogenetic tree is supposed to be known and is not estimated by the model. The model estimates the dates at which branches split, thus the dated tree requires $P - 2$ internal node ages that are free parameters, where P is the number of extant taxa (leaves of the tree). By definition, leaf ages are all set to 0. The root age is set arbitrarily to 1, but if fossils data are also available the dated tree can be rescaled into absolute time using cross-multiplication. A uniform prior is assumed over internal node ages $T^{(\mathrm{n})},\ \mathrm{n} \in \{P + 1, \ldots, 2P - 2\}$.

The duration $\Delta T^{(\mathrm{b})}$ represented by a given branch b, for $\mathrm{b} \in \{1, \ldots, 2P - 2\}$ is defined as the difference in ages between the oldest node at the tip of the branch $T^{(\mathrm{b}^{\uparrow})}$, and the youngest node $T^{(\mathrm{b}^{\downarrow})}$:

$$\Delta T^{(\mathrm{b})} = T^{(\mathrm{b}^{\uparrow})} - T^{(\mathrm{b}^{\downarrow})}. \tag{8.15}$$

8.5.4 Branch dependent traits

The effective population size N_{e} and mutation rate per unit of time μ are assumed to evolve along the phylogeny, and to be correlated. If quantitative life-history traits (LHTs) are also available for some nodes of the tree (leaves and/or internal nodes), they are also assumed to evolve along the phylogeny and to be correlated between them, and with N_{e} and μ. The total number of traits is noted L, when counting N_{e}, μ and all user-defined

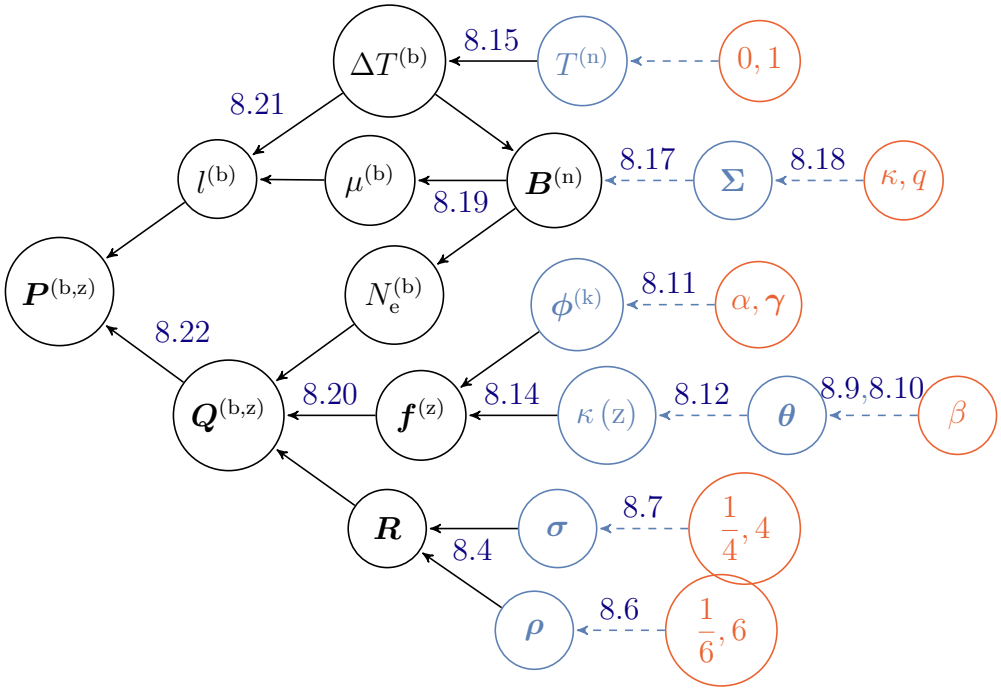


Figure 8.5: *Directed acyclic graph (DAG) of dependencies between variables. Nodes of the directed acyclic graph are the variables, and edges are the functions. Hyper-parameters are depicted in red circles, random variables in blue circles, and transformed variables in black. Blue dashed line denotes a drawing from a random distribution, and black solid lines denote a function. For a given node, all the nodes pointing toward him (upstream) are its dependencies which determines its distribution. The other way around, following the arrows in the DAG (downstream), simple prior distributions are combined together to form more complex joint prior distribution which ultimately defines the prior distribution of the model.*

LHT (denoted $\boldsymbol{X}$). Their variation through time is modelled by an L-dimensional log-Brownian process $\boldsymbol{B}$. By convention, the first component of the log-brownian corresponds to N_e, and the second component to μ. Thus:

$$\begin{cases} B_1(t) = \ln N_e(t) \\ B_2(t) = \ln \mu(t) \\ B_{k+2}(t) = \ln X_k(t), k \in \{1, \dots, L\} \end{cases} \tag{8.16}$$

The effective population size at the root is set to 1 for identifiability of the fitness profiles.

Along a branch $b \in \{1, \dots, 2P-2\}$ of the tree, a log-Brownian process starts at the oldest node at the tip of the branch ($b^{\uparrow}$), and ends at the youngest node ($b^{\downarrow}$). The rate of change of the log-Brownian process per unit of time is constant and determined by the positive semi-definite and symmetric covariance matrix $\boldsymbol{\Sigma}$. Thus the distribution at node $b^{\downarrow}$ of $\boldsymbol{B}^{(b^{\downarrow})}$ is multivariate Gaussian, with mean equals to the Brownian process sampled at the oldest node $\boldsymbol{B}^{(b^{\uparrow})}$, and variance $\Delta T^{(b)}\boldsymbol{\Sigma}$:

$$\boldsymbol{B}^{(b^{\downarrow})} \sim \mathcal{N}\left(\boldsymbol{B}^{(b^{\uparrow})}, \Delta T^{(b)}\boldsymbol{\Sigma}\right), \; b \in \{1, \dots, 2P-2\}. \tag{8.17}$$

The Brownian process at the root of the tree is uniformly distributed, except for the first component fixed to 0 for identifiability (see above). The prior on the covariance matrix is an inverse Wishart distribution, parameterized by $\kappa = 1$ and with $q = L + 1$

degrees of freedom:

$$\boldsymbol{\Sigma} \sim \text{Wishart}^{-1}(\kappa \boldsymbol{I}, q). \tag{8.18}$$

We are interested in approximating the expected substitution rates between codons over the branch. Ideally, under the Brownian process just described, the rates of substitution between codons are continuously changing through time. Also, even conditional on the value of N_e at both ends, the Brownian path along the branch entails a random component, leading to complicated integral expressions for substitution rates (Horvilleur and Lartillot, 2014). Here, a branchwise approximation is used (Lartillot and Poujol, 2011), which consists of first deriving an approximation for the mean N_e along the branch, conditional on the values of N_e at both ends, and then using this mean branchwise N_e to define the codon substitution rates.

In the case of log-Brownian process, the most likely path (or geodesic) from $\boldsymbol{B}^{(\text{b}^\uparrow)}$ to $\boldsymbol{B}^{(\text{b}^\downarrow)}$ is the straight line, and therefore, it would make sense to take the mean value of $\text{e}^{\boldsymbol{B}^{(\text{n})}}$ along this geodesic. We then have $N_\text{e}^{(\text{b})}$ and $\mu^{(\text{b})}$ for each branch $\text{b} \in \{1, \ldots, 2P-2\}$ of the tree:

$$\begin{cases} N_\text{e}^{(\text{b})} = \dfrac{\text{e}^{B_1^{(\text{b}^\downarrow)}} - \text{e}^{B_1^{(\text{b}^\uparrow)}}}{B_1^{(\text{b}^\downarrow)} - B_1^{(\text{b}^\uparrow)}} \\ \mu^{(\text{b})} = \dfrac{\text{e}^{B_2^{(\text{b}^\downarrow)}} - \text{e}^{B_2^{(\text{b}^\uparrow)}}}{B_2^{(\text{b}^\downarrow)} - B_2^{(\text{b}^\uparrow)}}. \end{cases} \tag{8.19}$$

8.5.5 Codon substitution rates

The mutation rate between codons i and j, denoted $\mu_{i,j}$ depends on the underlying nucleotide change between the codons. First, if codons i and j are not nearest-neighbours, $\mu_{i,j}$ is equal to 0. Second, if codons i and j are only one mutation away, $\mathcal{M}(i,j)$ denotes the nucleotide change (e.g. $\mathcal{M}(AAT, AAG) = TG$), and $\mu_{i,j}$ is given by the underlying nucleotide relative rate ($R_{\mathcal{M}(i,j)}$) scaled by the mutation rate per time (μ). Technically, the 4-dimensional nucleotide relative rate matrix ($\boldsymbol{R}$) is normalized such that we expect 1 substitution per unit of time, hence the scaling by μ.

For a given branch b and a given site z, the codon substitution rate (per unit of time) matrix $\boldsymbol{Q}^{(\text{b,z})}$ is given by:

$$\begin{cases} Q_{i,j}^{(\text{b,z})} = 0 \text{ if codons } i \text{ and } j \text{ are not neighbors,} \\ Q_{i,j}^{(\text{b,z})} = R_{\mathcal{M}(i,j)} \text{ if codons } i \text{ and } j \text{ are synonymous,} \\ Q_{i,j}^{(\text{b,z})} = R_{\mathcal{M}(i,j)} \dfrac{4N_\text{e}^{(\text{b})} \left(f_{\mathcal{A}(j)}^{(\text{z})} - f_{\mathcal{A}(i)}^{(\text{z})} \right)}{1 - \text{e}^{4N_\text{e}^{(\text{b})} \left(f_{\mathcal{A}(i)}^{(\text{z})} - f_{\mathcal{A}(j)}^{(\text{z})} \right)}} \text{ if } i \text{ and } j \text{ are non-synonymous,} \\ Q_{i,i}^{(\text{b,z})} = - \sum_{j \neq i, j=1}^{61} Q_{i,j}^{(\text{b,z})}. \end{cases} \tag{8.20}$$

We see from this equation that, f and N_e are confounded, such that increasing the effective population size while decreasing the fitnesses by the same factor leads to the same substitution rate.

The branch lengths $l^{(b)}$ are defined as the expected number of neutral substitutions per DNA site along a branch:

$$l^{(b)} = \mu^{(b)} \Delta T^{(b)}. \tag{8.21}$$

Together, the probability of transition between codons for a given branch b and site z is:

$$\boldsymbol{P}^{(b,z)} = e^{l^{(b)} \boldsymbol{Q}^{(b,z)}}, \tag{8.22}$$

which are the matrices necessary to compute the likelihood of the data (D) given the parameters of the model using the pruning algorithm.

8.5.6 Bayesian implementation

Bayesian inference was conducted using Markov Chain Monte Carlo (MCMC). Most phylogenetic MCMC samplers target the distribution over the model parameters given the sequence alignment, which means that they have to repeatedly invoke the pruning algorithm to recalculate the likelihood which is most often the limiting step of the MCMC. An alternative, which is used here, is to do the MCMC conditionally on the detailed substitution history $\mathcal{H}$, thus doing the MCMC over the augmented configuration ($\mathcal{H}$, D), under the target distribution obtained by combining the mapping-based likelihood with the prior over model parameters.

The key idea that makes this strategy efficient is that the mapping-based likelihood depends on compact summary statistics of $\mathcal{H}$, leading to very fast evaluation of the likelihood. On the other hand, this requires to implement more complex MCMC procedures that have to alternate between:

1. sampling $\mathcal{H}$ conditionally on the data and the current parameter configuration.

2. re-sampling the parameters conditionally on $\mathcal{H}$.

To implement the mapping-based MCMC sampling strategy, we first sample the detailed substitution history $\mathcal{H}$ for all sites along the tree. Several methods exist for doing this (Nielsen, 2002; Rodrigue *et al.*, 2008b), which are used here in combination (first trying the accept-reject method of Nielsen, then switching to the uniformization approach of Rodrigue *et al* if the first round has failed).

Then, we write down the probability of $\mathcal{H}$ given the parameters, and finally, we collect all factors that depend on some parameter of interest and make some simplifications. This ultimately leads to relatively compact sufficient statistics (see supplementary) allowing for fast numerical evaluation of the likelihood (Irvahn and Minin, 2014; Davydov *et al.*, 2016). As an example, making an MCMC move on the N_e at a given node of

the tree is faster since only the mapping-based likelihood (using path sufficient statistics) at the neighbouring branches of the node is necessary, instead of computing the likelihood for the entire tree.

Markov chain Monte Carlo (MCMC) are run for 4000 points and the first 1000 points are discarded as burn-in. Convergence is then assessed (see supplementary) by comparing two independent chains, checking that both site-specific fitness and branch N_e have the same posterior mean.

8.5.7 Correlation between traits

The correlation between trait a and trait $b \in \{1, \ldots, L\}$ can be obtained from the covariance matrix Σ:

$$\rho_{a,b} = \frac{\Sigma_{a,b}}{\sqrt{\Sigma_{a,a}\Sigma_{b,b}}}. \tag{8.23}$$

This correlation coefficient is then averaged over the posterior distribution, and statistical support is assessed based on the posterior probability of having a positive (or negative) value for the coefficient.

8.5.8 Simulations

To test the robustness of the model, four parameterized simulators were developed: `SimuDiv`, `SimuPoly`, `SimuFold` & `SimuGeo`. All four simulators use a log-Brownian multivariate process to model the changes in the mutation rate per generation, the generation time and N_e along the lineages. `SimuDiv`, `SimuFold` & `SimuGeo` all simulate point substitutions along the phylogenetic tree. The simulator starts from an initial sequence at equilibrium. The change in fitness is computed for all possible mutant, hence computing all strictly positive substitution rates. At each point, the next substitution is chosen proportional to these rates using in Gillespie's algorithm (Gillespie, 1977). At each node, the process is split, and finally stopped at the leaves of the tree. `SimuPoly` simulates explicitly each generation along the phylogeny under a Wright-Fisher population, consisting of three steps: mutation, selection and genetic drift of currently segregating alleles. Mutations are drawn randomly based on mutation rates. Drift is induced by the multinomial resampling of the currently segregating alleles. We assume that the DNA sequence is composed of exons, with no linkage between exons, and total linkage of sites within an exon. Moreover, in `SimuPoly`, the instant value of log-N_e can also be modelled as a sum of a log-Brownian process and an Ornstein-Uhlenbeck process. The log-Brownian motion accounts for long-term fluctuations, while the Ornstein-Uhlenbeck introduces short-term fluctuations. In `SimuDiv` and `SimuPoly`, each codon site contributes independently to the fitness depending on the encoded amino acids, through site-specific amino-acid fitness profiles experimentally determined (Bloom, 2017). In `SimuFold`, the fitness of a sequence is computed as the probability of the protein to be in the folded state. `SimuFold` is a `C++` adaptation of a Java code previously published (Goldstein

and Pollock, 2016, 2017), where we also allow for changes in N_e and μ along a phylogenetic tree. Supplementary materials describe the models in more details, as well as performance of the inference model against them.

8.5.9 Empirical data

For placental mammals, alignments were extracted from OrthoMam database (Ranwez *et al.*, 2007; Scornavacca *et al.*, 2019). Only highly conserved coding sequences are kept for the analysis, representing 226 CDS with $\leq 1\%$ of gaps in the alignment. Life-history traits (LHTs) for longevity, age at maturity and weight were obtained from AnAge database (De Magalhães and Costa, 2009; Tacutu *et al.*, 2012). We focused our analysis on 77 taxa for which information is available for at least one LHT.

8.6 Reproducibility - Supplementary Materials

Supplementary materials and figures are available in appendix, chapter 11. The scripts and instructions necessary to reproduce the simulated and empirical experiments are available at `https://github.com/ThibaultLatrille/MutationSelectionDrift`.

8.7 Author contributions

TL gathered and formatted the data, developed the new models in `BayesCode` and `SimuEvol` and conducted all analyses, in the context of a PhD work (Ecole Normale Superieure de Lyon). VL restructured and refactored the code sustaining the branch and site heterogeneous Bayesian Monte Carlo in `BayesCode`. TL and NL both contributed to the writing of the manuscript.

8.8 Acknowledgements

We wish to thank Tristan Lefébure for sharing the isopods phylogeny, alignments and life-history traits, and Annabelle Haudry & Théo Tricou for providing the Drosophila alignments, phylogeny and genome sizes. We thank Philippe Veber for insightful discussion on mutation-selection models and software development. We gratefully also acknowledge the help of Nicolas Rodrigue, Laurent Gueguen, Benoit Nahbolz and Laurent Duret for their advice and review concerning this manuscript. This work was performed using the computing facilities of the CC LBBE/PRABI.

9

A theoretical approach for quantifying the impact of changes in effective population size and expression level on the rate of coding sequence evolution

Thibault Latrille[1, 2], Nicolas Lartillot[1]

[1]Université de Lyon, Université Lyon 1, UMR CNRS 5558 Laboratoire de Biométrie et Biologie Évolutive, 69622 Villeurbanne, France
[2]École Normale Supérieure de Lyon, Universite de Lyon, 69007 Lyon, France

Contents

9.1 Introduction

Molecular sequences differ across species due to the particular history of nucleotide substitutions along their respective lineages. These substitutions in turn are the result of the interplay between evolutionary forces such as mutation and selection, whose relative forces are determined by the amount of random genetic drift. These forces have effects at different levels: mutations are carried by molecular sequences, selection is mediated at the level of individuals, while random genetic drift is a population sampling effect. Yet, they jointly contribute to the long-term molecular evolutionary process. Thus, the challenge of the study of molecular evolution is to tease out their respective contributions, based on comparative analyses.

One main aspect of this challenge is to correctly evaluate the role of random drift in the long term evolutionary process. Population genetics theory implies that the strength of drift, due to the stochastic sampling of mutations, is less pronounced in lineages with large effective population size (N_e), and as a consequence, the purification by selection of weakly deleterious mutations is more effective in large populations. This fundamental idea is at the core of the nearly-neutral theory of evolution. This theory posits that a substantial fraction of mutations are deleterious or weakly deleterious, and as a result, predicts that the substitution rate (relative to the neutral expectation), called ω, decreases along lineages with higher N_e (Ohta, 1972, 1992).

This prediction has been more quantitatively examined under the assumption that the selective effects of mutations are drawn from a fixed distribution of fitness effects (DFE) (Kimura, 1979; Welch *et al.*, 2008). Assuming a gamma distribution for the DFE, a key result obtained in this context is an approximate allometric scaling of ω as a function of N_e (i.e. $\omega \sim N_e^{-k}$), where k is the shape parameter of the DFE. In practice, DFEs are strongly leptokurtic, which thus predicts a weak negative relation between ω and N_e.

The study of protein-coding sequences evolution fostered another modelling approach, based on genotype-fitness maps instead of distribution of fitness effects. In this alternative approach, the selective effect of a mutation depends on the fitness of both the source and the target amino acids involved in the mutation event (Halpern and Bruno, 1998; Rodrigue *et al.*, 2010; Tamuri and Goldstein, 2012). Even though this modelling approach differs substantially from the one assuming a fixed DFE, it also predicts a negative correlation between ω and N_e, at least when the process is at equilibrium (Spielman and Wilke, 2015; Dos Reis, 2015).

Conversely, one striking theoretical result was the proof that ω is in fact predicted to be independent of N_e under relatively general circumstances, namely, whenever (i) the fitness is a log-concave function of a phenotype and (ii) the phenotype itself is equimutable. Equimutability states that the distribution of phenotypic changes due to mutation is independent of the current phenotype of individuals (Cherry, 1998). This general theoretical argument has been invoked in the context of *in silico* experiments of protein sequence evolution, assuming that proteins are under selection for their thermodynamic stability, with fitness being proportional to the folding probability of the protein (Goldstein, 2013).

Thermodynamic stability is itself computed using a 3D structural model of the protein. These computational experiments have led to the observation that ω is essentially independent of N_e. An explanation proposed for this result is that the distribution of changes in free energy of folding ($\Delta\Delta G$) due to mutations is approximately independent of the current free energy (ΔG), thus making the free energy of folding essentially equimutable.

However, the equimutability assumption is a relatively strong one, which also conflicts with combinatorial considerations about the relation between sequence and phenotype (Serohijos *et al.*, 2012). For example, if a protein sequence is already maximally stable, only destabilizing (or neutral) mutations can occur. More generally, assuming that the stability of a protein sequence reflects an underlying fraction of positions having already accepted destabilizing amino acids, then the probability of destabilizing mutational events is in turn expected to directly depend on the current stability of the protein.

Altogether, depending on the theoretical model mapping sequence to fitness, ω can be either independent or negatively correlated to N_e, or even positively if considering adaptive evolution and environmental changes (Lanfear *et al.*, 2014).

Empirically, variation in ω between lineages has been inferred using phylogenetic codon models applied to empirical sequences (Yang and Nielsen, 1998; Zhang and Nielsen, 2005). Confronting branch-specific ω estimates to life-history traits such as body mass or generation time uncovered a positive correlation (Popadin *et al.*, 2007; Nikolaev *et al.*, 2007). Subsequently, integrative inference methods combining molecular sequences and life-history traits have also found that ω correlates positively with traits such as longevity and body mass (Lartillot and Poujol, 2011; Figuet *et al.*, 2017). Since lineages with a large body size and extended longevity typically correspond to species with low N_e (Romiguier *et al.*, 2014), these empirical correlations suggest a negative correlation between ω and N_e, thus confirming the theoretical prediction of the nearly-neutral theory of evolution. However, the universality and robustness of the correlation between ω and life-history traits is still debated. Results have not been entirely consistent across independent studies. The correlation was found to be either not statistically significant (Lartillot and Delsuc, 2012), or even in the opposite direction depending on the specific clade under study or the potential biases taken into account (Lanfear *et al.*, 2010a; Nabholz *et al.*, 2013; Lanfear *et al.*, 2014; Figuet *et al.*, 2016).

If empirical evidence for a negative correlation of ω with N_e is still not totally convincing, another empirical correlation is known to be much more robust. Indeed, expression level or protein abundance is one of the best predictors of ω, with highly expressed proteins typically having lower ω values, a correlation clearly significant although relatively weak (Duret and Mouchiroud, 2000; Rocha and Danchin, 2004; Drummond *et al.*, 2005; Zhang and Yang, 2015; Song *et al.*, 2017). Theoretical models, also based on protein stability, have been invoked to explain this negative correlation between ω and expression level (Wilke and Drummond, 2006; Drummond and Wilke, 2008). According to this argument, selection against protein misfolding due to toxicity, which is stronger for more abundant proteins, induces abundant proteins to evolve toward greater stability, resulting in a more constrained and more slowly evolving protein coding se-

quence (Serohijos *et al.*, 2012).

The possibility that expression level and N_e might play similar roles in the evolution of proteins has already been noticed. More precisely, under models of selection against protein misfolding, the free energy of folding ΔG is predicted to vary similarly along a gradient of either N_e or expression level (Serohijos *et al.*, 2013). As a corollary, under strict equimutability of ΔG, these computational models imply that ω should also be independent of expression level (Serohijos *et al.*, 2012), akin to what is predicted with regards to changes in N_e.

Altogether, both theoretical results and empirical analyses are not yet conclusive about the question of how ω depends on N_e and expression level. In particular, the theoretical response of ω to changes in both N_e and expression level has not been quantified and, most importantly, has not been related to the specific map between genotype, phenotype and fitness. Such an analytical development would be useful to more decisively confront the theoretical predictions relating ω to both N_e and expression level to empirical data. Ultimately, relating proteins structural parameters to the response of ω would help to bridge the gap between protein thermodynamics on one side and comparative genomics on the other side.

Lastly, the theoretical results discussed so far are valid only at the mutation-selection-drift balance. In a non-equilibrium regime, however, and at least under a model assuming a site-independent genotype-fitness map, an increase in N_e first leads to an increase in ω caused by adaptive substitutions, and subsequently a decrease in ω due to stronger purifying selection in the long term (Jones *et al.*, 2016). Studying only equilibrium properties can thus be misleading. For this reason, the dynamic response of ω to changes in N_e must also be addressed, quantified, and its connection with the underlying selective landscape better characterized. Dynamic properties of ω to changes in N_e are of theoretical interest but are also empirically relevant, such that, if overlooked they could thwart the relation between theoretical expectations and empirical estimates.

In this context, the aim of the present study is to characterize the dynamics and equilibrium response of ω to changes in N_e and expression level, and to relate this response to structural parameters of the model. To this effect, we develop a general mathematical approach to derive a quantitative approximation of the response of ω to changes in N_e and expression level, in the context of a given genotype-phenotype-fitness map, as depicted in figure 9.1. In the light of previously published empirical estimates from protein thermodynamics and comparative genomics, we discuss the articulation between empirical data and our mechanistic model. We also discuss some of the alternative biophysical mechanisms that could determine the selective landscape on protein-coding sequences, and how they would modulate the response of ω to changes in N_e and expression level.

9.2 Results

9.2.1 Models of evolution

The results that are presented below are valid for a general category of models of sequence evolution, based on an additive trait x, such that the coding positions of the sequence contribute additively to the trait. The trait is under directional selection specified by a decreasing and log-concave fitness function $W(x)$. As a specific example, we more specifically consider a model of protein evolution under the constraint of thermodynamic stability, as depicted in the left panel of figure 9.1. This model is inspired from previous work (Williams *et al.*, 2006; Goldstein, 2011; Pollock *et al.*, 2012), except that we make several simplifying assumptions, allowing us to derive analytical equations.

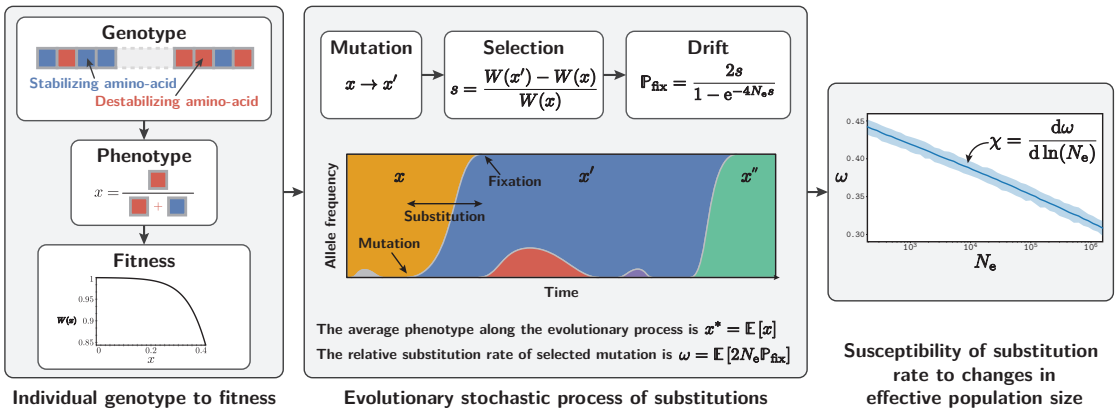


Figure 9.1: *Outline of the theoretical results. The genotype to fitness relationship is depicted in the left panel. The phenotype (x) is a real-valued function of the genotype (i.e. the amino-acid sequence), and is defined in our model as the fraction of destabilizing amino acids in the sequence. Fitness is a decreasing log-concave function of the phenotype, depending on structural parameters of the model. Once the relation from genotype to fitness is defined, the substitution process proceeds as presented in the middle panel. For a given effective population size N_e, the evolutionary process results in an average value of the phenotype x^* and an average substitution rate (relative to the neutral rate) ω. Averaging over time is equivalent to determining the statistical equilibrium, by ergodicity of the stochastic process. The slope of the scaling of the equilibrium ω as a function of log-N_e defines the susceptibility χ, which is a function of the structural parameters defined by the phenotype-fitness map.*

In the original biophysical model, protein stability is determined by the difference in free energy between the folded and unfolded conformations, called ΔG and measured in kcal/mol. Technically, free energy is computed based on the 3D conformation of the protein and using statistical potentials. As a result, the stabilizing or destabilizing effect of an amino acid at a particular site depends on amino acids present in the vicinity in 3D conformation, thus implementing what has been called specific epistasis (Starr and Thornton, 2016).

Here, we approximate this model such that the (de-)stabilizing effect at a particular site, such as measured by the $\Delta\Delta G$ of the mutation, does not depend on other neighbouring residues, thus disregarding specific epistasis (Dasmeh *et al.*, 2014). Instead, each site contributes independently and additively to ΔG. In addition, we assume that, for

each site of the sequence, only one amino acid is stabilizing the protein. All 19 other amino acids are equally destabilizing. Each site bearing a destabilizing amino acid contributes an excess of $\Delta\Delta G > 0$ (in kcal/mol) to the total ΔG. The smallest achievable value of ΔG, obtained when all amino acids of the sequence are stabilizing, is noted $\Delta G_{\min} < 0$. In this model, the most succinct phenotype of a given genotype (i.e. sequence) is just the proportion of destabilizing amino acids in the sequence, defined as $0 \leq x \leq 1$. Thus ΔG is a linear function of x:

$$\Delta G(x) = \Delta G_{\min} + n\Delta\Delta G x, \tag{9.1}$$

where n is the number of sites in the sequence.

For a given ΔG, thermodynamic equations allow one to derive the proportion of protein molecules that are in the native (folded) conformation in the cytoplasm. This fraction is assumed to be a proxy for fitness, motivated in part by the fact that a protein must be folded to perform its function. A slightly different model will be considered below, in order to take into account protein expression level (see section 9.2.3).

Analytically, the fitness function is given by the Fermi Dirac distribution and is typically close to 1, leading to a first-order approximation (Goldstein, 2011):

$$W(x) = \frac{1}{1 + e^{\beta(\Delta G_{\min} + n\Delta\Delta G x)}}, \tag{9.2}$$

$$\Rightarrow W(x) \simeq 1 - e^{\beta(\Delta G_{\min} + n\Delta\Delta G x)}, \tag{9.3}$$

$$\Rightarrow f(x) = \ln(W(x)) \simeq e^{\beta(\Delta G_{\min} + n\Delta\Delta G x)}, \tag{9.4}$$

where W is the Wrightian fitness for a given phenotype and f is the Malthusian fitness (or log-fitness). Here, $\Delta G_{\min}$ and $\Delta\Delta G$ are defined as above, and the parameter β is 1.686 mol/kcal at $25°C$ (or 298.2K).

Of note, even though the phenotypic effect of a mutation at a given site does not depend on the amino-acids that are present at other sites (i.e. the trait is additive), the fitness effect of a mutation still depends on other sites (i.e. the log-fitness is not additive). As a result, the molecular evolutionary process is site-interdependent, a property referred to as non-specific epistasis (Starr and Thornton, 2016; Dasmeh and Serohijos, 2018).

9.2.2 Response of ω to changes in N_e. Analytical approximation

For a given effective population size N_e, the evolutionary process reaches an equilibrium (figure 9.1, middle panel). This substitution rate at this equilibrium, normalized by the substitution rate of neutral of mutations to discard the influence of the underlying mutation rate, is denoted ω. This relative rate can also be interpreted as the mean fixation probability of mutations scaled by the fixation probability of neutral alleles $p = 1/2N_e$, the mean being weighted by the probability of occurrence of mutations in the population. As a result, an $\omega < 1$ indicates that mutations are negatively selected on average, and ω decreases with increasing strength of purifying selection.

In this section we present an analytical approximate solution for the response of ω after a change in N_e (in log space), as depicted in the right panel of figure 9.1. We call this response the susceptibility of ω to changes in N_e, and denote it as χ:

$$\chi = \frac{d\omega}{d\ln(N_e)} \tag{9.5}$$

Deriving χ is done in two steps. First, we determine the mean phenotype at equilibrium, when evolutionary forces of mutation, selection and genetic drift compensate each other. Subsequently, differential calculus is used to compute the response of the equilibrium phenotype to a change in N_e, which allows us to ultimately derive an equation for χ. The main results of our derivation are given both in the general case of any (log-concave) phenotype-fitness map, and in the specific case of the biophysical model introduced above. A more detailed derivation is available in the supplementary materials.

For a given genotype, mutations can have various effects: they can increase or decrease the proportion of destabilizing amino acids, or do nothing if the mutation is between two destabilizing amino acids. To derive the probabilities of such events to occur, we also make the simplifying assumption that all transitions between amino acids are equiprobable. Altogether, any mutation in the sequence can then have a phenotypic effect of 0 or $\delta x = 1/n$, with probabilities of transitions equal to:

$$\begin{cases} \delta x & \text{with probability } 1-x, \\ 0 & \text{with probability } \frac{18x}{19}, \\ -\delta x & \text{with probability } \frac{x}{19}. \end{cases} \tag{9.6}$$

In the extreme case of an optimal phenotype ($x = 0$), only destabilizing mutations are proposed. Moreover, the probability to propose a stabilizing mutation (effect $-\delta x$), or a neutral mutation (effect 0), is proportional to x. Conversely, the probability to propose a destabilizing mutation is equal to $(1-x)$. As a result, the mutation bias is proportional to $(1-x)/x$. This mutation bias fundamentally reflects a combinatorial effect, due to the number of mutational opportunities available in either direction.

Second, we need to determine the strength of selection acting on mutations. Destabilizing mutations are selected against with a negative selection coefficient which can be approximated by:

$$s \simeq \frac{1}{n}\frac{\partial f(x)}{\partial x} \tag{9.7}$$

$$\Rightarrow s \simeq -\beta \Delta\Delta G e^{\beta(\Delta G_{\min} + n\Delta\Delta G x)}, \tag{9.8}$$

where $f = \ln(W)$ is the log-fitness (or Malthusian fitness). Conversely, stabilizing mutations will be under positive selection with opposite sign but same absolute value. It is important to realize that the selective effect is dependent on x. Furthermore, because the fitness function is log-concave, the absolute value of s increases with x.

Based on these expressions for the mutational and selective pressures, one can then study the trajectory followed by the evolutionary process. Starting from an optimal sequence, mostly destabilizing mutations will occur, some of which may reach fixation and

accumulate until selection coefficients against new deleterious mutations is too strong, at which point the protein will reach a point of equilibrium called marginal stability (Taverna and Goldstein, 2002; Bloom *et al.*, 2007). Most importantly, the probability of fixation of mutations is affected by genetic drift, and thus depends on the effective population size (N_e). At the equilibrium between mutation, selection and drift, the process fluctuates through the occurrence of advantageous and deleterious substitutions compensating each other. This equilibrium can be determined by expressing the constraint that the selection coefficient of substitutions is expected to be null on average (Goldstein, 2013). Formally, and after simplification, the equilibrium phenotype denoted x^* is given in the general case by:

$$\ln\left(\frac{1-x^*}{x^*}\right) + \ln(19) \simeq -\frac{4N_e}{n}\frac{\partial f(x^*)}{\partial x^*} \tag{9.9}$$

$$\Rightarrow \ln\left(\frac{1-x^*}{x^*}\right) + \ln(19) \simeq 4N_e\beta\Delta\Delta G e^{\beta(\Delta G_{\min}+n\Delta\Delta G x^*)}, \tag{9.10}$$

in the more specific case of the biophysical model. This equation essentially expresses the mutation-selection equilibrium: the left-hand side of the equation is the log of the mutation bias at x, while the right-hand side is simply $4N_e s$, the scaled selection coefficient.

This equation cannot be solved explicitly for x^*, but a qualitative intuition on the consequences of change in N_e to the equilibrium phenotype x^* is given in figure 9.2. Intuitively, an increase in N_e results in a more optimal phenotype, closer to 0. The mutation bias (left-hand side of equation 9.10) decreases with x while the strength of selection (right-hand side of equation 9.10) increases with x, and the equilibrium phenotype is obtained at their intersection. An increase in N_e leads to shifting the selective response upward, which then results in a leftward shift of the equilibrium phenotype (i.e. closer to 0). The leftward shift is smaller for selective strengths characterized by a steeper curve, resulting in qualitatively weaker susceptibility of the equilibrium phenotype to changes in N_e

The results obtained thus far only relate the equilibrium phenotype (x^*) to N_e. To capture how ω varies with N_e, we also need to obtain an expression for ω as a function of x^*. At equilibrium we can derive (supplementary materials) the expected substitution rate of mutations, and thus ω, which simply approximates to:

$$\omega \simeq x^* \tag{9.11}$$

This simple approximation is due to the fact that the substitutions between two destabilizing amino acids (which are neutral) compose the largest proportion of proposed mutations having a substantial probability of fixation (equation 9.6). In contrast, stabilizing mutations are rare, while destabilizing mutations have a low probability of fixation. Since there is a fraction x^* of sites already occupied by a destabilizing amino-acid, these neutral substitutions occur at rate x^*.

Combined together, these analytical approximations yield the susceptibility (equa-

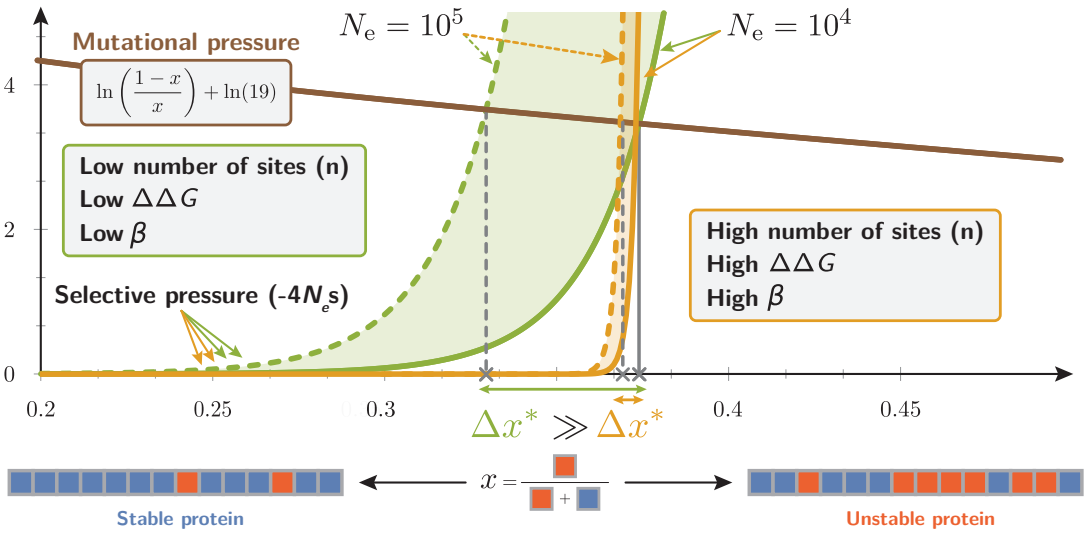


Figure 9.2: *Response of the equilibrium phenotype after a change in N_e. The equilibrium phenotype x^* is obtained when the selective pressure equals to the mutational pressure (equation 9.10). The selective pressure (right-hand side of eq. 9.10) increases exponentially with x where $\beta n\Delta\Delta G$ is the exponential growth rate (yellow and green curves). When $\beta n\Delta\Delta G$ is large, increasing N_e by an order of magnitude (yellow dotted curves) very moderately impacts the equilibrium phenotype (small Δx^*). In contrast, for small $\beta n\Delta\Delta G$ (green curves), the equilibrium phenotype is more strongly impacted by a change in N_e (large Δx^*). Finally, response of x^* to changes in N_e reflects the response of ω since both are approximately equal (equation 9.11).*

tion 9.5) of ω to a change in N_e:

$$\chi = \frac{\mathrm{d}\omega}{\mathrm{d}\ln(N_e)} \simeq -\frac{\frac{\partial f(x^*)}{\partial x^*}}{\frac{n}{4N_e}\frac{\partial \ln[(1-x^*)/x^*]}{\partial x^*} + \frac{\partial^2 f(x^*)}{\partial x^{*2}}}. \tag{9.12}$$

The two terms of the denominator correspond to the derivative of the mutational bias and the scaled selection coefficient, respectively. However, the mutational bias decreases weakly with x (blue curve on figure 9.2) while the strength of selection increases sharply with x (red and green curves). As a consequence, the derivative of the mutational bias is much lower than the derivative of the selection coefficient around the equilibrium point (i.e. the phenotype is *nearly* equimutable). The first term can therefore be ignored, which leads to a very compact equation for susceptibility χ in the general case:

$$\chi \simeq -\frac{\frac{\partial f(x^*)}{\partial x^*}}{\frac{\partial^2 f(x^*)}{\partial x^{*2}}} \tag{9.13}$$

The susceptibility is thus equal to the inverse of the relative curvature, i.e. the ratio of the second to the first derivatives, of the log-fitness function, taken at the equilibrium phenotype. Of note, this susceptibility is strictly negative for decreasing log-concave fitness functions, asserting that ω is a decreasing function of N_e. In addition, the susceptibility itself is low in absolute value (i.e. ω responds more weakly) for strongly concave log-fitness functions. This equation quantitatively captures the intuition developed in figure 9.2, namely that the response of ω is very weak if the selection curve is very steep around the equilibrium set point (red curve compared to green curve).

In the specific case of the biophysical model, the susceptibility (χ) further simplifies to:

$$\chi \simeq -\frac{1}{\beta n \Delta\Delta G}, \tag{9.14}$$

meaning that ω is linearly decreasing with N_e (in log scale) since χ is independent of x^*, or, in other words, that the exact value of the equilibrium phenotype has no impact on the slope. Moreover, only the compound parameter $\beta\Delta\Delta G n$ has an impact on the slope of the linear relationship. Thus, in particular, the slope of the linear relationship between ω and N_e is affected by $\Delta\Delta G$ but not by $\Delta G_{\min}$. Of note, empirically, only relative values of N_e (up to a multiplicative constant) are required to obtain an estimate of χ.

9.2.3 Response of ω to changes in protein expression level

Effective population size is not the sole predictor of ω, and expression level (or protein abundance) is also negatively correlated to ω. However, our previous model, which assumes that fitness is proportional to the folded fraction, and is thus independent of protein abundance, does not express the fact that selection is typically stronger for proteins characterized by higher levels of expression. An alternative biophysical model is to assume that each misfolded protein molecule has the same relative effect on fitness, caused by its toxicity for the cell (Drummond *et al.*, 2005; Wilke and Drummond, 2006; Drummond and Wilke, 2008; Serohijos *et al.*, 2012).

Our general derivation can be directly applied to this case. For a given protein with expression level y and a cost A representing the selective cost per misfolded molecule (positive constant), the fitness and selection coefficient can be defined as follows:

$$f(x) \simeq -Aye^{\beta(\Delta G_{\min} + n\Delta\Delta G x)} \tag{9.15}$$

$$\Rightarrow s \simeq -\beta\Delta\Delta G A y e^{\beta(\Delta G_{\min} + \Delta\Delta G n x)}. \tag{9.16}$$

Under this model, the total selective cost of a destabilizing mutation is now directly proportional to the total amount of misfolded proteins. This fitness function leads to the following expression for the mutation-selection-drift equilibrium:

$$\ln\left(\frac{1-x^*}{x^*}\right) + \ln(19) = 4N_e y A \beta \Delta\Delta G n e^{\beta(\Delta G_{\min} + \Delta\Delta G n x^*)}. \tag{9.17}$$

Importantly, in this equation, N_e and y are confounded factors appearing only as a product. This means that increasing either N_e or y leads to same change in equilibrium phenotype, and hence the same change in ω. In other words, the susceptibility of the response to changes in either N_e or expression level is the same:

$$\chi = \frac{\mathrm{d}\omega^*}{\mathrm{d}\ln(N_e)} = \frac{\mathrm{d}\omega^*}{\mathrm{d}\ln(y)} \simeq -\frac{1}{\beta n \Delta\Delta G}. \tag{9.18}$$

A similar result can be obtained under other models relating phenotype to fitness, for example if the selective cost is due to translational errors (supplementary materials). Alternatively if the protein is assumed to be regulated such as to reach a specific level of functional protein abundance under a general cost-benefit argument (Cherry, 2010;

Gout *et al.*, 2010), a multiplicative factor depending solely on the expression level is prefixed (supplementary materials). Altogether, we theoretically obtain the same linear decrease of ω with regards to either effective population size or expression level (in log space) under a broad variety of hypotheses.

9.2.4 Simulation experiments

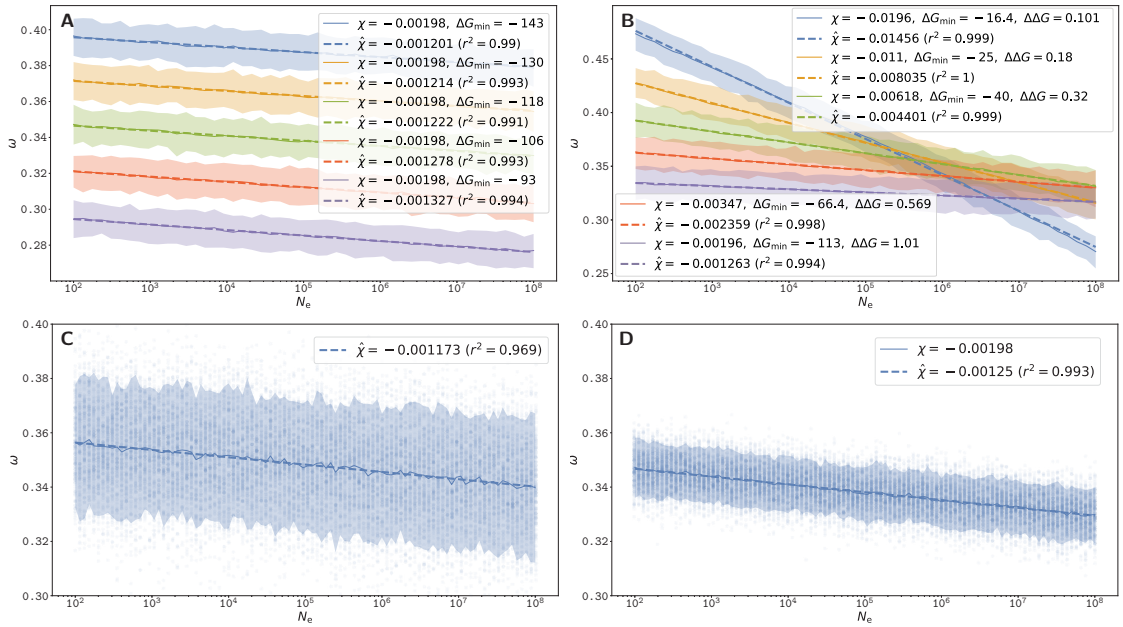


Figure 9.3: *Scaling of equilibrium ω as a function of log-Ne, under the additive phenotype model using Grantham distances (A,B,D) or the explicit biophysical model using a statistical potential (C), with $n = 300$ and $\beta = 1.686$. 200 replicates per N_e value are shown (dots). Solid lines are average over replicates, and shaded areas are 90% confidence interval. The slope (or susceptibility ($\hat{\chi}$), is estimated by linear regression (dashed lines). (A): ΔG_{min} are given in the legend, and $\Delta\Delta G = 1$. Decreasing ΔG_{min} (to more negative values) increases ω but does not impact the slope. (B): $\Delta\Delta G$ is increased and ΔG_{min} is changed accordingly such that the equilibrium value x^* is kept constant, by solving numerically equation 9.10. The estimated susceptibility ($\hat{\chi}$) decreases proportionally to the inverse of $\Delta\Delta G$, as predicted by our theoretical model. (C): Stability of the folded native state is computed using 3D structural conformations and pairwise contact potentials. (D): Additive model with $\Delta G_{min} = -118$ kcal/mol and $\Delta\Delta G = 1$ kcal/mol matches structural model shown in C (although with less variance).*

Our theoretical derivation of the susceptibility of ω to changes in N_e (and expression level) is based on several simplifying assumptions about the evolutionary model and makes multiple approximations. In order to test the robustness of our main result, we therefore conducted systematic simulation experiments, relaxing several of these assumptions. In each case, simulations were conducted under a broad range of values of N_e, monitoring the average ω observed at equilibrium and plotting the scaling of these measured equilibrium ω as a function of N_e.

Specifically, with respect to mutations, our derivation assumes that all amino-acid

transitions are equiprobable, or in other words, the complexity of the genetic code is not taken into account. Simulating evolution of DNA sequence and invoking a matrix of mutation rates between nucleotide allows us to test the robustness of our results to this assumption. Furthermore, with regard to the phenotypic effects of amino-acid changes, in our derivation, we assumed that all destabilizing amino acids have an identical impact on protein stability. In reality, one would expect conservative amino-acid replacements to be less destabilizing than radical changes. This assumption is relaxed in our simulation, such that destabilizing mutations in each position are now proportional to the Grantham distance (Grantham, 1974) between the optimal amino acid in this position and the amino acid proposed by the non-synonymous mutation. Finally, our derivation assumes that the number of sites in the sequence (n) is large, such that the selection coefficient is well approximated by the fitness derivative (equation 9.7). The robustness of this approximation was tested by conducting simulations with finite sequences of realistic length ($n = 300$ coding positions).

These simulation experiments demonstrate, first, that the relation between ω and log-N_e is indeed linear, at least in the range explored here, and that the slope of the linear regression matches the expected theoretical value (figure 9.3A). Secondly, we observe that the parameter $\Delta G_{\min}$ has virtually no effect on the slope of the linear regression, as also expected theoretically (figure 9.3B). Instead, decreasing $\Delta G_{\min}$ (to more negative values) merely results in an overall increase in ω over the whole range of N_e (i.e. has an impact on the intercept, not on the slope of the relation). This is due to the fact that decreasing $\Delta G_{\min}$ shifts the equilibrium to higher x^*, since more destabilizing sites can then reach fixation before reaching the point of marginal stability.

Finally, we relaxed our assumption that each site of the sequence contributes independently to ΔG, by taking into account the 3D structure of protein and using a statistical potential to estimate ΔG (supplementary materials). We implemented the original model considered in Williams *et al.* (2006), Goldstein (2011) and Pollock *et al.* (2012), in which the free energy is computed based on the 3D conformation using pairwise contact potential energies between neighbouring amino-acid residues (Miyazawa and Jernigan, 1985). The original works showed that under this model, ω is approximately independent of N_e (Goldstein, 2013). Using extensive simulations in order to obtain sufficient resolution, we observe that ω is in fact weakly dependent on N_e, being again approximately linear with log-N_e (figure 9.3C). Moreover, the observed slope ($\widehat{\chi} = -0.00117$) matches the slope obtained under the model of additive ΔG ($\widehat{\chi} = -0.00125$, figure 9.3D), considering an empirical $\Delta\Delta G = 1.0$ kcal/mol for destabilizing mutations and $n = 300$. In this experiment (figure 9.3D), $\Delta G_{\min}$ was set to -118 kcal/mol, which is the ΔG of the optimal (maximally stable) sequence of 300 sites (Goldstein, 2011).

9.2.5 Time to relaxation

Although the equilibrium value of ω after changes in N_e is an important feature of the ω-N_e relationship, another characteristic that is scarcely studied is the dynamic as-

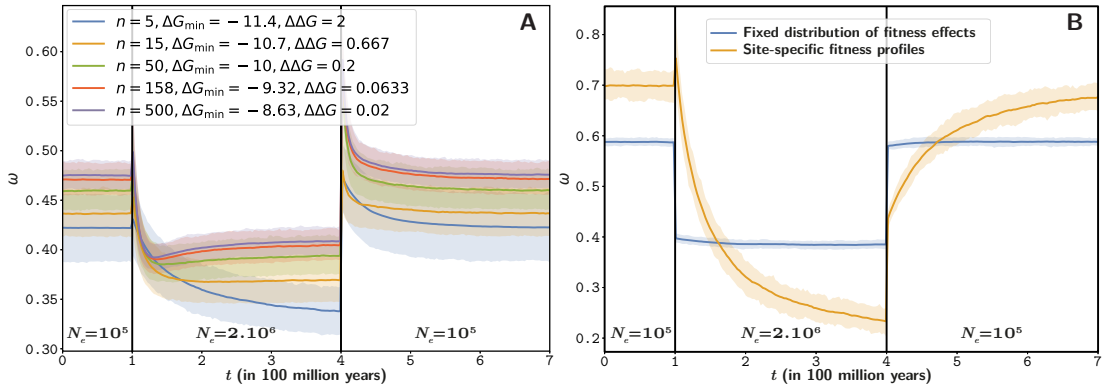


Figure 9.4: *Relaxation of ω after a change in N_e Solid line corresponds to the average over* 1000 *replicates and the shaded area corresponds to the* 90% *interval among replicates. The mutation rate (μ) is* 1e−8 *per year per site, and the total evolutionary period is 700 million years. (A): $\beta = 1.686$ for all simulations. The DNA sequence of* 500 *sites is divided into exons of equal size. However the number of sites per exon changes between simulations from $n = 5$ to $n = 500$. Moreover, $\Delta\Delta G$ is changed according to the exon size such that $n\Delta\Delta G$ (and as a result, the susceptibility) are kept constant, and ΔG_{min} is changed accordingly such that the equilibrium value x^* is kept constant, by solving numerically equation 9.10. Thus, regardless of exon size, x^* and χ are kept constant and thus the observed effect is due to the number of sites in the exon. We observe that increasing the number of sites leads to a reduced time to reach the new equilibrium. (B): In the context of a time-independent fitness landscape (yellow curve), where each amino acid has different fitness (site-specific profiles), the time taken to reach the new equilibrium value of ω after a change in N_e is long. In the context of a fixed distribution of fitness effects (blue curve), the relaxation time is non-existent and the new equilibrium value of ω is reached instantaneously.*

pect (Jones *et al.*, 2016), particularly the relaxation time to reach the new equilibrium ω. We observed in our simulations that the determining factor of the relaxation time is the number of sites n (figure 9.4A), such that the return to equilibrium is faster for longer sequences. This observation matches the theoretical prediction that more mutational opportunities are available for longer sequences, driving the trait close to equilibrium at a faster rate.

It may be useful to compare the relaxation pattern observed here with the predictions under two alternative models of sequence evolution, representing two extreme scenarios. On one hand, having fitness modelled at the level of sites, such as contemplated by many phylogenetic mutation-selection models (Halpern and Bruno, 1998; Rodrigue *et al.*, 2010; Tamuri and Goldstein, 2012), leads to a situation where every site has to adapt on its own to the new change in N_e. The relaxation time is then very long, on the order of the inverse of the per-site substitution rate. On the other hand, assuming a fixed distribution of fitness effect (DFE) as in Welch *et al.* (2008), the response of ω is instantaneous (figure 9.4B). Our model is effectively in between these two extreme scenarios.

Another characteristic observed in these non-equilibrium experiments is the discontinuity of ω after a change in N_e. Most importantly, both an increase and decrease in N_e lead to a discontinuity (figure 9.4A & 9.4B). These non-equilibrium behaviors can both be explained mechanistically. Under low N_e, the phenotype is far away from the

optimal phenotype because the efficacy of selection is weaker. A sudden increase in N_e results first in a short traction toward a more optimal phenotype, which results in a suddenly higher ω, caused by a transient adaptation of the protein toward a higher stability. Conversely, under high N_e the phenotype is closer to optimal and the purification of deleterious mutations is stronger. The reaction to a decrease in N_e is a relaxation of the purification and thus an ω closer to the neutral case, which results into higher ω until reaching the point of marginal stability. To note, an increase in N_e can theoretically and possibly lead to an ω that is temporarily greater than 1 due to adaptive evolution (Jones *et al.*, 2016), while a decrease in N_e always imply an $\omega < 1$, as it gives at most a neutral regime of relaxed selection.

9.3 Discussion

We provide a compact analytical result for the equilibrium response (which, by analogy with thermodynamics, we call the susceptibility) of ω to changes in N_e, and we relate this response to the parameterization of the genotype-phenotype-fitness map. An application to a model of selection against protein misfolding shows that the response of ω to variation in N_e (in log space) is linear, with a negative slope. Furthermore, this application demonstrates that effective population size and protein expression level are interchangeable with respect to their impact on the response of ω. Our compact theoretical results, which were obtained by making several simplifying assumptions, are supported by more complex simulations of protein evolution relaxing these assumptions. In particular, our theoretical predictions are verified under a numerical model of protein evolution in which the free energy is computed based on the 3D structure.

Overall, the susceptibility (χ) is a function of the structural parameters of the protein and takes a very simple analytical form, being inversely proportional to the product of three terms: the sequence size, the inverse temperature (β), and the average change in conformational energy of destabilizing mutations ($\Delta\Delta G$). Quantitatively, this product can be several orders of magnitude greater than 1 in practice, such that the susceptibility of ω, which is its inverse, is typically small. Previous studies using this model presented an apparent lack of response of ω to changes in N_e (Goldstein, 2013). We refine this result, by observing that there is in fact a very subtle and weak relation, which requires extensive computation to be detected, but which is well predicted by our theoretical derivation. Based on empirical estimates of the structural parameters $\beta = 1.686$, $n = 300$ sites and $\Delta\Delta G = 1.0$ kcal/mol for destabilizing mutations (Zeldovich *et al.*, 2007), the estimated susceptibility is $\hat{\chi} \simeq -0.002$. In other words, for a relative increase in N_e or expression level of 6 orders of magnitude, a factor approximately equal to 0.01 is subtracted from ω, a subtle relationship that requires laborious effort to be detected in simulated data.

9.3.1 Adequacy to empirical data

Empirically, variation in ω along the branches of phylogenetic trees has been inferred and correlated to proxies of N_e, such as body size or other life-history traits. These analyses showed mitigated support for a negative relation between ω and N_e (Lanfear *et al.*, 2014). More recently, phylogenetic integrative methods refined the estimate of covariation between ω and N_e along lineages by leveraging polymorphism data (Brevet and Lartillot, 2019). This approach gives an estimate of $\hat{\chi} \simeq 0.02$ in primates (supplementary materials) at least one order of magnitude greater than the quantitative estimate obtained above from the biophysical model. More empirical data across different clades would be required to robustly consolidate such empirical estimates, but as of yet, these results are challenging the idea of a very weak response.

The relation between ω and expression level provides an independent, and potentially more robust, source of empirical observation. Our theoretical results suggest that, under relatively general conditions, the response of ω to expression level should be of the same magnitude than the response to N_e. Empirically, the protein expression level is one of the best predictors of ω and the empirical estimation of χ in fungi, archaea and bacteria varies in the range $[-0.046; -0.021]$ (supplementary materials) extracted from Zhang and Yang (2015). Estimation in animals and plants gives somewhat lower estimates, in the range of $[-0.026; -0.004]$, although still higher (in absolute value) than -0.002.

Additionally, another empirical observation is the negative relation between the mean destabilizing effect of mutations (mean $\Delta\Delta G$) and the ΔG of the protein. Such a relation is empirically observed in Serohijos *et al.* (2012), where the slope of the linear regression is -0.13 ($r^2 = 0.04$). The slope of the linear correlation observed in our simulations is weaker, with an observed slope of -0.01 ($r^2 = 0.29$) under the 3D biophysical model, and -0.003 ($r^2 = 0.33$) under the model of additive phenotype parameterized by $\Delta\Delta G = 1$ and $n = 300$ (supplementary materials). This observation also sheds light on the correlation between ω and N_e in empirical data and in our model. Indeed, equimutability, or namely that the distribution of $\Delta\Delta G$ of mutations is independent of ΔG is a necessary condition to observe independence between ω and N_e (Cherry, 1998). In our model, the average $\Delta\Delta G$ of mutations at equilibrium depends on ΔG due to combinatorial considerations, but this dependence is weaker than empirically observed, which also translates into a weaker susceptibility of ω to changes in N_e or expression level than empirically observed.

Thus, overall, the response of ω to either N_e or expression level predicted by the biophysical model considered above seems lower than what is empirically observed. There are several possible explanations for this discrepancy. First, the biophysical model might be valid, but the numerical estimates used for n or $\Delta\Delta G$ could be inadequate. A $\Delta\Delta G$ of 1.0 kcal/mol for destabilizing mutations seems to correspond to empirical estimates (Zeldovich *et al.*, 2007). On the other hand, the effective number of positions implicated in the trait might be smaller than the total number of residues in the protein. In our model, all positions in the protein can in principle compensate for the destabilizing effect of a mutation at a particular position. In practice, the number of sites susceptible to compensate

each other is probably smaller, resulting in a stronger departure from equimutability.

Alternatively, the biophysical model considered here might be too restrictive. Recent empirical studies have provided evidence against the hypothesis that the rate of sequence evolution is driven solely by the toxicity effect of unfolded proteins (Plata and Vitkup, 2017; Razban, 2019; Biesiadecka *et al.*, 2020). Notably, the response of ω to changes in expression level has also been found theoretically to arise as a consequence of protein-protein interactions, where protein may either be in free form or engaged in non-specific interactions (Yang *et al.*, 2012; Zhang *et al.*, 2013). In non-specific interactions at the protein surface, stabilizing amino acids are hydrophilic and destabilizing amino acids are hydrophobic, sticking to hydrophobic residues at the surface of other proteins (Dixit and Maslov, 2013; Manhart and Morozov, 2015a).

Our theoretical results can be applied more broadly to protein-protein interactions using a mean-field argument (supplementary materials). Fitting this model with empirical structural estimates (Janin, 1995; Zhang *et al.*, 2008), we obtain a susceptibility of $\chi \simeq -0.2$ thus a much stronger response than under the model based on conformational stability. This much stronger response is due to fewer sites in the protein being involved in protein-protein interaction than for conformational stability, in addition to a lower free energy engaged in contact between residues.

Altogether, fitness based on protein stability is a compelling model of molecular evolution, but may not be a sufficiently comprehensive model to explain the amplitude of variation of ω empirically observed along a gradient of either effective population size or protein expression level. The net response of ω to changes in N_e or expression level could have several biophysical causes, which in the end would imply a weak but still empirically measurable response.

9.3.2 The statistical mechanics of molecular evolution

This study describes the signature imprinted on DNA sequences by an evolutionary process by merging equations from population genetics and from structural physicochemical first principles. More generally, it outlines a general approach for deriving quantitative predictions about the observable macroscopic properties of the molecular evolutionary process based on an underlying microscopic model of the detailed relation between sequence, phenotype and fitness. In this respect, it borrows from statistical mechanics, attempting approximations to derive analytically tractable results (Sella and Hirsh, 2005; Mustonen and Lässig, 2009; Bastolla *et al.*, 2012, 2017) The robustness of results can be assessed by computational implementations and simulations. Computational models offer a means to test the validity and robustness, while mathematical models offer an intuitive mechanistic mental analogy.

Ultimately, the approach could be generalized to other aspects of the evolutionary process. Beyond ω, other macroscopic observables could be of interest, for example site entropy, i.e. the effective number of observed amino acids per site at equilibrium (Goldstein and Pollock, 2016; Jimenez *et al.*, 2018; Jiang *et al.*, 2018), or the nucleotide or

amino-acid composition. In addition to N_e, other evolutionary forces could also be considered, for instance the mutational bias or GC-biased gene conversion. The susceptibility of the macroscopic observables to changes in the strength of these underlying forces could then more generally be investigated. As such, the framework outlined here could foster a better understanding of observable signatures of the long-term evolutionary process emerging from ecological parameters and molecular physico-chemical first principles, by carefully teasing out the combined effects of mutation, selection and drift.

9.4 Materials & Methods

Protein sequence evolution is simulated under an origin-fixation model (McCandlish and Stoltzfus, 2014), i.e. the whole population is considered monomorphic and only the succession of fixation events is modeled. Given the currently fixed sequence $\mathbb{S}$, we define $\mathcal{M}(\mathbb{S})$ as the set of all possible mutant that are one nucleotide away from $\mathbb{S}$. Non-sense mutants are not considered. For a protein of n amino-acid sites, $|\mathcal{M}(\mathbb{S})| \leq 9n$, since each codon has a maximum of 9 possible nearest neighbors that are not stop codons. For each mutant sequence $\mathbb{S}' \in \mathcal{M}(\mathbb{S})$, we compute its fitness and subsequently the selection coefficient of the mutant:

$$s(\mathbb{S}, \mathbb{S}') = \frac{W(\mathbb{S}') - W(\mathbb{S})}{W(\mathbb{S})}, \tag{9.19}$$

$$\Rightarrow s(\mathbb{S}, \mathbb{S}') \simeq f(\mathbb{S}') - f(\mathbb{S}'), \tag{9.20}$$

where W is the Wrightian fitness for a given phenotype and f is the Malthusian fitness (or log-fitness).

The waiting time before the next mutant invading the population, and the specific mutation involved in this event, are chosen using Gillespie's algorithm (Gillespie, 1977), according to the rates of substitution between $\mathbb{S}$ and each $\mathbb{S}' \in \mathcal{M}(\mathbb{S})$, which are given by:

$$Q_{\mathbb{S},\mathbb{S}'} = \mu_{\mathbb{S},\mathbb{S}'} \frac{4N_e s(\mathbb{S}, \mathbb{S}')}{1 - e^{-4N_e s(\mathbb{S},\mathbb{S}')}}, \tag{9.21}$$

where $\mu_{\mathbb{S},\mathbb{S}'}$ is the mutation rate between $\mathbb{S}$ and $\mathbb{S}'$, determined by the underlying $4x4$ nucleotide mutation rate matrix, and $Q_{\mathbb{S},\mathbb{S}'} = \mu_{\mathbb{S},\mathbb{S}'}$ in the case of synonymous substitutions. Various optimizations are implemented to reduce the computation time of mutant fitness. The simulation starts with a burn-in period to reach mutation-selection-drift equilibrium.

9.4.1 Models of the fitness function

Under the additive model for the free energy, the difference in free energy between folded and unfolded state is assumed to be given by:

$$\Delta G(\mathbb{S}) = \Delta G_{\min} + n\Delta\Delta G * x(\mathbb{S}),$$

where $0 \leq x(\mathbb{S}) \leq 1$ is the distance of $\mathbb{S}$ to the optimal sequence (i.e. the fraction of sites occupied by a destabilizing amino-acid). For each site of the sequence, the optimal

amino acids are chosen randomly at initialization, and the distance between the current amino acid and the optimal is scaled by the Grantham amino-acid distance (Grantham, 1974). The Wrightian fitness is defined as the probability of our protein to be in the folded state, given by the Fermi-Dirac distribution:

$$W(\mathbb{S}) = \frac{e^{-\beta \Delta G(\mathbb{S})}}{1 + e^{-\beta \Delta G(\mathbb{S})}} = \frac{1}{1 + e^{\beta \Delta G(\mathbb{S})}}, \tag{9.22}$$

where β is the inverse of the temperature ($\beta = 1/kT$).

For simulations under a 3D model of protein conformations, we adapted the model developed in Goldstein and Pollock (2017) to our C++ simulator (see supplementary materials).

For simulations under a site-independent fitness landscape, with site-specific fitness profiles, the protein log-fitness is computed as the sum of amino-acid log-fitness coefficients along the sequence. In this model, each codon site i has its own fitness profile, denoted $\boldsymbol{\phi}^{(i)} = \{\phi_{\mathrm{a}}^{(i)},\ 1 \leq \mathrm{a} \leq 20\}$, a vector of 20 amino-acid scaled (Wrightian) fitness coefficients. Since $\mathbb{S}[i]$ is the codon at site i, the encoded amino acid is $\mathcal{A}(\mathbb{S}[i])$, hence the fitness at site i is $\boldsymbol{\phi}^{(i)}_{\mathcal{A}(\mathbb{S}[i])}$. Altogether, the selection coefficient of the mutant $\mathbb{S}'$ is:

$$s(\mathbb{S}, \mathbb{S}') = \sum_{i=1}^{n} \ln \left(\frac{\boldsymbol{\phi}^{(i)}_{\mathcal{A}(\mathbb{S}'[i])}}{\boldsymbol{\phi}^{(i)}_{\mathcal{A}(\mathbb{S}[i])}} \right), \tag{9.23}$$

The fitness vectors $\boldsymbol{\phi}^{(i)}$ used in this study are extracted from Bloom (2017). They were experimentally determined by deep mutational scanning.

For simulations assuming a fixed distribution of fitness effects (DFE), the selection coefficient of the mutant $\mathbb{S}'$ is gamma distributed (shape $k > 0$):

$$-s(\mathbb{S}, \mathbb{S}') \sim \mathrm{Gamma}\left(|\bar{s}|, k\right) \tag{9.24}$$

9.4.2 Computing ω along the simulation

From the set of mutants $\mathcal{M}(\mathbb{S})$ that are one nucleotide away from $\mathbb{S}$, we define the subsets $\mathcal{N}(\mathbb{S})$ of non-synonymous and synonymous mutants ($\mathcal{N}(\mathbb{S}) \subseteq \mathcal{M}(\mathbb{S})$). The ratio of non-synonymous over synonymous substitution rates, given the sequence $\mathbb{S}$ at time t is defined as (Spielman and Wilke, 2015; Dos Reis, 2015; Jones *et al.*, 2016):

$$\omega(t) = \frac{\sum\limits_{\mathbb{S}' \in \mathcal{N}(\mathbb{S})} \mu_{\mathbb{S},\mathbb{S}'} \dfrac{4N_{\mathrm{e}} s(\mathbb{S}, \mathbb{S}')}{1 - \mathrm{e}^{-4N_{\mathrm{e}}(\mathbb{S}, \mathbb{S}')}}}{\sum\limits_{\mathbb{S}' \in \mathcal{N}(\mathbb{S})} \mu_{\mathbb{S},\mathbb{S}'}} \tag{9.25}$$

Averaged over all branches of the tree, the average ω is:

$$\omega = \langle \omega(t) \rangle, \tag{9.26}$$

$$= \int_t \omega(t) \mathrm{d}t, \tag{9.27}$$

where the integral is taken over all branches of the tree, while the integrand $\omega(t)$ is a piece-wise function changing after every point substitution event.

9.5 Reproducibility - Supplementary Materials

The mathematical developments under the general case of an arbitrary additive trait and an arbitrary log-concave fitness function, and the derived susceptibility under various fitness functions, as well as supplementary figures, are available in appendix, chapter 12. The scripts and instructions necessary to reproduce this study are available at `https://github.com/ThibaultLatrille/GenotypePhenotypeFitness`.

9.6 Author contributions

TL gathered and formatted the data, developed the new models in `SimuEvol` and conducted all analyses, in the context of a PhD work (Ecole Normale Superieure de Lyon). TL and NL both contributed to the writing of the manuscript.

9.7 Acknowledgements

We wish to thank Julien Joseph for whiteboard mathematical sessions. We gratefully acknowledge the help of Nicolas Rodrigue and Laurent Duret for their input on this work and their comments on the manuscript. This work was performed using the computing facilities of the CC LBBE/PRABI.

Part III

Conclusion

10 Discussion & perspectives

Contents

As a legacy of the nearly-neutral theory, the evolution of molecular sequences is seen as a stochastic process. One component of this process is creating diversity through mutation, while an antagonistic component is filtering out this diversity through selection, and finally the balance between these components is arbitrated by drift. In the long term, this stochastic process results in a history of substitution events along species trees, inducing complex patterns of molecular divergence between species. By analysing them, phylogenetic codon models aim at capturing the intrinsic parameters of evolution.

The focus of this thesis has been the development of new phylogenetic codon models and the modelling of the interplay between mutation, selection and drift in the evolutionary processes followed by protein-coding DNA sequences. In this conclusive chapter, I first recall the main results of this thesis in section 10.1. Subsequently, I attempt to discuss the limitations of my work. One main limitation concerns the problem of modelling site interdependence, which is discussed in section 10.2. Secondly, in section 10.3, I draw upon some important connections between the mechanistic models developed here and the problem of detecting adaptive evolutionary regimes. As a perspective, I discuss how phylogenetic mechanistic models could be unified with population genetics in sec-

tion 10.5, and the inference methodology that would be adapted to such an endeavour in section 10.5. Finally, before some concluding remarks, I discuss the question and the issue of reproducible sciences in evolutionary biology in section 10.6.

10.1 Summary of main results

In chapter 7, I developed a phenomenological codon model in which ω is seen not as a single parameter but as a tensor (95 free parameters). This sensor captures the small differences in fixation rate (or ω) in different directions, which gives an accurate representation of how mutation and selection oppose each other at equilibrium. This parameterization is the simplest one, in a phenomenological context, capable of correctly teasing apart mutation and selection. Thanks to this, this modelling approach yields a reliable estimate of the mutational process, while disentangling fixation probabilities in different directions.

In chapter 8, I developed an extended mechanistic mutation-selection model reconstructing site-specific fitness landscapes, long-term trends in effective population size and in the mutation rate along the phylogeny, from an alignment of DNA coding sequences. Simultaneously, the approach estimates the correlation between life-history traits, mutation rate and effective population size, intrinsically accounting for phylogenetic inertia. Our framework was tested against simulated data and then applied to empirical data in mammals, isopods, primates and Drosophila. Simulated and empirical evidence suggest that there is a persistent signal in substitution patterns that relates to the past history of N_e, whose trends correspond to the expected direction of correlation with life-history traits or ecological variables. However, the magnitude of inferred variation in N_e across the phylogeny is narrower than expected, which is probably a bias of the approach caused by the assumptions made on the structure of the fitness landscape.

As a way to further investigate this last question, the third manuscript in chapter 9 revisits the question of how the exact structure of the fitness landscape determines the quantitative response of the molecular evolutionary process, and in particular of d_N/d_S, to changes in N_e and in protein expression levels. Specifically, I derive a theoretical approximation for the quantitative response of d_N/d_S to changes in both N_e and expression level, under an explicit genotype-phenotype-fitness map. The development is generally valid for an additive trait under a log-concave fitness function, but was applied more specifically to a biophysical model in which proteins are under directional selection for maximizing their conformational stability. In this specific case, I predict a weak response of d_N/d_S to changes in either N_e or expression level (which are interchangeable), a result corroborated by simulations under more complex models. Based on this, I propose that fitness based on conformational stability might not provide a sufficient mechanism to explain the amplitude of the variation in the mean fixation probability which is observed empirically. Other aspects of protein biophysics could be explored such as protein-protein interactions, which could lead to a stronger response of the mean fixation probability to changes in N_e.

More globally, there is a remaining gap between quantitative predictions of biophys-

ical models and empirical observations relating the response of protein coding sequence evolution to changes in N_e and expression level.

10.2 Site interdependence and epistasis

One of the blind spots of the mechanistic codon model developed in chapter 8, and more generally of current mechanistic models of the Halpern-Bruno family (see section 3.3.1), is the assumption of site independence. This assumption is convenient, both computationally and statistically. Computationally, each site can be considered as an independent Markov process (see sections 2.2.1 and 4.1). Statistically, one can rely on mixture models to estimate site-specific amino-acid fitness profiles (see section 3.3.2) In contrast, from a modelling and inference perspective, accounting for epistasis is challenging both in terms of parametrization and in terms of computational complexity (see section 5.3.2) This complexity is the main reason why epistasis is generally ignored in phylogenetic models, and more particularly in codon models. Empirically, however, evolutionary biologists have many reasons to believe that this hypothesis of site independence is not adequate, especially given our knowledge about protein biophysics (see section 5.1.3). This approximation is therefore problematic, and raises multiple questions. What are the consequences of ignoring epistasis in the context of phylogenetic inference? Practically, how could we ultimately account for epistasis in the context of inference?

Previous studies have argued that models of molecular evolution should consider the importance of epistasis for its different roles: from its importance in speciation, in modulating the rate of adaptation, in interlocking between sites (Stokes shift), its downward impact on the d_N/d_S predicted by the mutation-selection models, and many other factors Goldstein and Pollock (2017); Miller *et al.* (2018). Based on our analysis presented in chapter 9, we argue that epistasis also has an important role in the response of d_N/d_S to changes in N_e, both in terms of its susceptibility and dynamics of the response. This is a conceptual point that, to my knowledge, had never been really identified until now.

More precisely, one key result of chapter 9 is that any model without (or ignoring) epistasis implies slow dynamics and a strong sensitivity of the mean fixation probability to changes in N_e. Intuitively, a model without epistasis exhibits a slow return to equilibrium upon a change in N_e due to the waiting time until the next substitution. Indeed, the evolutionary process is mainly mutation limited (see section 1.4.5) In addition, the mutation rate per site is very low, from 10^{-8} to 10^{-9} in mammals. As a result, for each site, the expected waiting time until the next mutation is between 100 to 1000 million years. As for the strong sensitivity of the mean fixation probability to changes in N_e, it originates in the fact that after a change in N_e, each site of the sequence has to adapt independently, and change its position in the fitness landscape.

In contrast, in the presence of epistasis, the burden of adapting to changes in N_e is shared by more sites, such that not all of them (and possibly, very few of them) have to switch their position in the fitness landscape, in order for the trait to return to equilibrium under the new N_e. As a result, adding epistasis to the model implies faster dynamics

and a weaker response of the mean fixation probability to changes in N_e.

These observations have several implications for empirical analyses. First, if epistasis implies a weak response of the d_N/d_S to changes in N_e, such as observed in chapter 9, for similar reasons, it may also explain the low magnitude of N_e variation estimated with site-specific mechanistic codon models in chapter 8. Empirically, it appears that the susceptibility of d_N/d_S to changes in N_e is between these two extremes, namely that of site-specific fitness landscapes, and the other extreme of a single univariate phenotype controlled by all sites. More probably, the ternary relation from sequence to phenotype to fitness implies several sites, but not all sites of the sequences, in a given phenotypic trait.

Second, the very long relaxation time implied by site-specific models are of the order of the depth of phylogeny (100 to 1000 My). As such, in the absence of epistasis, we should not even see d_N/d_S correlations with either LHTs or N_e because of it. Conversely, the fact that we see it is in itself an important indication of the presence of epistasis.

Ultimately, accounting for epistasis in mechanistic models of evolution is necessary but challenging from a computational and statistical perspective. Paths for statistical methods that can account for it are developed in section 10.5.

10.3 Adaptive landscape and positive selection

Another blind spot the mechanistic phylogenetic codon model is the absence of adaptive evolution (see section 1.4.4). Indeed, the mutation-selection equilibrium is essentially a nearly-neutral regime. As a result, at mutation-selection equilibrium, the sequence is close to the fitness optimum and therefore, most mutations are deleterious or compensate for previous deleterious mutations that reached fixation. Adaptation, on the other hand, can be seen as a process where the underlying fitness landscape is not fixed (i.e. time-independent) but is instead dynamic (i.e. time-dependent). In other words, it is not so much a fitness landscape than fitness seascapes (Mustonen and Lässig, 2009). Under a fitness seascape, the sequence is constantly running after a moving target, and as a result, there is net flux of adaptive substitutions (see section 3.4.3).

10.3.1 Mechanistic mutation-selection models under fitness seascapes

In the context of a mutation-selection framework, explicitly modelling adaptation in terms of fitness seascapes fluctuating along the phylogeny appears to be challenging. A first direction is to consider that adaptation consists in changes in the site-specific fitness profiles along some lineages. These changes can either be informed by experimental mutational scanning (Bloom, 2017), or estimated using a priori knowledge of phenotypic changes or ecological shifts (Tamuri *et al.*, 2009; Parto and Lartillot, 2017, 2018). Without knowledge of the drivers of adaptation, modulations of the fitness profiles through time could also be implemented as a Markov modulated process along the phylogeny. However, such an endeavour would be statistically and computationally challenging and might

require to first rethink the entire statistical approach (see section 10.5).

10.3.2 Hybrid mechanistic and phenomenological mutation-selection models

As an alternative to explicit models of adaptation through fitness seascapes, the current nearly-neutral mutation-selection framework can be leveraged as a null model. Deviation from this null model can be seen as a signal of adaptation (see section 3.4.3). In particular, if the sequences are under recurrent positive selection, the mean fixation probability (ν) of non-synonymous mutations will tend to be higher than predicted by the purely nearly-neutral model. This discrepancy between the mean fixation probability and the nearly-neutral expectation can be captured by a deviation parameter ω_*, as in Rodrigue and Lartillot (2016).

Thus far, however, this idea has been implemented only at the gene level (i.e. invoking a single ω_* for the whole gene). Yet, adaptation often occurs more locally, in specific domains of the protein. Accordingly, this gene-wide deviation parameter ω_* could be refined at the site-level. In this direction, in a work led by Nicolas Rodrigue, we developed a method to detect site-specific adaptation as a deviation from a null nearly-neutral model of evolution. This method has been developed in the generic programming environment provided by `BayesCode` (see section 4.2.7), and the manuscript accepted in the journal *Molecular Biology & Evolution* is available in the Appendix (page 217). These methods are relatively new, and must still be validated, more broadly applied to empirical data, and their predictions more extensively compared with those obtained from with classical codon models.

However, in its current form, phylogenetic models seeking adaptation as a deviation from near-neutrality assume a constant N_e across the tree. As already discussed, this assumption is of course not reasonable. A simple solution to this problem would be to add a deviation parameter ω_* in the mechanistic model developed in chapter 8. The resulting model would potentially be more effective at detecting positive selection.

Interestingly, ω_* could also be allowed to vary across branches, like N_e. This could have useful applications. For instance, bats are known to be reservoirs of pathogens. Recent results suggest they have a more efficient immune system (Baker *et al.*, 2013; Pavlovich *et al.*, 2018), such that proteins involved in host-pathogen interactions are under positive selection (Hawkins *et al.*, 2019; Vandewege *et al.*, 2020). But also, bats have large population sizes, compared to other mammals, and thus more efficient purifying selection. These two factors have opposing effects on d_N/d_S, and teasing them out might therefore be difficult using classical codon models. In contrast, the mutation-selection model with both N_e and ω_* varying across lineages could offer a way to estimate them separately, which would allow to address the problems mentioned above. In particular, this would allow us to answer if bats, compared to other mammals, have both stronger purifying selection (lower d_N/d_S) due to large N_e and stronger positive selection (larger ω_*).

10.3.3 Detecting adaptation with polymorphism

Phylogenetic codon models are only one of the methods currently available to detect adaptation. There are other approaches that are widely used in population genetics, and that make use of the signal contained in polymorphism data, such as originally pioneered by McDonald and Kreitman (1991). The idea behind the McDonald & Kreitman (MK) approach is to decompose the rate of selected substitutions (ω_{Tot}), as a mixture of both advantageous substitutions and non-adaptive (nearly-neutral) substitutions:

$$\omega_{\text{Tot}} = \omega_{\text{A}} + \omega_{\text{NA}}, \tag{10.1}$$

$$\iff \omega_{\text{A}} = \omega_{\text{Tot}} - \omega_{\text{NA}}, \tag{10.2}$$

where ω_{NA} is the rate of substitutions contributed by non-adaptive nearly-neutral evolution and ω_{A} is the total rate of substitutions contributed by adaptive evolution. In this context, under the assumption that adaptive mutations are rare, the ratio of non-synonymous over synonymous polymorphisms (π_N/π_S) mostly contains non-adaptive polymorphisms and is a measure of ω_{NA}:

$$\omega_{\text{NA}} \simeq \pi_N/\pi_S. \tag{10.3}$$

Moreover, the total ratio of non-synonymous over synonymous substitutions (d_N/d_S), estimated from divergence data, is a measure of ω_{Tot}:

$$\omega_{\text{Tot}} \simeq d_N/d_S. \tag{10.4}$$

Altogether, the rate of adaptive evolution, which contributes disproportionately to divergence is estimated as the difference between divergence and polymorphism:

$$\omega_{\text{A}} \simeq d_N/d_S - \pi_N/\pi_S. \tag{10.5}$$

However, estimation of the non-adaptive rate through π_N/π_S can be biased by moderately deleterious mutations and by the change in population size through time (Eyre-Walker, 2002). To overcome these biases, a method initially proposed by Eyre-Walker and Keightley (2009); Galtier (2016) relies on the synonymous and non-synonymous site-frequency spectra (SFS) to correct for demography and to estimate the distribution of fitness effects of mutations (DFE), modelled as a continuous distribution. This method, and subsequent developments which are reviewed in Moutinho *et al.* (2019b), provide more reliable estimates of ω_{NA} and, as a result, a better estimate of ω_{A}.

10.3.4 Confronting methods for detecting adaptation

The availability of independent phylogenetic methods (based on either phenomenological and mechanistic codon models) and population genetics approaches (using the McDonald & Kreitman ideas) for detecting adaptation raises the question whether they detect congruent signals of adaptation.

Empirically, phenomenological and mechanistic codon methods should be confronted to McDonald & Kreitman (MK) methods. In the case of the overlap between positively selected genes detected with phenomenological codon models and with MK methods, the set of genes detected does not seem to overlap beyond random expectations (He *et al.*, 2020). In contrast, at the site level, classical codon models are congruent with MK methods, with most adaptive mutations occurring at the surface of proteins in both methods (Moutinho *et al.*, 2019a). These results still need to be refined across clades and genes, and the availability of polymorphism and divergence DNA sequences that are aligned will make this comparison possible.

However, positively selected genes or sites detected by classical codon models and MK methods can theoretically be different, since the non-adaptive part of substitutions (ω_{NA}) is subtracted in the MK test while classical codon models do not account for it. Interestingly, another estimate of ω_{NA} in the context of phylogenetic codon models is what was referred to as ω_0 in the context of the mutation-selection null model (see section 3.4.1). As a result, mechanistic mutation-selection codon models (ω_0) and the MK test (π_N/π_S) should theoretically be more directly comparable. From the availability of divergence and polymorphism data, it is now possible to ask whether the rate of non-adaptive evolution measured by phylogenetic mutation-selection models and MK methods are congruent, and if not the reason for such discrepancy should be understood.

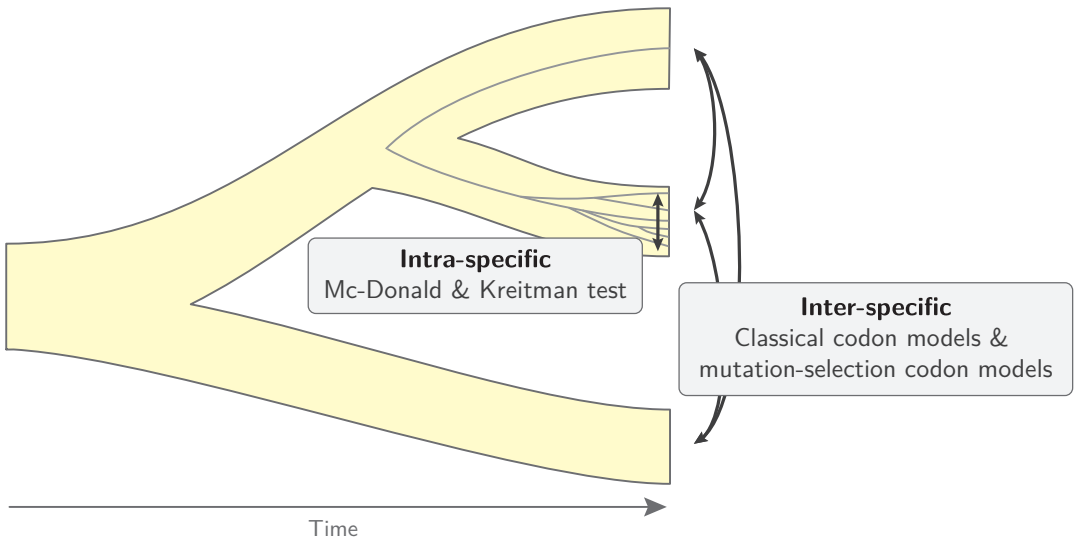


Figure 10.1: *Detecting adaptive evolution in coding sequences from inter- and intra-specific data*

10.4 Unifying phylogenetic and population-genetics models

Throughout this manuscript, the phylogenetic codon models that have been presented have ignored genetic diversity within species. As a result, all differences observed in the

alignment are assumed to be substitutions. However some of the differences observed in the alignment might in fact be polymorphisms segregating in the population. Moreover, substitutions are not instantaneous and ancestral polymorphisms can result in shared polymorphisms across species due to incomplete lineage sorting (Charlesworth, 2010).

Mistaking polymorphisms for substitutions is problematic, since both are not sensitive to mutation, selection and drift to the same extent (Mugal *et al.*, 2014). For example, the neutral diversity increases with N_e, while the rate of neutral substitutions is insensitive to N_e. Also, polymorphisms involving mildly deleterious mutations are very common but are normally filtered out by selection and thus, are not often seen as substitutions.

More generally, substitution rates are much more strongly influenced by fixation biases (such as gBGC) than polymorphisms, which are primarily reflecting mutation biases.

Interestingly, this suggests that polymorphism and divergence could be leveraged together to help disentangle mutation, selection and drift. In other words, phylogenetic and population-genetic approaches could be unified, in the context of a single modelling framework (Thorne *et al.*, 2012).

Such an integration between phylogenetic and population genetics has already been attempted in several studies. For example, Wilson *et al.* (2011) modelled codon evolution in a joint framework with 3 species, which allowed them to analyse the variation in selection pressure spatially along the genome and temporally between lineages. However, this methodology proved to be computationally intensive and does not scale well with the number of extant species. Alternatively, modelling substitutions as mutational events followed by a gradual fixation, using an explicit Wright-Fisher or Moran process along the phylogeny, makes it possible to estimate nucleotide mutation rates and mean fixation probabilities from genetic variation within and between species, while accounting for shared ancestral polymorphisms and incomplete lineage sorting (De Maio *et al.*, 2013; Schrempf *et al.*, 2016; Bergman *et al.*, 2018; Schrempf *et al.*, 2019). In particular, this methodology was used to disentangle gBGC and the mutational bias (Borges *et al.*, 2019; Borges and Kosiol, 2020). However, this methodology does not scale well with the number of states of the models. For that reason, it would be particularly difficult to translate the approach from nucleotides (4 states) to codons (61 states).

Because mechanistic codon models are based on population-genetic first principles, they can theoretically be extended to account for within species diversity. One strategy would be to augment molecular divergence data between species with information about molecular polymorphism within species. Such an approach was attempted during the first year of my PhD training. The formalism that was used is based on Poisson Random fields (details can be found in appendices page 215). This extension was rather straightforward to implement in `BayesCode`, in the context of site-specific mechanistic codon models.

A first implementation was tested against simulations under a Wright-Fisher model of evolution along the phylogeny. It yielded an accurate estimation of diversity ($\theta = 4N_e u$) in extant species, and was able to better tease out mutation and selection by leveraging both divergence and polymorphism signals. However, it was found to be computationally intensive, even with the use of sufficient statistics to accelerate the

computation (see section 4.2.6). Moreover, the assumption of a constant N_e along the phylogeny in mutation-selection codon models was arguably the strongest restriction to relax in this context (Rousselle *et al.*, 2018). Indeed, it makes limited sense to integrate empirical data about genetic diversity in extant species that generally have quite different levels of diversity if N_e is considered constant along the phylogeny. This was historically the reason why I decided to first extend phylogenetic site-specific mutation-selection codon models by incorporating branch-specific N_e, such as presented in chapter 8.

Once incorporating branch specific N_e, the original goal was then to add polymorphism data in the context of this improved mutation-selection codon model. As it turns out, however, there are other issues that needed to be addressed, before achieving this integration, in particular, the fact that the range of N_e inferred by the model turns out to be too narrow (see section 8), as well as computational issues. Indeed, with branch-specific N_e and extant polymorphism, the computing time to reach convergence of the MCMC became prohibitive.

Retrospectively, even though site- and branch-specific mutation-selection phylogenetic codon models can be extended by incorporating empirical data about polymorphism, as I have started to do in `BayesCode`, I believe it is not yet the path forward to build a unified phylogenetic and population-genetic model. Before doing this, I believe phylogenetic codon models and population-genetics method should first be confronted, and the discrepancy should be understood, such as presented in the previous section. The impact of epistasis should also be better understood and characterized. Only in a second step, subsequently to this confrontation, could phylogenetic models in principle accommodate extant polymorphism. However, this will probably require another approach of inference, more computationally reasonable than the one that I have explored in my work (in chapter 8).

10.5 Mechanistic and phenomenological models

Models of inference are classified broadly into phenomenological and mechanistic (Rodrigue and Philippe, 2010). Mechanistic models dissect the detailed causal chain of events responsible for each substitution event and then use this to construct a detailed model from first principles. By doing this, they relate structural, population genetics and ecological parameters to the likelihood function (see chapter 8). As such, mechanistic inference models are suitable to construct an integrative framework, for example relating the signal available in molecular sequences to structural parameters, expression level across genes and varying effective population size across lineages.

Once such models have been fitted to empirical data, the estimated parameters can then be confronted with independent estimations, which allow one to robustly test the model since independent estimates of biological and ecological parameters should of course be congruent (Dasmeh *et al.*, 2014). However mechanistic models are computationally very intensive, to a point where they can reach the current limits in available

computing power (personal computers or clusters)[1]. Moreover, increased complexity of the models bears another consequence: the liability of the code and software decreases, compromising the reproducibility of the results obtained with such models. For these reasons, mechanistic models tend to make a number of strong simplifying assumptions (such as no epistasis), which can have detrimental effects on the robustness of the inference (see chapter 8).

In contrast, phenomenological models are formulated in terms of aggregate parameters, capturing the average rate of synonymous or non-synonymous substitutions, or their ratio. Their aim is to determine the statistical distribution of these aggregate quantities across the tree, across genes, or across sites, but without deriving them from first principles. Compared to mechanistic models, they are computationally much more efficient. On the other hand, they do not give direct access to the population-genetic parameters.

The distinction between phenomenological models and mechanistic models is useful to frame different models in this context. However, that there is no such thing as a pure phenomenological or pure mechanistic model. The most phenomenological models are still based on the overall process of sequence change, and the fact that there is, for instance, a difference between synonymous and non-synonymous substitutions. The most mechanistic models are still highly coarse-grained, abstracting away a certain amount of biological, chemical, and physical phenomena and replacing it with descriptions of how things are observed to behave at a higher level. This raises the question of how to benefit from the advantages of the two approaches. Observations and experiments done throughout this thesis led me to crystallize the conception that models of inference should be mechanistic in essence, in the sense that they should be parameterized by variables that are derived from first principles, but should be phenomenological in practice, in the sense that these variables should nevertheless be aggregate parameters.

The first manuscript presented in this thesis (chapter 7) gives some preliminary directions in this regard. A mean-field argument was used to derive a phenomenological model based on an underlying mechanistic site-specific model. As a result, the mean-field parameters of the phenomenological model capture aggregate quantities that are averages across sites. The phenomenological model that was obtained using this approach is easier to fit to the data. Nonetheless, it captures essential parameters that are easy to interpret mechanistically, after the fact.

Altogether, hybrid models based on mechanistic first principles but obtained by deriving aggregate parameters are avoiding the pitfalls of both approaches, being based on independently identifiable parameters and at the same time being computationally parsimonious. Such hybrid models can be developed with the following procedure:

1. Define a mechanistic microscopic model from molecular first principles (see chapter 9). This model can be potentially complex, modelling all kinds of variations, for example, incorporating site interdependence or polymorphism within species. Such model is meant to be implemented in simulations, but never in inference.

[1]Apart from the physical limit of resource available, the use of computing resources bears ecological consequences on environmental degradation and $C0_2$ emissions.

2. Use a mean-field argument and calculate aggregate quantities emerging from the microscopic model, leveraging population genetics first principles. Possibly, use theoretical developments such as presented in chapter 9 to approximate the response of the aggregate quantities to changes in the underlying mechanistic parameters.

3. Implement the inference phenomenological model whose parameters correspond to the mean-field aggregates. Such phenomenological models are meant to be confronted with simulations under the mechanistic model and subsequently applied to empirical data in order to estimate the parameters of interest and their covariation with variables that might change across species or across genes.

Such endeavour, however, requires mathematical work to derive the relationship between parameters of interest (such as $\Delta\Delta G, N_e$, expression level, ...) and aggregate parameters of evolution that are extractable from the data (ν). Chapter 9 represents one such mathematical development.

10.6 Reproducible science

This thesis is based on a combination of analytical developments, computational simulations and inference models, which I argue are complementary, but more importantly, they are all jointly required. Theoretical modelling allows one to understand the principles, while simulations are crucial to verify the soundness. Inference makes it possible to extract and test the theoretical results using empirical data, which are verified and tested against simulations. Simulations have thus a dual role, testing the robustness of both inference procedures and theoretical results, outside of their comfort zone and assumptions. However, this assumes we are confident enough to write reproducible programs. In this direction, the next section is dedicated to my experience in reproducing results.

First, I stand firmly on the ground that data, codes and scripts should be rendered open access for any published and peer reviewed paper. Practically, the availability of the data and source code should simply be enforced upon submission to any journal, which is currently not the case for all journals even within the fields of bioinformatics and genomics. It is true that such enforcement imposes a heavier burden on scientists upon publishing. However, it avoids the bloating of the technical debt, or research debt resulting from building theories on the ground of a dangerous and possibly shaky basement. It also encourages peer collaboration, both helping the team or person(s) who made the code available, and the community as a whole. A straightforward approach is to provide a `git` versioned repository, with the advantage that collaboration is facilitated through web hosted repositories such as `GitLab` (hosted by institutions) or `GitHub` (hosted by a private company, Microsoft, at the moment of writing).

Nonetheless, code availability is a necessary condition, but not the sole requirement of reproducible research. Specific instructions to reproduce the results should also be made available (Wilson *et al.*, 2014; Darriba *et al.*, 2018). The first step to reproduce a code is to

have the required environment, meaning the necessary libraries and dependencies for the code and scripts. For script and code written in `Python`, the package manager `Anaconda` (or `Conda`) provides a readily available environment to configure the necessary libraries with their versions. More complex environments requiring code compilation or system-level packages can leverage containerization technology such as `Docker` or `Singularity` for example, but any other containers implementing system-level virtualization is very helpful to provide the necessary libraries. Once the environment is specified, the documentation can be made available as a README with the necessary instructions.

More generally, notebooks such as `Jupyter Notebook`, `RMarkdown` or `Org-mode` to name a few also provide an environment for knitting together code and instructions, allowing anyone to follow step-by-step experiments, analysis and results, in a similar fashion as laboratory notebooks are required in wet labs. It is important to note that notebooks can run code from a variety of languages (`C++`, `Haskell`, `Java`, `Julia`, `Python`, `Wolfram Language`, `Matlab`, `Ruby`, ...). These tools are emerging in the community, as well as workflow management system (`Nextflow`, `Snakemake`, ...) allowing one to create reproducible and scalable data analyses running on computing clusters.

Using this range of tools helps other scientists who might want to understand, test or build upon published work. Moreover, they are also very helpful for the person or team implementing them, since a more rigorous and reproducible environment makes it possible to more easily track down bugs and test programs under different conditions or datasets[2]. During the development period, continuous integration pipelines are valuable to increase the reliability of code generation, which should be used whether working alone or inside a team, but is, of course, more critical for collaborative code where one cannot control all of the code that is written.

Collaborative coding practices such as peer-coding sessions are really useful to implement critical code at the core of the program under development. I argue that efficient peer-coding sessions can be organized by dividing the tasks into a group focused in the detailed implementation while the others are free to focus on edge cases and on the overall implications of different implementations. Moreover, peer-coding sessions provide a convenient and structured place for learning good practices, for expanding technical knowledge while correcting bad habits.

Another remarkable practice is to write two independent versions of the program, using if possible different algorithms and languages but with the same functionality, but most importantly, by a different person. Subsequent testing of the programs against each other under the same conditions and datasets should result in the same outcomes[3]. Such a model of reproducible computing experiments and analyses is laborious and demanding, but I argue this is the definition of reproducibility we collectively should aim for, namely where one can independently reproduce the same experiment. If the two (or more)

[2]Notebooks are very useful to present work and data analysis, but should not be used during development since they often offer poor integration with debugger and code inspection tools, enforce awkward software design patterns, and often result in bloated versioned repository.

[3]An extreme version is adversarial coding (or chaos engineering), where the goal is to find conditions on which the program written by someone else fails.

programs result in different outcome, one can run the code with different conditions to pinpoint which one is the failing code (which might actually be both versions). Personally, having practised this method, I strongly believe it pervasively reduces our research debt that we might inadvertently burden others with whenever not realizing the program is bugged, and that it also ultimately saves us time on debugging and conducting research. Finally, explaining to others our choices of algorithms, implementation and data structure forces us to express intelligibly our ideas and, therefore, to better understand them, while gaining from others insights and new algorithmic expertise.

10.7 Concluding remarks

To conclude, this work is an encouraging, although still far from complete, attempt to build integrated models of the evolution of protein-coding DNA sequences. I think my work has contributed to consolidating the idea that the patterns of substitutions inform us on the long-term fluctuations in genetic drift along branches and selection along sequences. This thesis also further emphasizes that the assumptions made on the structure of the fitness landscape have a critical importance in the sensitivity of changes in substitution rates to changes in ecological (N_e) or molecular variables (protein expression level). Conversely, empirical observations of the patterns of substitutions in response to changes in molecular or ecological variables inform us about the underlying structure of the fitness landscape.

Altogether, this work can be seen as a building block toward bridging phylogeny and population genetics. Constructing an integrated framework is theoretically possible but with a still limited scope so far. However, confronting the estimation of phylogenetic codon models and population-genetics approaches, for example, on the question of the rate of adaptive substitution, is a path forward toward an integrated view of protein-coding DNA sequence evolution. Finally, I believe this thesis is not providing disruptive results, but instead is consolidating theoretical models on which molecular evolution is based, and points out some of the pitfalls to avoid.

Part IV
Appendices

11

Inferring long-term population size - Supplementary Materials

Contents

11.1 Summary statistics

11.1.1 Partial correlation coefficient

The correlation coefficient $\rho_{a,b}$ give the total regression between two variables. Partial-correlation coefficient account for the entire covariance matrix, and measure the correlation between 2 traits, knowing the values of all the other traits:

$$\rho_{a,b|c\in\{1,\dots,L\}\setminus\{a,b\}} = -\frac{\Omega_{a,b}}{\sqrt{\Omega_{a,a}\Omega_{b,b}}}, \tag{11.1}$$

where the precision matrix $\mathbf{\Omega}$ is the inverse of the covariance matrix:

$$\mathbf{\Omega} = \mathbf{\Sigma}^{-1} \tag{11.2}$$

11.1.2 Fitness profile entropy

For a category k, the Shannon's entropy (Ω) of the fitness profile ($\boldsymbol{\phi}$) is defined as:

$$\Omega^{(\mathrm{k})} = -\sum_{\mathrm{a}=1}^{20} \phi_{\mathrm{a}}^{(\mathrm{k})} \ln\left(\phi_{\mathrm{a}}^{(\mathrm{k})}\right) \tag{11.3}$$

The Shannon's entropy measures the flatness of the fitness profile, with a value of 0 corresponding to a single peak fitness landscape (only one amino acid is present), and a value of $log(20) \simeq 3$ corresponding to a neutral landscape, where each amino acid has the same fitness.

The Shannon's entropy can be averaged over all sites as:

$$\langle\Omega\rangle = \frac{1}{\mathrm{Z}}\sum_{\mathrm{z}=1}^{\mathrm{Z}} \Omega^{\kappa(\mathrm{z})} \tag{11.4}$$

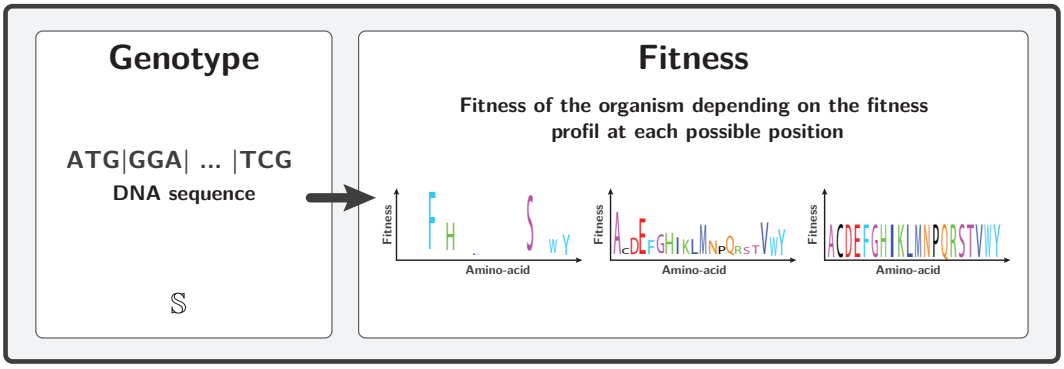
11.2 Simulations

11.2.1 Site-specific fitness profiles (`SimuDiv`)

For simulations under a site-independent fitness landscape, with site-specific fitness profiles, the protein log-fitness is computed as the sum of amino-acid log-fitness coefficients along the sequence. In this model, each codon site z has its own fitness profile, denoted $\boldsymbol{\phi}^{(\mathrm{z})} = \{\phi_{\mathrm{a}}^{(\mathrm{z})},\ 1 \leq \mathrm{a} \leq 20\}$, a vector of 20 amino-acid scaled (Wrightian) fitness coefficients. Since $\mathbb{S}[\mathrm{z}]$ is the codon at site z, the encoded amino acid is $\mathcal{A}(\mathbb{S}[\mathrm{z}])$, hence the fitness at site z is $\boldsymbol{\phi}^{(\mathrm{z})}_{\mathcal{A}(\mathbb{S}[\mathrm{z}])}$. Altogether, the selection coefficient of the mutant $\mathbb{S}'$ is:

$$s\left(\mathbb{S},\mathbb{S}'\right) = \sum_{\mathrm{z}=1}^{\mathrm{Z}} \ln\left(\frac{\boldsymbol{\phi}^{(\mathrm{z})}_{\mathcal{A}(\mathbb{S}'[\mathrm{z}])}}{\boldsymbol{\phi}^{(\mathrm{z})}_{\mathcal{A}(\mathbb{S}[\mathrm{z}])}}\right), \tag{11.5}$$

The fitness vectors $\boldsymbol{\phi}^{(\mathrm{z})}$ used in this study are extracted from Bloom (2017). They were experimentally determined by deep mutational scanning.



The next change in the protein coding DNA and the time to next the event is chosen using Gillespie's algorithm (Gillespie, 1977), according to the rates of substitution between codons:

$$Q_{i,j} = \mu_{i,j} \frac{4N_e s\left(\mathbb{S}^t, \mathbb{S}^{t+1}\right)}{1 - e^{-4N_e s(\mathbb{S}^t, \mathbb{S}^{t+1})}}, \tag{11.6}$$

where $Q_{i,j} = \mu_{i,j}$ in the case of synonymous substitutions.

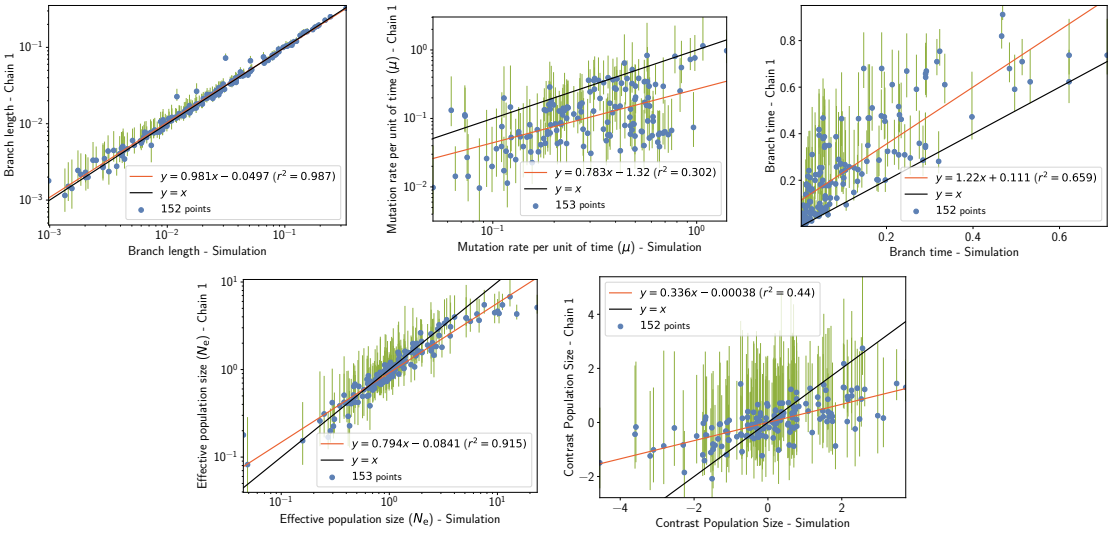


Figure 11.1: *Inferred branch parameters under simulations accounting for site-specific amino-acid profiles, long term fluctuation of N_e, mutation rate per generation and generation time. Estimation is obtained with the mechanistic inference model developed in this paper of site-specific amino-acid fitness profiles and log-Brownian process for N_e, μ and life-history traits.*

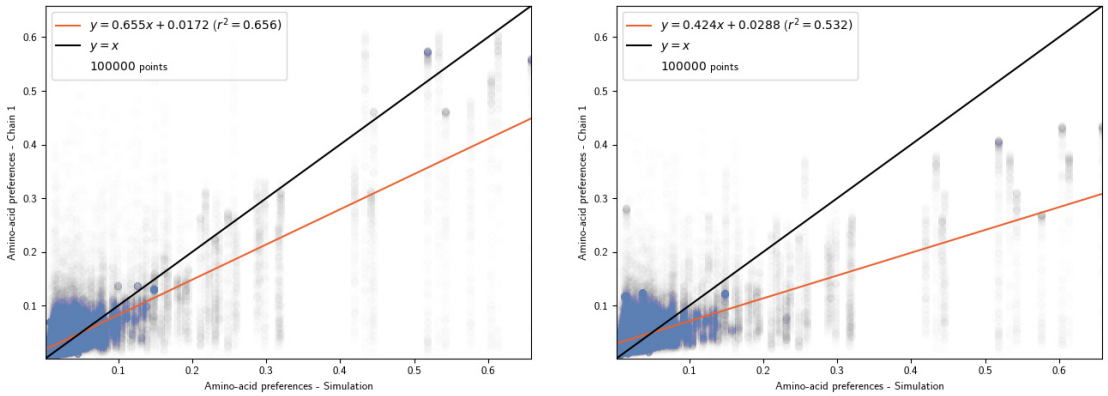


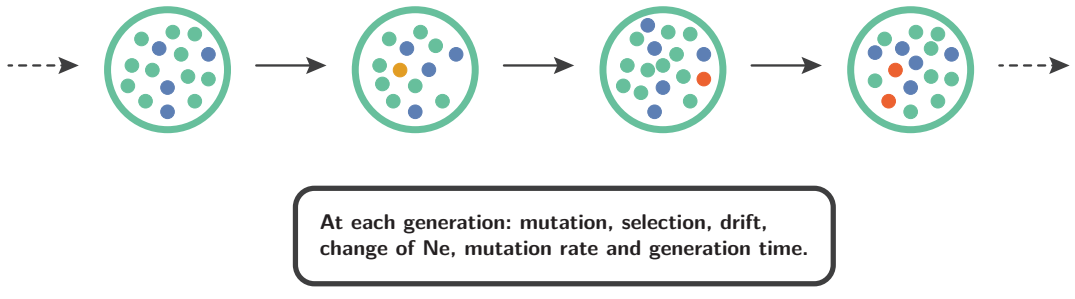
Figure 11.2: *Inferred and simulated site-specific amino-acid profiles under simulation accounting for long term fluctuation of N_e, mutation rate per generation and generation time. Estimation is obtained with the mechanistic inference model developed in this paper of site-specific amino-acid fitness profiles and log-Brownian process for N_e, μ and life-history traits (in the left panel), or under the assumption of constant N_e (in the right panel).*

Experiment	$\langle\Omega\rangle$ (**branch** N_e)	$\langle\Omega\rangle$ (**constant** N_e)
`SimuDiv`, chain 1	2.30 ± 0.04	2.45 ± 0.02
`SimuDiv`, chain 2	2.30 ± 0.04	2.45 ± 0.02

Table 11.1: *Estimated amino-acid entropy under simulations accounting for long term fluctuation of N_e, mutation rate per generation and generation time. Estimation is obtained with the mechanistic inference model developed in this paper of site-specific amino-acid fitness profiles and log-Brownian process for N_e, μ and life-history traits (in the left column), or under the assumption of constant N_e (in the right column).*

11.2.2 Wright-Fisher with polymorphism (`SimuPoly`)

The evolutionary dynamics was formalized as a Wright-Fisher model with mutation, selection and drift The population is assumed to be panmictic, with effective population size N_e and with non-overlapping generations.



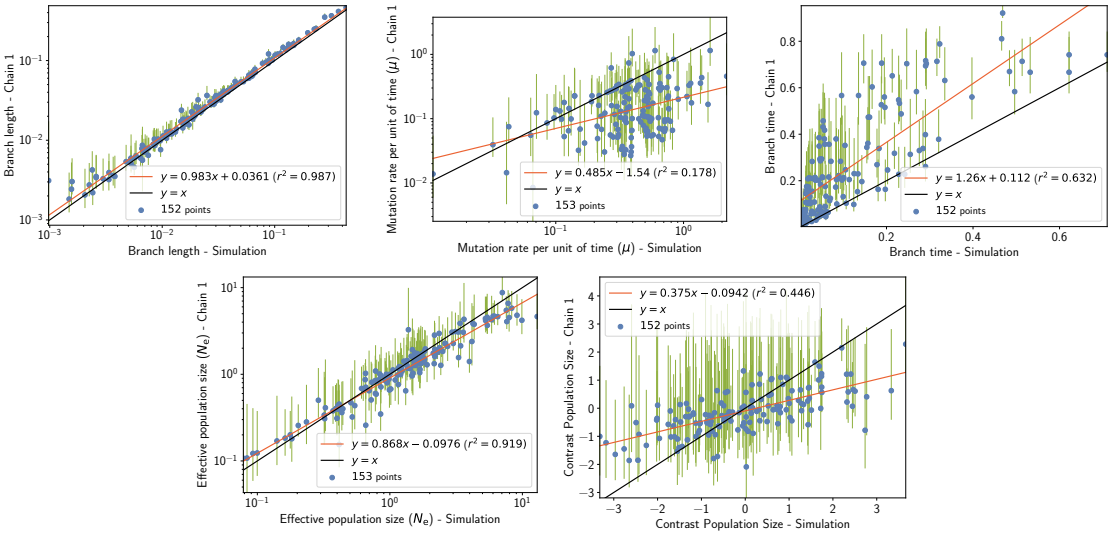


Figure 11.3: *Inferred branch parameters under simulation accounting for finite population effects, site linkage and short term fluctuation of N_e. Estimation is obtained with the mechanistic inference model developed in this paper of site-specific amino-acid fitness profiles and log-Brownian process for N_e, μ and life-history traits.*

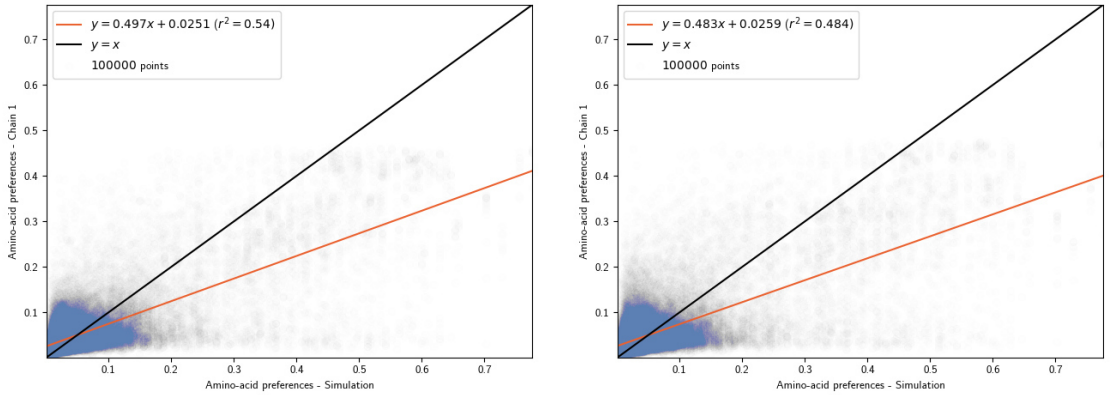


Figure 11.4: *Inferred and simulated site-specific amino-acid profiles under simulation accounting for finite population effects, site linkage and short term fluctuation of N_e. Estimation is obtained with the mechanistic inference model developed in this paper of site-specific amino-acid fitness profiles and log-Brownian process for N_e, μ and life-history traits (in the left panel), or under the assumption of constant N_e (in the right panel).*

Experiment	$\langle\Omega\rangle$ (**branch** N_e)	$\langle\Omega\rangle$ (**constant** N_e)
`SimuPoly`, chain 1	2.47 ± 0.03	2.37 ± 0.02
`SimuPoly`, chain 2	2.47 ± 0.03	2.37 ± 0.02

Table 11.2: *Estimated amino-acid entropy under simulation accounting for finite population effects, site linkage and short term fluctuation of N_e. Estimation is obtained with the mechanistic inference model developed in this paper of site-specific amino-acid fitness profiles and log-Brownian process for N_e, μ and life-history traits (in the left column), or under the assumption of constant N_e (in the right column).*

11.2.3 Fisher geometric landscape (`SimuGeo`)

We simulated substitutions in a protein using an adaptation of Fisher's geometric landscape (Tenaillon, 2014; Blanquart and Bataillon, 2016). In the original context, the phenotype is a vector ($\boldsymbol{P}$) in a multidimensional space, where the number of dimensions is often termed complexity. From a phenotype, the fitness is a monotonously decreasing function of the phenotype distance to 0. The exact functional phenotype-fitness map depends on 2 external parameters controlling for strength (α) and epistasis (β). If the phenotype-fitness map is explicit, the genotype-phenotype map is more pervasive. Mutations are seen has displacement of the phenotype in the multidimensional space. Beneficial mutations are moving the phenotype closer to 0, whereas deleterious mutations are moving the phenotype further away. In such original context, the distribution of mutational effects is not dependent on the current genotype, but this can be relaxed using a genotype-phenotype map.

In a protein context, the genotype-phenotype map can be defined by assigning to each of the 20 amino acid a vector in the multidimensional space. Since different sites of the protein do not have the same physico-chimical properties, we can define a specific genotype-phenotype map for each position of the sequence. Overall, the protein phenotype is computed as the sum of site-specific multidimensional vectors, obtained by accessing the amino acid present at each site of the protein. From a DNA sequence $\mathbb{S}^t$ after t substitutions, the protein's phenotype is given by:

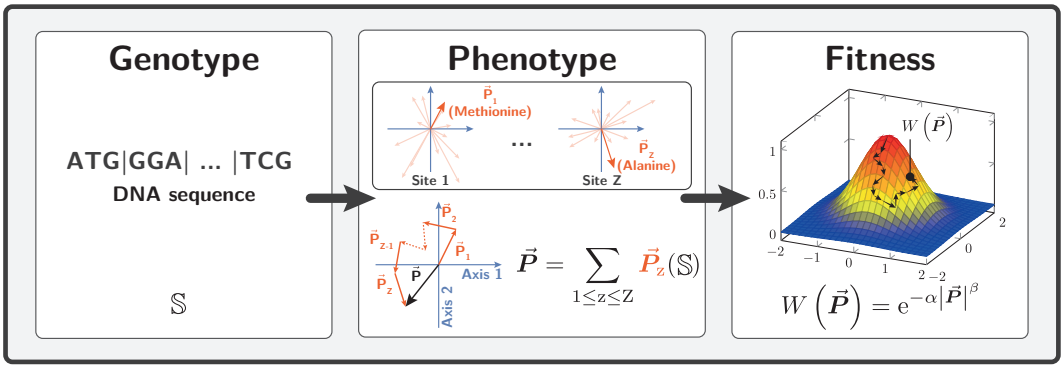
$$\boldsymbol{P}\left(\mathbb{S}^t\right) = \sum_{\mathrm{z}=1}^{\mathrm{Z}} \boldsymbol{P}_{\mathrm{z}}\left(\mathbb{S}^t(\mathrm{z})\right), \tag{11.7}$$

where $\boldsymbol{P}_{\mathrm{z}}$ is the genotype-phenotype map at site z.

And the Wrightian fitness of $\mathbb{S}^t$ is :

$$W\left(\boldsymbol{P}\left(\mathbb{S}^t\right)\right) = e^{-\alpha\left|\boldsymbol{P}\left(\mathbb{S}^t\right)\right|^{\beta}}, \tag{11.8}$$

where strength ($\alpha > 0$) and epistasis (β) are parameters of the fitness function.



For each possible mutant (at time $t+1$ substitutions), we compute $\boldsymbol{P}\left(\mathbb{S}^{t+1}\right)$ from the

updated sequence $\mathbb{S}^{t+1}$, and subsequently the selection coefficient of the mutant:

$$s\left(\mathbb{S}^{t}, \mathbb{S}^{t+1}\right) = \frac{W\left(\boldsymbol{P}\left(\mathbb{S}^{t+1}\right)\right) - W\left(\boldsymbol{P}\left(\mathbb{S}^{t}\right)\right)}{W\left(\boldsymbol{P}\left(\mathbb{S}^{t}\right)\right)}. \tag{11.9}$$

The next change in the protein coding DNA and the time to next the event is chosen using Gillespie's algorithm (Gillespie, 1977), according to the rates of substitution between codons:

$$Q_{i,j} = \mu_{i,j} \frac{4N_{\mathrm{e}} s\left(\mathbb{S}^{t}, \mathbb{S}^{t+1}\right)}{1 - \mathrm{e}^{-4N_{\mathrm{e}} s\left(\mathbb{S}^{t}, \mathbb{S}^{t+1}\right)}}, \tag{11.10}$$

where $Q_{i,j} = \mu_{i,j}$ in the case of synonymous substitutions.

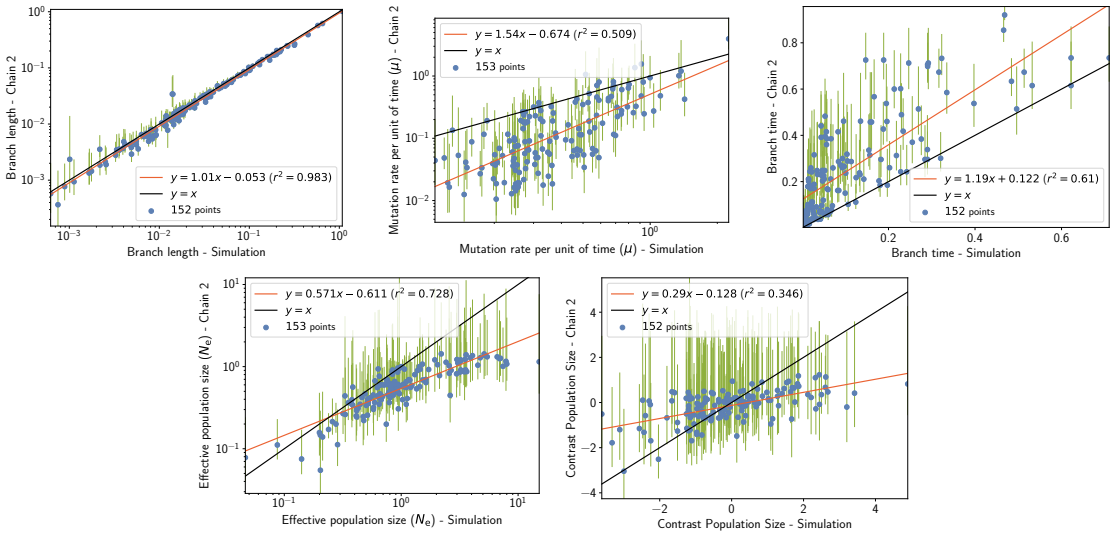


Figure 11.5: *Inferred branch parameters under simulation accounting for site epistasis in geometric landscape, thus fluctuation of the selection coefficient along the phylogeny. Estimation is obtained with the mechanistic inference model developed in this paper of site-specific amino-acid fitness profiles and log-Brownian process for N_e, μ and life-history traits.*

Experiment	$\langle\Omega\rangle$ (**branch** N_{e})	$\langle\Omega\rangle$ (**constant** N_{e})
`SimuGeo`, chain 1	2.27 ± 0.02	2.46 ± 0.02
`SimuGeo`, chain 2	2.23 ± 0.04	2.46 ± 0.02

Table 11.3: *Estimated amino-acid entropy under simulation accounting for site epistasis (geometric landscape), thus fluctuation of the selection coefficient along the phylogeny. Estimation is obtained with the mechanistic inference model developed in this paper of site-specific amino-acid fitness profiles and log-Brownian process for N_e, μ and life-history traits (in the left column), or under the assumption of constant N_e (in the right column).*

11.2.4 Protein folding probability (`SimuFold`)

We simulated substitutions in the protein phosphatase (Z = 300 codon sites) as in Goldstein and Pollock (2017). From a DNA sequence $\mathbb{S}^t$ after t substitutions, we compute the

free energy of the folded state $G_{\mathrm{F}}\left(\mathbb{S}^t\right)$, using the 3-dimensional structure of the folded state and pair-wise contact energies between neighboring amino-acid residues:

$$G_{\mathrm{F}}\left(\mathbb{S}^t\right) = \sum_{\mathrm{z}=1}^{\mathrm{Z}} \sum_{r \in \mathcal{V}(\mathrm{z})} I\left(\mathbb{S}^t(\mathrm{z}), \mathbb{S}^t(r)\right), \tag{11.11}$$

where $I(a,b)$ is the pair-wise contact energies between amino acid a and b, using contact potentials estimated by Miyazawa and Jernigan (1985), and $\mathcal{V}(\mathrm{z})$ are the neighbor residues of site z (closer than 7Å) in the 3D structure.

The free energy of unfolded states $G_{\mathrm{U}}\left(\mathbb{S}^t\right)$ is approximated using 55 decoy 3D structures that supposedly represent a sample of possible unfolded states:

$$G_{\mathrm{U}}\left(\mathbb{S}^t\right) = \left\langle G\left(\mathbb{S}^t\right)\right\rangle - kT\ln(1.0\mathrm{E}^{160}) - \frac{2\left[\left\langle G\left(\mathbb{S}^t\right)^2\right\rangle - \left\langle G\left(\mathbb{S}^t\right)\right\rangle^2\right]}{kT} \tag{11.12}$$

where the average $\langle . \rangle$ runs other the 55 decoy 3D structures, and k is the Boltzmann constant and T the temperature in Kelvin.

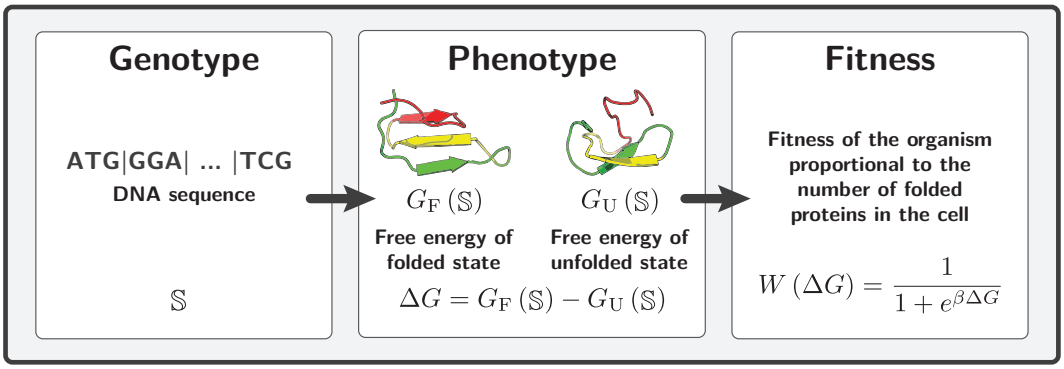
From the energy of folded and unfolded states, we can compute the difference in free energy between the states:

$$\Delta G\left(\mathbb{S}^t\right) = G_{\mathrm{F}}\left(\mathbb{S}^t\right) - G_{\mathrm{U}}\left(\mathbb{S}^t\right) \tag{11.13}$$

Wrightian fitness is defined as the probability of our protein to be in the folded state:

$$W\left(\Delta G\left(\mathbb{S}^t\right)\right) = \mathbb{P}_{\mathrm{F}}\left(\mathbb{S}^t\right) = \frac{e^{-\beta G_{\mathrm{F}}\left(\mathbb{S}^t\right)}}{e^{-\beta G_{\mathrm{F}}\left(\mathbb{S}^t\right)} + e^{-\beta G_{\mathrm{U}}\left(\mathbb{S}^t\right)}} = \frac{1}{1 + e^{\beta \Delta G\left(\mathbb{S}^t\right)}}, \tag{11.14}$$

where β is the inverse of the temperature ($\beta = 1/kT$).



For each possible mutant (at time $t+1$ substitutions), we compute ΔG^{t+1} from the updated sequence $\mathbb{S}^{t+1}$, and subsequently the selection coefficient of the mutant:

$$s\left(\mathbb{S}^t, \mathbb{S}^{t+1}\right) = \frac{W\left(\Delta G\left(\mathbb{S}^{t+1}\right)\right) - W\left(\Delta G\left(\mathbb{S}^t\right)\right)}{W\left(\Delta G\left(\mathbb{S}^t\right)\right)}. \tag{11.15}$$

The next change in the protein coding DNA and the time to next the event is chosen using Gillespie's algorithm (Gillespie, 1977), according to the rates of substitution between codons:

$$Q_{i,j} = \mu_{i,j} \frac{4N_{\mathrm{e}} s\left(\mathbb{S}^t, \mathbb{S}^{t+1}\right)}{1 - \mathrm{e}^{-4N_{\mathrm{e}} s\left(\mathbb{S}^t, \mathbb{S}^{t+1}\right)}}, \tag{11.16}$$

where $Q_{i,j} = \mu_{i,j}$ in the case of synonymous substitutions.

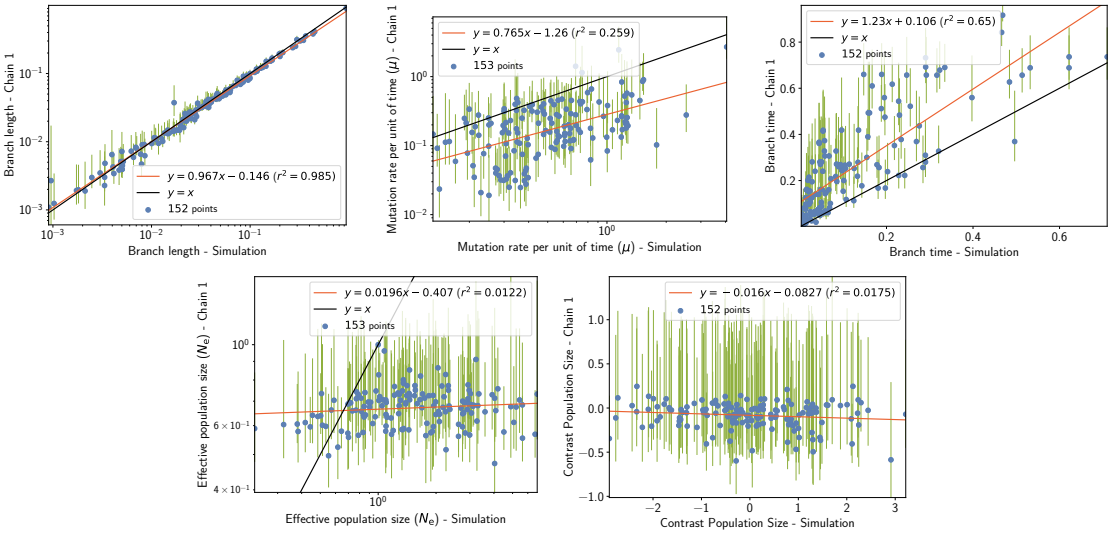


Figure 11.6: *Inferred branch parameters under simulation accounting for site epistasis (folding stability model), thus fluctuation of the selection coefficient along the phylogeny. Estimation is obtained with the mechanistic inference model developed in this paper of site-specific amino-acid fitness profiles and log-Brownian process for N_e, μ and life-history traits.*

Experiment	$\langle\Omega\rangle$ (**branch** N_e)	$\langle\Omega\rangle$ (**constant** N_e)
`SimuFold`, chain 1	1.31 ± 0.05	1.61 ± 0.03
`SimuFold`, chain 2	1.30 ± 0.04	1.60 ± 0.03

Table 11.4: *Estimated amino-acid entropy under simulation accounting for site epistasis (folding stability model), thus fluctuation of the selection coefficient along the phylogeny. Obtained with the mechanistic inference model developed in this paper of site-specific amino-acid fitness profiles and log-Brownian process for N_e, μ and life-history traits (in the left column), or under the assumption of constant N_e (in the right column).*

11.3 Empirical data in mammals

11.3.1 Chain convergence

Obtained with the mechanistic inference model developed in this paper of site-specific amino-acid fitness profiles and log-Brownian process for N_e, μ and life-history traits.

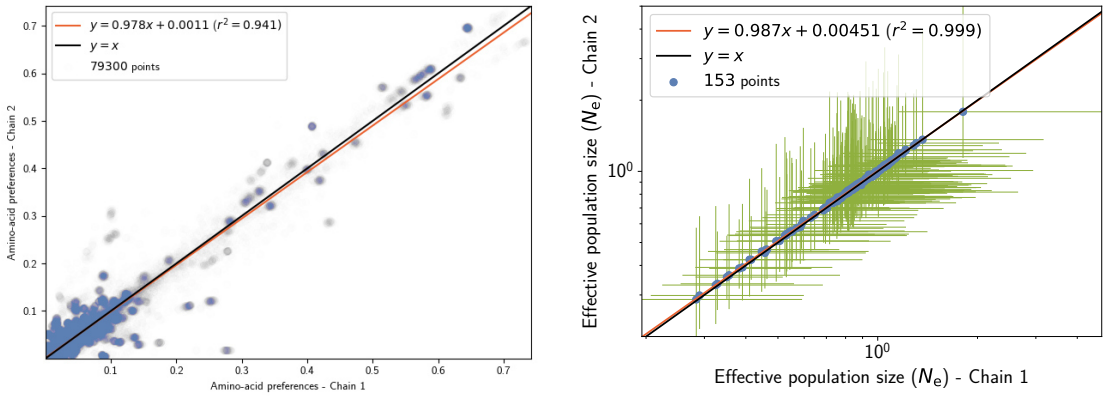


Figure 11.7: *Chain convergence of site amino-acid preferences (left panel) and branch N_e (right panel).*

11.3.2 Traits estimation & correlation (replicate 1, chain 1)

Obtained with the mechanistic inference model developed in this paper of site-specific amino-acid fitness profiles and log-Brownian process for N_e, μ and life-history traits.

Covariance (Σ)	N_e	μ	Maximum longevity	Adult weight	Female maturity
N_e	0.281**	0.324**	−0.268**	−1.29**	−0.308**
μ	-	1.93**	−1.12**	−5.19**	−1.43**
Maximum longevity	-	-	0.934**	3.58**	1.01**
Adult weight	-	-	-	19.9**	4.48**
Female maturity	-	-	-	-	1.53**

Table 11.5: *Covariance coefficient between effective population size (N_e), mutation rate per site per unit of time (μ), and life-history traits (maximum longevity, adult weight and female maturity) were computed in placental mammals. Asterisks indicate strength of support (*$pp > 0.95$, **$pp > 0.975$).*

Partial coefficient	N_e	μ	Maximum longevity	Adult weight	Female maturity
N_e	-	−0.146	−0.177	−0.265*	−0.0223
μ	-	-	−0.283*	−0.396**	−0.327**
Maximum longevity	-	-	-	0.236*	0.383**
Adult weight	-	-	-	-	0.179
Female maturity	-	-	-	-	-

Table 11.6: *Partial correlation coefficient between effective population size (N_e), mutation rate per site per unit of time (μ), and life-history traits (maximum longevity, adult weight and female maturity) were computed in placental mammals. Asterisks indicate strength of support (*$pp > 0.95$, **$pp > 0.975$).*

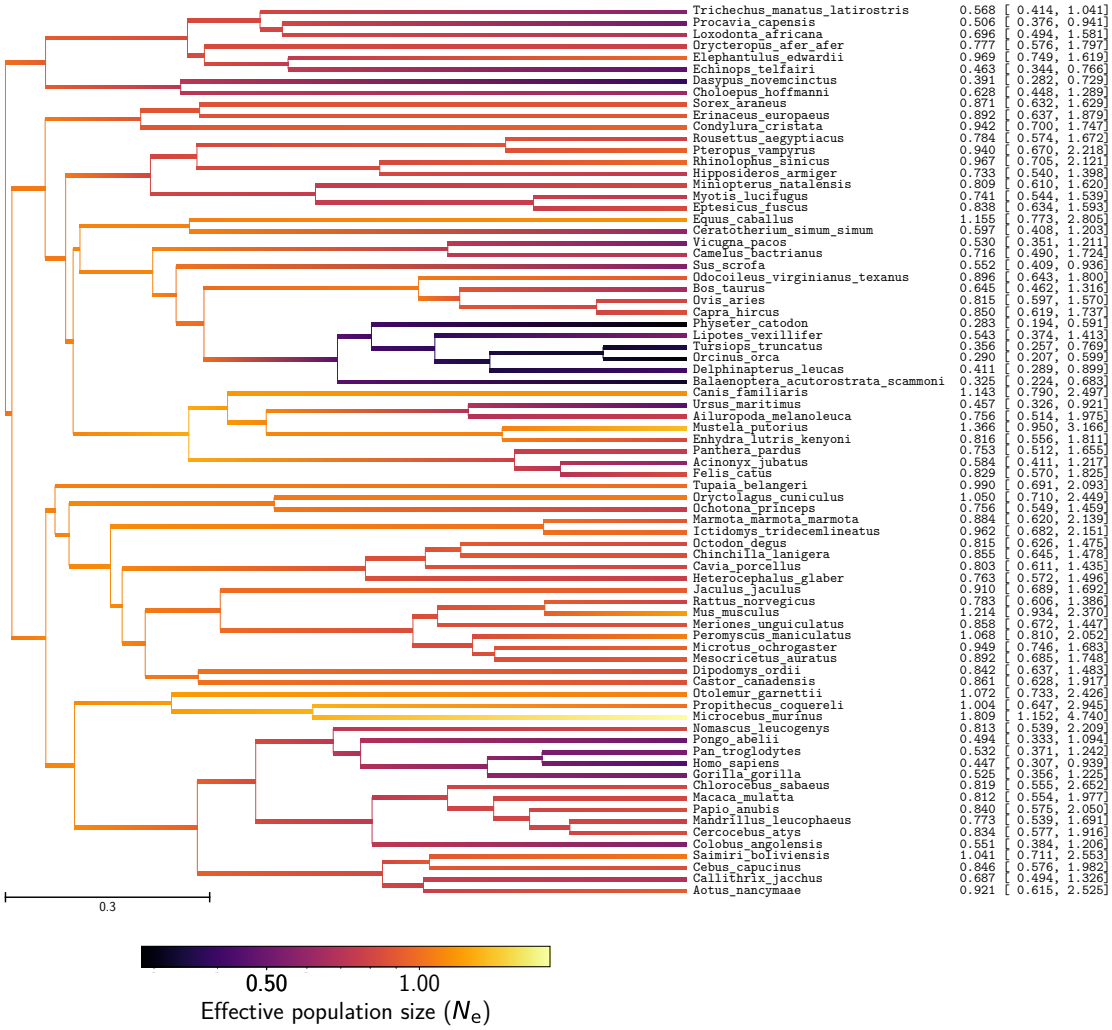


Figure 11.8: *Effective population size (N_e) estimation in mammals*

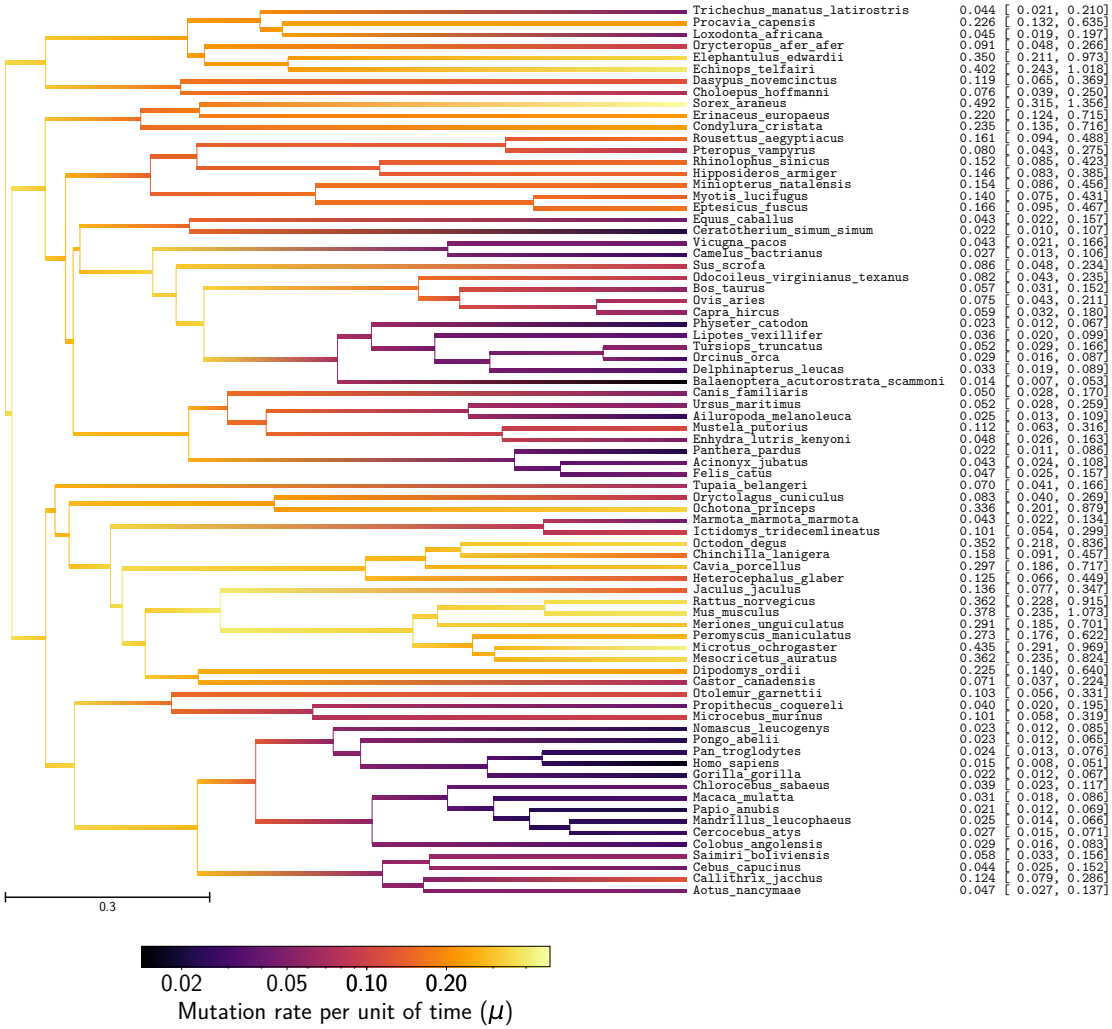


Figure 11.9: *Mutation rate (μ) estimation in mammals*

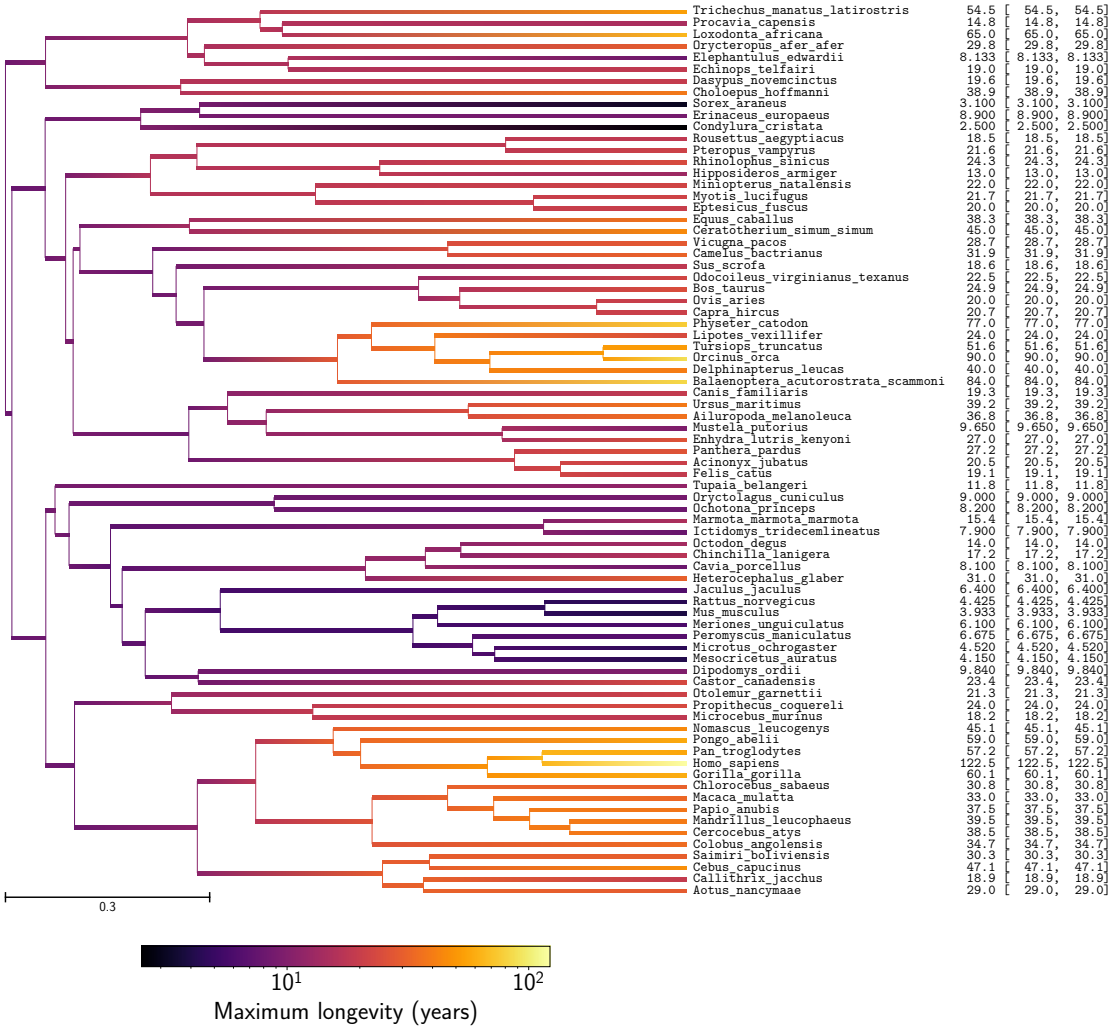


Figure 11.10: *Maximum longevity estimation in mammals*

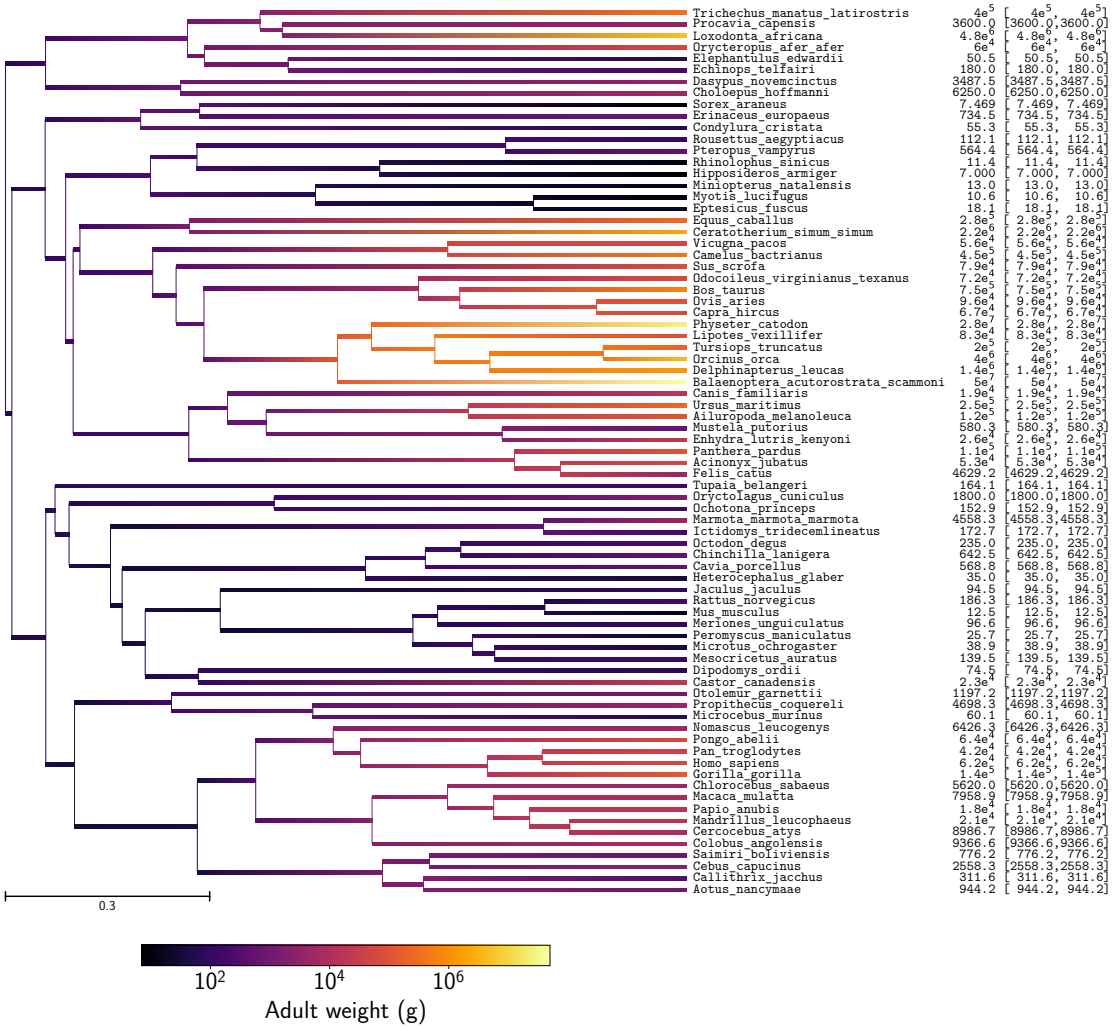


Figure 11.11: *Adult weight estimation in mammals*

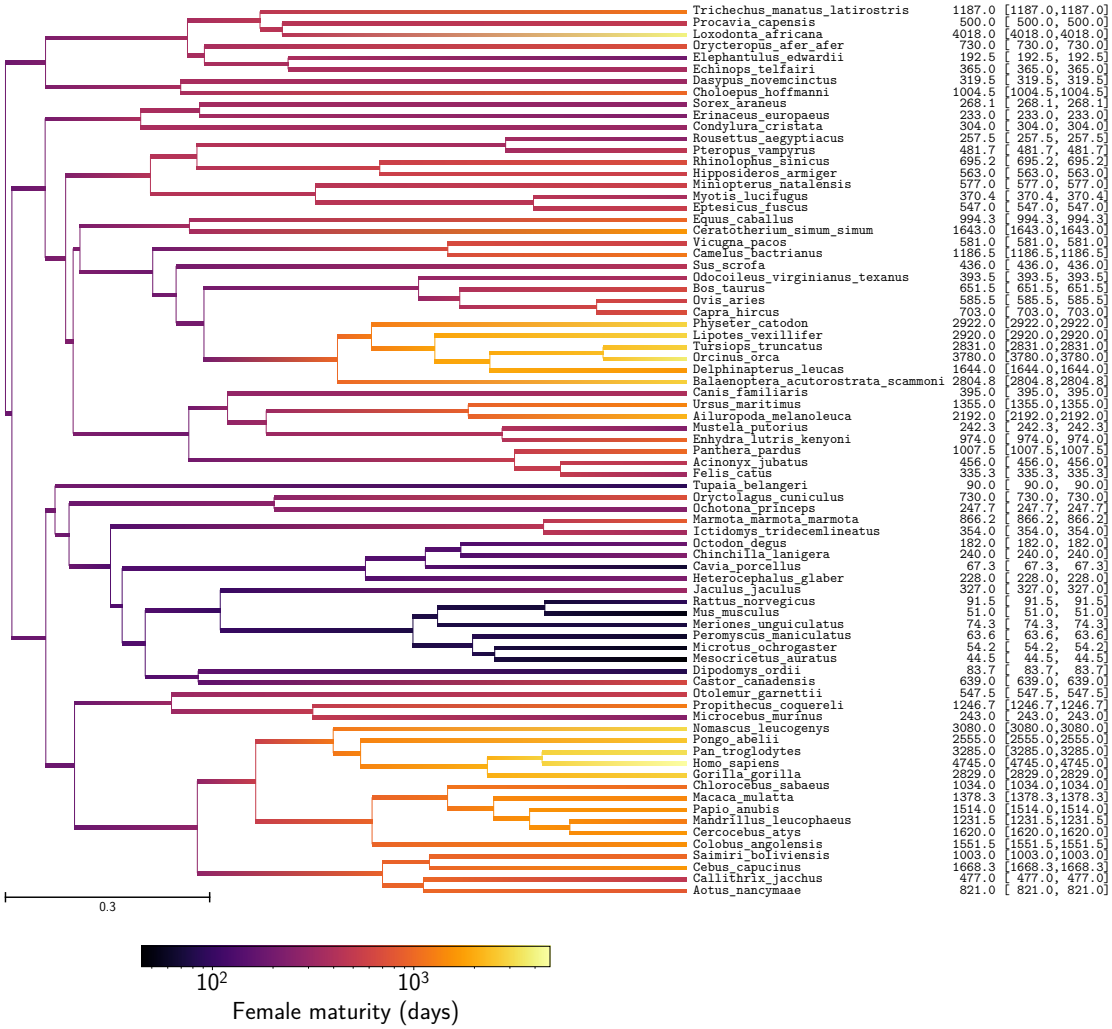


Figure 11.12: *Female maturity estimation in mammals*

11.3.3 Repeatability of experiments

4 independent inferences were performed on a randomly chosen set of 18 coding sequences (CDS) out of 226. Obtained with the mechanistic inference model developed in this paper of site-specific amino-acid fitness profiles and log-Brownian process for N_e, μ and life-history traits. Each plot is a correlation between a pair of experiments for a given parameter. For each node (or branch) of the tree, the mean posterior of the parameter over the MCMC (after burn-in) is represented in blue dots, green solid lines are the 90% confidence interval of the MCMC. Solid red line is the regression line between replicates.

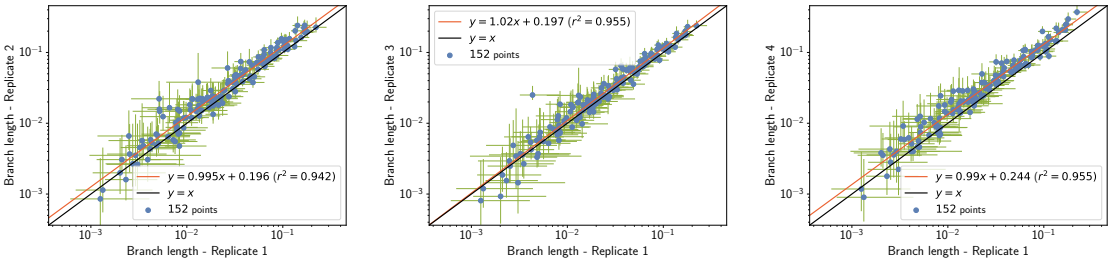


Figure 11.13: *Repeatability of branch length (l) estimation in mammals*

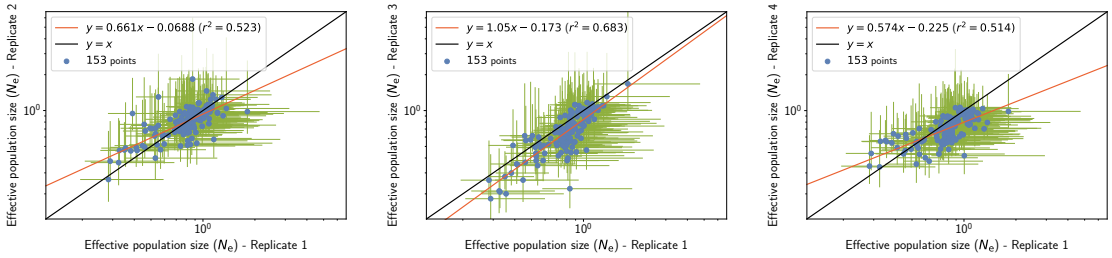


Figure 11.14: *Repeatability of effective population size (N_e) estimation in mammals*

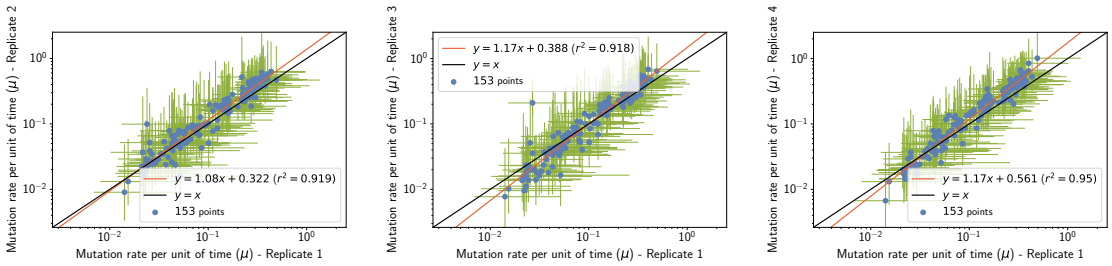


Figure 11.15: *Repeatability of mutation rate (μ) estimation in mammals*

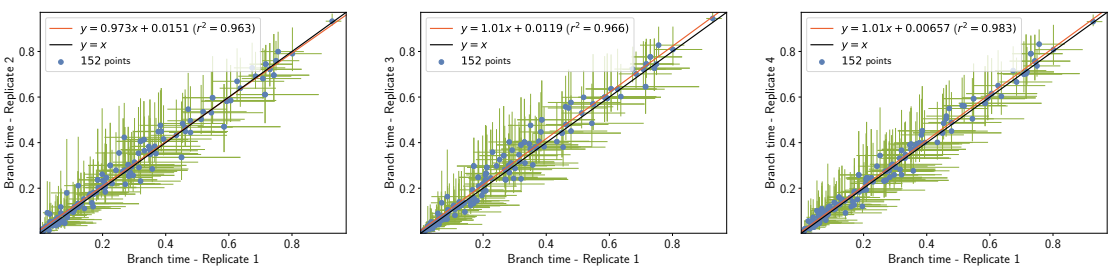


Figure 11.16: *Repeatability of branch time (ΔT) estimation in mammals*

Rep. 1	Rep. 2	Rep. 3	Rep. 4	Taxon
0.568	0.469	0.489	0.548	*Trichechus manatus latirostris*
0.506	0.706	0.615	0.65	*Procavia capensis*
0.696	0.799	0.532	0.595	*Loxodonta africana*
0.777	0.812	0.651	0.717	*Orycteropus afer afer*
0.969	0.904	0.68	0.949	*Elephantulus edwardii*
0.463	0.673	0.56	0.586	*Echinops telfairi*
0.391	0.945	0.51	0.639	*Dasypus novemcinctus*
0.628	0.621	0.52	0.376	*Choloepus hoffmanni*
0.871	1.84	0.745	0.819	*Sorex araneus*
0.892	0.833	1.12	1.06	*Erinaceus europaeus*
0.942	1.15	1.1	0.916	*Condylura cristata*
0.784	0.679	0.488	0.535	*Rousettus aegyptiacus*
0.94	0.838	0.662	0.604	*Pteropus vampyrus*
0.967	0.823	0.586	0.636	*Rhinolophus sinicus*
0.733	0.98	0.876	0.746	*Hipposideros armiger*
0.809	0.934	0.742	0.738	*Miniopterus natalensis*
0.741	0.504	0.442	0.53	*Myotis lucifugus*
0.838	0.849	0.588	0.753	*Eptesicus fuscus*
1.16	0.573	0.846	0.711	*Equus caballus*
0.597	0.524	0.438	0.402	*Ceratotherium simum simum*
0.53	0.399	0.438	0.356	*Vicugna pacos*
0.716	0.68	0.418	0.432	*Camelus bactrianus*
0.552	1.3	0.531	0.43	*Sus scrofa*
0.896	0.861	0.761	0.568	*Odocoileus virginianus texanus*
0.645	0.844	0.583	0.69	*Bos taurus*
0.815	0.649	0.747	0.473	*Ovis aries*
0.85	0.723	0.742	0.538	*Capra hircus*
0.283	0.264	0.261	0.342	*Physeter catodon*
0.543	0.517	0.345	0.486	*Lipotes vexillifer*
0.356	0.484	0.2	0.549	*Tursiops truncatus*
0.29	0.376	0.182	0.437	*Orcinus orca*
0.411	0.491	0.356	0.488	*Delphinapterus leucas*
0.325	0.366	0.211	0.337	*Balaenoptera acutorostrata scammoni*
1.14	1.35	1	0.842	*Canis familiaris*
0.457	0.766	0.615	0.461	*Ursus maritimus*
0.756	0.716	0.569	0.53	*Ailuropoda melanoleuca*
1.37	1.04	1.31	0.795	*Mustela putorius*
0.816	0.557	0.7	0.657	*Enhydra lutris kenyoni*
0.753	0.881	0.683	0.556	*Panthera pardus*
0.584	0.61	0.381	0.561	*Acinonyx jubatus*
0.829	0.761	0.655	0.602	*Felis catus*
0.99	0.738	1.04	0.627	*Tupaia belangeri*
1.05	1.46	1.17	0.897	*Oryctolagus cuniculus*
0.756	0.751	0.514	0.902	*Ochotona princeps*
0.884	0.746	0.541	0.862	*Marmota marmota marmota*
0.962	0.933	0.773	0.977	*Ictidomys tridecemlineatus*
0.815	0.733	0.688	0.874	*Octodon degus*
0.855	0.979	0.691	0.645	*Chinchilla lanigera*
0.803	1.08	0.684	0.898	*Cavia porcellus*
0.763	0.76	0.702	0.518	*Heterocephalus glaber*
0.91	0.655	0.449	0.865	*Jaculus jaculus*
0.783	0.956	0.91	0.883	*Rattus norvegicus*
1.21	0.963	1.01	0.839	*Mus musculus*
0.858	0.856	0.818	0.828	*Meriones unguiculatus*
1.07	1.11	0.877	0.757	*Peromyscus maniculatus*
0.949	1.16	1.01	1.06	*Microtus ochrogaster*
0.892	0.949	1.12	0.788	*Mesocricetus auratus*
0.842	1.07	1.01	0.695	*Dipodomys ordii*
0.861	0.583	0.494	0.575	*Castor canadensis*
1.07	1.13	0.812	0.821	*Otolemur garnettii*
1	0.945	0.741	0.418	*Propithecus coquereli*
1.81	0.98	1.67	0.985	*Microcebus murinus*
0.813	0.512	0.399	0.582	*Nomascus leucogenys*
0.494	0.71	0.568	0.475	*Pongo abelii*
0.532	0.713	0.381	0.513	*Pan troglodytes*
0.447	0.508	0.261	0.433	*Homo sapiens*
0.525	0.611	0.493	0.528	*Gorilla gorilla*
0.819	0.754	0.782	0.71	*Chlorocebus sabaeus*
0.812	0.816	0.538	0.679	*Macaca mulatta*
0.84	0.8	0.555	0.676	*Papio anubis*
0.773	0.813	0.501	0.628	*Mandrillus leucophaeus*
0.834	0.823	0.221	0.631	*Cercocebus atys*
0.551	0.749	0.599	0.706	*Colobus angolensis*
1.04	0.93	0.466	0.859	*Saimiri boliviensis*
0.846	0.519	0.444	0.667	*Cebus capucinus*
0.687	0.658	0.659	0.805	*Callithrix jacchus*
0.921	0.532	0.614	0.794	*Aotus nancymaae*
6.38	**6.96**	**9.19**	**3.16**	**Maximum range**

Table 11.7: *Repeatability of effective population size (N_e) estimation in mammals, for the extant taxa.*

Rep. 1	Rep. 2	Rep. 3	Rep. 4	Taxon
0.0436	0.039	0.0645	0.0422	*Trichechus manatus latirostris*
0.226	0.281	0.369	0.229	*Procavia capensis*
0.0455	0.0656	0.0435	0.0342	*Loxodonta africana*
0.0909	0.0995	0.144	0.0862	*Orycteropus afer afer*
0.35	0.614	0.434	0.457	*Elephantulus edwardii*
0.402	0.478	0.684	0.848	*Echinops telfairi*
0.119	0.194	0.104	0.0861	*Dasypus novemcinctus*
0.0764	0.0713	0.109	0.0915	*Choloepus hoffmanni*
0.492	0.468	0.64	1.01	*Sorex araneus*
0.22	0.521	0.281	0.317	*Erinaceus europaeus*
0.235	0.228	0.282	0.352	*Condylura cristata*
0.161	0.173	0.186	0.266	*Rousettus aegyptiacus*
0.08	0.0942	0.0565	0.103	*Pteropus vampyrus*
0.152	0.222	0.157	0.158	*Rhinolophus sinicus*
0.146	0.232	0.208	0.164	*Hipposideros armiger*
0.154	0.166	0.135	0.23	*Miniopterus natalensis*
0.14	0.164	0.174	0.158	*Myotis lucifugus*
0.166	0.154	0.166	0.217	*Eptesicus fuscus*
0.0428	0.0657	0.0345	0.0404	*Equus caballus*
0.0217	0.0368	0.0158	0.017	*Ceratotherium simum simum*
0.0431	0.0512	0.0335	0.0788	*Vicugna pacos*
0.027	0.0244	0.0357	0.0244	*Camelus bactrianus*
0.0859	0.0431	0.0402	0.0497	*Sus scrofa*
0.0818	0.0677	0.0525	0.0805	*Odocoileus virginianus texanus*
0.057	0.0696	0.0575	0.0424	*Bos taurus*
0.0753	0.0883	0.0886	0.101	*Ovis aries*
0.0586	0.0764	0.0653	0.0641	*Capra hircus*
0.0234	0.0291	0.0161	0.0181	*Physeter catodon*
0.0355	0.0534	0.0373	0.038	*Lipotes vexillifer*
0.0524	0.0474	0.0481	0.0393	*Tursiops truncatus*
0.0292	0.0254	0.0184	0.0171	*Orcinus orca*
0.0331	0.0347	0.019	0.0318	*Delphinapterus leucas*
0.0141	0.00903	0.00772	0.0067	*Balaenoptera acutorostrata scammoni*
0.0505	0.0795	0.0329	0.065	*Canis familiaris*
0.0518	0.0158	0.0165	0.0256	*Ursus maritimus*
0.0255	0.0517	0.0513	0.0368	*Ailuropoda melanoleuca*
0.112	0.0774	0.101	0.163	*Mustela putorius*
0.0479	0.0513	0.0456	0.0528	*Enhydra lutris kenyoni*
0.0218	0.0175	0.0144	0.0194	*Panthera pardus*
0.0426	0.0261	0.039	0.0486	*Acinonyx jubatus*
0.0465	0.0235	0.038	0.038	*Felis catus*
0.0703	0.0841	0.0683	0.127	*Tupaia belangeri*
0.0829	0.127	0.107	0.112	*Oryctolagus cuniculus*
0.336	0.364	0.28	0.398	*Ochotona princeps*
0.0429	0.0489	0.0222	0.0388	*Marmota marmota marmota*
0.101	0.102	0.131	0.136	*Ictidomys tridecemlineatus*
0.352	0.5	0.416	0.447	*Octodon degus*
0.158	0.143	0.195	0.183	*Chinchilla lanigera*
0.297	0.39	0.357	0.283	*Cavia porcellus*
0.125	0.114	0.069	0.116	*Heterocephalus glaber*
0.136	0.193	0.139	0.18	*Jaculus jaculus*
0.362	0.466	0.366	0.401	*Rattus norvegicus*
0.378	0.383	0.409	0.451	*Mus musculus*
0.291	0.256	0.271	0.443	*Meriones unguiculatus*
0.273	0.186	0.224	0.343	*Peromyscus maniculatus*
0.435	0.633	0.473	0.528	*Microtus ochrogaster*
0.362	0.561	0.541	0.538	*Mesocricetus auratus*
0.225	0.221	0.181	0.336	*Dipodomys ordii*
0.0708	0.0903	0.0716	0.0659	*Castor canadensis*
0.103	0.197	0.0849	0.151	*Otolemur garnettii*
0.0403	0.0785	0.0421	0.0504	*Propithecus coquereli*
0.101	0.0509	0.0524	0.144	*Microcebus murinus*
0.0229	0.0197	0.0232	0.0201	*Nomascus leucogenys*
0.0231	0.0187	0.0112	0.0145	*Pongo abelii*
0.0239	0.0992	0.0197	0.0292	*Pan troglodytes*
0.0154	0.0131	0.0102	0.0131	*Homo sapiens*
0.0222	0.0158	0.00956	0.0145	*Gorilla gorilla*
0.0392	0.0311	0.018	0.0295	*Chlorocebus sabaeus*
0.0314	0.029	0.0164	0.0242	*Macaca mulatta*
0.0211	0.0224	0.0137	0.0201	*Papio anubis*
0.0245	0.0231	0.0138	0.0232	*Mandrillus leucophaeus*
0.0269	0.0343	0.209	0.0233	*Cercocebus atys*
0.0291	0.0259	0.014	0.0208	*Colobus angolensis*
0.0581	0.0503	0.0673	0.117	*Saimiri boliviensis*
0.0445	0.0498	0.038	0.039	*Cebus capucinus*
0.124	0.0856	0.0996	0.114	*Callithrix jacchus*
0.0472	0.0469	0.0366	0.0677	*Aotus nancymaae*
34.9	**70.1**	**88.6**	**151**	**Maximum range**

Table 11.8: *Repeatability of mutation rate (μ) estimation in mammals, for the extant taxa.*

Correlation (ρ)	N_e	μ	Maximum longevity	Adult weight	Female maturity
N_e	-	0.439^{**}	-0.523^{**}	-0.544^{**}	-0.47^{**}
μ	-	-	-0.832^{**}	-0.835^{**}	-0.833^{**}
Maximum longevity	-	-	-	0.827^{**}	0.845^{**}
Adult weight	-	-	-	-	0.809^{**}
Female maturity	-	-	-	-	-

Correlation (ρ)	N_e	μ	Maximum longevity	Adult weight	Female maturity
N_e	-	0.51^{**}	-0.591^{**}	-0.496^{**}	-0.465^{**}
μ	-	-	-0.771^{**}	-0.722^{**}	-0.679^{**}
Maximum longevity	-	-	-	0.802^{**}	0.812^{**}
Adult weight	-	-	-	-	0.764^{**}
Female maturity	-	-	-	-	-

Correlation (ρ)	N_e	μ	Maximum longevity	Adult weight	Female maturity
N_e	-	0.497^{**}	-0.643^{**}	-0.577^{**}	-0.627^{**}
μ	-	-	-0.803^{**}	-0.795^{**}	-0.739^{**}
Maximum longevity	-	-	-	0.836^{**}	0.843^{**}
Adult weight	-	-	-	-	0.805^{**}
Female maturity	-	-	-	-	-

Correlation (ρ)	N_e	μ	Maximum longevity	Adult weight	Female maturity
N_e	-	0.707^{**}	-0.687^{**}	-0.638^{**}	-0.611^{**}
μ	-	-	-0.85^{**}	-0.865^{**}	-0.83^{**}
Maximum longevity	-	-	-	0.839^{**}	0.851^{**}
Adult weight	-	-	-	-	0.817^{**}
Female maturity	-	-	-	-	-

Table 11.9: *In all four replicates, covariance coefficient between effective population size (N_e), mutation rate per site per unit of time (μ), and life-history traits (maximum longevity, adult weight and female maturity) were computed in placental mammals. Asterisks indicate strength of support ($^{*}pp > 0.95$, $^{**}pp > 0.975$).*

11.3.4 Amino-acid preferences entropy

Experiment	$\langle\Omega\rangle$ (**branch** N_e)	$\langle\Omega\rangle$ (**constant** N_e)
Mammals 18 CDS, replicate 1, Chain 1	1.07 ± 0.10	1.14 ± 0.10
Mammals 18 CDS, replicate 2, Chain 2	1.07 ± 0.09	1.14 ± 0.10
Mammals 18 CDS, replicate 2, Chain 1	1.06 ± 0.10	1.12 ± 0.09
Mammals 18 CDS, replicate 2, Chain 2	1.06 ± 0.09	1.11 ± 0.10
Mammals 18 CDS, replicate 3, Chain 1	1.08 ± 0.12	1.15 ± 0.11
Mammals 18 CDS, replicate 3, Chain 2	1.04 ± 0.10	1.18 ± 0.11
Mammals 18 CDS, replicate 4, Chain 1	0.94 ± 0.11	1.02 ± 0.12
Mammals 18 CDS, replicate 4, Chain 2	0.89 ± 0.11	1.02 ± 0.11
Mammals 36 CDS, replicate 1, Chain 1	1.02 ± 0.06	1.07 ± 0.10
Mammals 36 CDS, replicate 1, Chain 2	0.91 ± 0.07	1.03 ± 0.07
Mammals 36 CDS, replicate 2, Chain 1	0.92 ± 0.09	0.96 ± 0.09
Mammals 36 CDS, replicate 2, Chain 2	1.01 ± 0.09	1.02 ± 0.11
Mammals 36 CDS, replicate 3, Chain 1	0.93 ± 0.00	1.05 ± 0.09
Mammals 36 CDS, replicate 3, Chain 2	1.02 ± 0.07	1.05 ± 0.11
Mammals 36 CDS, replicate 4, Chain 1	1.04 ± 0.07	1.10 ± 0.08
Mammals 36 CDS, replicate 4, Chain 2	1.03 ± 0.10	1.08 ± 0.08
Mammals 36 CDS, replicate 5, Chain 1	1.03 ± 0.10	1.03 ± 0.08
Mammals 36 CDS, replicate 5, Chain 2	0.99 ± 0.10	1.04 ± 0.08
Mammals 36 CDS, replicate 6, Chain 1	1.05 ± 0.10	1.10 ± 0.08
Mammals 36 CDS, replicate 6, Chain 2	0.97 ± 0.11	1.10 ± 0.10

Table 11.10: *Estimated amino-acid entropy in mammals. Obtained with the mechanistic inference model developed in this paper of site-specific amino-acid fitness profiles and log-Brownian process for N_e, μ and life-history traits (in the left column), or under the assumption of constant N_e (in the right column).*

11.3.5 Traits estimation with branch ω (replicate 1, chain 1)

Obtained with the phenomenological inference model of log-Brownian process for the μ and the relative non-synonymous substitution rate (ω), as in Lartillot and Poujol (2011).

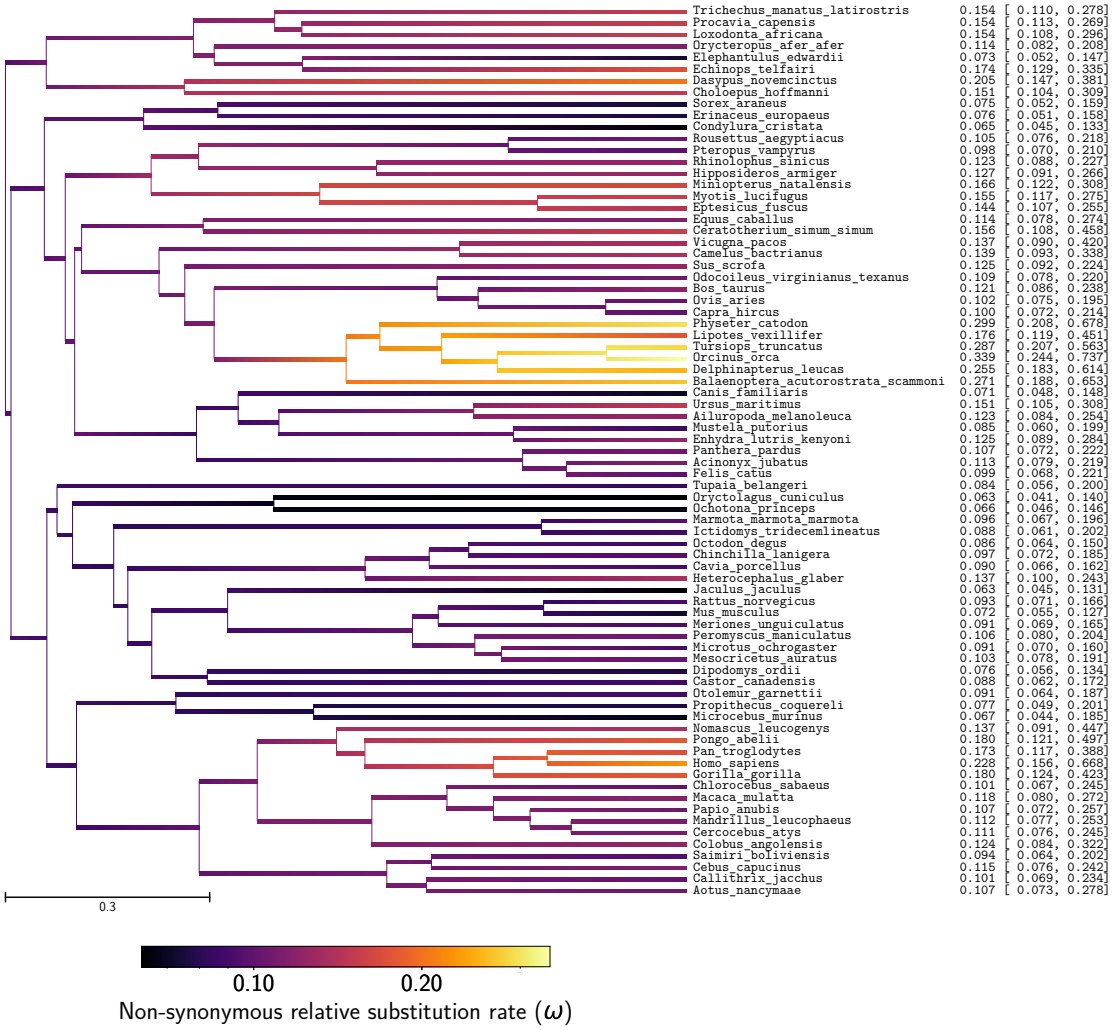


Figure 11.17: *Non-synonymous substitution rate (ω) estimation in mammals*

Correlation (ρ)	ω	μ	Maximum longevity	Adult weight	Female maturity
ω	-	-0.374^{**}	0.544^{**}	0.43^{**}	0.433^{**}
μ	-	-	-0.807^{**}	-0.781^{**}	-0.824^{**}
Maximum longevity	-	-	-	0.801^{**}	0.83^{**}
Adult weight	-	-	-	-	0.785^{**}
Female maturity	-	-	-	-	-

Table 11.11: *Correlation coefficient between non-synonymous substitution rate (ω), mutation rate per site per unit of time (μ), and life-history traits (maximum longevity, adult weight and female maturity) were computed in placental mammals. Asterisks indicate strength of support ($^{*}pp > 0.95$, $^{**}pp > 0.975$).*

Covariance (Σ)	ω	μ	Maximum longevity	Adult weight	Female maturity
ω	0.215^{**}	-0.236^{**}	0.231^{**}	0.828^{**}	0.242^{**}
μ	-	1.82^{**}	-0.998^{**}	-4.38^{**}	-1.34^{**}
Maximum longevity	-	-	0.837^{**}	3.04^{**}	0.917^{**}
Adult weight	-	-	-	17.1^{**}	3.93^{**}
Female maturity	-	-	-	-	1.45^{**}

Table 11.12: *Correlation coefficient between non-synonymous substitution rate (ω), mutation rate per site per unit of time (μ), and life-history traits (maximum longevity, adult weight and female maturity) were computed in placental mammals. Asterisks indicate strength of support ($^{*}pp > 0.95$, $^{**}pp > 0.975$).*

Partial coefficient	ω	μ	Maximum longevity	Adult weight	Female maturity
ω	-	0.15	0.369^{**}	0.0468	0.0223
μ	-	-	-0.299^{*}	-0.272	-0.382^{**}
Maximum longevity	-	-	-	0.283^{**}	0.338^{**}
Adult weight	-	-	-	-	0.21^{*}
Female maturity	-	-	-	-	-

Table 11.13: *Partial correlation coefficient between non-synonymous substitution rate (ω), mutation rate per site per unit of time (μ), and life-history traits (maximum longevity, adult weight and female maturity) were computed in placental mammals. Asterisks indicate strength of support ($^{*}pp > 0.95$, $^{**}pp > 0.975$).*

11.4 Empirical data in Isopods

11.4.1 Traits estimation (replicate 1, chain 1)

Obtained with the mechanistic inference model developed in this paper of site-specific amino-acid fitness profiles and log-Brownian process for N_{e}, μ and life-history traits.

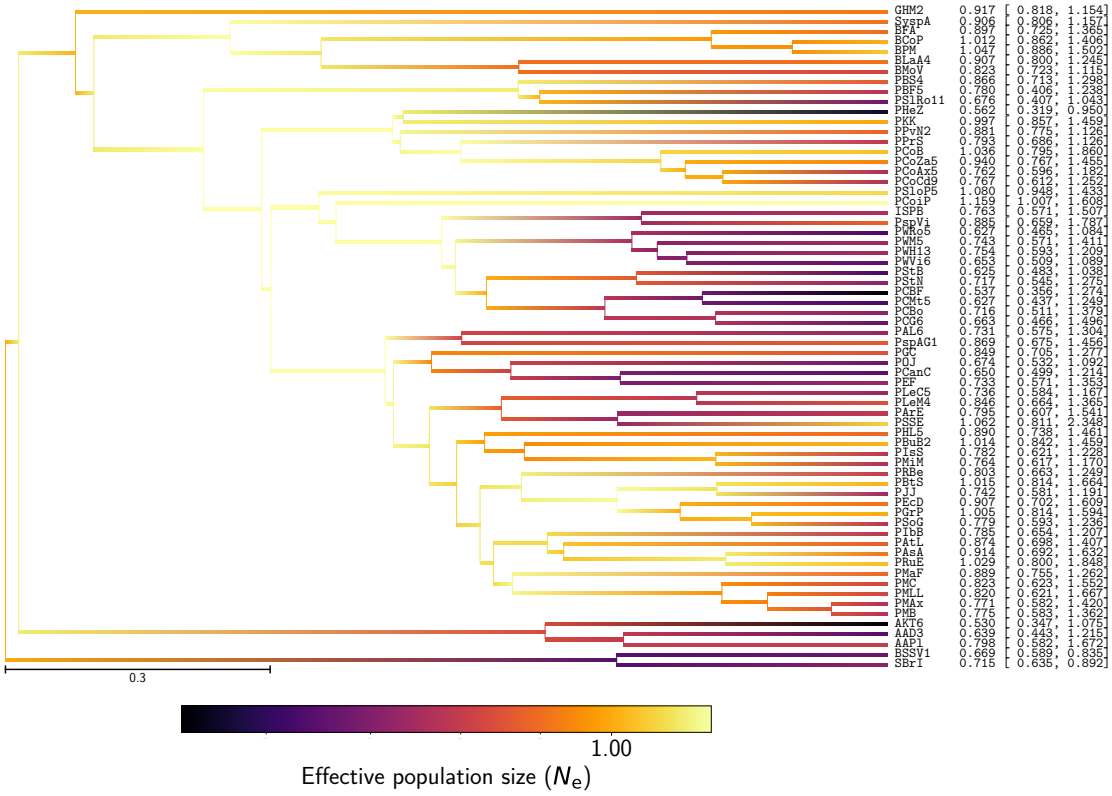


Figure 11.18: *Effective population size (N_e) estimation in isopods*

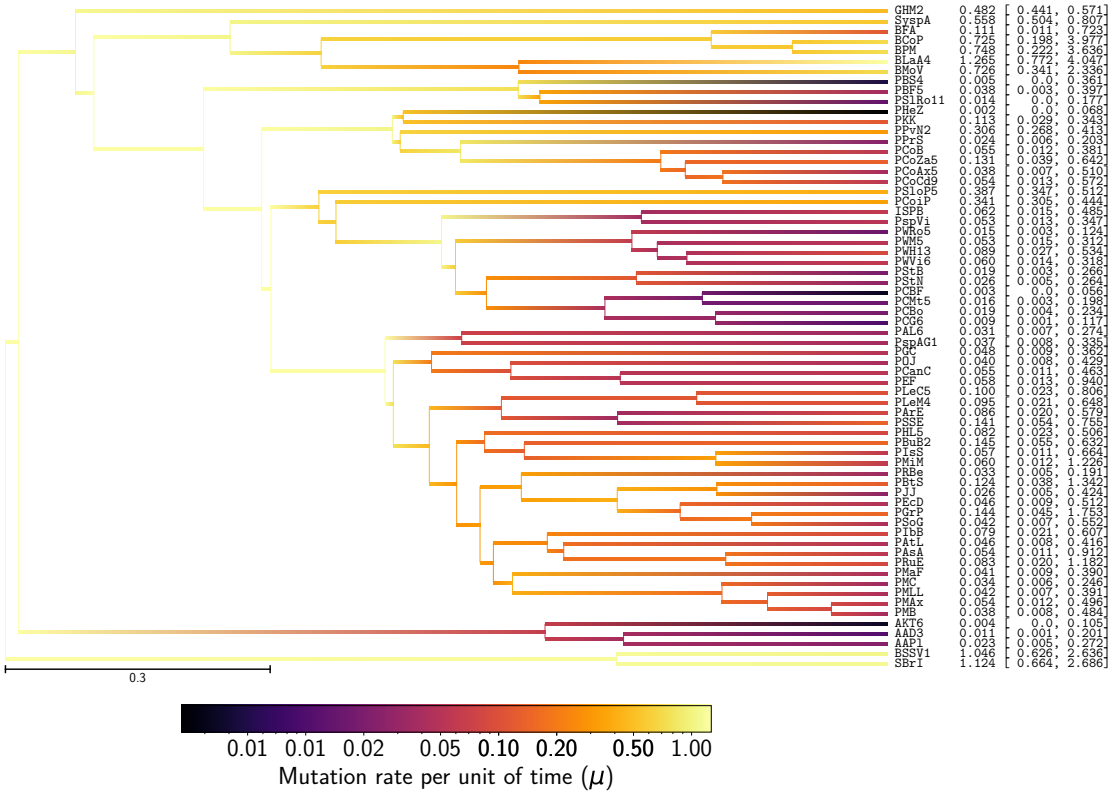


Figure 11.19: *Mutation rate (μ) estimation in isopods*

11.4.2 Repeatability of experiments

6 independent inferences were performed on a randomly chosen set of 12 coding sequences (CDS) out of 135. Obtained with the mechanistic inference model developed in this paper of site-specific amino-acid fitness profiles and log-Brownian process for N_e, μ. Each plot is a correlation between a pair of experiments for a given parameter. For each node (or branch) of the tree, the mean posterior of the parameter over the MCMC (after burn-in) is represented in blue dots, green solid lines are the 90% confidence interval of the MCMC. Solid red line is the regression line between replicates.

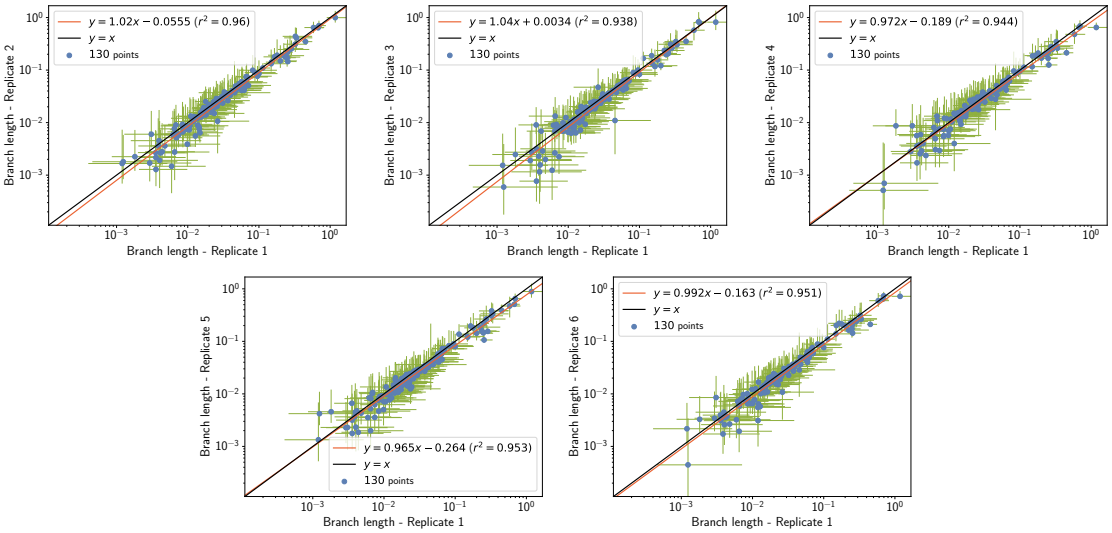


Figure 11.20: *Repeatability of branch length (l) estimation in isopods*

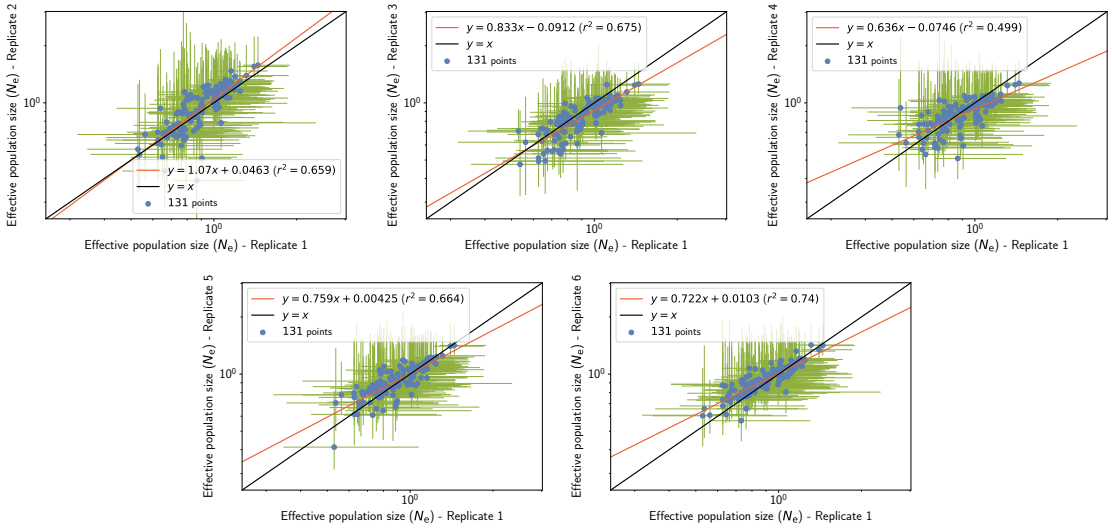


Figure 11.21: *Repeatability of effective population size (N_e) estimation in isopods*

Rep. 1	Rep. 2	Rep. 3	Rep. 4	Rep. 5	Rep. 6	Habitat	Pigmentation	Ocular structure	Code	Taxon
0.917	1.21	0.91	1.09	1.1	0.997	Underground	Depigmented	Anophthalmia	GHM2	*Gallasellus heyli*
0.906	0.511	0.876	0.907	0.948	0.964	Underground	Depigmented	Anophthalmia	SyspA	*Synasellus sp*
0.897	0.954	0.838	0.819	0.721	0.859	Underground	Depigmented	Anophthalmia	BFA	*Bragasellus frontellum*
1.01	1.15	0.78	0.894	0.78	0.945	Surface	Pigmented	Ocular	BCoP	*Bragasellus cortesi*
1.05	0.95	0.881	0.815	0.87	1.08	Surface	Pigmented	Ocular	BPM	*Bragasellus peltatus*
0.907	0.891	0.769	0.831	0.873	0.907	Underground	Depigmented	Anophthalmia	BLaA4	-
0.823	0.797	0.839	0.949	0.852	1.06	Underground	Depigmented	Anophthalmia	BMoV	*Bragasellus molinai*
0.866	0.39	0.61	0.51	0.773	0.871	Underground	Depigmented	Anophthalmia	PBS4	*Proasellus boui*
0.78	0.637	0.702	0.808	0.919	0.884	Underground	Depigmented	Anophthalmia	PBF5	*Proasellus boui*
0.676	0.657	0.718	0.895	0.777	0.884	Underground	Depigmented	Anophthalmia	PSlRo11	*Proasellus slavus*
0.562	0.681	0.62	0.616	0.778	0.608	Underground	Depigmented	Anophthalmia	PHeZ	*Proasellus hercegovinensis*
0.997	1.15	0.962	1.08	0.907	1.06	Surface	Pigmented	Ocular	PKK	*Proasellus karamani*
0.881	0.897	0.753	0.746	0.877	0.901	Underground	Depigmented	Anophthalmia	PPvN2	*Proasellus pavani*
0.793	0.66	0.876	0.711	0.813	0.808	Underground	Depigmented	Anophthalmia	PPrS	*Proasellus parvulus*
1.04	1.06	1.06	0.968	1.05	0.957	Surface	Pigmented	Ocular	PCoB	*Proasellus coxalis*
0.94	1.17	1.06	0.937	1.18	1.14	Surface	Pigmented	Ocular	PCoZa5	*Proasellus coxalis*
0.762	0.595	0.792	0.731	0.902	0.657	Underground	Depigmented	Anophthalmia	PCoAx5	*Proasellus coxalis*
0.767	0.705	0.711	0.848	0.847	0.761	Underground	Depigmented	Microphthalmia	PCoCd9	*Proasellus coxalis*
1.08	1.09	0.893	0.975	1.1	0.968	Underground	Depigmented	Anophthalmia	PSloP5	*Proasellus slovenicus*
1.16	1.28	1.02	1.07	1.21	1.32	Surface	Pigmented	Ocular	PCoiP	*Proasellus coiffaiti*
0.763	0.888	0.753	0.709	0.766	0.86	Underground	Depigmented	Anophthalmia	ISPB	*Proasellus nsp*
0.885	0.789	0.765	0.675	0.67	0.857	Underground	Depigmented	Anophthalmia	PspVi	*Proasellus nsp*
0.627	0.636	0.493	0.771	0.754	0.732	Underground	Depigmented	Anophthalmia	PWRo5	*Proasellus walteri*
0.743	0.671	0.558	1.03	0.834	0.761	Underground	Depigmented	Anophthalmia	PWM5	*Proasellus walteri*
0.754	0.718	0.54	0.79	0.656	0.824	Underground	Depigmented	Anophthalmia	PWH13	*Proasellus walteri*
0.653	0.689	0.538	0.774	0.667	0.764	Underground	Depigmented	Anophthalmia	PWVi6	*Proasellus walteri*
0.625	0.701	0.682	0.703	0.796	0.875	Underground	Depigmented	Anophthalmia	PStB	*Proasellus strouhali*
0.717	0.693	0.682	0.582	0.839	0.734	Underground	Depigmented	Anophthalmia	PStN	*Proasellus strouhali*
0.537	0.535	0.475	0.941	0.705	0.656	Underground	Depigmented	Anophthalmia	PCBF	*Proasellus cavaticus*
0.627	0.611	0.512	0.659	0.614	0.68	Underground	Depigmented	Anophthalmia	PCMt5	*Proasellus cavaticus*
0.716	0.761	0.632	0.603	0.844	0.747	Underground	Depigmented	Anophthalmia	PCBo	*Proasellus cavaticus*
0.663	0.437	0.495	0.535	0.689	0.793	Underground	Depigmented	Anophthalmia	PCG6	*Proasellus cavaticus*
0.731	0.668	0.778	0.805	0.608	0.568	Underground	Depigmented	Anophthalmia	PAL6	*Proasellus albigensis*
0.869	0.737	0.729	0.92	0.827	0.896	Underground	Depigmented	Anophthalmia	PspAG1	*Proasellus n*
0.849	0.839	0.95	0.966	0.931	0.94	Underground	Depigmented	Anophthalmia	PGC	*Proasellus grafi*
0.674	0.79	0.719	0.772	0.784	0.826	Surface	Part. dep.	Microphthalmia	POJ	*Proasellus ortizi*
0.65	0.517	0.732	0.613	0.711	0.729	Underground	Depigmented	Anophthalmia	PCanC	*Proasellus cantabricus*
0.733	0.659	0.684	0.579	0.795	0.802	Surface	Part. dep.	Microphthalmia	PEF	*Proasellus ebrensis*
0.736	0.743	0.8	0.685	0.805	0.935	Underground	Depigmented	Anophthalmia	PLeC5	-
0.846	0.711	0.785	0.864	0.957	0.889	Underground	Depigmented	Anophthalmia	PLeM4	-
0.795	1.03	0.846	0.87	0.869	0.851	Surface	Part. dep.	Microphthalmia	PArE	*Proasellus aragonensis*
1.06	0.785	0.694	0.756	0.893	0.804	Underground	Depigmented	Anophthalmia	PSSE	*Proasellus spelaeus*
0.89	0.774	0.727	0.766	0.926	0.992	Underground	Depigmented	Anophthalmia	PHL5	-
1.01	0.994	0.783	0.789	1.14	1.17	Underground	Depigmented	Anophthalmia	PBuB2	-
0.782	1.14	0.991	0.853	1	1.07	Surface	Pigmented	Ocular	PIsS	*Proasellus istrianus*
0.764	0.878	0.853	0.736	0.887	0.966	Surface	Pigmented	Ocular	PMiM	*Proasellus micropectinatus*
0.803	1.08	0.819	0.96	1.1	0.823	Surface	Part. dep.	Microphthalmia	PRBe	*Proasellus racovitzai*
1.01	1.1	0.905	0.931	0.884	1.01	Surface	Pigmented	Ocular	PBtS	*Proasellus beticus*
0.742	0.896	0.842	0.826	0.84	0.882	Underground	Depigmented	Microphthalmia	PJJ	*Proasellus jaloniacus*
0.907	1.04	0.792	0.594	0.859	0.836	Underground	Depigmented	Anophthalmia	PEcD	*Proasellus escolai*
1.01	1.02	0.86	0.786	1.06	0.922	Surface	Part. dep.	Microphthalmia	PGrP	*Proasellus granadensis*
0.779	0.738	0.707	0.731	0.812	0.808	Underground	Depigmented	Anophthalmia	PSoG	*Proasellus solanasi*
0.785	0.918	0.854	0.788	1.04	0.986	Surface	Pigmented	Ocular	PIbB	*Proasellus ibericus*
0.874	0.836	0.764	0.815	0.866	1.05	Underground	Depigmented	Anophthalmia	PAtL	*Proasellus arthrodilus*
0.914	0.951	0.888	0.834	0.881	0.886	Surface	Part. dep.	Microphthalmia	PAsA	*Proasellus assaforensis*
1.03	0.994	0.939	0.854	0.971	1	Underground	Depigmented	Anophthalmia	PRuE	*Proasellus rectus*
0.889	0.764	0.761	0.651	0.698	0.847	Underground	Depigmented	Anophthalmia	PMaF	*Proasellus margalefi*
0.823	1.1	0.961	0.914	1.03	0.838	Surface	Pigmented	Ocular	PMC	*Proasellus meridianus*
0.82	1.05	0.806	0.914	0.856	0.799	Surface	Pigmented	Ocular	PMLL	*Proasellus meridianus*
0.771	0.889	0.796	0.907	0.839	0.863	Underground	Part. dep.	Microphthalmia	PMAx	*Proasellus meridianus*
0.775	0.982	0.892	1.02	1.05	0.882	Surface	Pigmented	Ocular	PMB	*Proasellus meridianus*
0.53	0.57	0.71	0.679	0.413	0.603	Underground	Depigmented	Anophthalmia	AKT6	*Asellus kosswigi*
0.639	0.936	0.732	0.859	0.859	0.866	Surface	Pigmented	Ocular	AAD3	*Asellus aquaticus*
0.798	0.882	0.795	0.89	0.641	0.875	Surface	Pigmented	Ocular	AAPl	*Asellus aquaticus*
0.669	0.704	0.684	0.655	0.805	0.711	Underground	Depigmented	Anophthalmia	BSSV1	*Balkanostenasellus skopljensis*
0.715	0.682	0.685	0.614	0.707	0.77	Underground	Depigmented	Anophthalmia	SBrI	*Stenasellus breuili*
2.19	**3.29**	**2.24**	**2.13**	**2.93**	**2.33**	-	-	-	-	**Maximum range**

Table 11.14: *Repeatability of effective population size (N_e) estimation in isopods, for the extant taxa.*

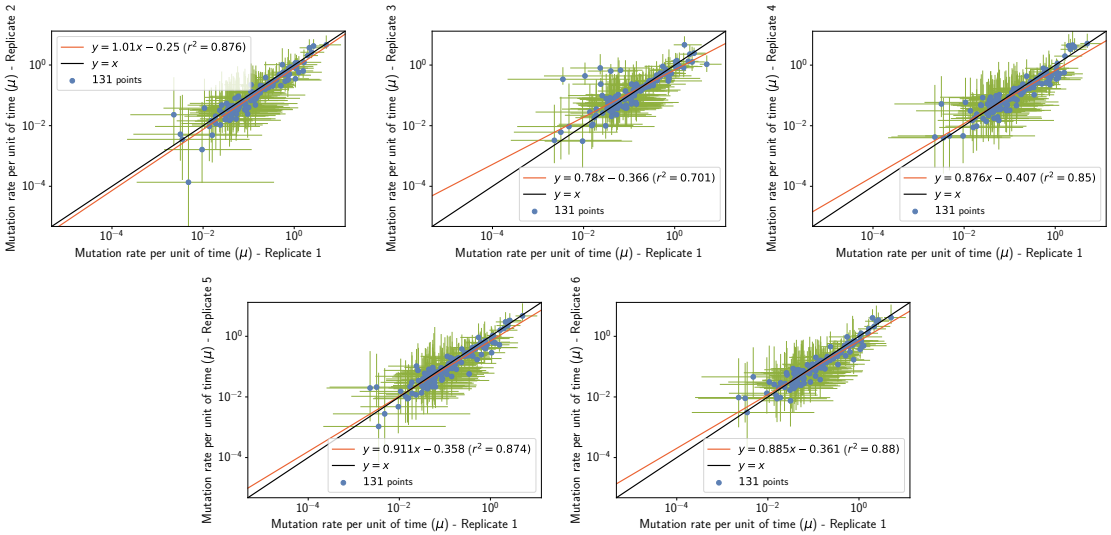


Figure 11.22: *Repeatability of mutation rate (μ) estimation in isopods*

Rep. 1	Rep. 2	Rep. 3	Rep. 4	Rep. 5	Rep. 6	Habitat	Pigmentation	Ocular structure	Code	Taxon
0.482	0.488	0.54	0.462	0.456	0.497	Underground	Depigmented	Anophthalmia	GHM2	*Gallasellus heyli*
0.558	0.563	0.536	0.52	0.503	0.59	Underground	Depigmented	Anophthalmia	SyspA	*Synasellus sp*
0.111	0.0674	0.0703	0.0781	0.0555	0.0852	Underground	Depigmented	Anophthalmia	BFA	*Bragasellus frontellum*
0.725	0.32	0.3	0.377	0.421	0.579	Surface	Pigmented	Ocular	BCoP	*Bragasellus cortesi*
0.748	0.337	0.364	0.331	0.427	0.692	Surface	Pigmented	Ocular	BPM	*Bragasellus peltatus*
1.27	0.579	0.893	0.516	0.607	0.49	Underground	Depigmented	Anophthalmia	BLaA4	-
0.726	0.462	0.442	0.463	0.428	0.519	Underground	Depigmented	Anophthalmia	BMoV	*Bragasellus molinai*
0.00478	0.000136	0.00937	0.00462	0.00277	0.046	Underground	Depigmented	Anophthalmia	PBS4	*Proasellus boui*
0.038	0.0347	0.0822	0.0873	0.048	0.0628	Underground	Depigmented	Anophthalmia	PBF5	*Proasellus boui*
0.0137	0.00942	0.0281	0.031	0.0105	0.0259	Underground	Depigmented	Anophthalmia	PSlRo11	*Proasellus slavus*
0.00228	0.0229	0.00329	0.00422	0.0197	0.00957	Underground	Depigmented	Anophthalmia	PHeZ	*Proasellus hercegovinensis*
0.113	0.224	0.235	0.103	0.218	0.202	Surface	Pigmented	Ocular	PKK	*Proasellus karamani*
0.306	0.25	0.257	0.225	0.229	0.23	Underground	Depigmented	Anophthalmia	PPvN2	*Proasellus pavani*
0.0238	0.0105	0.235	0.0472	0.104	0.0275	Underground	Depigmented	Anophthalmia	PPrS	*Proasellus parvulus*
0.0545	0.0359	0.0712	0.0458	0.073	0.0453	Surface	Pigmented	Ocular	PCoB	*Proasellus coxalis*
0.131	0.121	0.213	0.149	0.198	0.141	Surface	Pigmented	Ocular	PCoZa5	*Proasellus coxalis*
0.0381	0.015	0.0518	0.0305	0.0522	0.0356	Underground	Depigmented	Anophthalmia	PCoAx5	*Proasellus coxalis*
0.0544	0.0332	0.0673	0.0728	0.0931	0.0672	Underground	Depigmented	Microphthalmia	PCoCd9	*Proasellus coxalis*
0.387	0.328	0.364	0.348	0.3	0.317	Underground	Depigmented	Anophthalmia	PSloP5	*Proasellus slovenicus*
0.341	0.28	0.296	0.246	0.0573	0.253	Surface	Pigmented	Ocular	PCoiP	*Proasellus coiffaiti*
0.062	0.0427	0.0495	0.0556	0.0217	0.0281	Underground	Depigmented	Anophthalmia	ISPB	*Proasellus nsp*
0.0533	0.0419	0.0549	0.0516	0.0228	0.0245	Underground	Depigmented	Anophthalmia	PspVi	*Proasellus nsp*
0.0151	0.0144	0.0101	0.015	0.00874	0.0115	Underground	Depigmented	Anophthalmia	PWRo5	*Proasellus walteri*
0.0531	0.0167	0.0231	0.0911	0.0218	0.0243	Underground	Depigmented	Anophthalmia	PWM5	*Proasellus walteri*
0.0886	0.0317	0.0384	0.0491	0.0402	0.0934	Underground	Depigmented	Anophthalmia	PWH13	*Proasellus walteri*
0.0597	0.093	0.0527	0.106	0.0297	0.0383	Underground	Depigmented	Anophthalmia	PWVi6	*Proasellus walteri*
0.019	0.0117	0.0353	0.0242	0.0131	0.0291	Underground	Depigmented	Anophthalmia	PStB	*Proasellus strouhali*
0.0263	0.0262	0.026	0.0173	0.0345	0.0534	Underground	Depigmented	Anophthalmia	PStN	*Proasellus strouhali*
0.00317	0.00521	0.00612	0.0523	0.0211	0.00902	Underground	Depigmented	Anophthalmia	PCBF	*Proasellus cavaticus*
0.0159	0.00485	0.0112	0.00879	0.00919	0.00876	Underground	Depigmented	Anophthalmia	PCMt5	*Proasellus cavaticus*
0.0188	0.0205	0.0221	0.00963	0.0302	0.00951	Underground	Depigmented	Anophthalmia	PCBo	*Proasellus cavaticus*
0.0095	0.00162	0.00309	0.00461	0.00475	0.0132	Underground	Depigmented	Anophthalmia	PCG6	*Proasellus cavaticus*
0.0314	0.021	0.0791	0.0263	0.0238	0.00749	Underground	Depigmented	Anophthalmia	PAL6	*Proasellus albigensis*
0.0372	0.0313	0.073	0.0375	0.0454	0.0191	Underground	Depigmented	Anophthalmia	PspAG1	*Proasellus n*
0.0477	0.0403	0.0817	0.0666	0.0594	0.0425	Underground	Depigmented	Anophthalmia	PGC	*Proasellus grafi*
0.0404	0.03	0.0467	0.0547	0.0285	0.0343	Surface	Part. dep.	Microphthalmia	POJ	*Proasellus ortizi*
0.0551	0.0385	0.0367	0.0343	0.0369	0.052	Underground	Depigmented	Anophthalmia	PCanC	*Proasellus cantabricus*
0.0578	0.0337	0.0446	0.0149	0.0349	0.0499	Surface	Part. dep.	Microphthalmia	PEF	*Proasellus ebrensis*
0.1	0.0399	0.0968	0.0322	0.0743	0.104	Underground	Depigmented	Anophthalmia	PLeC5	-
0.095	0.0243	0.0793	0.0904	0.0706	0.12	Underground	Depigmented	Anophthalmia	PLeM4	-
0.0856	0.0426	0.0472	0.0317	0.0361	0.0311	Surface	Part. dep.	Microphthalmia	PArE	*Proasellus aragonensis*
0.141	0.0482	0.0583	0.0461	0.0474	0.0372	Underground	Depigmented	Anophthalmia	PSSE	*Proasellus spelaeus*
0.0822	0.0338	0.0387	0.0355	0.0522	0.059	Underground	Depigmented	Anophthalmia	PHL5	-
0.145	0.0777	0.052	0.0705	0.0868	0.132	Underground	Depigmented	Anophthalmia	PBuB2	-
0.0573	0.0857	0.0794	0.0706	0.0643	0.0843	Surface	Pigmented	Ocular	PIsS	*Proasellus istrianus*
0.0599	0.0387	0.0584	0.0471	0.0606	0.0728	Surface	Pigmented	Ocular	PMiM	*Proasellus micropectinatus*
0.0328	0.0543	0.0365	0.052	0.0391	0.0289	Surface	Part. dep.	Microphthalmia	PRBe	*Proasellus racovitzai*
0.124	0.0797	0.15	0.168	0.128	0.12	Surface	Pigmented	Ocular	PBtS	*Proasellus beticus*
0.0256	0.0475	0.0985	0.0846	0.0751	0.0621	Underground	Depigmented	Microphthalmia	PJJ	*Proasellus jaloniacus*
0.0455	0.03	0.0441	0.0175	0.0339	0.0279	Underground	Depigmented	Anophthalmia	PEcD	*Proasellus escolai*
0.144	0.102	0.234	0.163	0.185	0.161	Surface	Part. dep.	Microphthalmia	PGrP	*Proasellus granadensis*
0.0423	0.028	0.0529	0.0454	0.0455	0.0635	Underground	Depigmented	Anophthalmia	PSoG	*Proasellus solanasi*
0.0788	0.0552	0.0724	0.0547	0.0686	0.0646	Surface	Pigmented	Ocular	PIbB	*Proasellus ibericus*
0.0458	0.0447	0.0517	0.0443	0.0601	0.0791	Underground	Depigmented	Anophthalmia	PAtL	*Proasellus arthrodilus*
0.0543	0.0351	0.0573	0.0478	0.065	0.0503	Surface	Part. dep.	Microphthalmia	PAsA	*Proasellus assaforensis*
0.083	0.0536	0.139	0.0608	0.0918	0.0674	Underground	Depigmented	Anophthalmia	PRtE	*Proasellus rectus*
0.0413	0.0168	0.0202	0.0369	0.0292	0.0359	Underground	Depigmented	Anophthalmia	PMaF	*Proasellus margalefi*
0.0343	0.034	0.0559	0.0219	0.0199	0.0235	Surface	Pigmented	Ocular	PMC	*Proasellus meridianus*
0.0416	0.0215	0.0337	0.0312	0.0187	0.0231	Surface	Pigmented	Ocular	PMLL	*Proasellus meridianus*
0.0542	0.0241	0.044	0.0194	0.0247	0.0486	Underground	Part. dep.	Microphthalmia	PMAx	*Proasellus meridianus*
0.0385	0.0252	0.0807	0.0782	0.0541	0.06	Surface	Pigmented	Ocular	PMB	*Proasellus meridianus*
0.00351	0.00353	0.34	0.00401	0.00106	0.00307	Underground	Depigmented	Anophthalmia	AKT6	*Asellus kosswigi*
0.0107	0.0378	0.438	0.0589	0.0152	0.031	Surface	Pigmented	Ocular	AAD3	*Asellus aquaticus*
0.0232	0.0321	0.802	0.048	0.0118	0.0209	Surface	Pigmented	Ocular	AAPl	*Asellus aquaticus*
1.05	0.901	1.31	0.585	0.841	0.711	Underground	Depigmented	Anophthalmia	BSSV1	*Balkanostenasellus skopljensis*
1.12	0.784	1.11	0.479	0.641	0.701	Underground	Depigmented	Anophthalmia	SBrI	*Stenasellus breuili*
554	**6.63e+03**	**423**	**146**	**792**	**231**	-	-	-	-	**Maximum range**

Table 11.15: *Repeatability of mutation rate (μ) estimation in isopods, for the extant taxa.*

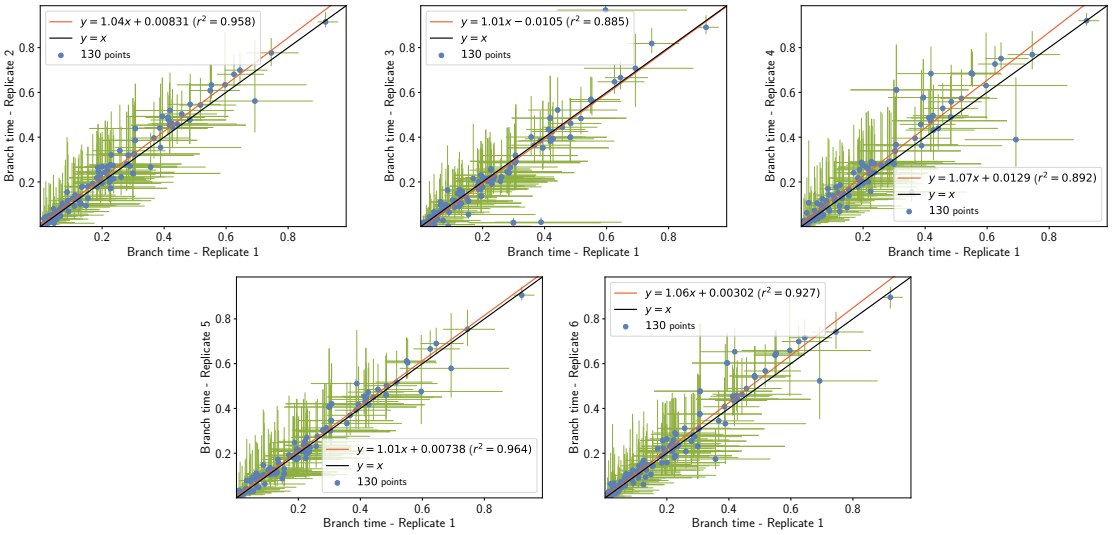


Figure 11.23: *Repeatability of branch time (ΔT) estimation in isopods*

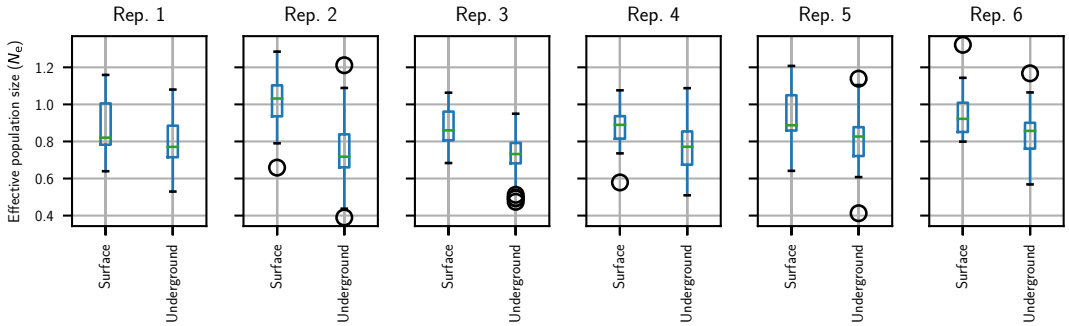


Figure 11.24: *N_e as a function of habitat in isopods.*

```
Analysis of Variance Table

Response: PopulationSize
           Df Sum Sq Mean Sq F value    Pr(>F)
Habitat     1 1.6777 1.67769  89.506 < 2.2e-16 ***
rep         5 0.4226 0.08452   4.509 0.0005236 ***
Residuals 389 7.2913 0.01874
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

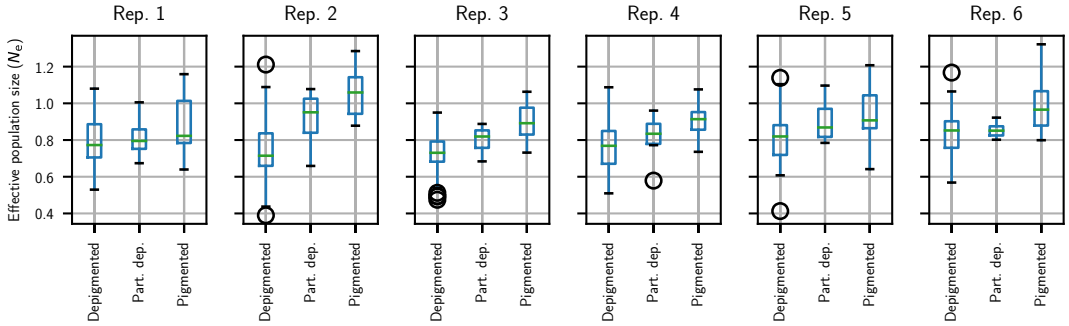


Figure 11.25: *N_e as a function of pigmentation in isopods*

```
Analysis of Variance Table

Response: PopulationSize
              Df Sum Sq Mean Sq F value    Pr(>F)
Pigmentation   2 1.9442 0.97210 53.6917 < 2.2e-16 ***
rep            5 0.4226 0.08452  4.6681 0.0003764 ***
Residuals    388 7.0248 0.01811
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

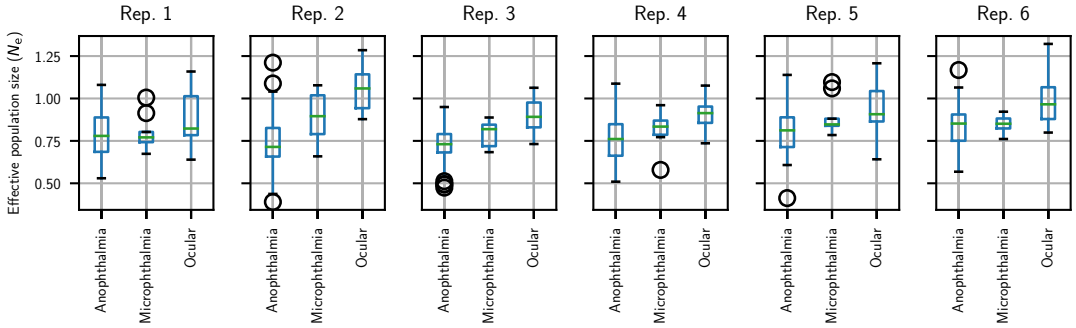


Figure 11.26: *N_e as a function of ocular structure in isopods*

```
Analysis of Variance Table

Response: PopulationSize
                 Df Sum Sq Mean Sq F value    Pr(>F)
Ocular.structure  2 1.9335 0.96676  53.316 < 2.2e-16 ***
rep               5 0.4226 0.08452   4.661  0.000382 ***
Residuals       388 7.0355 0.01813
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

11.5 Empirical data in Primates

11.5.1 Chain convergence

Obtained with the mechanistic inference model developed in this paper of site-specific amino-acid fitness profiles and log-Brownian process for N_e, μ and life-history traits.

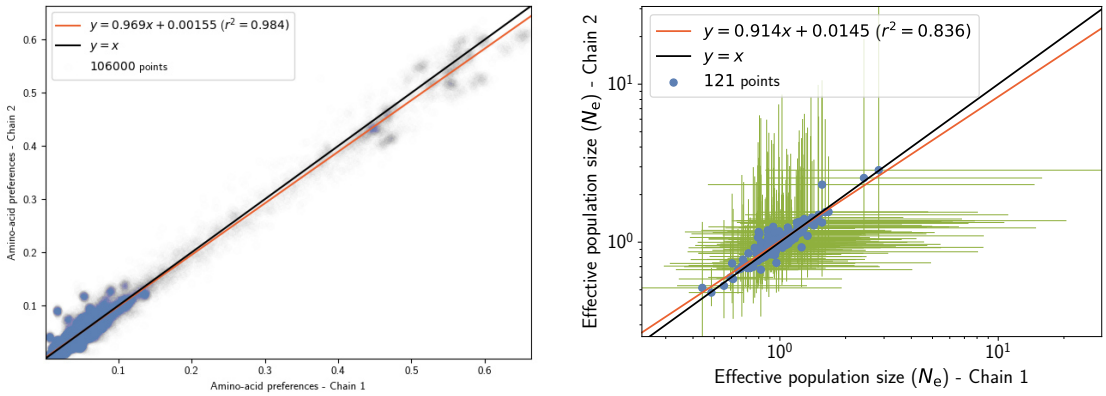


Figure 11.27: *Chain convergence of site amino-acid preferences (left panel) and branch N_e (right panel).*

11.5.2 Traits estimation (chain 1)

Obtained with the mechanistic inference model developed in this paper of site-specific amino-acid fitness profiles and log-Brownian process for N_e, μ and life-history traits.

Correlation (ρ)	N_e	μ	maturity	mass	longevity	π_S	π_N/π_S	generation time
N_e	-	-0.433^{**}	0.155	0.166	0.157	-0.133	0.104	0.16
μ	-	-	-0.792^{**}	-0.791^{**}	-0.773^{**}	0.62^{**}	-0.59	-0.78^{**}
maturity	-	-	-	0.986^{**}	0.985^{**}	-0.8^{**}	0.746	0.991^{**}
mass	-	-	-	-	0.977^{**}	-0.737^{**}	0.695	0.981^{**}
longevity	-	-	-	-	-	-0.819^{**}	0.752	0.999^{**}
π_S	-	-	-	-	-	-	-0.86^{**}	-0.816^{**}
π_N/π_S	-	-	-	-	-	-	-	0.752
generation time	-	-	-	-	-	-	-	-

Table 11.16: *Correlation coefficient between effective population size (N_e), mutation rate per site per unit of time (μ), and life-history traits (maximum longevity, adult weight and female maturity) were computed in primates. Asterisks indicate strength of support ($^*pp > 0.95$, $^{**}pp > 0.975$).*

Covariance ($\boldsymbol{\Sigma}$)	N_e	μ	maturity	mass	longevity	π_S	π_N/π_S	generation time
N_e	1.08^{**}	-1.39^{**}	0.66	1.18	0.414	-0.251	0.0898	0.452
μ	-	9.86^{**}	-10.1^{**}	-17.5^{**}	-6.44^{**}	3.42^{**}	-1.28	-6.96^{**}
maturity	-	-	16.9^{**}	28.4^{**}	10.6^{**}	-5.39^{**}	1.9	11.5^{**}
mass	-	-	-	49.8^{**}	18.1^{**}	-8.89^{**}	3.29	19.5^{**}
longevity	-	-	-	-	6.99^{**}	-3.75^{**}	1.31	7.47^{**}
π_S	-	-	-	-	-	3.26^{**}	-0.986^{**}	-3.96^{**}
π_N/π_S	-	-	-	-	-	-	0.419^{**}	1.39
generation time	-	-	-	-	-	-	-	8.02^{**}

Table 11.17: *Correlation coefficient between effective population size (N_e), mutation rate per site per unit of time (μ), and life-history traits (maximum longevity, adult weight and female maturity) were computed in primates. Asterisks indicate strength of support ($^{*}pp > 0.95$, $^{**}pp > 0.975$).*

Partial coefficient	N_e	μ	maturity	mass	longevity	π_S	π_N/π_S	generation time
N_e	-	-0.411	-0.0622	0.0184	-0.0436	-0.0482	-0.00476	0.0333
μ	-	-	0.0548	-0.101	0.146	-0.0134	-0.102	-0.124
maturity	-	-	-	0.292	-0.793^{**}	-0.167	0.0547	0.824^{**}
mass	-	-	-	-	-0.0589	0.43	-0.195	0.101
longevity	-	-	-	-	-	-0.159	-0.148	0.991^{**}
π_S	-	-	-	-	-	-	-0.573^{**}	0.11
π_N/π_S	-	-	-	-	-	-	-	0.144
generation time	-	-	-	-	-	-	-	-

Table 11.18: *Partial correlation coefficient between Neffective population size (N_e), mutation rate per site per unit of time (μ), and life-history traits (maximum longevity, adult weight and female maturity) were computed in primates. Asterisks indicate strength of support ($^{*}pp > 0.95$, $^{**}pp > 0.975$).*

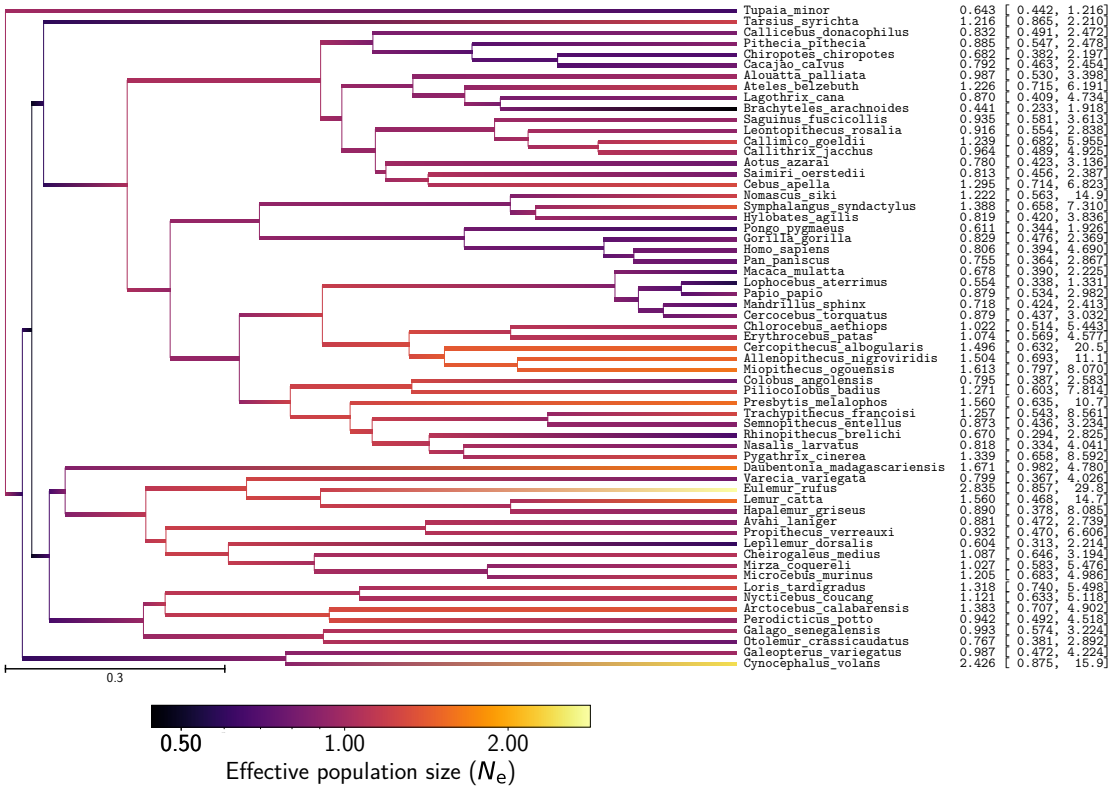


Figure 11.28: *Effective population size (N_e) estimation in primates*

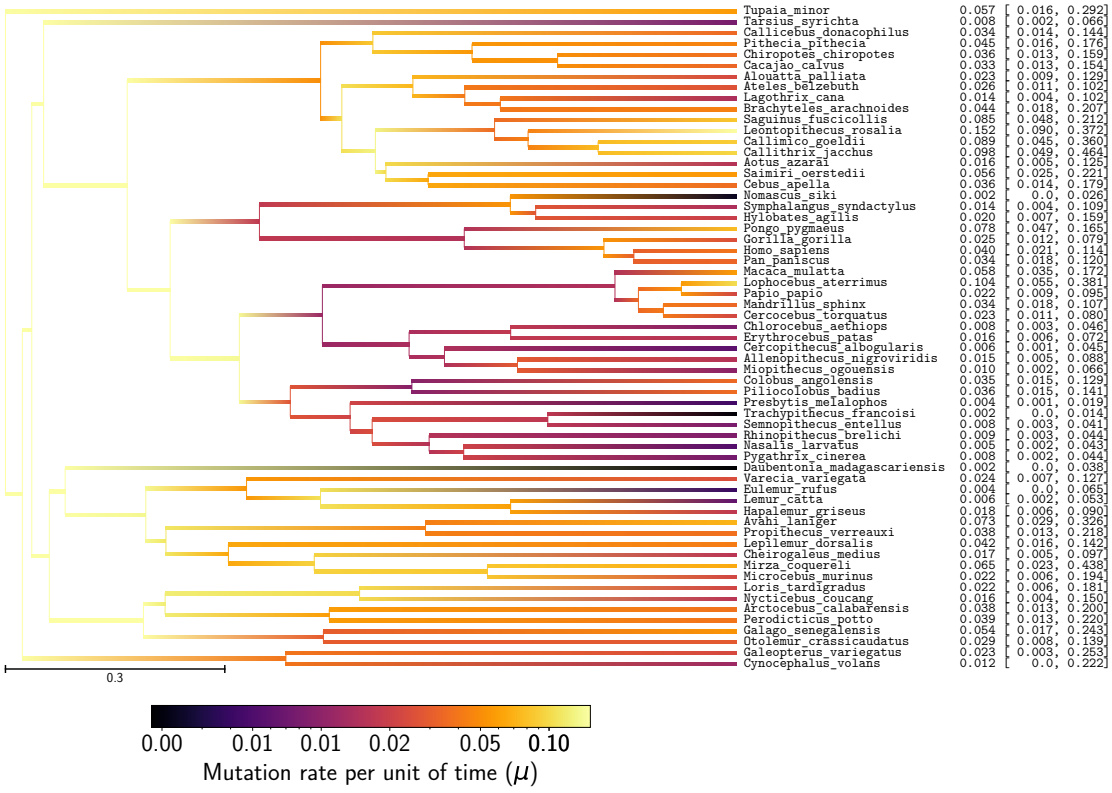


Figure 11.29: *Mutation rate (μ) estimation in primates*

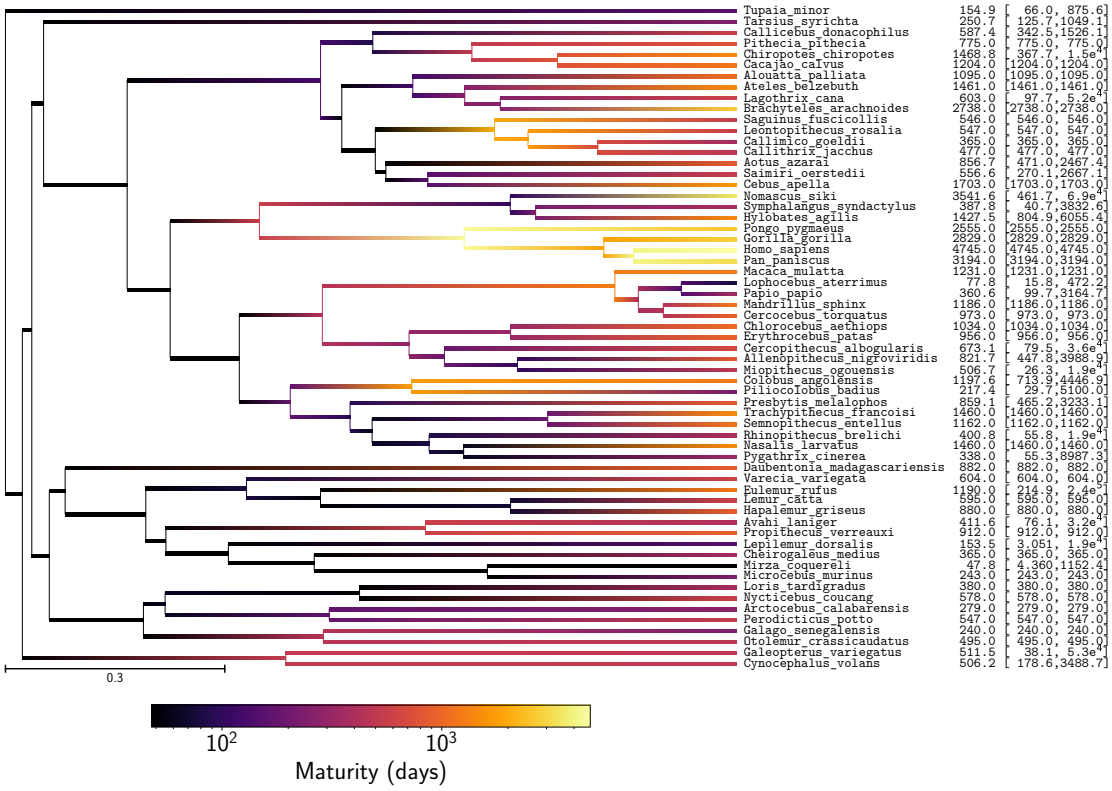


Figure 11.30: *Female maturity estimation in primates*

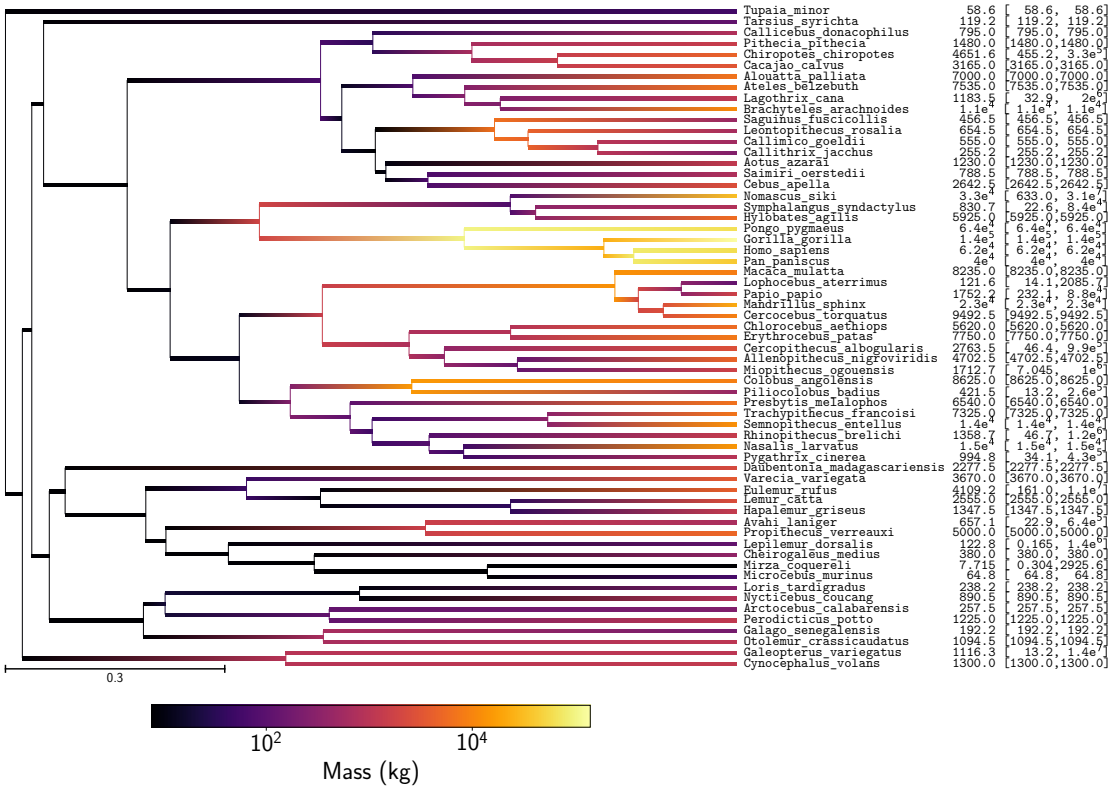


Figure 11.31: *Mass estimation in primates*

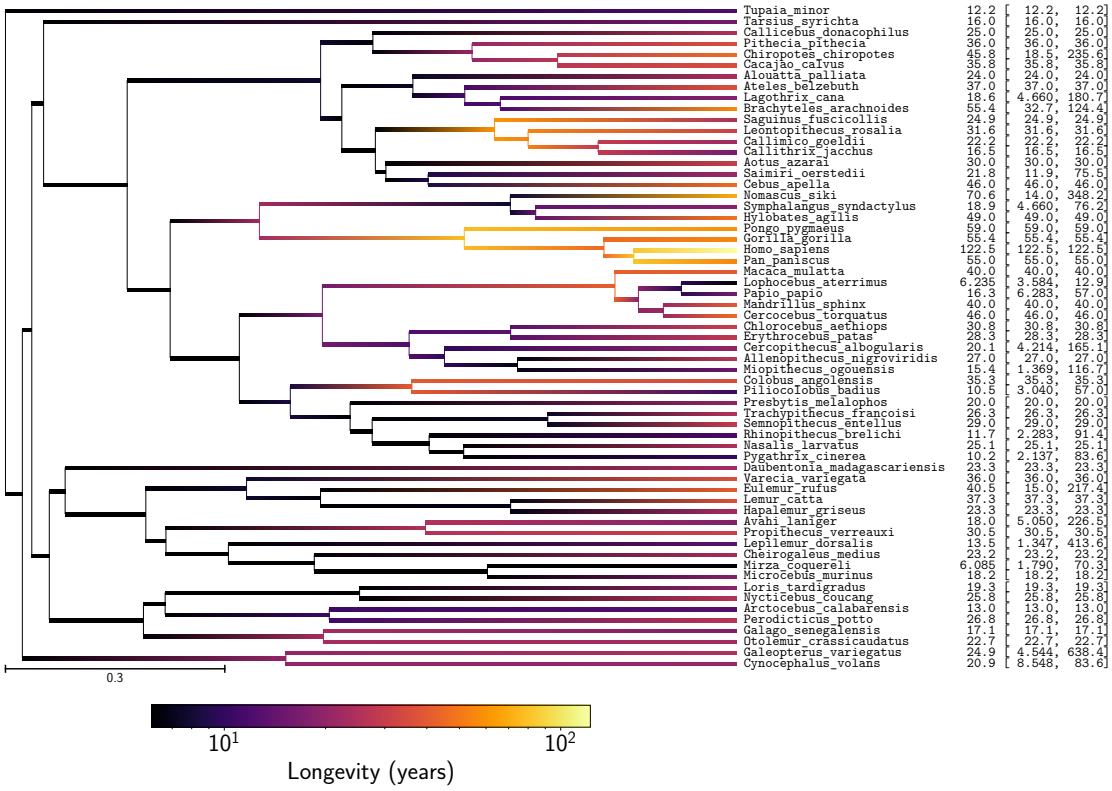


Figure 11.32: *Longevity estimation in primates*

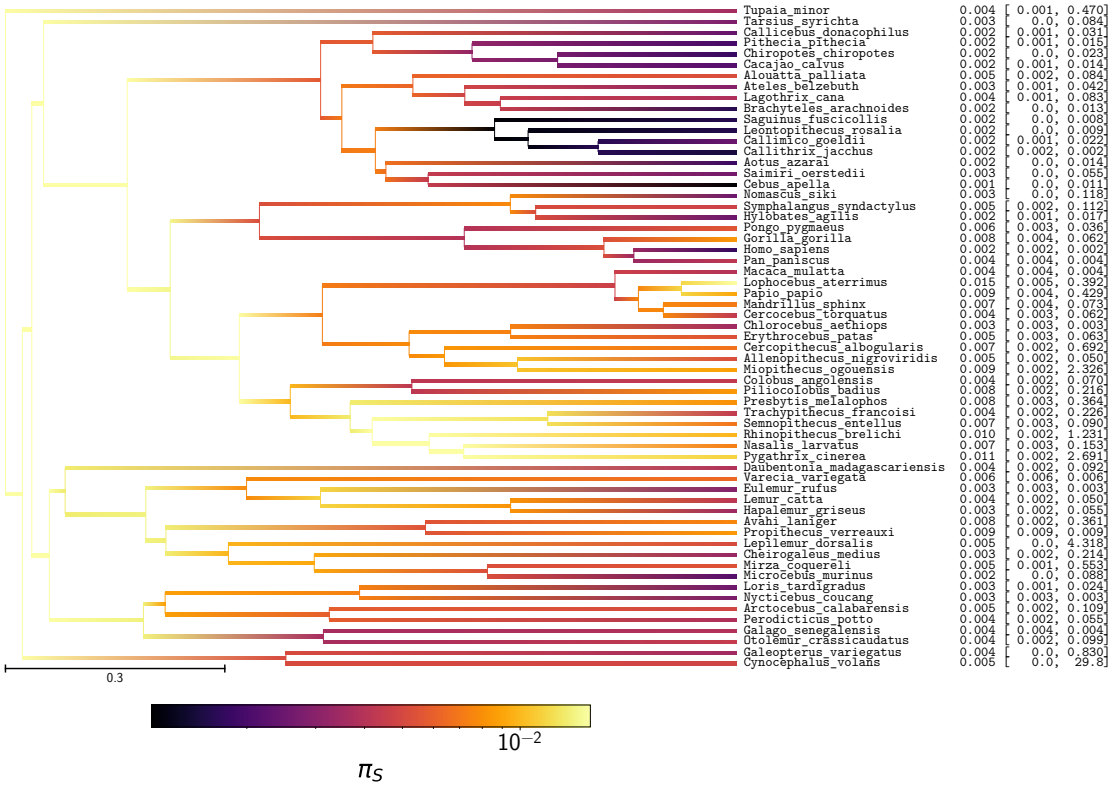


Figure 11.33: *π_S estimation in primates*

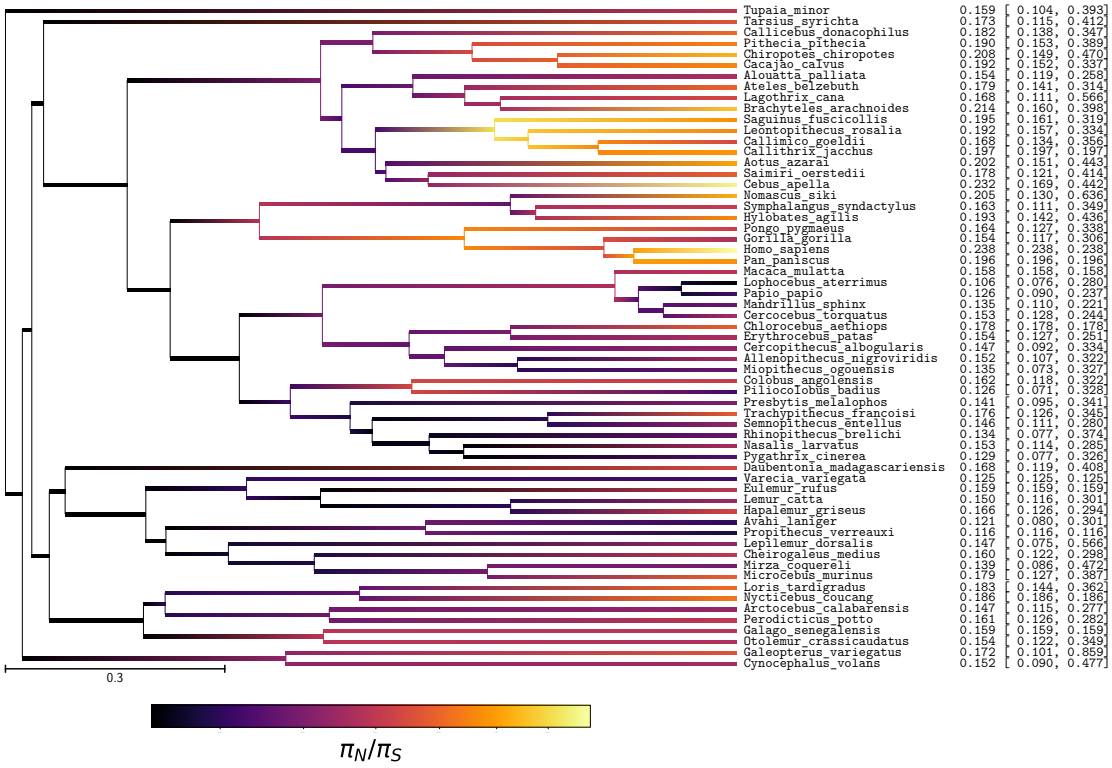


Figure 11.34: *π_N/π_S estimation in primates*

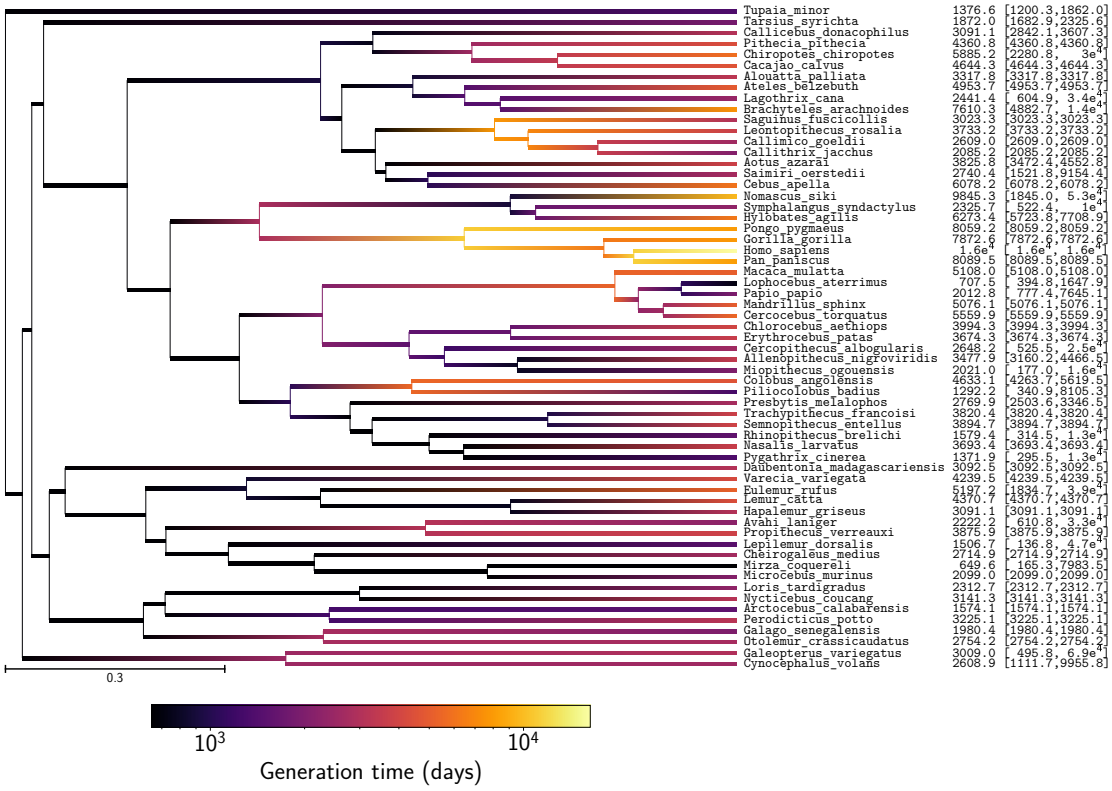


Figure 11.35: *Generation time estimation in primates*

11.5.3 Amino-acid preferences entropy

Experiment	$\langle\Omega\rangle$ (**branch** N_e)	$\langle\Omega\rangle$ (**constant** N_e)
Primates, chain 1	1.41 ± 0.10	1.49 ± 0.08
Primates, chain 2	1.40 ± 0.10	1.48 ± 0.08

Table 11.19: *Estimated amino-acid entropy in primates. Obtained with the mechanistic inference model developed in this paper of site-specific amino-acid fitness profiles and log-Brownian process for N_e, μ and life-history traits (in the left column), or under the assumption of constant N_e (in the right column).*

11.5.4 Traits estimation with branch ω (chain 1)

Obtained with the phenomenological inference model of log-Brownian process for the μ and the relative non-synonymous substitution rate (ω), as in Lartillot and Poujol (2011).

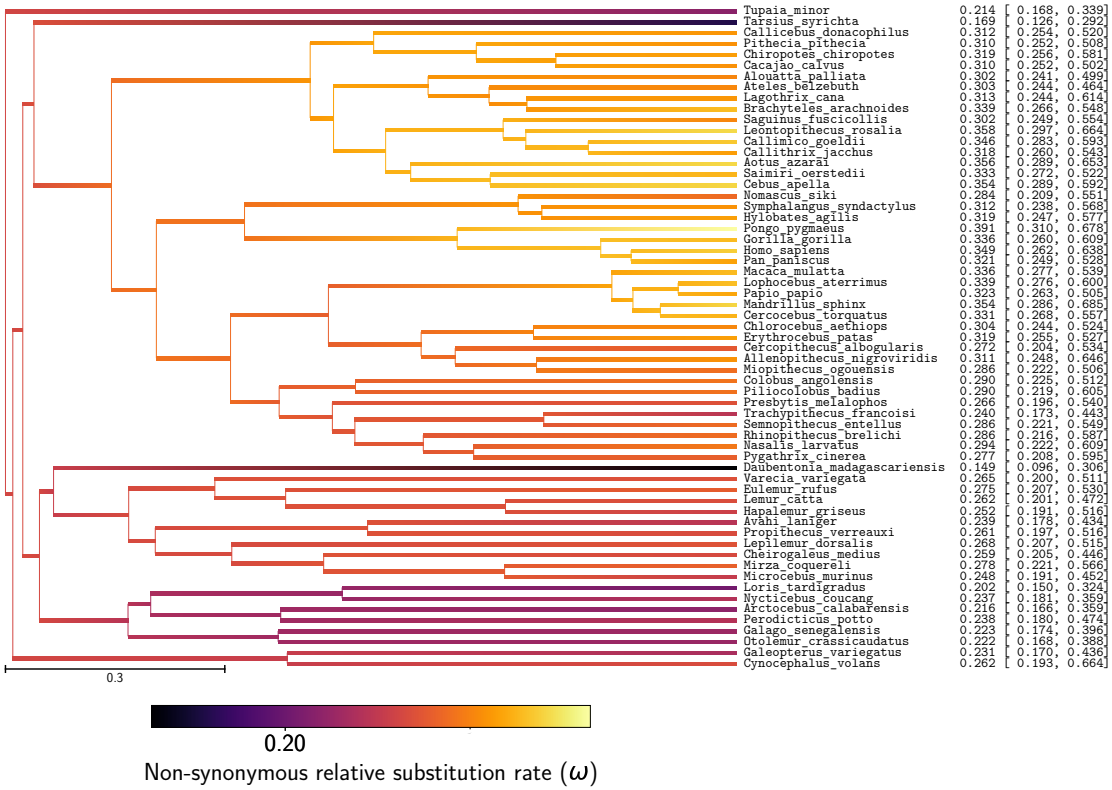


Figure 11.36: *Non-synonymous substitution rate (ω) estimation in primates*

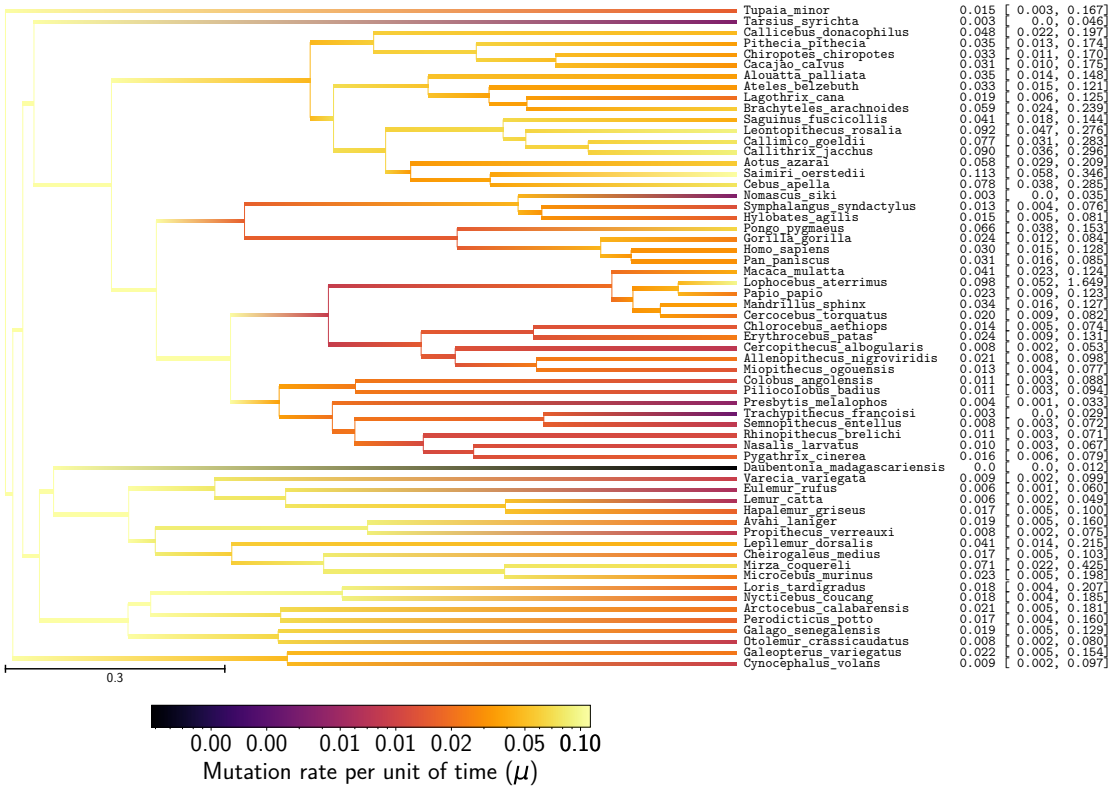


Figure 11.37: *Mutation rate (μ) estimation in primates*

Correlation (ρ)	ω	μ	maturity	mass	longevity	π_S	π_N/π_S	generation time
ω	-	0.294	0.000316	0.0361	0.0155	-0.197	0.145	0.0111
μ	-	-	-0.804^{**}	-0.798^{**}	-0.817^{**}	-0.0201	0.031	-0.823^{**}
maturity	-	-	-	0.952^{**}	0.957^{**}	-0.166	0.162	0.97^{**}
mass	-	-	-	-	0.933^{**}	-0.0437	0.0427	0.943^{**}
longevity	-	-	-	-	-	-0.223	0.165	0.999^{**}
π_S	-	-	-	-	-	-	-0.664	-0.212
π_N/π_S	-	-	-	-	-	-	-	0.162
generation time	-	-	-	-	-	-	-	-

Table 11.20: *Correlation coefficient between non-synonymous substitution rate (ω), mutation rate per site per unit of time (μ), and life-history traits (maximum longevity, adult weight and female maturity) were computed in primates. Asterisks indicate strength of support ($^{*}pp > 0.95$, $^{**}pp > 0.975$).*

Covariance ($\boldsymbol{\Sigma}$)	ω	μ	maturity	mass	longevity	π_S	π_N/π_S	generation time
ω	0.0674^{**}	0.231	-0.0106	0.0149	-0.00138	-0.0435	0.0101	-0.00314
μ	-	8.71^{**}	-4.8^{**}	-9.22^{**}	-3.97^{**}	0.188	0.0483	-4.08^{**}
maturity	-	-	4.95^{**}	8.37^{**}	3.29^{**}	-1.01	0.000924	3.53^{**}
mass	-	-	-	16.3^{**}	6.14^{**}	-0.932	-0.0741	6.45^{**}
longevity	-	-	-	-	2.76^{**}	-0.577	0.0919	2.82^{**}
π_S	-	-	-	-	-	1.3^{**}	-0.148	-0.637
π_N/π_S	-	-	-	-	-	-	0.182^{**}	0.0775
generation time	-	-	-	-	-	-	-	2.92^{**}

Table 11.21: *Correlation coefficient between non-synonymous substitution rate (ω), mutation rate per site per unit of time (μ), and life-history traits (maximum longevity, adult weight and female maturity) were computed in primates. Asterisks indicate strength of support ($^{*}pp > 0.95$, $^{**}pp > 0.975$).*

Partial coefficient	ω	μ	maturity	mass	longevity	π_S	π_N/π_S	generation time
ω	-	0.463	-0.0461	0.248	-0.027	-0.193	-0.0681	0.0319
μ	-	-	0.0649	-0.000258	0.0374	-0.128	0.115	-0.075
maturity	-	-	-	0.228	-0.834^{**}	-0.0991	0.0491	0.854^{**}
mass	-	-	-	-	-0.038	0.435	-0.123	0.0851
longevity	-	-	-	-	-	-0.184	-0.145	0.994^{**}
π_S	-	-	-	-	-	-	-0.553^{*}	0.125
π_N/π_S	-	-	-	-	-	-	-	0.136
generation time	-	-	-	-	-	-	-	-

Table 11.22: *Partial correlation coefficient between non-synonymous substitution rate (ω), mutation rate per site per unit of time (μ), and life-history traits (maximum longevity, adult weight and female maturity) were computed in primates. Asterisks indicate strength of support ($^{*}pp > 0.95$, $^{**}pp > 0.975$).*

11.6 Sufficient statistics

A sequence of length Z evolves by point substitutions, according to a random process defined by the substitution matrices $\boldsymbol{Q}^{(\mathrm{b,z})}$, over a phylogenetic tree. A realization of the random process along a branch b, and at a particular site z results in a detailed substitution history, which will be denoted by $\mathcal{H}^{(\mathrm{b,z})}$.

11.6.1 Path sufficient statistics

All sites owning to the same category of fitness profile share the same substitution rate matrix. Hence, $\mathcal{H}^{(\mathrm{b,z})}$ can be gathered across all sites owing to a specific category k, denoted $\mathcal{H}^{(\mathrm{b})}$. If we express the probability of the substitution mapping ($\mathcal{H}^{(\mathrm{b,k})}$) as a function of the codon substitution process for this category k, we get the following expression:

$$\mathbb{P}(\mathcal{H}^{(\mathrm{b,k})}|l^{(\mathrm{b})},\boldsymbol{Q}^{(\mathrm{b,k})}) \propto \left[\prod_{i=1}^{61}\left[\pi_i^{(\mathrm{b,k})}\right]^{n_i^{(\mathrm{b,k})}}\right]\cdot\left[\prod_{1\leq i,j\leq 61}\left[Q_{i,j}^{(\mathrm{b,k})}\right]^{m_{i,j}^{(\mathrm{b,k})}}\right]\cdot\left[\prod_{i=1}^{61}\mathrm{e}^{-\left|Q_{i,i}^{(\mathrm{b,k})}\right|a_i^{(\mathrm{b,k})}}\right], \tag{11.17}$$

where we define the sufficient statistics:

- $m_{i,j}^{(\mathrm{b,k})}$ is the total number of substitutions from codon i to codon j

- $n_i^{(\mathrm{b,k})}$ is the number of sites starting with codon i at the tip of the branch.
- $a_i^{(\mathrm{b,k})}$ is the total waiting time in codon i.

Once these sufficient statistics have been computed, the parameters of the substitution matrix $\boldsymbol{Q}^{(\mathrm{b,k})}$ can be resampled conditional on $\mathcal{H}^{(\mathrm{b,k})}$, using equation 11.17 each time the likelihood needs to be recomputed. This leads to relatively fast MCMC strategy.

11.6.2 Length sufficient statistics

$\mathcal{H}^{(\mathrm{b,z})}$ can also be gathered across all sites along a specific branch, giving $\mathcal{H}^{(\mathrm{b})}$. Then the probability of the substitution history given the branch lengths ($l^{(\mathrm{b})} = \mu^{(\mathrm{b})}\Delta T^{(\mathrm{b})}$), takes a very simple form:

$$\mathbb{P}(\mathcal{H}^{(\mathrm{b})}|L^{(\mathrm{b})}) \propto \left[L^{(\mathrm{b})}\right]^{u^{(\mathrm{b})}} \mathrm{e}^{-r^{(\mathrm{b})}L^{(\mathrm{b})}}, \tag{11.18}$$

where we define the sufficient statistics:

- $u^{(\mathrm{b})}$ is the total number of substitutions over branch b, summed over all sites.
- $r^{(\mathrm{b})}$ is the mean rate away from current codon state (averaged over the entire substitution history).

Thus, formally, the probability of the substitution mapping can be summarized by saying that the total number of substitutions along a given branch over all sites, $u^{(\mathrm{b})}$, is Poisson distributed, of mean $r^{(\mathrm{b})}L^{(\mathrm{b})}$.

11.6.3 Scatter sufficient statistics

From the independent contrast $\boldsymbol{C}^{(\mathrm{b})}$ of the Brownian process $\boldsymbol{B}^{(\mathrm{n})}$, we can define the 2×2 scatter sufficient statistic matrix, $\boldsymbol{A}$ as:

$$\boldsymbol{A} = \sum_{b=1}^{2P-2} \boldsymbol{C}^{(\mathrm{b})} \cdot \left[\boldsymbol{C}^{(\mathrm{b})}\right]^{\mathrm{T}} \tag{11.19}$$

By Bayes theorem, the posterior on $\boldsymbol{\Sigma}$, conditional on a particular realization of B (and thus of $\boldsymbol{C}$) is an invert Wishart distribution, of parameter $\kappa\boldsymbol{I} + \boldsymbol{A}$ and with $2P - 2 + 3$ degrees of freedom.

$$\boldsymbol{\Sigma} \sim \text{Wishart}^{-1}\left(\kappa\boldsymbol{I} + \boldsymbol{A}, 2P - 2 + 3\right) \tag{11.20}$$

This invert Wishart distribution can be obtained by sampling $2P - 2 + 3$ independent and identically distributed multivariate normal random variables $\boldsymbol{Z}^{(a)}$ defined by

$$\boldsymbol{Z}^{(a)} \sim \mathcal{N}\left(\boldsymbol{0}, \left[\kappa\boldsymbol{I} + \boldsymbol{A}\right]^{-1}\right). \tag{11.21}$$

And from these multivariate samples, $\boldsymbol{\Sigma}$ is Gibbs sampled as:

$$\boldsymbol{\Sigma} = \left(\sum_{k=1}^{2P-2+3} \boldsymbol{Z}^{(a)} \cdot \left[\boldsymbol{Z}^{(a)}\right]^{\mathrm{T}}\right)^{-1} \tag{11.22}$$

12

Substitution rate susceptibility - Supplementary Materials

Contents

Notations in the main manuscript and the supplementary file are identical, expect that $\Delta\Delta G$ is simplified to γ and $\Delta G_{\min}$ is simplified to α in the supplementary (hereby) for formula readability and developments. $\Delta\Delta G$ and $\Delta G_{\min}$ are kept here to refer to the empirical estimations.

12.1 ω response after a change in N_e

12.1.1 Genotype to phenotype map

Define n as the number of sites in the genotype sequence. Each site can be in one of $\mathrm{K} \geq 2$ states, where only 1 state is defined the stable state, and $\mathrm{K}-1$ states are unstable. For a given genotype sequence, define phenotype $0 \leq x \leq 1$ as the current proportion of sites in the unstable state. After a mutation, given that only one site can change at a time, the absolute change of x is either 0 or $\delta x = 1/n$. Define $\rho_x(\delta x)$ as the probability to get a change of phenotype equal to δx, if the current phenotype is x:

$$\begin{cases} \delta x & \text{with probability } \rho_x(\delta x) = 1 - x, \\ 0 & \text{with probability } \rho_x(0) = x\left[1 - \frac{1}{\mathrm{K}-1}\right], \\ -\delta x & \text{with probability } \rho_x(-\delta x) = \frac{x}{\mathrm{K}-1}. \end{cases} \tag{12.1}$$

12.1.2 Selection coefficient

$s(x, \delta x)$ is the selection coefficient of an effect δx if the current phenotype is x:

$$s(x,\delta x) = \frac{W(x+\delta x) - W(x)}{W(x)}, \tag{12.2}$$

$$\simeq \frac{1}{W(x)}\frac{\partial W(x)}{\partial x}\delta x, \tag{12.3}$$

$$\simeq \frac{\partial \ln(W(x))}{\partial x}\delta x, \tag{12.4}$$

$$\simeq \frac{\partial f(x)}{\partial x}\delta x, \tag{12.5}$$

where $W(x)$ is the Wrightian fitness of phenotype x, and $f = \ln(W)$ is the log-fitness (or Malthusian fitness). And the selective effect of the opposite change $(-\delta x)$ is the opposite selection coefficient:

$$s(x,-\delta x) \simeq -s(x,\delta x) \text{ from eq. 12.5}, \tag{12.6}$$

$$\iff S(x,-\delta x) \simeq -S(x,\delta x), \tag{12.7}$$

where $S(x^*, \delta x) = 4N_\mathrm{e}s(x^*, \delta x)$ is the scaled selection coefficient.

12.1.3 Probability of fixation

The probability of fixation of a mutation with effect δx, for a resident phenotype x is :

$$\mathbb{P}_\mathrm{fix}(x,\delta x) = \frac{1 - \mathrm{e}^{-2s(x,\delta x)}}{1 - \mathrm{e}^{-4N_\mathrm{e}s(x,\delta x)}}, \tag{12.8}$$

$$\simeq \frac{2s(x,\delta x)}{1 - \mathrm{e}^{-4N_\mathrm{e}s(x,\delta x)}}, \tag{12.9}$$

$$= \frac{2s(x,\delta x)}{1 - \mathrm{e}^{-S(x,\delta x)}}. \tag{12.10}$$

And in the case of neutral mutations, the probability of fixation is:

$$\mathbb{P}_{\text{fix}}(x, 0) = \frac{1}{2N_{\text{e}}}. \tag{12.11}$$

And the ratio of probability of fixation between selected and neutral mutations is:

$$\frac{\mathbb{P}_{\text{fix}}(x, \delta x)}{\mathbb{P}_{\text{fix}}(x, 0)} = \frac{2N_{\text{e}}2s(x, \delta x)}{1 - \mathrm{e}^{-S(x,\delta x)}} \text{ from eq. 12.10 and 12.11,} \tag{12.12}$$

$$= \frac{S(x, \delta x)}{1 - \mathrm{e}^{-S(x,\delta x)}}. \tag{12.13}$$

12.1.4 Equilibrium phenotype

At equilibrium phenotype x^*, the expected selection coefficient of mutation that reached fixation must be 0:

$$0 = \mathbb{E}_{\delta x}\left[s(x^*, \delta x)\mathbb{P}_{\text{fix}}(x^*, \delta x)\right], \tag{12.14}$$

$$\iff 0 = \frac{2s(x^*, \delta x)^2}{1 - \mathrm{e}^{-S(x^*,\delta x)}}\rho_{x^*}(\delta x) + s(x^*, 0)\frac{\rho_{x^*}(0)}{2N_{\text{e}}} + \frac{2s(x^*, -\delta x)^2}{1 - \mathrm{e}^{-S(x^*,-\delta x)}}\rho_{x^*}(-\delta x) \text{ from eq. 12.10 and 12.11,} \tag{12.15}$$

$$\implies \frac{2s(x^*, \delta x)^2}{1 - \mathrm{e}^{-S(x^*,\delta x)}}\rho_{x^*}(\delta x) \simeq \frac{-2s(x^*, \delta x)^2}{1 - \mathrm{e}^{S(x^*,\delta x)}}\rho_{x^*}(-\delta x) \text{ from eq. 12.7,} \tag{12.16}$$

$$\iff \frac{\rho_{x^*}(\delta x)}{\rho_{x^*}(-\delta x)} \simeq \mathrm{e}^{-S(x^*,\delta x)}\frac{\mathrm{e}^{-S(x^*,\delta x)} - 1}{\mathrm{e}^{-S(x^*,\delta x)}\left(1 - \mathrm{e}^{S(x^*,\delta x)}\right)}, \tag{12.17}$$

$$\iff \ln\left(\frac{1 - x^*}{x^*}\right) + \ln(\mathrm{K} - 1) \simeq -S(x^*, \delta x) \text{ from eq. 12.1,} \tag{12.18}$$

$$\iff \lambda_K(x^*) \simeq -S(x^*, \delta x), \tag{12.19}$$

where $\lambda_K(x^*) = \ln\left(\frac{1-x^*}{x^*}\right) + \ln(\mathrm{K} - 1)$.

12.1.5 Relative substitution rate (ω) at equilibrium

The substitution rate of all selected relative to the substitution rate of neutral mutations is denoted ω, which can also be interpreted as the mean fixation probability of mutations scaled by the fixation probability of neutral mutations $p = 1/2N_{\text{e}}$.

$$\omega = \mathbb{E}_{\delta x}\left[\frac{\mathbb{P}_{\text{fix}}(x, \delta x)}{\mathbb{P}_{\text{fix}}(x, 0)}\right], \tag{12.20}$$

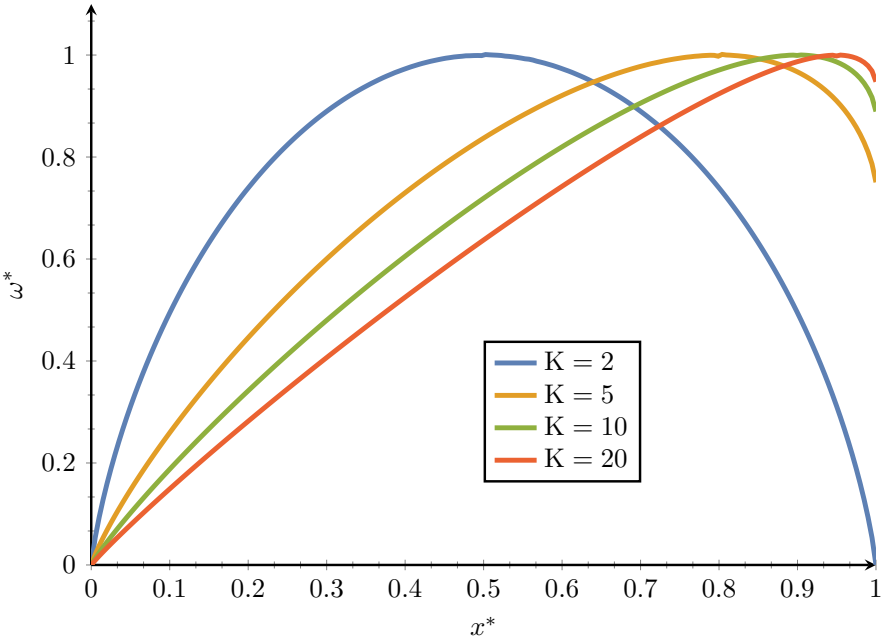
$$= (1 - x)\frac{S(x, \delta x)}{1 - \mathrm{e}^{-S(x,\delta x)}} + x\left(\frac{\mathrm{K} - 2}{\mathrm{K} - 1}\right) + \frac{x}{\mathrm{K} - 1}\frac{S(x, -\delta x)}{1 - \mathrm{e}^{-S(x,-\delta x)}} \text{ from eq. 12.1, 12.10 and 12.11,} \tag{12.21}$$

$$= (1 - x)\frac{S(x, \delta x)}{1 - \mathrm{e}^{-S(x,\delta x)}} - \frac{x}{\mathrm{K} - 1}\frac{S(x, \delta x)}{1 - \mathrm{e}^{S(x,\delta x)}} + x\left(\frac{\mathrm{K} - 2}{\mathrm{K} - 1}\right) \text{ from eq. 12.7.} \tag{12.22}$$

ω^* at equilibrium is then determined by the phenotype at equilibrium x^*:

$$\omega^* = (1 - x^*)\frac{S(x^*, \delta x)}{1 - \mathrm{e}^{-S(x^*,\delta x)}} - \frac{x^*}{\mathrm{K} - 1}\frac{S(x^*, \delta x)}{1 - \mathrm{e}^{S(x^*,\delta x)}} + x^*\left(\frac{\mathrm{K} - 2}{\mathrm{K} - 1}\right), \tag{12.23}$$

$$= x^*\left[\frac{2(x^* - 1)\lambda_K(x^*)}{\mathrm{K}(x^* - 1) + 1} + \frac{\mathrm{K} - 2}{\mathrm{K} - 1}\right] \text{ from eq. 12.18.} \tag{12.24}$$



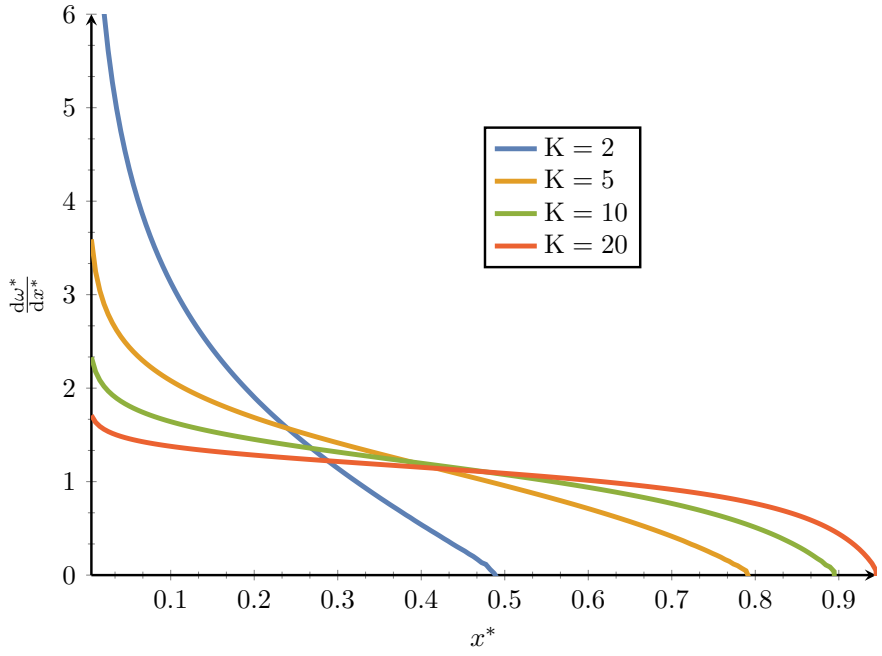
Moreover, given that the number of state if large enough $\mathrm{K} \gg 1$, the equilibrium ω can be approximated as:

$$\omega^* = x^* \left[\frac{2(x^* - 1)\lambda_K(x^*)}{\mathrm{K}(x^* - 1) + 1} + \frac{\mathrm{K} - 2}{\mathrm{K} - 1} \right], \tag{12.25}$$

$$\simeq x^* \tag{12.26}$$

And the derivative of ω^* w.r.t to x^* is:

$$\frac{\mathrm{d}\omega^*}{\mathrm{d}x^*} = 2 \left[\frac{\mathrm{K}(x^* - 1) + 1 + \left[\mathrm{K}(x^* - 1)^2 + 2x^* - 1\right] \lambda_K(x^*)}{(\mathrm{K}(x^* - 1) + 1)^2} \right] + \frac{\mathrm{K} - 2}{\mathrm{K} - 1}. \tag{12.27}$$



Moreover, given that the number of state if large enough $\mathrm{K} \gg 1$, the response in equilibrium ω due to change in phenotype can be approximated as:

$$\frac{\mathrm{d}\omega^*}{\mathrm{d}x^*} = 2\left[\frac{\mathrm{K}(x^*-1)+1+\left[\mathrm{K}(x^*-1)^2+2x^*-1\right]\lambda_K(x^*)}{(\mathrm{K}(x^*-1)+1)^2}\right] + \frac{\mathrm{K}-2}{\mathrm{K}-1}, \tag{12.28}$$

$$\simeq \frac{2\lambda_K(x^*)}{\mathrm{K}} + 1, \tag{12.29}$$

$$\simeq 1. \tag{12.30}$$

12.1.6 ω response after a change in N_e

Define the function $G(x, N_\mathrm{e})$ as:

$$G(x, N_\mathrm{e}) \equiv \lambda_K(x^*) + 4N_\mathrm{e}s(x, \delta x), \tag{12.31}$$

The equilibrium equation (eq. 12.18) states that $G(x^*, N_\mathrm{e}) = 0$, meaning that x^* is implicitly a function of N_e:

$$G(x^*(N_\mathrm{e}), N_\mathrm{e}) = 0, \tag{12.32}$$

$$\implies \frac{\partial G(x^*, N_\mathrm{e})}{\partial x^*}\frac{\mathrm{d}x^*}{\mathrm{d}N_\mathrm{e}} + \frac{\partial G(x^*, N_\mathrm{e})}{\partial N_\mathrm{e}} = 0, \tag{12.33}$$

$$\iff \left[\frac{\partial \lambda_K(x^*)}{\partial x^*} + 4N_\mathrm{e}\frac{\partial s(x^*, \delta x)}{\partial x^*}\right]\frac{\mathrm{d}x^*}{\mathrm{d}N_\mathrm{e}} + 4s(x^*, \delta x) = 0, \tag{12.34}$$

$$\iff \left[\frac{\partial \lambda_K(x^*)}{\partial x^*} + 4N_\mathrm{e}\frac{\partial^2 f(x^*)}{\partial {x^*}^2}\delta x\right]\frac{\mathrm{d}x^*}{\mathrm{d}N_\mathrm{e}} = -4\frac{\partial f(x^*)}{\partial x^*}\delta x \text{ from eq. 12.5}, \tag{12.35}$$

$$\iff 4\delta x\left[\frac{1}{4\delta x N_\mathrm{e}}\frac{\partial \lambda_K(x^*)}{\partial x^*} + \frac{\partial^2 f(x^*)}{\partial {x^*}^2}\right]N_\mathrm{e}\frac{\mathrm{d}x^*}{\mathrm{d}N_\mathrm{e}} = -4\delta x\frac{\partial f(x^*)}{\partial x^*}, \tag{12.36}$$

$$\iff \frac{\mathrm{d}x^*}{\mathrm{d}\ln(N_\mathrm{e})} = -\frac{\frac{\partial f(x^*)}{\partial x^*}}{\frac{1}{4\delta x N_\mathrm{e}}\frac{\partial \lambda_K(x^*)}{\partial x^*} + \frac{\partial^2 f(x^*)}{\partial {x^*}^2}}. \tag{12.37}$$

Giving the equation for the response of phenotype at equilibrium after a change of effective population size. Together, the response of substitution rate at equilibrium, after a change of effective population size can be obtained as:

$$\frac{\mathrm{d}\omega^*}{\mathrm{d}\ln(N_\mathrm{e})} = \frac{\mathrm{d}\omega^*}{\mathrm{d}x^*}\frac{\mathrm{d}x^*}{\mathrm{d}\ln(N_\mathrm{e})}, \tag{12.38}$$

$$= -\frac{\mathrm{d}\omega^*}{\mathrm{d}x^*}\frac{\frac{\partial f(x^*)}{\partial x^*}}{\frac{1}{4\delta x N_\mathrm{e}}\frac{\partial \lambda_K(x^*)}{\partial x^*} + \frac{\partial^2 f(x^*)}{\partial {x^*}^2}} \text{ from eq. 12.37}. \tag{12.39}$$

Moreover, with the approximation that $\left|4N_\mathrm{e}\frac{\partial s(x^*, \delta x)}{\partial x^*}\right| \gg \left|\frac{\partial \lambda_K(x^*)}{\partial x^*}\right|$, meaning that a change in phenotype causes a higher change in scaled selection coefficient than mutational bias, we have:

$$\frac{\mathrm{d}x^*}{\mathrm{d}\ln(N_\mathrm{e})} = -\frac{\frac{\partial f(x^*)}{\partial x^*}}{\frac{1}{4\delta x N_\mathrm{e}}\frac{\partial \lambda_K(x^*)}{\partial x^*} + \frac{\partial^2 f(x^*)}{\partial {x^*}^2}}, \tag{12.40}$$

$$\implies \frac{\mathrm{d}x^*}{\mathrm{d}\ln(N_\mathrm{e})} \simeq -\frac{\frac{\partial f(x^*)}{\partial x^*}}{\frac{\partial^2 f(x^*)}{\partial {x^*}^2}}. \tag{12.41}$$

Together, these approximations leads to the following response in equilibrium ω after change in N_e as:

$$\frac{d\omega^*}{d\ln(N_e)} \simeq -\frac{\frac{\partial f(x^*)}{\partial x^*}}{\frac{\partial^2 f(x^*)}{\partial x^{*2}}} \tag{12.42}$$

12.2 Models for the log-fitness function

12.2.1 Folded fraction

All phenotype-fitness functions considered below are log-concave, and as a result, $\frac{\partial f(x^*)}{\partial x^*}$ is a decreasing function of x; the less stable the protein already is, the stronger the purifying selection against additional destabilizing mutations. More precisely, fitness functions depends on the folded fraction of the protein of interest, which is given by the Fermi-Dirac distribution:

$$\mathbb{P}_F(x) = \frac{1}{1+e^{\beta(\alpha+\gamma n x)}}, \tag{12.43}$$

where x is the fraction of destabilizing mutations, each contributing to γ in free energy of folding (also denoted empirically $\Delta\Delta G$), and $\beta = 1/kT$. Thus, $\alpha < 0$ is the difference in free energy between folded and unfolded state when all sites are stable (also denoted empirically $\Delta G_{\min}$). As a result, $n\gamma$ is thus the expected change in ΔG when all sites are considered unstable. The misfolded fraction is $\mathbb{P}_U = 1 - \mathbb{P}_F$. In addition, $\mathbb{P}_F$ is typically close to 1 (or $\mathbb{P}_U \ll 1$), so that we can use a first-order approximation:

$$\mathbb{P}_F(x) = 1 - \mathbb{P}_U(x) \tag{12.44}$$

$$\simeq 1 - e^{\beta(\alpha+\gamma n x)} \tag{12.45}$$

or equivalently

$$\mathbb{P}_U(x) \simeq e^{\beta(\alpha+\gamma n x)} \tag{12.46}$$

12.2.2 Fitness equal to folded fraction

A first model is to assume that the fitness is equal to the folded fraction (Goldstein, 2013):

$$W(x) = \frac{1}{1+\mathrm{e}^{\beta(\alpha+n\gamma x)}}. \tag{12.47}$$

The derivative of fitness w.r.t to phenotype is:

$$\frac{\partial f(x)}{\partial x} = -\frac{\partial \ln\left(1+\mathrm{e}^{\beta(\alpha+n\gamma x)}\right)}{\partial x} \text{ from eq. 12.47,} \tag{12.48}$$

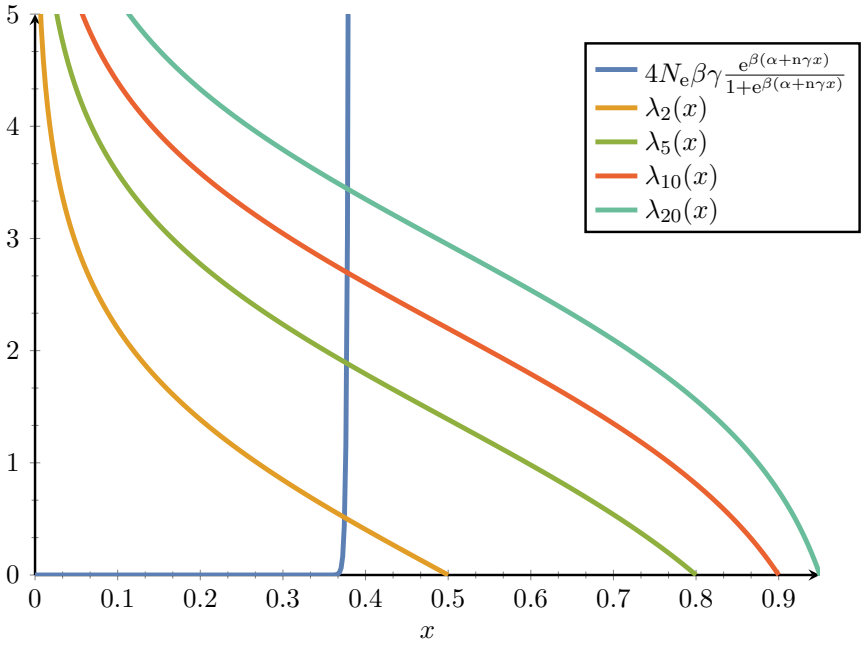
$$= -\beta n\gamma \frac{\mathrm{e}^{\beta(\alpha+n\gamma x)}}{1+\mathrm{e}^{\beta(\alpha+n\gamma x)}}, \tag{12.49}$$

$$\simeq -\beta n\gamma \mathrm{e}^{\beta(\alpha+n\gamma x)}. \tag{12.50}$$

The equilibrium phenotype (x^*) is :

$$\lambda_K(x^*) = 4N_e\beta\gamma\frac{e^{\beta(\alpha+n\gamma x^*)}}{1+e^{\beta(\alpha+n\gamma x^*)}} \text{ from eq. 12.19 and 12.49.} \tag{12.51}$$

Using $N_e = 10^4$, $\beta = 1.686$, $\alpha = \Delta G_{\min} = -118$, $n = 300$, $\gamma = \Delta\Delta G = 1$, we have the following :



Where in this example we can visually appreciate that the a change in phenotype causes a higher change in scaled selection coefficient than mutational bias (eq. 12.41). And the second derivative of fitness w.r.t to phenotype is:

$$\frac{\partial^2 f(x)}{\partial x^2} = -\beta n\gamma\frac{\partial}{\partial x}\left(\frac{e^{\beta(\alpha+n\gamma x)}}{1+e^{\beta(\alpha+n\gamma x)}}\right) \text{ from eq. 12.49,} \tag{12.52}$$

$$= -\beta n\gamma\beta n\gamma\frac{e^{\beta(\alpha+n\gamma x)}}{\left(1+e^{\beta(\alpha+n\gamma x)}\right)^2}, \tag{12.53}$$

$$= \frac{\beta n\gamma}{1+e^{\beta(\alpha+n\gamma x)}}\frac{\partial f(x)}{\partial x} \text{ from eq. 12.49,} \tag{12.54}$$

$$\simeq \beta n\gamma\frac{\partial f(x)}{\partial x} \tag{12.55}$$

Finally, ω response after a change in N_e is simply:

$$\frac{d\omega^*}{d\ln(N_e)} \simeq -\frac{1}{\beta n\gamma} \text{from eq. 12.54 and 12.30,} \tag{12.56}$$

which is independent of x^*, meaning ω is linearly decreasing with N_e in log space. This model, however, does not express the fact that selection is typically stronger for proteins characterized by higher levels of expression.

12.2.3 Selective cost proportional to amount of misfolded protein

A slight variation is to assume that the selective cost itself is proportional to the total amount of misfolded protein (Drummond *et al.*, 2005; Wilke and Drummond, 2006; Drummond and Wilke, 2008; Serohijos *et al.*, 2012). For a given protein with expression level y:

$$f(x) = -Ay\mathbb{P}_{\mathrm{U}}(x), \tag{12.57}$$

where A is the cost per misfolded macromolecule. Then,

$$\frac{\partial f(x)}{\partial x} \simeq -Ay\beta\gamma n e^{\beta(\alpha+\gamma n x)}. \tag{12.58}$$

Under this model, the phenotype at equilibrium is given by:

$$\lambda_K(x^*) = 4N_{\mathrm{e}}yA\beta\gamma n e^{\beta(\alpha+\gamma n x^*)} \text{ from eq. 12.19 and 12.49.} \tag{12.59}$$

And the response of ω after a change in N_{e} is the same as before:

$$\frac{\mathrm{d}\omega^*}{\mathrm{d}\ln(N_{\mathrm{e}})} \simeq -\frac{1}{\beta n\gamma}. \tag{12.60}$$

Since N_{e} and y are confounded factors, meaning they only appear in the equation as a product between the two, implicit derivation leads to the same result whenever the derivation is w.r.t N_{e} or y, leading to same compact equation:

$$\frac{\mathrm{d}\omega^*}{\mathrm{d}\ln(y)} = \frac{\mathrm{d}\omega^*}{\mathrm{d}\ln(N_{\mathrm{e}})} \simeq -\frac{1}{\beta n\gamma}. \tag{12.61}$$

12.2.4 Translational errors

Another variant account for translational errors. Translational errors occur at a rate ρ per residue. These errors contribute additional destabilizing mutations, each with effect size $\delta x = 1/n$. The total number of translational errors per macromolecule is approximately Poisson distributed:

$$\pi_k = e^{-\rho n}\frac{(\rho n)^k}{k!} \tag{12.62}$$

and the total selective cost is now an average over all possible values of k:

$$f(x) = -Ay\sum_k \pi_k e^{\beta(\alpha+\gamma n x+\gamma k)} \tag{12.63}$$

$$= -Aye^{\beta(\alpha+\gamma n x)}\sum_k e^{-\rho n}\frac{(\rho n)^k}{k!}e^{\beta\gamma k} \tag{12.64}$$

$$= -Aye^{\beta(\alpha+\gamma n x)+\rho n(e^{\beta\gamma}-1)} \tag{12.65}$$

$$\simeq -Aye^{\beta(\alpha+\gamma n x)+\rho\beta\gamma n} \tag{12.66}$$

$$= -Aye^{\beta(\alpha+\gamma n(x+\rho))} \tag{12.67}$$

In words, the fitness function is the same the previous model, except that the trait x (fraction of destabilizing mutations) is shifted by ρ, the mean fraction of additional mutations contributed by translation errors. This additional factor is independent of x, and as a result, the scaled selection strength is essentially the same, up to a proportionality constant (contributed by the shift):

$$4N_{\mathrm{e}}\frac{\partial f(x)}{\partial x} \propto -4N_{\mathrm{e}}ye^{\beta(\alpha+\gamma n(x+\rho))} \tag{12.68}$$

$$\propto -4N_{\mathrm{e}}ye^{\beta(\alpha+\gamma nx)} \tag{12.69}$$

Moreover, ω response after a change in N_{e} is again the same as before:

$$\frac{\mathrm{d}\omega^*}{\mathrm{d}\ln(N_{\mathrm{e}})} \simeq -\frac{1}{\beta n\gamma}. \tag{12.70}$$

12.2.5 Cost-benefit argument

The cost-benefit argument (Beaulieu *et al.*, 2018) is based on two assumptions

1. the expression level is regulated so that the total number of *functional* macromolecules is maintained at a target level y;
2. the log-fitness is proportional to the ratio of the *total* cost of expression over the benefit contributed by the protein.

Specifically, the protein is assumed to be regulated so as to reach a level of expression of functional proteins of y, and contributes a total benefit B (which depends on its specific function). Given that only a fraction $\mathbb{P}_{\mathrm{F}}(x) = 1 - \mathbb{P}_{\mathrm{U}}(x)$ of the total amount of protein expressed by the cell is functional, the total cost of expression C is then equal to:

$$C(x) = \frac{y}{\mathbb{P}_{\mathrm{F}}(x)} \tag{12.71}$$

$$\simeq y(1 + \mathbb{P}_{\mathrm{U}}(x)) \tag{12.72}$$

Then, the log-fitness is given by:

$$f(x) = -A\frac{y}{B}\left(1 + e^{\beta(\alpha+\gamma nx)}\right) \tag{12.73}$$

$$= -by(1 + e^{\beta(\alpha+\gamma nx)}), \tag{12.74}$$

where $b = A/B$. Compared to models 2 (section 12.2.3) and 3 (section 12.2.4), the log-fitness now has an additional term that depends on the target expression level y, but not on trait x. The scaled strength of selection on mutations affecting x has thus the same functional form as for the two previous models:

$$4N_{\mathrm{e}}\frac{\partial f(x)}{\partial x} \propto -4N_{\mathrm{e}}ye^{\beta(\alpha+\gamma nx)} \tag{12.75}$$

Alternative cost-expression models could also be used, allowing for a non-linear cost function for expression or for some susceptibility of the realized equilibrium expression

level, as a function of the number of mutations. Under these models, the strength of selection is still expected to be an increasing function of y, although not linear:

$$4N_{\mathrm{e}}\frac{\partial f(x)}{\partial x} \propto -4N_{\mathrm{e}}g(y)e^{\beta(\alpha+\gamma nx)}, \tag{12.76}$$

where g is some function of y. Moreover, ω response after a change in N_{e} is again the same as before:

$$\frac{\mathrm{d}\omega^*}{\mathrm{d}\ln(N_{\mathrm{e}})} \simeq -\frac{1}{\beta n\gamma}. \tag{12.77}$$

12.3 Model of protein-protein interactions

The proteome is assumed to be composed of m protein species, all with same abundance C. Each macromolecule may either be in free form or engaged in a non-specific interaction. Only pairwise interactions are considered, and higher-order interactions are ignored. The equilibrium is characterized by:

$$[ij] = \frac{[i][j]}{C_0}\, e^{\beta E_{ij}}, \tag{12.78}$$

where $[i]$ and $[j]$ are the concentrations of protein species i and j, and $[ij]$ is the concentration of their (non-specific) dimer. Here, E_{ij} is the interaction free energy, which can itself be decomposed as a sum of three terms:

$$E_{ij} = \alpha + E_i + E_j \tag{12.79}$$

$$= \alpha + \gamma n(x_i + x_j), \tag{12.80}$$

where we assume that each protein has $n = 100$ residues at its surface, x_i stands for the fraction of hydrophobic residues at the surface of protein i, and each hydrophobic residue makes an additive contribution of $\Delta\Delta G$ to the total.

By conservation of the total number of molecules:

$$C = [i] + \sum_{j\neq i}[ij] \tag{12.81}$$

$$= [i] + \sum_{j\neq i}\frac{[i][j]}{C_0}\, e^{\beta E_{ij}} \tag{12.82}$$

and we note:

$$\epsilon_i = \sum_j [ij] \tag{12.83}$$

the fraction of protein i sequestered in non-specific interactions. We assume that the log fitness is proportional to the total amount of protein sequestered in non-specific interactions:

$$f(x) = -b\sum_i \epsilon_i, \tag{12.84}$$

where $b > 0$ is a parameter determining the overall stringency of selection against non-specific interactions.

12.3.1 Mean field, weak-interaction limit

To make the model tractable and compact, we assume that non-specific interactions are weak, i.e. $\epsilon_i \ll 1$ for all i. We then make a first-order approximation in the ϵ_i's. In addition, we use a mean-field approximation, such that, when considering a specific protein species i, we assume that all other proteins have the same fraction $\bar{x}$ of hydrophobic residues at their surface. The value of $\bar{x}$ could in principle be found using a self-consistent argument, essentially by (1) explicitly calculating the net substitution flux for protein i with fraction x_i, under mean field $\bar{x}$, and (2) expressing the constraint that this substitution process for protein i is stationary at $x_i = \bar{x}$. This derivation is not conducted here, as it is not needed. Using these approximations, we can re-express the conservation of total mass as:

$$C = [i] + (m-1)[i]\frac{C}{C_0}\, e^{\beta(\alpha+\gamma n(\bar{x}+x_i))} \tag{12.85}$$

Here, we have used the fact that $[j] = C(1-\epsilon_j)$ can be approximated as $[j] \simeq C$ since it is involved in a term already of the order of ϵ_i. As a result, all $m-1$ terms of the sum over $j \neq i$ are identical. Next, solving for $[i]$ gives:

$$[i] = \frac{C}{1 + (m-1)\frac{C}{C_0}\, e^{\beta(\alpha+\gamma n(\bar{x}+x_i))}} \tag{12.86}$$

$$\simeq C\left(1 - m\frac{C}{C_0}\, e^{\beta(\alpha+\gamma n(\bar{x}+x_i))}\right) \tag{12.87}$$

$$= C(1-\epsilon_i) \tag{12.88}$$

and thus ϵ_i can be identified with:

$$\epsilon_i = m\frac{C}{C_0}\, e^{\beta(\alpha+\gamma n(\bar{x}+x_i))} \tag{12.89}$$

Now, assume that the system is at equilibrium (thus $x_i = \bar{x}$). The strength of selection acting on mutations occurring at the surface of protein i, of effect size $\delta x = \pm 1/n$, is given by $s = \kappa \delta x$ where:

$$\kappa_i = b\frac{\mathrm{d}\epsilon_i}{\mathrm{d}x_i} \tag{12.90}$$

$$= b\beta\gamma n m\frac{C}{C_0}\, e^{\beta(\alpha+\gamma n(\bar{x}+x_i))} \tag{12.91}$$

and thus:

$$\ln(\kappa_i) = \ln\left(b\beta\gamma n m\frac{C}{C_0}\right) + \beta(\alpha + \gamma n\bar{x}) + \beta\gamma n x_i, \tag{12.92}$$

where only the last term depends on x_i. Finally, applying the main result of this work to the present case allows us to express the response of ω as a function of N_e as:

$$\chi = \frac{\mathrm{d}\omega}{\mathrm{d}\ln(N_\mathrm{e})} \tag{12.93}$$

$$= 2(\lambda-1)\,\frac{\mathrm{d}\ln\kappa_i}{\mathrm{d}x_i} \tag{12.94}$$

$$= 2(\lambda-1)\,\frac{1}{\beta\gamma n} \tag{12.95}$$

Note that, here, we have used $K = 2$ (hydrophobic and polar residues are roughly equally likely to occur by mutation), and assumed $x^* \ll 1$. A more accurate formula could be used without this latter assumption. In any case, χ is now dependent on x^*, through λ.

12.3.2 Empirical calibration

Based on empirical estimates found in Zhang *et al.* (2008). The mean fraction of hydrophobic residues at the surface of proteins is 0.22 ± 0.06. With $n = 100$ residues, this makes 22 ± 6. The mean value for E_{ij} is $7kT$, with a standard deviation of $\sigma = 1.8kT$. Assuming that this standard deviation of $\pm 1.8kT$ is contributed by ± 6 mutations gives $\Delta\Delta G = 1.8/6 = 0.3$ kT or 0.18 kcal per mole. Also, with $x = 0.22$, $\lambda \simeq 4$, and thus $\chi = 6/30 = 0.2$, thus a much stronger response than under the model based on conformational stability.

12.4 Empirical estimation

Type	**Specie**	$\hat{\chi}$	r^2
Plant	Oryza sativa	-0.008	0.047
Plant	Arabidopsis thaliana	-0.012	0.128
Archaea	Sulfolobus solfataricus	-0.037	0.097
Archaea	Thermococcus kodakarensis	-0.026	0.058
Fungi	Saccharomyces cerevisiae	-0.029	0.211
Fungi	Aspergillus nidulans	-0.034	0.124
Bacteria	Escherichia coli	-0.021	0.151
Bacteria	Bacillus subtilis	-0.046	0.151
Animal	Caenorhabditis elegans	-0.026	0.039
Animal	Drosophila melanogaster	-0.005	0.021
Animal	Mus musculus	-0.008	0.085
Animal	Homo sapiens	-0.004	0.031

Table 12.1: *Substitution rate as a function of expression level compiled by Zhang and Yang (2015).*

In Brevet and Lartillot (2019), the covariance matrix with $\ln(N_e)$ and $\ln(\omega)$ as entries allows to approximate χ:

$$\widehat{\frac{\mathrm{d}\ln(\omega)}{\mathrm{d}\ln(N_e)}} = \frac{\mathrm{Cov}[\ln(\omega), \ln(N_e)]}{\mathrm{Var}[\ln(N_e)]}, \tag{12.96}$$

$$\Rightarrow \widehat{\frac{\mathrm{d}\omega}{\omega \mathrm{d}\ln(N_e)}} = \frac{\mathrm{Cov}[\ln(\omega), \ln(N_e)]}{\mathrm{Var}[\ln(N_e)]}, \tag{12.97}$$

$$\Rightarrow \widehat{\chi} \simeq \widehat{\omega} \frac{\mathrm{Cov}[\ln(\omega), \ln(N_e)]}{\mathrm{Var}[\ln(N_e)]}, \tag{12.98}$$

$$\Rightarrow \widehat{\chi} \simeq 0.2 \frac{-0.45}{4.45}, \tag{12.99}$$

$$\Rightarrow \widehat{\chi} \simeq -0.02 \tag{12.100}$$

12.5 Simulation using the 3D structure of protein

We simulated substitutions in the protein phosphatase (Z = 300 codon sites) as in Goldstein and Pollock (2017). From a DNA sequence $\mathbb{S}$ after t substitutions, we compute the free energy of the folded state $G_{\mathrm{F}}(\mathbb{S})$, using the 3-dimensional structure of the folded state and pair-wise contact energies between neighboring amino-acid residues:

$$G_{\mathrm{F}}(\mathbb{S}) = \sum_{\mathrm{z}=1}^{\mathrm{Z}} \sum_{r \in \mathcal{V}(\mathrm{z})} I(\mathbb{S}(\mathrm{z}), \mathbb{S}(r)), \tag{12.101}$$

where $I(a, b)$ is the pair-wise contact energies between amino acid a and b, using contact potentials estimated by Miyazawa and Jernigan (1985), and $\mathcal{V}(\mathrm{z})$ are the neighbor residues of site z (closer than 7Å) in the 3D structure.

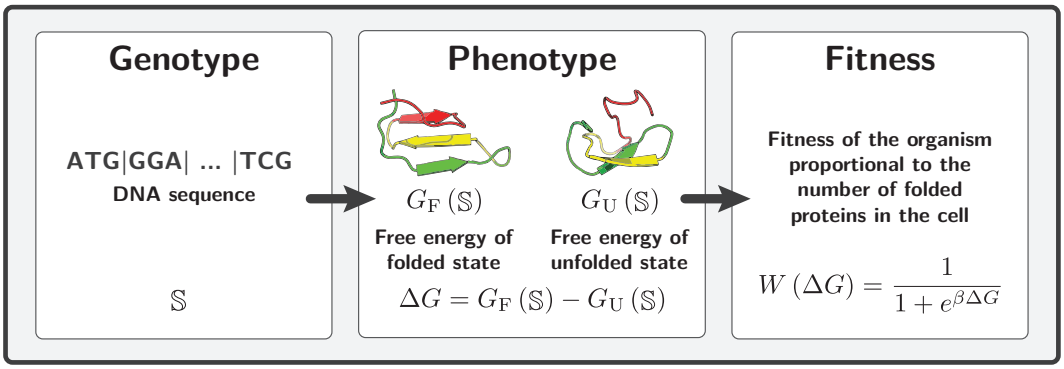
The free energy of unfolded states $G_{\mathrm{U}}(\mathbb{S})$ is approximated using 55 decoy 3D structures that supposedly represent a sample of possible unfolded states:

$$G_{\mathrm{U}}(\mathbb{S}) = \langle G(\mathbb{S}) \rangle - kT \ln(1.0\mathrm{E}^{160}) - \frac{2\left[\left\langle G(\mathbb{S})^2 \right\rangle - \langle G(\mathbb{S}) \rangle^2\right]}{kT} \tag{12.102}$$

where the average $\langle . \rangle$ runs other the 55 decoy 3D structures, and k is the Boltzmann constant and T the temperature in Kelvin.

From the energy of folded and unfolded states, we can compute the difference in free energy between the states:

$$\Delta G(\mathbb{S}) = G_{\mathrm{F}}(\mathbb{S}) - G_{\mathrm{U}}(\mathbb{S}) \tag{12.103}$$



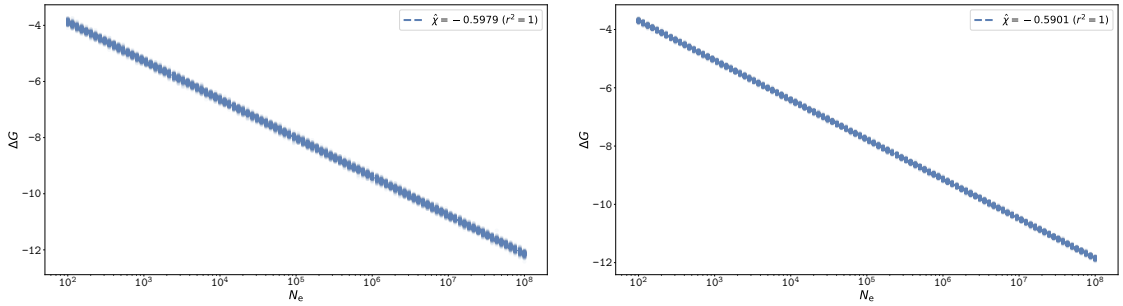


Figure 12.1: *ΔG response to changes in N_e. Left Panel: Model of folding free energy computed using 3D structural conformations and pairwise contact potential energies between neighbouring amino-acid residues. Right Panel: Additive phenotype model, where for each non-optimal amino acid, γ is scaled by the Grantham distance to the optimal amino acid. Scaling experiment simulating sequence evolution and recording the average ΔG (y-axis) observed at equilibrium as a function of N_e (x-axis). Along the x-axis, 200 replicate simulations are performed for each different N_e, the average (solid lines) and 90% confidence interval (shaded area) of ω are shown. ΔG is linearly dependent on log-N_e, with a slope equal to $1/\beta = 0.593$.*

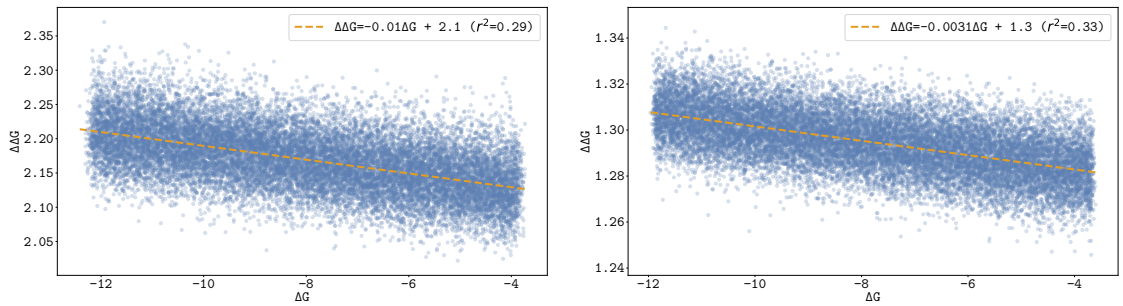


Figure 12.2: *$\Delta\Delta G$ correlation to ΔG. Left Panel: Model of folding free energy computed using 3D structural conformations and pairwise contact potential energies between neighbouring amino-acid residues. Right Panel: Additive phenotype model, where for each non-optimal amino acid, γ is scaled by the Grantham distance to the optimal amino acid. Simulations are performed for N_e varying from 10^2 to 10^8, where each dot is an independent simulation at equilibrium. Along each simulation, the average $\Delta\Delta G$ of all proposed mutations is recorded (y-axis), and represented as a function of the average ΔG (x-axis). $\Delta\Delta G$ is negatively correlated to Δ, which is expected since protein under higher N_e are more stable (lower ΔG, see above), and mutations are more destabilizing on average. To be more precise, the negative correlation between $\Delta\Delta G$ and ΔG is observed empirically with a linear fit of $\Delta\Delta G = -0.13\Delta G + 0.23$ (Serohijos et al., 2012). This correlation is a necessary condition for observing a response of ω to changes in N_e (Goldstein, 2013) and protein expression level (Serohijos et al., 2012).*

12.6 Simulated ω response to changes in N_e

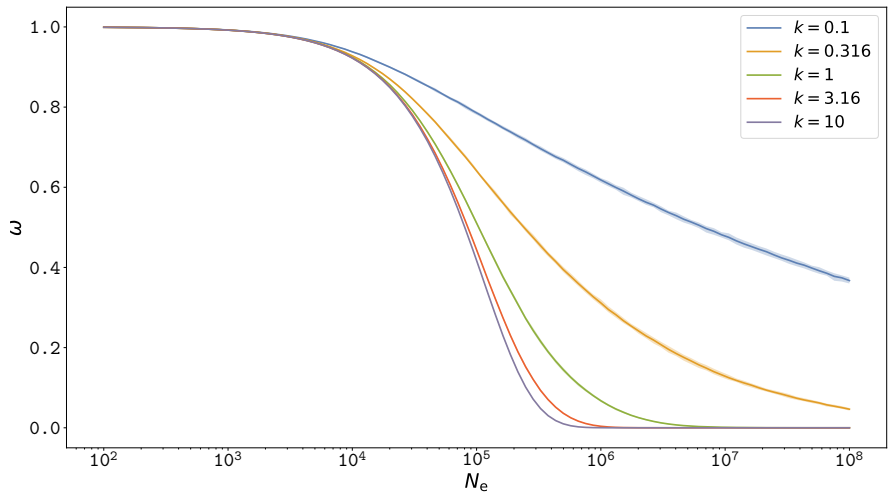


Figure 12.3: *ω at equilibrium as a function of N_e (log scale), under a model of gamma distributed selection coefficient. For each population size, 200 simulations were performed and the average (solid line) and 90% confidence interval (shaded area) are shown. In the model of gamma distributed fitness effect, ω at equilibrium is strongly dependent on log-N_e where the slope correlation is proportional to the inverse of the shape parameter of the gamma distribution (Welch et al., 2008).*

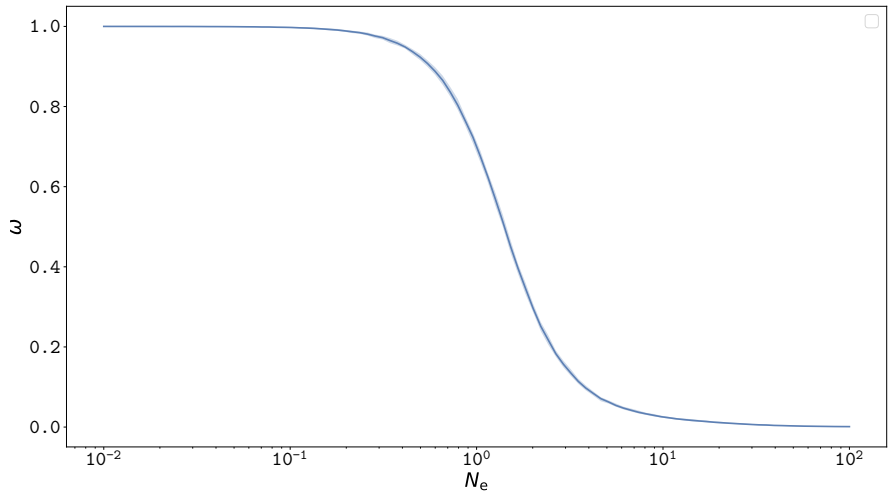


Figure 12.4: *ω at equilibrium as a function of N_e (relative), under a model of amino-acid fitness profiles. For each population size, 200 simulations were performed and the average (solid line) and 90% confidence interval (shaded area) are shown. In the model of site-wise amino-acid fitness profiles taken from (Bloom, 2017), ω at equilibrium is strongly dependent on log-N_e.*

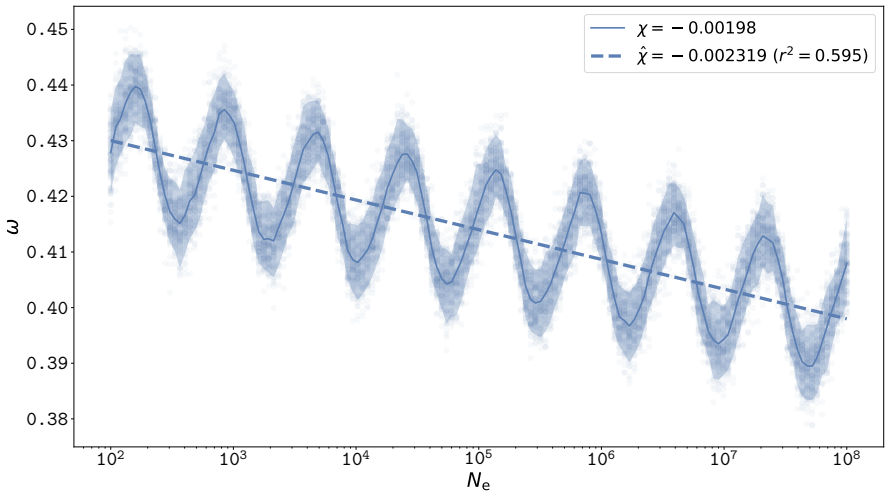


Figure 12.5: *ω at equilibrium as a function of N_e (log scale), for a model of additive free energy of folding. For each population size, 200 simulations were performed and the average (solid line) and 90% confidence interval (shaded area) are shown. The fixed parameters are $\alpha = -118$, $\gamma = 1$, $n = 300$, $\beta = 1.686$. The simulations of our additive free energy model match the theoretical prediction that the slope of the linear relation (dashed line) is equal to $\beta n \gamma)^{-1} = 0.00198 \simeq= 0.00199$. The non-monotony is suspected to be due to the discrete number of sites and states, such that the changes in ΔG after a mutation is either -1, 0 or 1. Such non-monotony is not observed with the Grantham model, in which the ω is lower and the slope of the response is lower, closer to the empirical 3D model.*

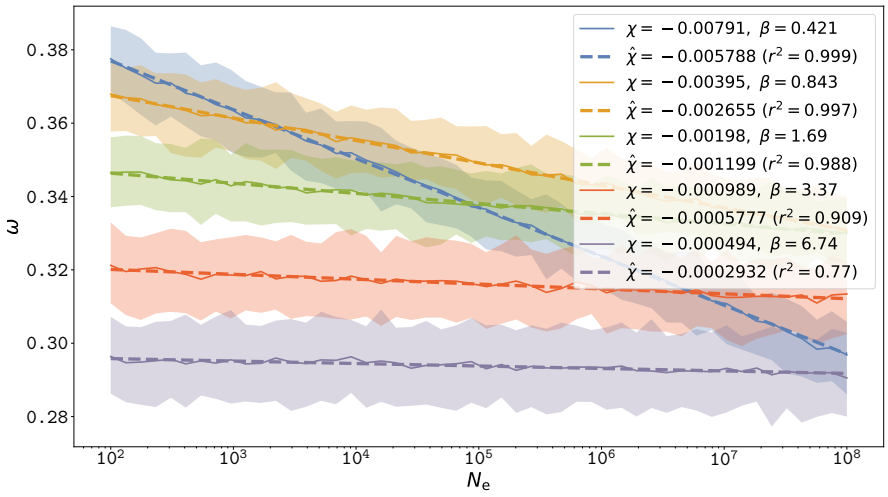


Figure 12.6: *ω at equilibrium as a function of N_e (log scale), for various parameter β. For each population size, 200 simulations were performed and the average (solid line) and 90% confidence interval (shaded area) are shown. The fixed parameters are $\alpha = -118$, $\gamma = 1$, $n = 300$, and for each non-optimal amino acid, γ is scaled by the Grantham distance to the optimal amino acid. β are given in the legend. Increasing β decreases the slope of the ω-N_e relationship, as predicted in our theoretical model.*

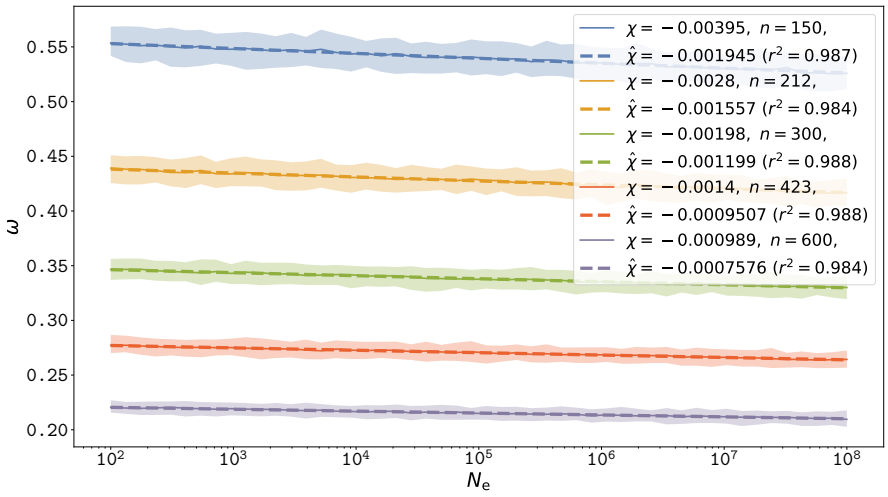


Figure 12.7: *ω at equilibrium as a function of N_e (log scale), for various sequence size. For each population size, 200 simulations were performed and the average (solid line) and 90% confidence interval (shaded area) are shown. The fixed parameters are $\alpha = -118$, $\gamma = 1$, $\beta = 1.686$, and for each non-optimal amino acid, γ is scaled by the Grantham distance to the optimal amino acid. n are given in the legend. Increasing n decreases the slope of the ω-N_e relationship, as predicted in our theoretical model.*

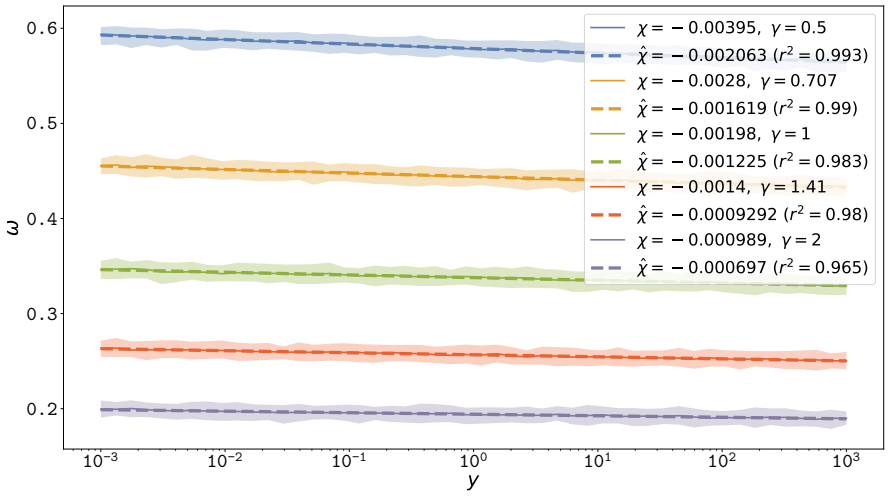


Figure 12.8: *ω at equilibrium as a function of the expression level y (log scale), for various value of γ. For each population size, 200 simulations were performed and the average (solid line) and 90% confidence interval (shaded area) are shown. The fixed parameters are $\alpha = -118$, $\beta = 1.686$, $n = 300$, and for each non-optimal amino acid, γ is scaled by the Grantham distance to the optimal amino acid. γ are given in the legend. Increasing γ increases the slope of the ω-y relationship, as predicted in our theoretical model.*

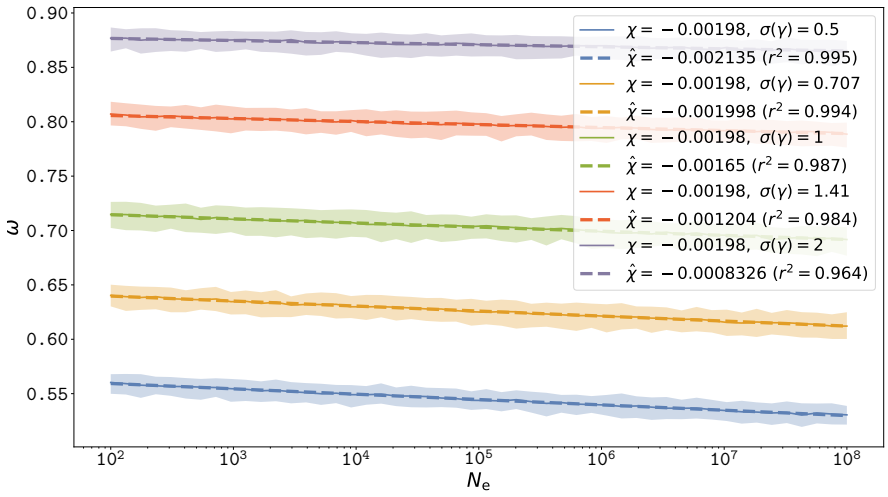


Figure 12.9: *ω at equilibrium as a function of N_e (log scale), for various between site variance. For each population size, 200 simulations were performed and the average (solid line) and 90% confidence interval (shaded area) are shown. The parameters are $\alpha = -118$, $n = 300$, $\beta = 1.686$ and each site has its own gamma distributed γ with mean 1 and standard deviation given in the legend. γ is scaled by the Grantham distance to the optimal amino acid. Increasing the variance of γ increases ω, by shifting the equilibrium to higher x^* since more unstable sites with low γ are fixed before reaching sensible deleterious selection coefficient against unstable mutations. Once many sites are unstable, the ω is higher since non-synonymous mutations between unstable states are effectively neutral. However the slope of the ω-N_e relationship is not sensibly changed.*

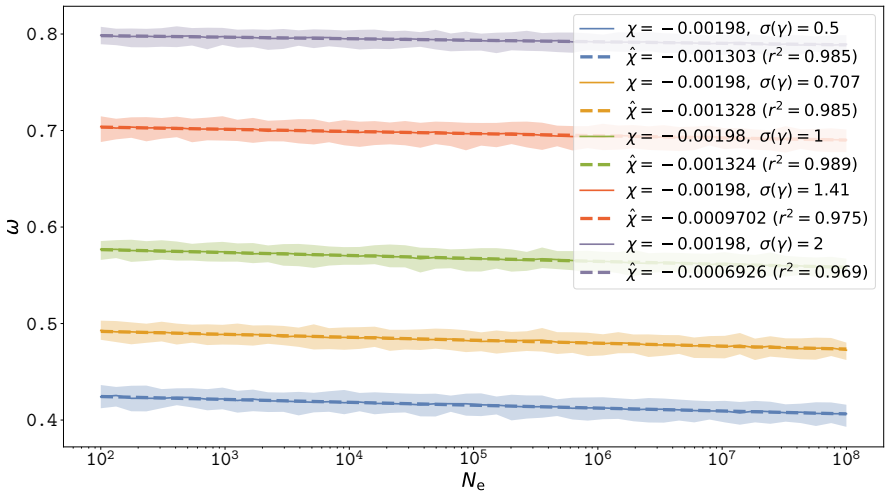


Figure 12.10: *ω at equilibrium as a function of N_e (log scale), for various between site variance under a model considering Grantham distances. For each population size, 200 simulations were performed and the average (solid line) and 90% confidence interval (shaded area) are shown. The parameters are $\alpha = -118$, $n = 300$, $\beta = 1.686$ and each site has its own gamma distributed γ with mean 1 and standard deviation given in the legend. Increasing the variance of γ increases ω, by shifting the equilibrium to higher x^* since more unstable sites with low γ are fixed before reaching sensible deleterious selection coefficient against unstable mutations. Once many sites are unstable, the ω is higher since non-synonymous mutations between unstable states are effectively neutral. However the slope of the ω-N_e relationship is not sensibly changed.*

12.7 Simulated relaxation time of ω

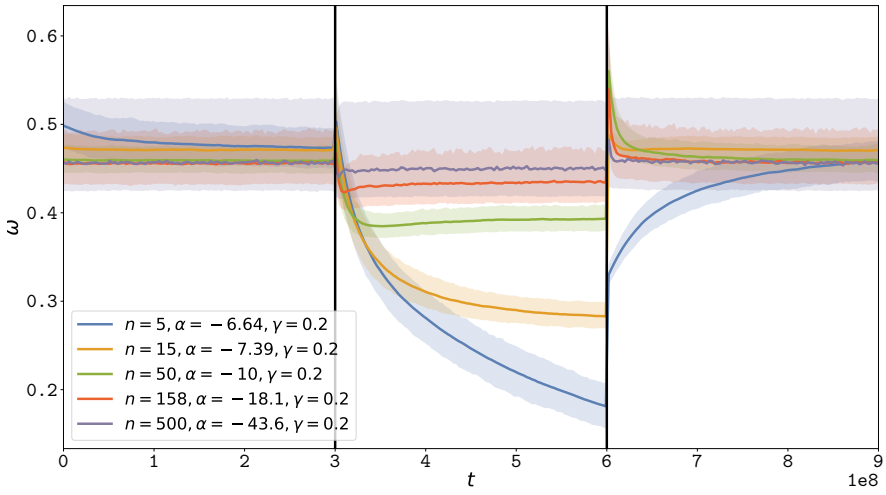


Figure 12.11: *ω Relaxation after a brutal change in N_e, for various n while correcting for α. The left and right panel correspond to low N_e ($1e^5$) and the middle panel corresponds to high N_e ($2e^6$). Solid line corresponds to the average over replicates (r) and the shaded area correspond to the 90% interval among replicates. The mutation rate (μ) is 1e−8 per year per site, and the total time of the computation is 900 million years. $\beta = 1.686$, $\gamma = 0.2$ for all simulations. The number of sites is changed from $n = 15$ to $n = 158$, and the number of replicates is changed accordingly such that the total number of sites ($n*r$) is kept constant. Moreover, α is changed according to n and γ such that the equilibrium value x^* is kept constant, by solving numerically equation 12.18. Increasing n implies a higher rate of relaxation.*

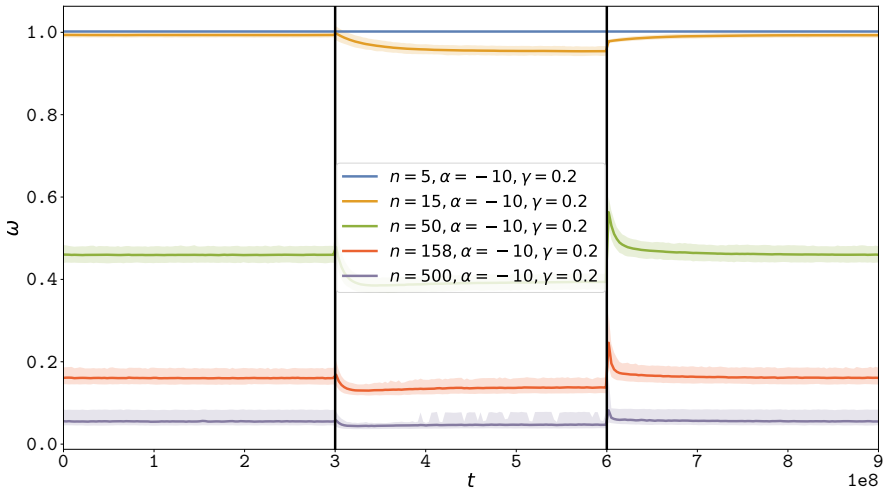


Figure 12.12: *ω Relaxation after a brutal change in N_e, for various n. The left and right panel correspond to low N_e ($1e^5$) and the middle panel corresponds to high N_e ($2e^6$). Solid line corresponds to the average over replicates (r) and the shaded area correspond to the 90% interval among replicates. The mutation rate (μ) is 1e−8 per year per site, and the total time of the computation is 900 million years. $\beta = 1.686$, $\gamma = 0.2$ and $\alpha = -10$ for all simulations. The number of sites is changed from $n = 15$ to $n = 158$, and the number of replicates is changed accordingly such that the total number of sites ($n * r$) is kept constant. Increasing n implies a higher ω at equilibrium, a lower response of the ω to changes in N_e and a higher rate of relaxation.*

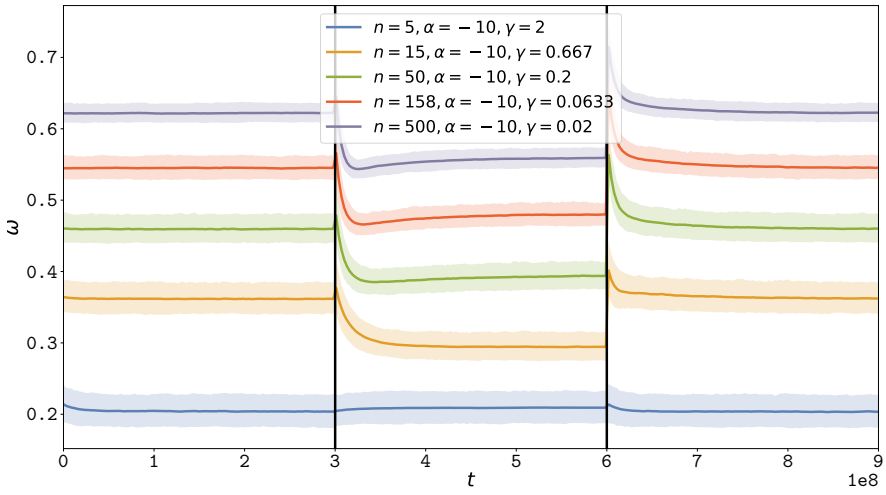


Figure 12.13: *ω Relaxation after a brutal change in N_e, for various n while correcting for γ. The left and right panel correspond to low N_e ($1e^5$) and the middle panel corresponds to high N_e ($2e^6$). Solid line corresponds to the average over replicates (r) and the shaded area correspond to the 90% interval among replicates. The mutation rate (μ) is 1e−8 per year per site, and the total time of the computation is 900 million years. $\beta = 1.686$, $\alpha = -10$ for all simulations. The number of sites is changed from $n = 15$ to $n = 158$, and the number of replicates is changed accordingly such that the total number of sites ($n * r$) is kept constant. Moreover, γ is changed according to n such that the product γn is kept constant, thus the response of the ω to changes in N_e is kept constant. Increasing n implies a higher ω at equilibrium, and a higher rate of relaxation.*

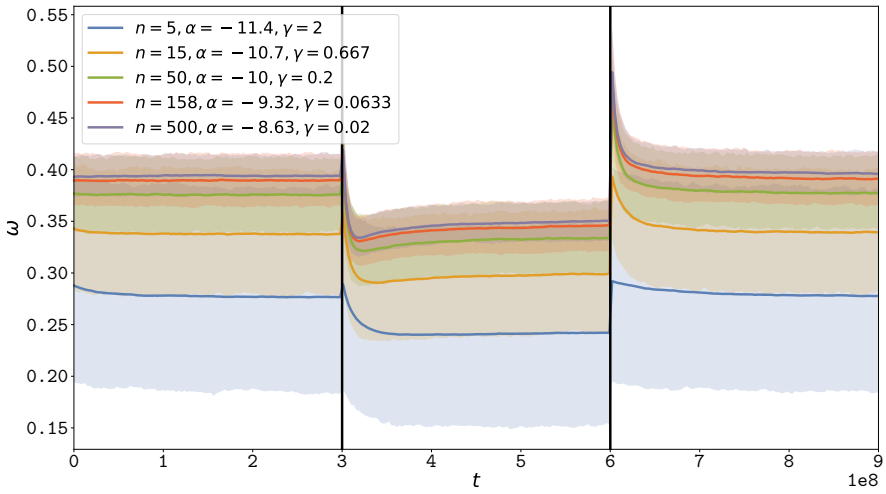


Figure 12.14: *ω Relaxation after a brutal change in N_e, under a Grantham model. The left and right panel correspond to low N_e ($1e^5$) and the middle panel corresponds to high N_e ($2e^6$). Solid line corresponds to the average over replicates (r) and the shaded area correspond to the 90% interval among replicates. The mutation rate (μ) is $1e{-}8$ per year per site, and the total time of the computation is 900 million years. $\beta = 1.686$, $\gamma = -10$ for all simulations. The number of sites is changed from $n = 15$ to $n = 158$, and the number of replicates is changed accordingly such that the total number of sites ($n * r$) is kept constant. Moreover, γ is changed according to n such that the product γn is kept constant, thus the response of the ω to changes in N_e is kept constant. Finally, α is changed according to n and γ such that the equilibrium value x^* is kept constant, by solving numerically equation 12.18. Increasing n implies a higher rate of relaxation.*

12.8 Distribution of fitness effects

DNA mutations changing a genotype can result in a change of phenotype, and ultimately a change in fitness. From a specific genotype, all the possible mutations thus result in a distribution of phenotypic effect (DPE) and fitness effects (DFE). The DPE and DFE are not known a priori, but are the resulting consequence of the mutation-selection-drift balance. Empirically, these distributions are of particular importance since they can be obtained experimentally or inferred with other data. As an example, DFE can be inferred from polymorphism dataset (Eyre-walker and Keightley, 2007; Galtier, 2016). Moreover, the distribution of $\Delta\Delta G$ for novel mutations can be obtained experimentally.

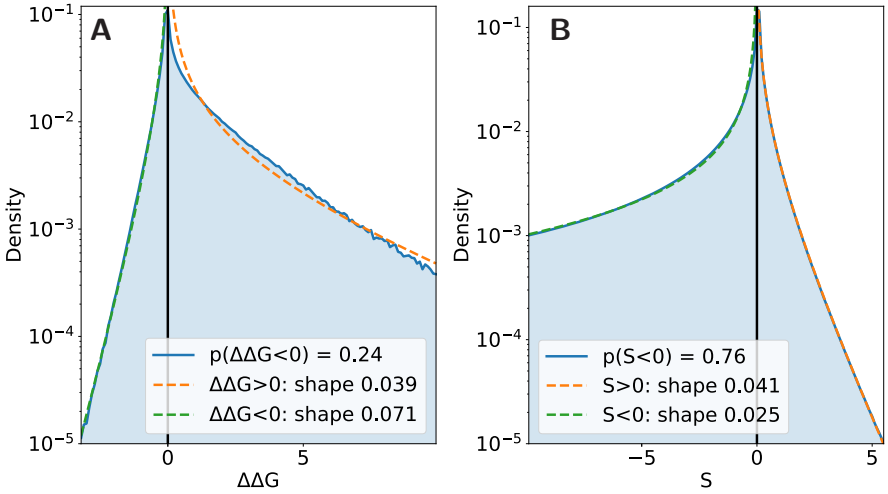


Figure 12.15: *Distribution of fitness effects and phenotypic effect for novels non-synonymous mutations observed along a simulation at the mutation-selection balance. $\alpha = -118$, $\gamma = 1$, $n = 300$, $\beta = 1.686$, and for each non-optimal amino acid, γ is scaled by the Grantham distance to the optimal amino acid. Each side of the distribution is fitted to a gamma distribution, shown in dotted line. Panel A. Distribution of observed $\Delta\Delta G$, which fit adequately the gamma distribution for negative $\Delta\Delta G$ (stabilizing mutations). Panel A. Distribution of observed selection coefficient, which fit adequately the gamma distribution for both positive and negative selection coefficient. However the shape parameter estimated is not the same for positive and negative selection coefficients.*

Mutation-selection-drift as bridge between phylogeny and population-genetics

The first strategy is to augment information about interspecies conservation with information about genetic polymorphisms. $g(x,S)\mathrm{d}x$ is the expected time for which the population frequency of the derived allele, at the site, is in the range $(x, x+\mathrm{d}x)$ before eventual absorption:

$$g(x,S) \approx \frac{2\left[1-\mathrm{e}^{-S(1-x)}\right]}{(1-\mathrm{e}^{-S})x(1-x)} \tag{12.104}$$

Sawyer and Hartl (1992) expanded the modeling of site evolution to multiple sites. The model makes the following assumptions:

- Mutations arise at Poisson times (rate u per site per generation)
- Each mutation occurs at a new site (infinite sites, irreversible)
- Each mutant follows an independent Wright-Fisher process (no linkage)

In a sample of size n, the expected number of sites with k (which ranges from 1 to $n-1$) copies of the derived allele is defined as a function of $g(x)$:

$$\begin{aligned} G(k,n,\theta,S) &= 2N_{\mathrm{e}}u\int_0^1 g(x,S)\binom{n}{k}x^k(1-x)^{n-k}\mathrm{d}x \\ &= \theta\int_0^1 \frac{1-\mathrm{e}^{-S(1-x)}}{(1-\mathrm{e}^{-S})x(1-x)}\binom{n}{k}x^k(1-x)^{n-k}\mathrm{d}x, \text{ where } \theta = 4N_{\mathrm{e}}u \\ &= \binom{n}{k}\frac{\theta}{1-\mathrm{e}^{-S}}\int_0^1 \left(1-\mathrm{e}^{-S(1-x)}\right)x^{k-1}(1-x)^{n-k-1}\mathrm{d}x \end{aligned} \tag{12.105}$$

In the mutation selection-framework developed, the fitness of a given genotype is a function of the encoded amino-acid through the site-wise amino-acid fitness profiles ($\boldsymbol{f}^{(\mathrm{z})}$ at site z). Thus the coefficient ($S = 4N_{\mathrm{e}}s$) associated to a mutation is a function of the amino acids encoded by the ancestral (i) and derived (j) codon. Altogether the selection coefficient from i to j at site z is:

$$\begin{aligned} S_{i,j}(N_{\mathrm{e}}, \boldsymbol{f}^{(\mathrm{z})}) &= 4N_{\mathrm{e}}(f_j^{(\mathrm{z})} - f_i^{(\mathrm{z})}) \\ &= F_j^{(\mathrm{z})} - F_i^{(\mathrm{z})} \end{aligned} \tag{12.106}$$

Similarly, the mutation rate between by the ancestral (i) and derived (j) codon is a function of the nucleotide changes between the codons. If the codons are not neighbor, meaning a single mutation is not sufficient to jump from i) to j, the mutation rate is

equal to 0. If the codons are neighbors, the mutation rate is given by the nucleotide rate matrix ($\boldsymbol{u}$). Altogether, the scaled mutation rate $\theta_{i,j}$ from codon i to j is:

$$\theta_{i,j}(N_{\mathrm{e}}, u, \boldsymbol{R}) = 4N_{\mathrm{e}}uR_{\mathcal{M}(i,j)}, \tag{12.107}$$

where $\mathcal{M}(i,j)$ denotes the nucleotide change between neighbors codon i and j (e.g. $\mathcal{M}(AAT, AAG) = TG$). If a site is polymorphic and the ancestral (i) and derived (j) codons are neighbors, the probability of observing i copies ($n \geq i > 0$) of the derived codon (j), in a sample of size n, at site z, is given by:

$$\mathbb{P}(i = n - k, j = i \mid N_{\mathrm{e}}, u, \boldsymbol{R}, \boldsymbol{f}^{(\mathrm{z})}) = G\left(k, n, \theta_{i,j}(N_{\mathrm{e}}, u, \boldsymbol{R}), S_{i,j}(N_{\mathrm{e}}, \boldsymbol{f}^{(\mathrm{z})})\right) \tag{12.108}$$

Altogether, the probability that a site is monomorphic is given by:

$$\mathbb{P}(i = n \mid N_{\mathrm{e}}, u, \boldsymbol{R}, \boldsymbol{f}^{(\mathrm{z})}) = 1 - \sum_{j \in \mathcal{V}(i)} \sum_{k=1}^{n} G\left(k, n, \theta_{i,j}(N_{\mathrm{e}}, u, \boldsymbol{R}), S_{i,j}(N_{\mathrm{e}}, \boldsymbol{f}^{(\mathrm{z})})\right), \tag{12.109}$$

where $\mathcal{V}(i)$ is the set of codons neighbors of codon i (i.e. one mutation away). And all other probabilities equal to 0.0.

A Bayesian Mutation–Selection Framework for Detecting Site-Specific Adaptive Evolution in Protein-Coding Genes

Nicolas Rodrigue,[1,*] Thibault Latrille,[2] and Nicolas Lartillot[2]

[1]Department of Biology, Institute of Biochemistry, and School of Mathematics and Statistics, Carleton University, Ottawa, Canada
[2]Université de Lyon, Université Lyon 1, CNRS; UMR 5558, Laboratoire de Biométrie et Biologie Évolutive, Villeurbanne, F-69622, France

***Corresponding author**: E-mail: nicolas.rodrigue@carleton.ca

Associate editor: Nicolas Rodrigue

Abstract

In recent years, codon substitution models based on the mutation–selection principle have been extended for the purpose of detecting signatures of adaptive evolution in protein-coding genes. However, the approaches used to date have either focused on detecting global signals of adaptive regimes—across the entire gene—or on contexts where experimentally derived, site-specific amino acid fitness profiles are available. Here, we present a Bayesian site-heterogeneous mutation–selection framework for site-specific detection of adaptive substitution regimes given a protein-coding DNA alignment. We offer implementations, briefly present simulation results, and apply the approach on a few real data sets. Our analyses suggest that the new approach shows greater sensitivity than traditional methods. However, more study is required to assess the impact of potential model violations on the method, and gain a greater empirical sense its behavior on a broader range of real data sets. We propose an outline of such a research program.

Key words: **nearly neutral evolution, fitness landscape, Dirichlet process, Markov chain Monte Carlo.**

Introduction

Codon substitution models (Goldman and Yang 1994; Muse and Gaut 1994) are among the important modern tools used for uncovering potential signals of molecular adaptation from protein-coding gene alignments. One set of broadly used models focuses on estimating the ratio of rates of nonsynonymous (*dN*) and synonymous (*dS*) substitutions. These models introduce a multiplicative parameter, denoted ω, to entries in a codon substitution matrix corresponding to nonsynonymous events. Because ω is the only distinction between the rate specification of nonsynonymous and synonymous events, it directly corresponds to $\omega = dN/dS$.

Fitting a model with a single (global) nonsynonymous multiplicative parameter almost always leads to $\omega < 1$ (Yang 2006), given the pervasive purifying selection that operates at most codon sites over most of evolutionary history. Many efforts were thus made to develop codon substitution models with distributions of ω values across sites and/or across the branches of a phylogeny (reviewed in Yang 2019). A common objective of such developments is to uncover specific sites having evolved under an adaptive regime (e.g., with $\omega > 1$), perhaps along a particular branch of the phylogeny.

Meanwhile, another set of codon substitution models was proposed, with a focus on accounting for purifying selection at the amino acid level in a site-heterogeneous manner (Halpern and Bruno 1998). Having nucleotide-level parameters controlling a mutational process, and amino acid fitness profiles controlling selection, they have come to be known as *mutation–selection* models (e.g., Yang and Nielsen 2008; Rodrigue et al. 2010). In these models, the *dN*/*dS* ratio is not explicitly parameterized. Instead, it is an emerging quantity, induced by the interplay between mutation, selection, and drift. Spielman and Wilke (2015) have shown how to calculate the *dN*/*dS* induced by the mutation–selection framework—which we denote ω_0 (Rodrigue and Lartillot 2017)—and found that, under specific conditions (i.e., a substitution process at equilibrium, without selection on synonymous variants), it is always true that $\omega_0 \leq 1$, as expected from a model focused on purifying selection.

In the last few years, the mutation–selection framework has been extended for the purpose of detecting genes having evolved under an adaptive regime, in either a global (Rodrigue and Lartillot 2017) or site-specific (Bloom 2017) manner. Like their traditional predecessors, these recent mutation–selection models introduce a multiplicative parameter on nonsynonymous rates. However, because amino acid profiles are also involved in modulating nonsynonymous rates, such a multiplicative parameter—which we denote as ω_* (Rodrigue and Lartillot 2017)—cannot be interpreted as the *dN*/*dS* ratio; we chose to emphasize this distinction with an asterisk in the notation. Given that the mutation–selection formulation itself induces a certain *dN*/*dS* ratio, ω_0, the net overall *dN*/*dS* ratio, ω, can be thought of as $\omega = \omega_0 \times \omega_*$, which can be rearranged to $\omega_* = \omega/\omega_0$. The latter equation helps clarify the interpretation of ω_* as a measure of the deviation in nonsynonymous rates from the expectation

under the pure mutation–selection equilibrium; in particular, $\omega_* > 1$ indicates that nonsynonymous rates are higher than expected, even though they might not be so high as to lead to $\omega > 1$.

New Approaches

Here, we conduct the first exploration of a Bayesian mutation–selection model with site-heterogeneous amino acid fitness profiles and site-heterogeneous ω_* values. The Bayesian nature of the model qualifies it as a *random-effects* approach, in contrast to the *fixed-effects* approach utilized to date in maximum-likelihood versions of mutation–selection models (Halpern and Bruno 1998; Holder et al. 2008; Tamuri et al. 2014; Bloom 2017).

Results and Discussion

Models with Global ω or ω_*

We first contrasted the difference in behavior between a traditional codon substitution model inspired from Muse and Gaut (1994), with a global ω parameter (a traditional model we denote MG-M0, described in detail in the Materials and Methods section), and a mutation–selection model with a Dirichlet process prior on amino acid profiles across sites and a global ω_* parameter (a model presented in Rodrigue and Lartillot, 2017, which we denote here as MutSel-M0$_*$, and also described in the Materials and Methods section).

Simulations

Figure 1 shows results of the two models on data generated through a simulation approach explicitly allowing for fluctuating selection at some sites; for these sites, amino acid fitness profiles change along the branches of the phylogeny, as described in Rodrigue and Lartillot (2017) and in the Materials and Methods section. The simulation system is an attempt at mimicking an adaptive substitution process, where the simulated substitution history tracks a changing amino acid fitness optimum along the branches of the tree, and thus accrues more nonsynonymous substitutions than expected under a pure nearly neutral regime (i.e., mutation–selection balance). An important distinction with Rodrigue and Lartillot (2017) is that here the simulated data set contains only 10% of codon sites generated under adaptive regimes, and 90% of codon sites generated under a pure nearly neutral mutation–selection formulation (Rodrigue et al. 2010). We produced alignments of 300 codons in length, repeating the simulation thrice, with different sets of empirically inferred amino acid profiles (see Lowe and Rodrigue 2020, and the Materials and Methods section).

Results under the traditional MG-M0 model (red) reflect the overall purifying selection governing most of the data-generating processes, with posterior mean ω values at 0.14, 0.15, 0.13 in three replicates displayed in panels 1A, 1B, and 1C, respectively. The fact that 10% of sites where produced under an adaptive regime is underwhelming to the MG-M0 model, and indeed little is generally expected of it in practice. Results under the MutSel-M0$_*$ model (blue) show a posterior distribution for ω_* situated above 1, with $p(\omega_* > 1|D) \geq 0.99$ (where D refers to the data set) for the first two replicates (fig. 1*A* and *B*); indeed, the second replicate has a posterior mean that surpasses 2. For the third replicate, we find a slightly lower probability, at $p(\omega_* > 1|D) \sim 0.93$, still highly suggestive of a signal for adaptive evolution.

Previous studies (Rodrigue and Lartillot 2017; Lowe and Rodrigue 2020) have shown that a simulation conducted with 100% of sites under the pure nearly neutral mutation–selection formulation leads to a posterior distribution of ω_* situated around 1 (while the ω parameter inferred under the MG-M0 model on such simulated data tends to be closer to 0 than to 1, as shown in Rodrigue and Lartillot 2017). Here, however, 10% of sites have evolved with higher than expected nonsynonymous rates, which pulls the distribution of ω_* to the right. Already with the use of single additional parameter, ω_*, the mutation–selection framework allows us to detect adaptation where the traditional framework with a single ω parameter would not. Note that these results are under ideal conditions, however, free of the numerous potential model violations present in real data that could sway inferences of ω_*.

Real Data

Figure 2 shows the results of these models on a hand-full of real alignments. Figure 2*A* displays the results on the well-known β-Globin alignment sampled across 17 vertebrates (Yang et al. 2000). As in the simulation, the MG-M0 model indicates that $\omega < 1$. In contrast, under MutSel-M0$_*$, the posterior mean of ω_* is around 1.8, with a high posterior probability in favor of a value greater than 1, $p(\omega_* > 1|D) > 0.99$, suggesting the presence of adaptive evolution in this gene. As described with the simulation experiment presented above, and assuming negligible effects of potential model violations, adaptive evolution on even a relatively small fraction of the sites of the gene could be sufficient to induce such a rightward shift in the posterior distribution of ω_*.

Figure 2*B* displays results on an alignment of the alcohol dehydrogenase (Adh) gene sampled across 23 species of *Drosophila*. Here again, the MG-M0 model indicates that $\omega < 1$, with a posterior mean $\sim$0.13. In contrast, with the MutSel-M0$_*$ model, we find a posterior mean $\omega_* \sim 1.2$, and $p(\omega_* > 1|D) > 0.95$. As for the β-Globin alignment, and again assuming no major effects from potential model violations, this result could be explained by a fraction of sites evolving under adaptive evolution regimes. No previous phylogenetic approach has found signals of adaptive evolution in this gene, in spite of the fact that population-genetic approaches have long suggested adaptation in many instances (e.g., McDonald and Kreitman 1991; Matzkin and Eanes 2003; Matzkin 2003). While a specific scenario of Adh adaptation in specific species has been refuted by Siddiq and Thornton (2019), their study nonetheless provides strong experimental evidence of major fitness effects of some mutations, suggesting adaptive opportunities across *Drosophila*.

The four remaining panels of figure 2 (*C–F*) show results on four genes sampled across placental mammals (Lartillot and Delsuc 2012). Again, $\omega < 1$ in all four genes, whereas ω_* is

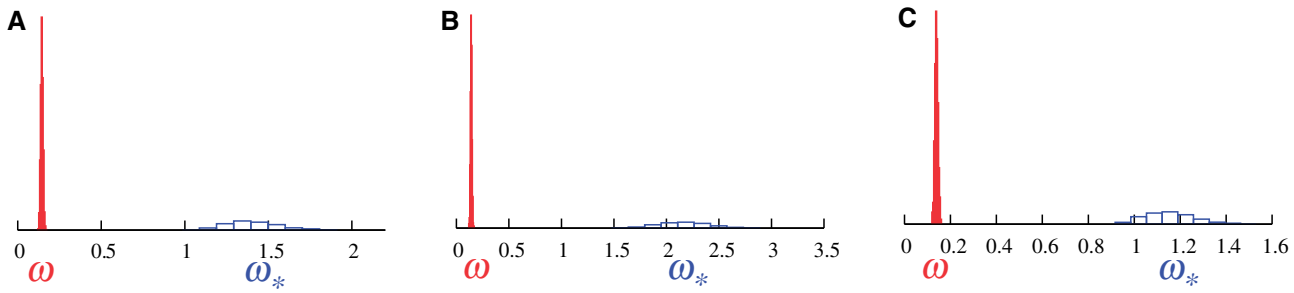


Figure 1:. Posterior distributions of ω (red, under MG-M0) and ω_* (blue, under MutSel-M0$_*$) on simulated data sets with 10% of sites evolved under adaptive evolution (see Materials and Methods section).

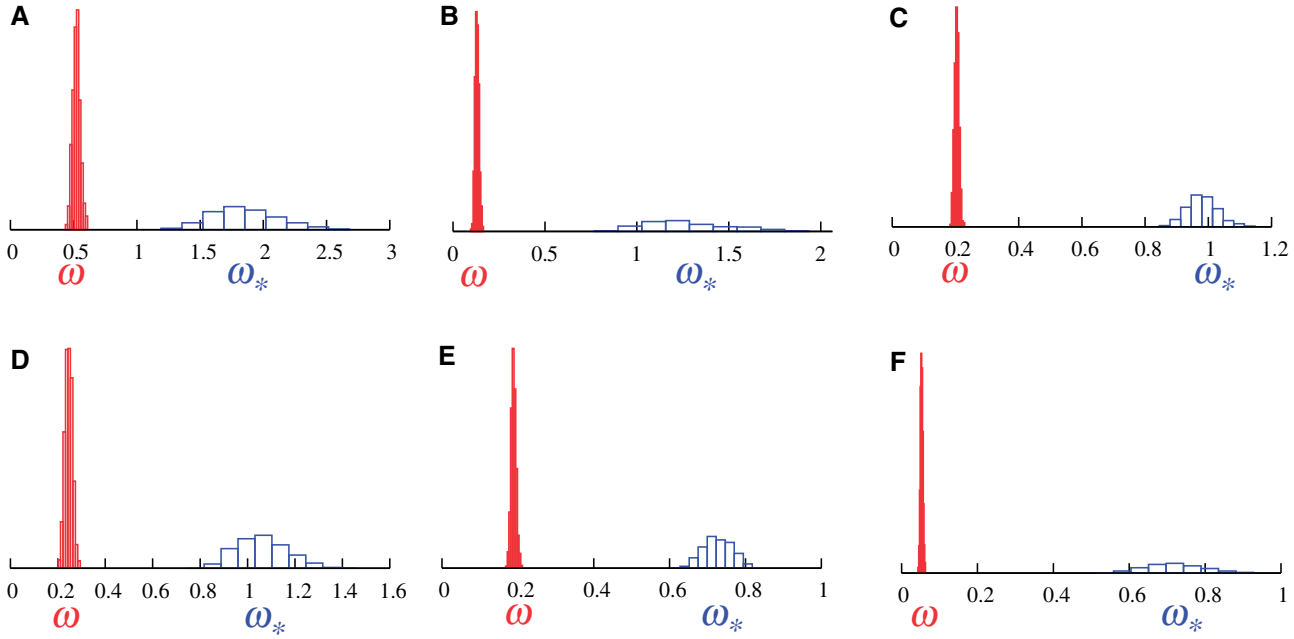


Figure 2:. Posterior distributions of ω (red, under MG-M0) and ω_* (blue, under MutSel-M0$_*$) on β-Globin, Adh, Vwf, Adora3, Rbp3, S1pr1 data sets (see Materials and Methods section).

either around 1, or slightly below, which does not suggest adaptive evolution in these genes across placental mammals. This does not rule out the possibility that some of these genes have some sites under adaptive evolution, but perhaps these sites are too few and/or too mildly adaptive to raise ω_* beyond 1. Previous simulations studies have pointed to epistasis (Rodrigue and Lartillot 2017) or weak evolutionary signal (Lowe and Rodrigue 2020) as potential reasons for $\omega_* < 1$. In the absence of major effects from model violations, these are conditions that tend to make the model conservative in the detection of adaptive regimes.

Models with Heterogeneous ω or ω_*

In spite of the potential of the MutSel-M0$_*$ model—able to capture relatively subtle signals of adaptive evolution—it still does not directly allow us to pinpoint which sites are most responsible for such signals. This is one of the motivations of *site-models*. Classical site-models (Nielsen and Yang 1998; Yang et al. 2000; Yang and Swanson 2002) consider alignment sites as having been produced from a distribution of possible ω values. They are typically used in the context of an empirical Bayes approach for identifying sites with a strong statistical support for a $\omega > 1$; and they are more efficient at detecting positive selection than the simple MG-M0 model with a single ω for all sites. For instance, they do find sites under positive selection in the case of the β-Globin gene (detailed below, but also see Yang et al. 2000). On the other hand, site-models might still miss those sites under weaker positive selection. In particular, an adaptive regime at a site could be sufficiently strong to increase the *dN*/*dS* ratio, but not to the point of driving it well above 1. In other words, at least in their current version, these models might present the same limitation as the classical MG-M0 model, as compared with MutSel-M0$_*$ model, although now at the level of the single site. This in turn suggests that the rationale of estimating ω_* in the context of a mutation–selection model should be explored not just globally over the whole gene (Rodrigue and Lartillot 2017), but as a distribution across sites of the gene (Bloom 2017).

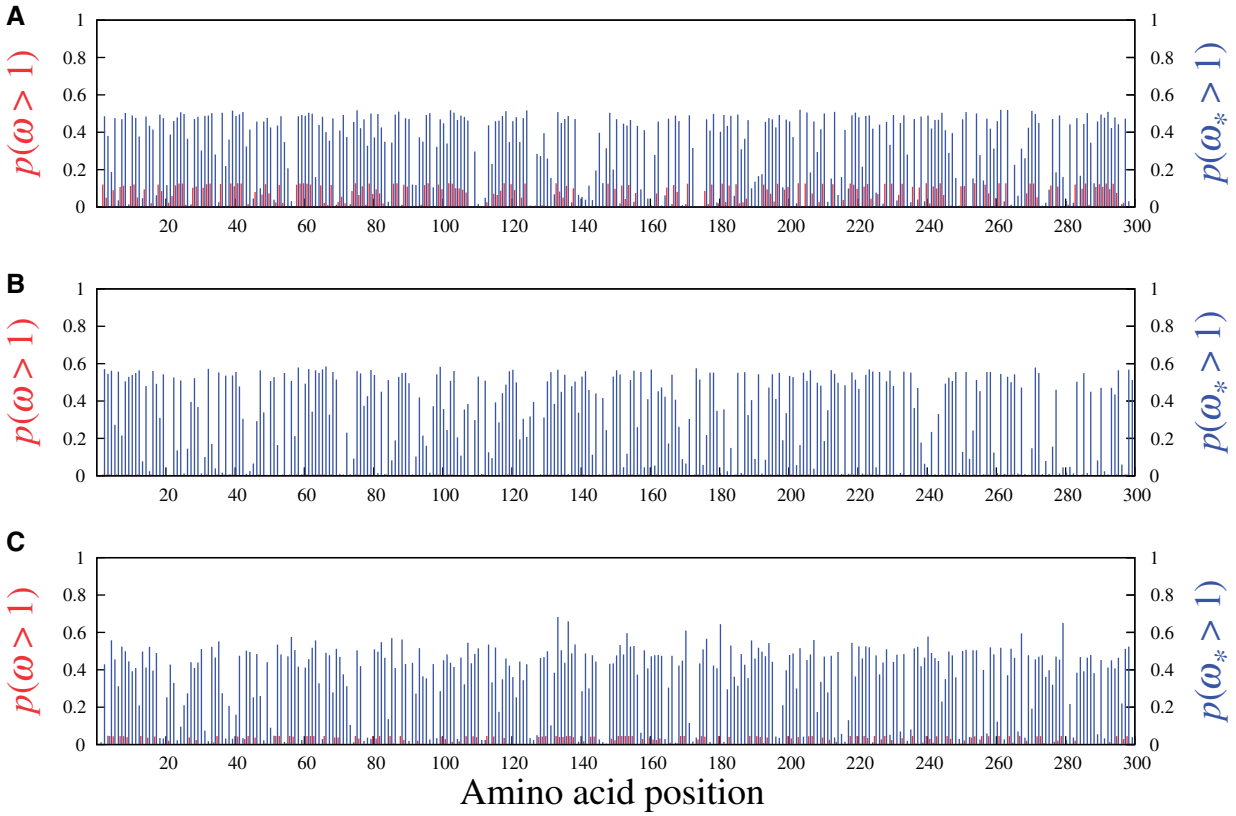


Figure 3:. Site-specific posterior probabilities of ω (red, under MG-M3) and ω_* (blue, under MutSel-M3$_*$) being greater than 1 on data sets simulated under the pure mutation–selection framework.

To illustrate this point, and for simplicity here, we work with the classical MG-M3 model, inspired from Muse and Gaut (1994) and Yang et al. (2000), which invokes a finite mixture of three ω values—with their respective weights—jointly estimated with all other parameters given the data. We also study a new model referred to as MutSel-M3$_*$, which is built from a finite mixture of three ω_* values, and respective weights, combined with the Dirichlet process prior on amino acid profiles across sites, and global mutational parameters. The two forms of across-site heterogeneity are independent in the model construction, in that each site draws its amino acid profile and its ω_* independently from the two corresponding mixtures.

Simulations

As a verification, figure 3 shows the results under the MG-M3 (red) and MutSel-M3$_*$ (blue) models on three simulated data sets, this time generated entirely under the pure mutation–selection framework (i.e., no adaptive regimes within the data-generating processes). In accordance with the simulation, no sites have high probabilities of having $\omega_* > 1$ (or $\omega > 1$). Most sites have posterior probabilities of $\omega_* > 1$ ranging from 0 to 0.5, or not much more, suggesting that the MutSel-M3$_*$ model tends to mildly underestimate some site-specific ω_* values. One possible reason for such underestimates is the fact that, in its current form, the mutation–selection apparatus utilized tends to overestimate ω_0 (the nonsynonymous to synonymous rate ratio *induced* by the amino acid fitness profiles), as shown by Spielman and Wilke (2015). Overall, however, if the data-generating process does not depart too drastically from the model's assumptions, this behavior tends to make MutSel-M3$_*$ conservative vis-à-vis inferences of adaptive evolution.

These simulations also highlight an inherent risk built into the MutSelM3$_*$ model's construction, in comparison with MG-M3: the threshold for a site to considered of interest—in terms of potential adaptive evolution—is much closer to the value expected under the null (of no adaptive regime) under MutSelM3$_*$ than under MG-M3, with the latter reporting site-specific probabilities of having $\omega > 1$ that are close to 0; for the second replicate in particular (fig. 3*B*), $p(\omega > 1|D)$ never surpasses 0.007. In other words, finding a site with $p(\omega > 1|D) > 0.95$ under the MG-M3 model represents a dramatic increase in nonsynonyomous rate, compared to finding one with $p(\omega_* > 1|D) > 0.95$ under the MutSelM3$_*$ model, which could make MutSelM3$_*$ more vulnerable to false positives from stochastic effects, or from the effects of model violations.

Figure 4 shows the results on the three simulated data sets studied in figure 1 (i.e., with 10% of sites simulated with an adaptive regime). The panels include vertical marks at the top, showing the 30 codon sites simulated under adaptive regimes. Sites evolving under an adaptive regime tend to accrue more nonsynonymous substitutions than under a nearly neutral regime, which would shift ω_* to the right of the unit. With a threshold posterior probability of 0.95 for

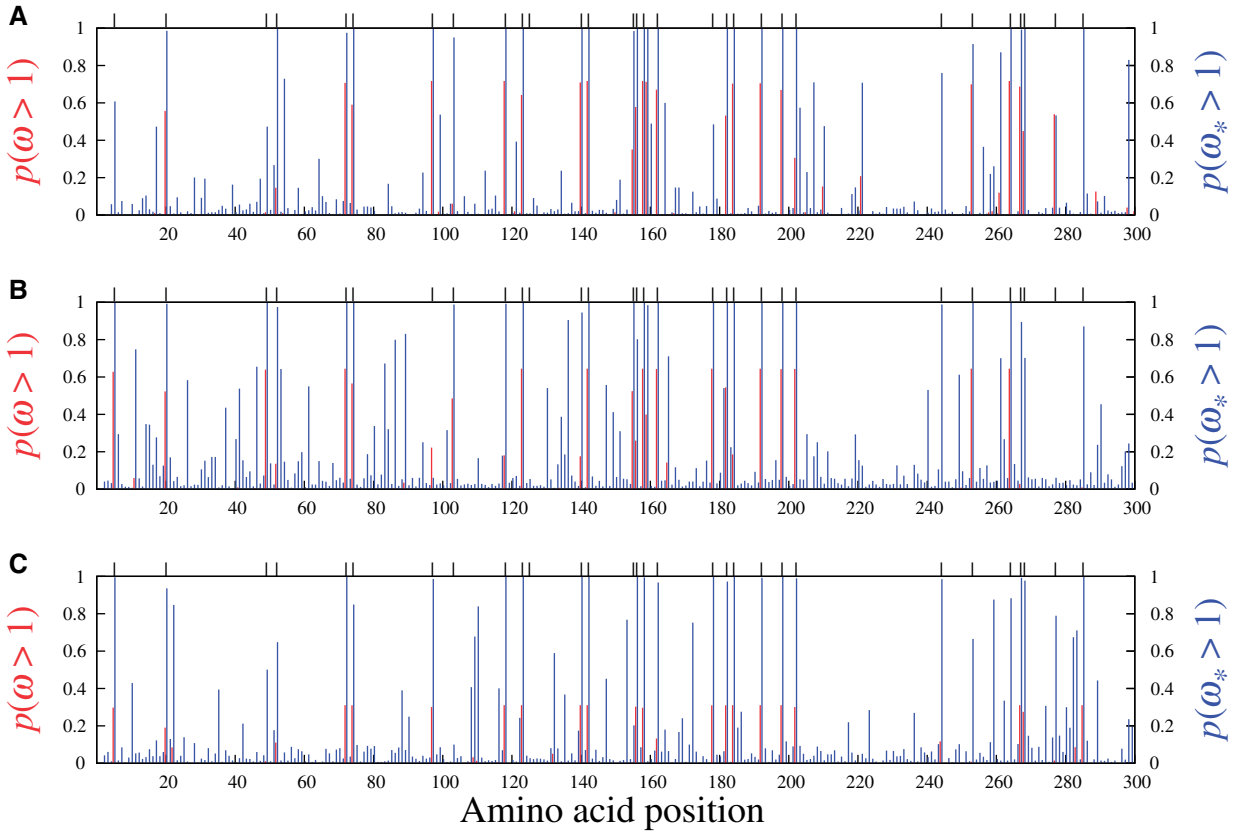


FIGURE 4:. Site-specific posterior probabilities of ω (red, under MG-M3) and ω_* (blue, under MutSel-M3$_*$) being greater than 1 on data sets simulated with 30 sites (marked with at top of panels) under an adaptive regime, and the remaining 270 sites under the pure mutation–selection framework.

$p(\omega_* > 1|D)$, the MutSel-M3$_*$ model correctly identifies 23/30 sites (76%), calls 1 false positive, and misses 7 sites for the first and second replicates, whereas for the third replicate it correctly identifies 20/30, with no false positives. Of note, a single false positive out of 24 discoveries, using a threshold of 0.95, corresponds to an accuracy of ∼96%, thus suggesting that the posterior probabilities are reasonably well-calibrated, reflecting our actual rate of true discovery. The MG-M3 models identify no sites at this threshold, although the plot suggests that it nonetheless faintly detects some adaptive signal. Interestingly, the sites leading to false positives under the MutSel-M3$_*$ model also tempt the MG-M3 model; the simulations are stochastic processes, and can, from time to time, accumulate a disproportionately high number of nonsynonymous substitutions, even when the configuration of the simulating model is one of pure mutation–selection balance. In other words, false positives may not come about solely as a result of a problem with MutSel-M3$_*$ model itself, but rather, at least partly, from a chance occurrence in the simulation. Still, this demonstrates the increased risk of the MutSelM3$_*$ model over MG-M3. However, the MG-M3 model also clearly lacks sensitivity; the sure way of having no false positives is to have no positives at all. It is particularly noteworthy that some of the sites correctly identified by MutSel-M3$_*$ show virtually no signal under MG-M3 (e.g., sites 52, 103, 285 in the first replicate, fig. 4A). In contrast, all of the sites simulated with an adaptive regime but missing the 0.95 threshold under MutSel-M3$_*$ nonetheless have relatively high probabilities of having $\omega_* > 1$. Overall, under ideal conditions, the MutSel-M3$_*$ model seems to have considerably greater sensitivity than the traditional-style MG-M3, at the cost of a mildly increased risk of false positives.

Real Data

Figures 5 and 6 display the results obtained from analyzing the six real data sets mentioned above with the MG-M3 and MutSel-M3$_*$ models. For the β-GLOBIN alignment (fig. 5A), our Bayesian version of the classic MG-M3 model leads to the same set of sites identified with these traditional models in the maximum likelihood context (Yang et al. 2000): at the 95% threshold, the sites are 7, 11, 42, 48, 50, 54, 67, 85, and 123. Under the MutSel-M3$_*$ model, these same sites are also found, and the following three are added: 10, 74, and 84. (The complete lists of sites identified at different thresholds are reported in table 1.) It is interesting to note that the MG-M3 model found $p(\omega > 1|D) = 0.381$ for site 10, $p(\omega > 1|D) = 0.244$ for site 74, and $p(\omega > 1|D) = 0.074$ for site 84. These last three sites, and site 84 in particular, yield results compatible with the interpretation of having evolved under a mild adaptive regime, of changing amino acid fitness profiles over time, leading to an increase in nonsynonymous rate; the increase is not to the point where $\omega > 1$ at a site in question, although it is enough for $\omega_* > 1$. Sites 10 and 74 are known

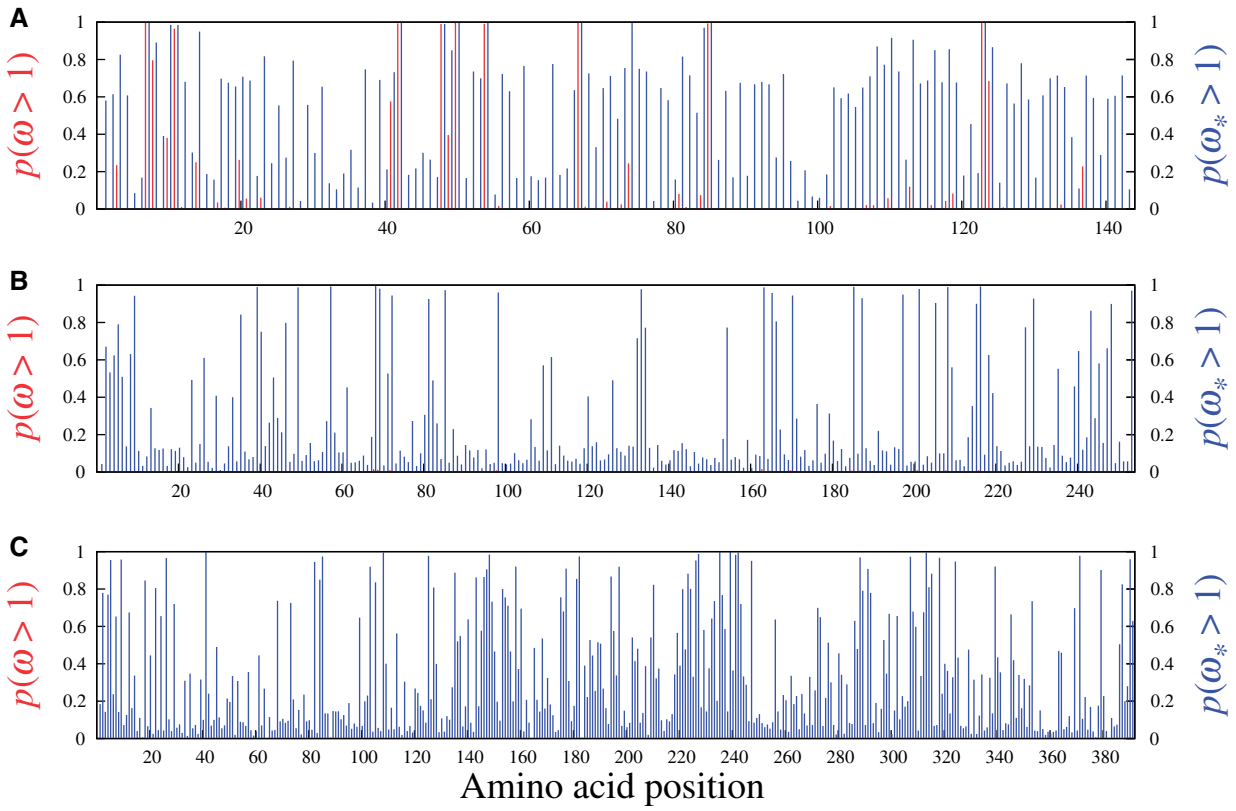


Figure 5:. Site-specific posterior probabilities of ω (red, under MG-M3) and ω_* (blue, under MutSel-M3$_*$) being greater than 1 on β-Globin, Adh, and Vwf.

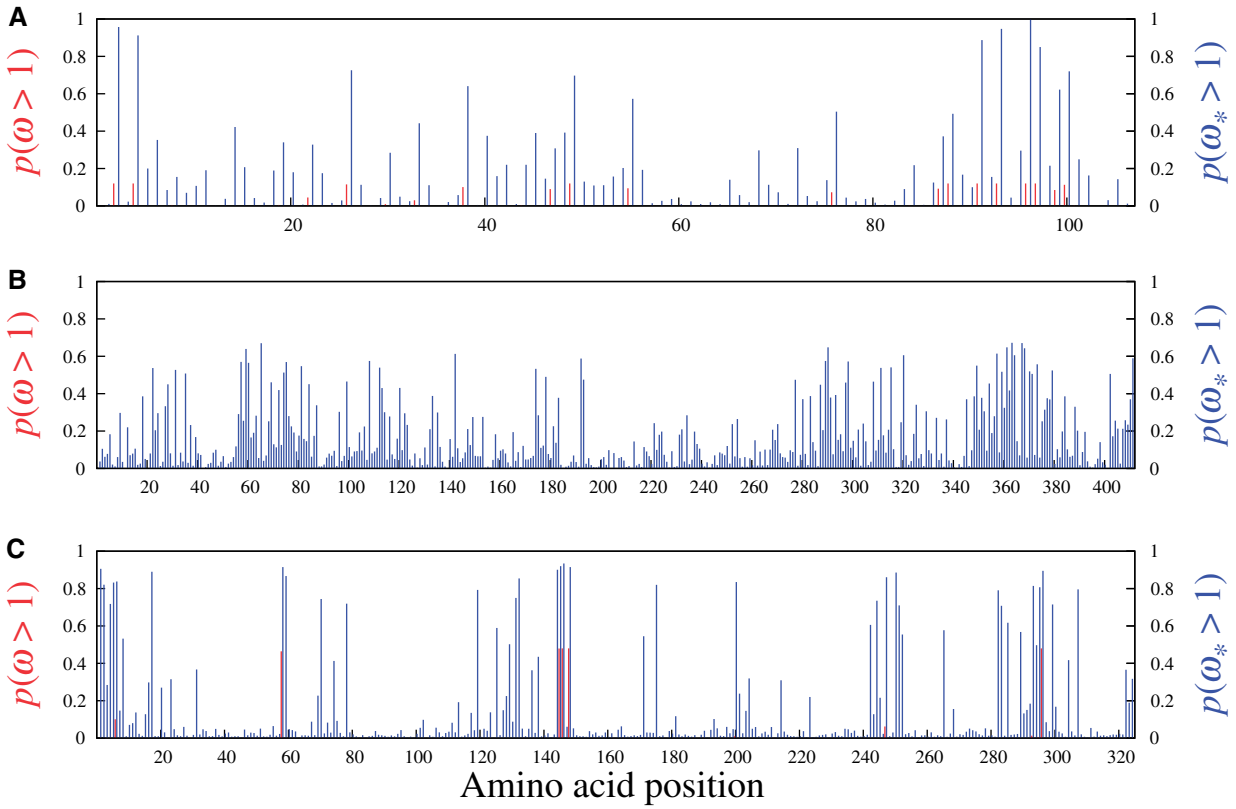


Figure 6:. Site-specific posterior probabilities of ω (red, under MG-M3) and ω_* (blue, under MutSel-M3$_*$) being greater than 1 on Adora3, Rbp3, and S1pr1.

Table 1. Amino acid sites under positive selection.

Data	Model	Sites
β−Globin	MG-M3	**7**, 11, **42**, **48**, **50** **54**, **67**, **85**, **123**
	MutSel-M3$_*$	**7**, 10, 11, *14*, **42**, **48**, **50** **54**, **67**, **74**, 84, **85**, *110*, *113*, **123**
Adh	MG-M3	–
	MutSel-M3$_*$	9, 39, 49, **57**, **68**, 69, *72*, 81, 85, 98, 133, 163, 165, *170*, **185**, 187, 197, 201, 205, 208, 216, 229, 253
Vwf	MG-M3	–
	MutSel-M3$_*$	5, 9, 26, **41**, *82*, 85, *103*, **108**, 125, *147*, 148, *158*, *177*, 182, *197*, 226, 227, **235**, **239**, 241, **242**, 247, 288, *291*, 307, **313**, 318, *324*, 339, 371, *379*, 390
Adora3	MG-M3	–
	MutSel-M3$_*$	2, *4*, *93*, **96**
Rbp3	MG-M3	–
	MutSel-M3$_*$	–
S1pr1	MG-M3	–
	MutSel-M3$_*$	*1*, *58*, *144*, *145*, *146*, *148*

Note.—Numbers in italic font are at the 0.9 level, in plain font at the 0.95 level, and in bold font at 0.99 level.

to be involved in oxygen affinity, which could indeed make them a target for adaptive evolution.

The sites uncovered by MutSel-M3$_*$ on the β-Globin data set are conditional on the overall construction of the model, which makes many oversimplified assumptions. As such, the list of sites should be considered provisional, in need of more thorough investigation by external means, and in the context of a larger scale application of the model. Some of the model violations potentially at play here, and that have mislead other types of approaches to detecting adaptive evolution, include variable effective population size (Rousselle et al. 2018), biased gene-conversion (Ratnakumar et al. 2010), multinucleotide mutations (Venkat et al. 2018), and nonhomogeneous/nonneutral synonymous substitution rates (Wisotsky et al. forthcoming). Richer simulation studies will be needed to better understand how the MutSel-M3$_*$ model reacts to such violations, and the extent to which they could be responsible for false positives.

Results of the analysis of Adh (fig. 5*B*, table 1) suggest several sites under adaptive evolution under the MutSel-M3$_*$ model, whereas the MG-M3 yields posterior probabilities of $\omega > 1$ at all sites that are numerically indistinguishable from 0. Given that most studies suggesting adaptation in this gene have relied on population-genetic methodologies, which pool the statistics across all sites, a comparison of sites uncovered by the MutSel-M3$_*$ model with previous results is not possible.

As with the analyses of the β-Globin data set, much more work will be required to determine the plausibility of these new results on the Adh data set. In addition to the aforementioned potential model violations, with a sampling across *Drosophila*, which have high effective population sizes, features such as uneven codon usage can become highly pronounced (Powell and Moriyama 1997), potentially misleading inferences of site-specific adaptation as well. As a hypothetical example, suppose that the codon TTG is used almost exclusively for encoding leucine, and that GTG is similarly strongly favored for encoding valine. Also suppose that leucine and valine are of equivalent fitness at a given site. In such a context, nonsynonymous substitutions between TTG and GTG accumulate more readily than synonymous substitutions. If this feature were to be present to a high extent, it could mislead the MutSel-M3$_*$ model into inferring $\omega_* > 1$, thus suggesting adaptive evolution where the regime is in fact one of strict purifying selection on codon usage. Simulations should eventually be used to study effects relevant to high effective population sets of taxa—such as codon usage—on the inferences of MutSel-M3$_*$.

Our analysis of the mammalian-level alignment of the gene Vwf also suggests several sites with adaptive signatures under the MutSel-M3$_*$ model, and none under the MG-M3 model (fig. 5*C*, table 1). A previous study, utilizing branch-heterogeneous models, has suggested adaptive evolution conferring venom resistance to opposoms that prey on pitvipers (Jansa and Voss, 2011). Moreover, variants of this gene have been found to have dramatic effects on its own expression levels in mice (Lemmerhirt et al. 2006), and hence with high potential for strong fitness effects.

While these latter studies are precedents to finding sites with signatures of adaptive evolution under the MutSel-M3$_*$ model, many of the model violations mentioned above could apply here as well. At the mammalian scale of this Vwf data set, a mutation–selection-based test of selection on codon usage has been shown to be misled by the effect of CpG hypermutability (Laurin-Lemay et al. 2018). This context-dependent mutational feature could have the effect of inflating ω_* values beyond 1 at sites where there is in fact no adaptive evolution (Suzuki et al. 2009). Again, however, more simulation work is required to better understand how such issues play out with the MutSel-M3$_*$.

Of the remaining mammalian gene alignments studied with the MutSel-M3$_*$ model, two suggest very few sites having evolved under adaptive regimes (Adora3 and S1pr1, in fig. 6*A* and *C*, respectively), and one (Rbp3, fig. 6*B*) with none. The traditional MG-M3 model suggests no sites under adaptive evolution for these data sets. These three genes may be typical of results under the MutSel-M3$_*$ model at the mammalian scale (i.e., few, if any sites with high $p(\omega_* > 1|D)$), but broader empirical studies evaluating the relative proportion of genes with several sites having high probabilities of $\omega_* > 1$ are pressing.

Future Directions

The traditional codon models based on ω have become increasingly well understood thanks to decades of empirical applications and simulation studies. A similar project should be considered within the mutation–selection framework. We

have already suggested several lines of research meriting further attention, and we expand on these themes below.

Simulation Studies

A flurry of recent research has shown how a variety of approaches are highly susceptible to model violations, with many instances of purported signals of molecular adaptation being the result of unaccounted features of the evolutionary process (e.g., Ratnakumar et al. 2010; Rousselle et al. 2018; Venkat et al. 2018; Laurin-Lemay et al. 2018; Wisotsky et al. forthcoming). From the codon substitution modeling perspective, this raises important questions regarding the mutation–selection-based approach we propose here: whereas the biological expectation under traditional models is for ω values closer to 0 than to 1, such that $\omega > 1$ is a drastic threshold, representing a very pronounced increase in nonsynonymous rates, the biological expectation under the new approach is for ω_* values closer to 1, and thus naturally approaching threshold of $\omega_* > 1$. This could make the mutation–selection-based methods highly susceptible to model violations that mildly increase nonsynonymous rates for reasons other than adaptive evolution. We plan to use richer simulations to study how the new approach reacts to such model violations, and if expanding the model to recognize features such as variable effective population size, CpG hypermutability, codon usage and gene conversion biases, could introduce greater robustness to inferences of adaptive evolution.

Empirical Studies

A more detailed examination, ideally combined with experimental corroborations, of the sites uncovered by the model is pressing, and hopefully based on far more than the hand-full of data sets of the present study. This would help build our empirical understanding how the model behaves in a variety of different contexts (Moutinho et al. 2019; Slodkowicz and Goldman 2020). We hope to apply the model on a few thousand genes from the OrthoMamm database (Scornavacca et al. 2019) in a first step, before engaging broader applications across varied taxanomic contexts.

Model Variations

While we have outlined the modeling strategy with a three-component finite mixture of ω_* values, in combination with a Dirichlet process prior on amino acid profiles, many other possibilities could be considered: various parametric families on ω_* (as did Yang et al. 2000, with ω), nonparametric approaches on ω_* (as proposed for ω by Huelsenbeck et al. 2006), grids of predetermined ω_* values (in the spirit of Murrell et al. 2013), along with similar choices on modeling amino acid fitness heterogeneity (e.g., Rodrigue et al. 2010; Rodrigue 2013; Rodrigue and Lartillot 2014). The potentially complex interactions between the numerous combinations also entail a large study.

Applications

We propose these modeling ideas in two independent software packages (see below). One of our Markov chain Monte Carlo implementations can run under fixed topology as well as sample over trees, and thus enable studies of the impact of phylogenetic uncertainty in inferences of adaptive evolution, utilizing both traditional and mutation–selection codon substitution models; this also suggests more extensive studies on the potential of such models for phylogenetic inference *per se*. Another implementation we offer lends itself to integrative modeling objectives, with a wide suite of potential research avenues utilizing the mutation–selection-based approaches. Foreseeable directions in the short-term with the latter implementation include capturing the evolution of effective population size over the phylogeny, along with joint inferences of continuous-trait evolution, as formalized by Lartillot and Poujol (2011).

Materials and Methods

Simulated Data

We used the simulation system described in Rodrigue and Lartillot (2017) to generate artificial data sets using a mutation–selection framework with global mutation parameters and site-specific amino acid fitness profiles. The mutation-level parameters (which assume no selection on synonymous variants) are as given in Rodrigue and Lartillot (2017), as is the phylogenetic tree (with 38 tips). With nearly neutral simulations (i.e., with the pure mutation–selection formulation, such as detailed below), the amino acid fitness profile used to simulate a codon site is chosen at random from a set of empirically derived profiles. We obtained such profiles by running the pure Dirichlet process-based mutation–selection model (Rodrigue et al. 2010) on a multigene data set at the scale of placental mammals (Lartillot and Delsuc 2012), and calculating the posterior mean amino acid profile at each site. The simulation draws at random (with replacement) one such site-specific posterior mean profile to run the evolutionary process along the tree at one codon site, repeating to produce alignments of 300 codons. For simulations with adaptive evolutionary regimes, the starting profiles are altered along the branches of the phylogeny as detailed in Rodrigue and Lartillot (2017), with the *Red Queen* parameter set to 0.01. In contrast to the simulations in Rodrigue and Lartillot (2017), however, the adaptive simulations herein are applied to only 10% of sites of the alignment; these 30 sites were chosen at random, that is, they were spread out randomly across the alignment. The remaining 270 codon sites are simulated with the Red Queen parameter set to 0, thus constituting pure mutation–selection regimes.

Real Data

We used previously studied alignments of protein-coding genes provided by the authors of earlier works:

- β-Globin: 17 vertebrate sequences of β-globin gene, 144 codons in length, taken from Yang et al. (2000);
- Adh: 23 *Drosophila* sequences of the alcohol dehydrogenase gene, 254 codons in length, taken from Yang et al. (2000);
- Vwf: 62 sequences, at the scale of placental mammals, of the von Willbrand factor gene, 392 codons in length,

taken from Lartillot and Delsuc (2012), as were the next three alignments;

- ADORA3: 67 sequences of the adenosine receptor A3 gene, 107 codons in length;
- RBP3: 54 sequences of the retinol-binding protein 3, 412 codons in length;
- S1PR1: 67 sequences of the sphingosine-1-phosphate receptor 1 gene, 325 codons in length.

Substitution Models

The MG-M0 codon substitution model, inspired from Muse and Gaut (1994), but with a single ω parameter distinguishing nonsynonymous events, has entries given as:

$$Q_{ij} = \begin{cases} \mu_{ij}, & \text{if } i \text{ and } j \text{ are synonymous,} \\ \mu_{ij}\omega, & \text{if } i \text{ and } j \text{ are nonsynonymous.} \end{cases} \quad (1)$$

Here, μ_{ij} is the mutational parameterization, which we set as a *general-time reversible* nucleotide-level model (Lanave et al. 1984), with six exchangeability parameters (five degrees of freedom) and four frequency parameters (three degrees of freedom). The MG-M3 model has the same form, but rather than a single ω parameter, it invokes three different values (with their respective weights), and has a likelihood function consisting of the a weighted average of likelihood scores under each of the three ω values (Yang et al. 2000).

The MutSel-M0$_*$ model, presented in Rodrigue and Lartillot (2017), is given as:

$$Q_{ij}^{(n)} = \begin{cases} \mu_{ij}, & \text{if } i \text{ and } j \text{ are synonymous,} \\ \mu_{ij}\omega_* \dfrac{S_{ij}^{(n)}}{1 - e^{-S_{ij}^{(n)}}}, & \text{if } i \text{ and } j \text{ are nonsynonymous,} \end{cases} \quad (2)$$

where $S_{ij}^{(n)} = F_j^{(n)} - F_i^{(n)} = 4N_e s_{ij} = 4N_e f_j^{(n)} - f_i^{(n)}$ is the *scaled selection coefficient* (scaled by the effective population size N_e and a ploidy-dependent constant, in this example set at 4 Yang and Nielsen, 2008), calculated from the difference in fitness associated with a mutant protein with the amino acid encoded by codon j at site n, denoted $F_j^{(n)}$, with that of the wild-type population where the amino-acid encoded by i is fixed at that position, $F_i^{(n)}$. Site-specific fitness profiles are treated as random effects within a Dirichlet process system (Rodrigue et al. 2010; Rodrigue and Lartillot 2014). As with the MG-M3 model, the MutSel-M3$_*$ model invokes three distinct ω_* values, with their respective weights, as a finite mixture model of heterogeneity across sites.

Priors

Branch lengths are endowed with an exponential prior of mean controlled by a hyperprior, itself endowed with an exponential prior of mean 1. Nucleotide exchangeabilities and frequencies are each endowed with flat Dirichlet priors, whereas ω and ω_* have priors following a gamma law, controlled by two hyperparameters, each endowed with exponential priors of mean 1. Weights of finite mixture on ω or ω_* follow with flat Dirichlet prior. Amino acid fitness profiles follow a Dirichlet process prior (Rodrigue et al. 2010), implemented under a stick-breaking representation (Lartillot et al. 2013; Rodrigue and Lartillot 2014).

Data Availability

For convenience, all data sets (simulated and real) studied herein are included in the Supplementary Material.

The models presented have been implemented in an experimental version (2) of PhyloBayes-MPI (https://github.com/bayesiancook/pbmpi2), allowing for a joint sampling across parameter space, auxiliary variables, and tree topology space. We have also implemented the models in a new software called BayesCode, which is focused on integrative comparative methods under fixed topology (https://github.com/bayesiancook/bayescode). Example scripts demonstrating the use of the software are provided in the Supplementary Material.

Supplementary Material

Supplementary data are available at *Molecular Biology and Evolution* online.

Acknowledgments

The work was funded by the Natural Sciences and Engineering Research Council of Canada (N.R.), and the French National Research Agency, Grant ANR-15-CE12-0010-01/DASIRE (T.L., N.L.).

References

Bloom JD. 2017. Identification of positive selection in genes is greatly improved by using experimentally informed site-specific models. *Biol Direct*. 12:1.

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

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

Holder MT, Zwickl DJ, Dessimoz C. 2008. Evaluating the robustness of phylogenetic methods to among-site variability in substitution processes. *Philos Trans R Soc B*. 363(1512):4013–4021.

Huelsenbeck JP, Jain S, Frost SWD, Pond SLK. 2006. A Dirichlet process model for detecting positive selection in protein-coding DNA sequences. *Proc Natl Acad Sci U S A*. 103(16):6263–6268.

Jansa SA, Voss RS. 2011. Adaptive evolution of the venom-targeted vwf protein in opossums that eat pitvipers. *PLoS One* 6(6):e20997.

Lanave C, Preparata G, Sacone C, Serio G. 1984. A new method for calculating evolutionary substitution rates. *J Mol Evol*. 20(1):86–93.

Lartillot N, Delsuc F. 2012. Joint reconstruction of divergence times and life-history evolution in placental mammals using a phylogenetic covariance model. *Evolution* 66(6):1773–1787.

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

Lartillot N, Rodrigue N, Stubbs D, Richer J. 2013. PhyloBayes-MPI: phylogenetic reconstruction with infinite mixtures of profiles in a parallel environment. *Syst Biol*. 62(4):611–615.

Laurin-Lemay S, Philippe H, Rodrigue N. 2018. Multiple factors confounding phylogenetic detection of selection on codon usage. *Mol Biol Evol*. 35(6):1463–1472.

Lemmerhirt HL, Shavit JA, Levy GG, Cole SM, Long JC, Ginsburg D. 2006. Enhanced VWF biosynthesis and elevated plasma VWF due to a natural variant in the murine Vwf gene. *Blood* 108(9):3061–3067.

Lowe C, Rodrigue N. 2020. Detecting adaptation from multi-species protein-coding DNA sequence alignments. *Phylogenet Genomic Era*. 4–5.

Matzkin LM. 2003. Population genetics and geographic variation of alcohol dehydrogenase (Adh) paralogs and glucose-6-phosphate dehydrogenase (G6pd) in *Drosophila mojavensis*. *Mol Biol Evol*. 21(2):276–285.

Matzkin LM, Eanes WF. 2003. Sequence variation of alcohol dehydrogenase (adh) paralogs in cactophilic *Drosophila*. *Genetics* 163(1):181–194.

McDonald JH, Kreitman M. 1991. Adaptive protein evolution at the *adh* locus in *Drosophila*. *Nature* 351(6328):652–654.

Moutinho AF, Trancoso FF, Dutheil JY. 2019. The impact of protein architecture on adaptive evolution. *Mol Biol Evol*. 36(9):2013–2028.

Murrell B, Moola S, Mabona A, Weighill T, Sheward D, Pond SLK, Scheffler K. 2013. Fubar: a fast, unconstrained Bayesian approximation for inferring selection. *Mol Biol Evol*. 30(5):1196–1205.

Muse SV, Gaut BS. 1994. A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates, with application to the chloroplast genome. *Mol Biol Evol*. 11(5):715–724.

Nielsen R, Yang Z. 1998. Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. *Genetics* 148(3):929–936.

Powell JR, Moriyama EN. 1997. Evolution of codon usage bias in *Drosophila*. *Proc Natl Acad Sci U S A*. 94(15):7784–7790.

Ratnakumar A, Mousset S, Glémin S, Berglund J, Galtier N, Duret L, Webster MT. 2010. Detecting positive selection within genomes: the problem of biased gene conversion. *Phil Trans R Soc B*. 365(1552):2571–2580.

Rodrigue N. 2013. On the statistical interpretation of site-specific variables in phylogeny-based substitution models. *Genetics* 193(2):557–564.

Rodrigue N, Lartillot N. 2014. Site-heterogeneous mutation-selection models within the phylobayes-mpi package. *Bioinformatics* 30(7):1020–1021.

Rodrigue N, Lartillot N. 2017. Detecting adaptation in protein-coding genes using a Bayesian site-heterogeneous mutation-selection codon substitution model. *Mol Biol Evol*. 34(1):204–214.

Rodrigue N, Philippe H, Lartillot N. 2010. 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.

Rousselle M, Mollion M, Nabholz B, Bataillon T, Galtier N. 2018. Overestimation of the adaptive substitution rate in fluctuating populations. *Biol Lett*. 14(5):20180055.

Scornavacca C, Belkhir K, Lopez J, Dernat R, Delsuc F, Douzery EJP, Ranwez V. 2019. OrthoMaM v10: scaling-up orthologous coding sequence and exon alignments with more than one hundred mammalian genomes. *Mol Biol Evol*. 36(4):861–862.

Siddiq MA, Thornton JW. 2019. Fitness effects but no temperature-mediated balancing selection at the polymorphic *adh* gene of *Drosophila melanogaster*. *Proc Natl Acad Sci U S A*. 116(43):21634–21640.

Slodkowicz G, Goldman N. 2020. Integrated structural and evolutionary analysis reveals common mechanisms underlying adaptive evolution in mammals. *Proc Natl Acad Sci U S A*. 117(11):5977–5986.

Spielman SJ, Wilke CO. 2015. The relationship between dN/dS and scaled selection coefficients. *Mol Biol Evol*. 32(4):1097–1108.

Suzuki Y, Gojobori T, Kumar S. 2009. 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.

Tamuri AU, Goldman N, dos Reis M. 2014. A penalized likelihood method for estimating the distributionof selection coefficients from phylogenetic data. *Genetics* 197(1):257–271.

Venkat A, Hahn MW, Thornton JW. 2018. Multinucleotide mutations cause false inferences of lineage-specific positive selection. *Nat Ecol Evol*. 2(8):1280–1288.

Wisotsky SR, Kosakovsky Pond SL, Shank SD, Muse SV. Forthcoming. Synonymous site-to-site substitution rate variation dramatically inflates false positive rates of selection analyses: ignore at your own peril. *Mol Biol Evol*. 37(8):2430–2439.

Yang Z. 2006. *Computational Molecular Evolution*. Oxford Series in Ecology and Evolution. Oxford: Oxford University Press.

Yang Z. 2019. Adaptive molecular evolution. In: Balding DJ, Moltke I, and Marioni J, editors. Handbook of statistical genomics. Vol. i. Hoboken (NJ): Wiley.

Yang Z, Nielsen R. 2008. Mutation-selection models of codon substitution and their use to estimate selective strengths on codon usage. *Mol Biol Evol*. 25(3):568–579.

Yang Z, Nielsen R, Goldman N, Pedersen AM. 2000. Codon-substitution models for heterogeneous selection pressure at amino acid sites. *Genetics* 155:431–449.

Yang Z, Swanson WJ. 2002. Codon-substitution models to detect adaptive evolution that account for heterogeneous selective pressures among site classes. *Mol Biol Evol*. 19(1):49–57.

Bibliography

Akashi, H. 1999. Inferring the Fitness Effects of DNA Mutations From Polymorphism and Divergence Data: Statistical Power to Detect Directional Selection Under Stationarity and Free Recombination. *Genetics*, 151(1): 221 LP – 238. *Cited at page 9*

Albery, W. J. and Knowles, J. R. 1976. Evolution of Enzyme Function and the Development of Catalytic Efficiency. *Biochemistry*, 15(25): 5631–5640. *Cited at page 9*

Araya, C. L., Fowler, D. M., Chen, W., Muniez, I., Kelly, J. W., and Fields, S. 2012. A fundamental protein property, thermodynamic stability, revealed solely from large-scale measurements of protein function. *Proceedings of the National Academy of Sciences of the United States of America*, 109(42): 16858–16863. *Cited at page 65*

Arenas, M. 2015. Trends in substitution models of molecular evolution. *Frontiers in Genetics*, 6(October). *Cited at page 63*

Arenas, M., Sánchez-Cobos, A., and Bastolla, U. 2015. Maximum-Likelihood Phylogenetic Inference with Selection on Protein Folding Stability. *Molecular Biology and Evolution*, 32(8): 2195–2207. *Cited at page 71*

Arenas, M., Weber, C. C., Liberles, D. A., and Bastolla, U. 2017. ProtASR: An Evolutionary Framework for Ancestral Protein Reconstruction with Selection on Folding Stability. *Systematic Biology*, 66(6): 1054–1064. *Cited at page 71*

Aris-Brosou, S. and Yang, Z. 2002. Effects of Models of Rate Evolution on Estimation of Divergence Dates with Special Reference to the Metazoan 18S Ribosomal RNA Phylogeny. *Systematic Biology*, 51(5): 703–714. *Cited at page 11*

Auer, P., Cesa-Bianchi, N., and Fischer, P. 2002. Finite-time analysis of the multiarmed bandit problem. *Machine Learning*, 47(2-3): 235–256. *Cited at page 30*

Avery, O. T., Macleod, C. M., and McCarty, M. 1944. 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. *Journal of Experimental Medicine*, 79(2): 137–158. *Cited at page 5*

Baker, M. L., Schountz, T., and Wang, L. F. 2013. Antiviral Immune Responses of Bats: A Review. *Zoonoses and Public Health*, 60(1): 104–116. *Cited at page 144*

Bastolla, U., Bruscolini, P., and Velasco, J. L. 2012. Sequence determinants of protein folding rates: Positive correlation between contact energy and contact range indicates selection for fast folding. *Proteins: Structure, Function and Bioinformatics*, 80(9): 2287–2304. *Cited at page 135*

Bastolla, U., Dehouck, Y., and Echave, J. 2017. What evolution tells us about protein physics, and protein physics tells us about evolution. *Cited at pages 63, 135*

Bazykin, G. A. 2015. Changing preferences: deformation of single position amino acid fitness landscapes and evolution of proteins. *Biology letters*, 11(10): 20150315. *Cited at page 46*

Beaulieu, J. M., O'Meara, B. C., Zaretzki, R., Landerer, C., Chai, J., and Gilchrist, M. A. 2018. Population Genetics Based Phylogenetics Under Stabilizing Selection for an Optimal Amino Acid Sequence: A Nested Modeling Approach. *Molecular Biology and Evolution*, 36(4): 834–851. *Cited at page 200*

Bellman, R. 1966. Dynamic programming. *Science*, 153(3731): 34–37. *Cited at page 29*

Berger-Tal, O., Nathan, J., Meron, E., and Saltz, D. 2014. The Exploration-Exploitation Dilemma: A Multidisciplinary Framework. *PLoS ONE*, 9(4). *Cited at page 30*

Bergman, J., Schrempf, D., Kosiol, C., and Vogl, C. 2018. Inference in population genetics using forward and backward, discrete and continuous time processes. *Journal of Theoretical Biology*, 439: 166–180. *Cited at page 147*

Biesiadecka, M. K., Sliwa, P., Tomala, K., and Korona, R. 2020. An Overexpression Experiment Does Not Support the Hypothesis That Avoidance of Toxicity Determines the Rate of Protein Evolution. *Genome Biology and Evolution*, 12(5): 589–596. *Cited at page 135*

Blanquart, F. and Bataillon, T. 2016. Epistasis and the structure of fitness landscapes: Are experimental fitness landscapes compatible with fisher's geometric model? *Genetics*, 203(2): 847–862. *Cited at pages 110, 159*

Bloom, J. D. 2014a. An experimentally determined evolutionary model dramatically improves phylogenetic fit. *Molecular Biology and Evolution*, 31(8): 1956–1978. *Cited at pages 46, 70*

Bloom, J. D. 2014b. An experimentally informed evolutionary model improves phylogenetic fit to divergent lactamase homologs. *Molecular Biology and Evolution*, 31(10): 2753–2769. *Cited at pages 46, 70, 80, 88*

Bloom, J. D. 2017. Identification of positive selection in genes is greatly improved by using experimentally informed site-specific models. *Biology Direct*, 12(1): 1. *Cited at pages 10, 70, 88, 101, 103, 111, 118, 137, 143, 155, 206*

Bloom, J. D., Raval, A., and Wilke, C. O. 2007. Thermodynamics of Neutral Protein Evolution. *Genetics*, 175(1): 255 LP – 266. *Cited at page 127*

Bolívar, P., Guéguen, L., Duret, L., Ellegren, H., and Mugal, C. F. 2019. GC-biased gene conversion conceals the prediction of the nearly neutral theory in avian genomes. *Genome Biology*, 20(1): 5. *Cited at pages 44, 81, 100*

Borges, R. and Kosiol, C. 2020. Consistency and identifiability of the polymorphism-aware phylogenetic models. *Cited at page 147*

Borges, R., Szöllosi, G. J., and Kosiol, C. 2019. Quantifying GC-biased gene conversion in great ape genomes using polymorphism-aware models. *Genetics*, 212(4): 1321–1336. *Cited at page 147*

Bowler, P. J. 2003. *Evolution: The History of an Idea*. University of California Press. *Cited at page 4*

Brevet, M. and Lartillot, N. 2019. Reconstructing the history of variation in effective population size along phylogenies. *bioRxiv*, page 793059. *Cited at pages 44, 108, 111, 134, 203*

Bulmer, M., Wolfe, K. H., and Sharp, P. M. 1991. Synonymous nucleotide substitution rates in mammalian genes: Implications for the molecular clock and the relationship of mammalian orders. *Proceedings of the National Academy of Sciences of the United States of America*, 88(14): 5974–5978. *Cited at page 11*

Bustamante, C. D., Fledel-Alon, A., Williamson, S., Nielsen, R., Hubisz, M. T., Glanowski, S., Tanenbaum, D. M., White, T. J., Sninsky, J. J., Hernandez, R. D., Civello, D., Adams, M. D., Cargill, M., and Clark, A. G. 2005. Natural selection on protein-coding genes in the human genome. *Nature*, 437(7062): 1153–1157. *Cited at page 10*

Capderrey, C., Kaufmann, B., Jean, P., Malard, F., Konecny-Dupré, L., Lefébure, T., and Douady, C. J. 2013. Microsatellite Development and First Population Size Estimates for the Groundwater Isopod Proasellus walteri. *PLoS ONE*, 8(9): e76213. *Cited at pages 107, 109*

Cargill, M., Altshuler, D., Ireland, J., Sklar, P., Ardlie, K., Patil, N., Lane, C. R., Lim, E. P., Kalyanaraman, N., Nemesh, J., Ziaugra, L., Friedland, L., Rolfe, A., Warrington, J., Lipshutz, R., Daley, G. Q., and Lander, E. S. 1999. Characterization of single-nucleotide polymorphisms in coding regions of human genes. *Nature Genetics*, 22(3): 231–238. *Cited at page 9*

Cavalli-Sforza, L. L. and Edwards, A. W. 1967. Phylogenetic analysis. Models and estimation procedures. *American Journal of Human Genetics*, 19(3): 233. *Cited at page 56*

Chan, P. P. and Lowe, T. M. 2008. GtRNAdb: a database of transfer RNA genes detected in genomic sequence. *Nucleic Acids Research*, 37(suppl_1): D93–D97. *Cited at page 35*

Chargaff, E., Zamenhof, S., and Green, C. 1950. Human desoxypentose nucleic acid: Composition of human desoxypentose nucleic acid. *Nature*, 165(4202): 756–757. *Cited at page 5*

Charlesworth, B. 1994. The effect of background selection against deleterious mutations on weakly selected, linked variants. *Genetical Research*, 63(3): 213–227. *Cited at page 10*

Charlesworth, D. 2010. Don't forget the ancestral polymorphisms. *Heredity*, 105(6): 509–510. *Cited at page 147*

Chen, J., Glémin, S., and Lascoux, M. 2017. Genetic diversity and the efficacy of purifying selection across plant and animal species. *Molecular Biology and Evolution*, 34(6): 1417–1428. *Cited at page 10*

Cherry, J. L. 1998. Should We Expect Substitution Rate to Depend on Population Size? *Genetics*, 150(2). *Cited at pages 10, 69, 100, 121, 134*

Cherry, J. L. 2010. Expression Level, Evolutionary Rate, and the Cost of Expression. *Genome Biology and Evolution*, 2: 757–769. *Cited at page 129*

Chi, P. B., Kim, D., Lai, J. K., Bykova, N., Weber, C. C., Kubelka, J., and Liberles, D. A. 2018. A new parameter-rich structure-aware mechanistic model for amino acid substitution during evolution. *Proteins: Structure, Function, and Bioinformatics*, 86(2): 218–228. *Cited at page 71*

ChungWu, I. and Wen-Hsiung Li 1985. Evidence for higher rates of nucleotide substitution in rodents than in man. *Proceedings of the National Academy of Sciences of the United States of America*, 82(6): 1741–1745. *Cited at page 11*

Cornish-Bowden, A. 1976. The effect of natural selection on enzymic catalysis. *Journal of Molecular Biology*, 101(1): 1–9. *Cited at page 9*

Crick, F. 1958. On protein synthesis. *Symposia of the Society for Experimental Biology*, 12(138-63): 8. *Cited at page 5*

Crick, F. 1970. Central dogma of molecular biology. *Nature*, 227(5258): 561–563. *Cited at page 5*

Darriba, D., Flouri, T., and Stamatakis, A. 2018. The State of Software for Evolutionary Biology. *Molecular Biology and Evolution*, 35(5): 1037–1046. *Cited at page 150*

Dasmeh, P. and Serohijos, A. W. R. 2018. Estimating the contribution of folding stability to nonspecific epistasis in protein evolution. *Proteins: Structure, Function, and Bioinformatics*, 86(12): 1242–1250. *Cited at page 125*

Dasmeh, P., Serohijos, A. W., Kepp, K. P., and Shakhnovich, E. I. 2014. The Influence of Selection for Protein Stability on dN/dS Estimations. *Genome Biology and Evolution*, 6(10): 2956–2967. *Cited at pages 124, 148*

Davydov, I. I., Robinson-Rechavi, M., and Salamin, N. 2016. State aggregation for fast likelihood computations in molecular evolution. *Bioinformatics*, 33(3): btw632. *Cited at pages 61, 117*

De Koning, A. P., Gu, W., and Pollock, D. D. 2010. Rapid likelihood analysis on large phylogenies using partial sampling of substitution histories. *Molecular Biology and Evolution*, 27(2): 249–265. *Cited at page 61*

De Magalhães, J. P. and Costa, J. 2009. A database of vertebrate longevity records and their relation to other life-history traits. *Journal of Evolutionary Biology*, 22(8): 1770–1774. *Cited at page 119*

De Maio, N., Schlötterer, C., and Kosiol, C. 2013. Linking Great Apes Genome Evolution across Time Scales Using Polymorphism-Aware Phylogenetic Models. *Molecular Biology and Evolution*, 30(10): 2249–2262. *Cited at page 147*

Dixit, P. D. and Maslov, S. 2013. Evolutionary Capacitance and Control of Protein Stability in Protein-Protein Interaction Networks. *PLoS Computational Biology*, 9(4). *Cited at pages 67, 135*

Dobzhansky, T. 1974. Chance and creativity in evolution. In F. J. Ayala and T. Dobzhansky, editors, *Studies in the Philosophy of Biology: Reduction and Related Problems*, pages 307–338. Macmillan Education UK, London. *Cited at page 5*

Dos Reis, M. 2015. How to calculate the non-synonymous to synonymous rate ratio of protein-coding genes under the fisher-wright mutation-selection framework. *Biology Letters*, 11(4). *Cited at pages 48, 49, 84, 101, 121, 137*

Doud, M. B., Ashenberg, O., and Bloom, J. D. 2015. Site-Specific Amino Acid Preferences Are Mostly Conserved in Two Closely Related Protein Homologs. *Molecular Biology and Evolution*, 32(11): 2944–2960. *Cited at page 70*

Drummond, A. J., Ho, S. Y. W., Phillips, M. J., and Rambaut, A. 2006. Relaxed Phylogenetics and Dating with Confidence. *PLoS Biology*, 4(5): e88. *Cited at page 11*

Drummond, D. A. and Wilke, C. O. 2008. Mistranslation-Induced Protein Misfolding as a Dominant Constraint on Coding-Sequence Evolution. *Cell*, 134(2): 341–352. *Cited at pages 68, 122, 129, 199*

Drummond, D. A., Bloom, J. D., Adami, C., Wilke, C. O., and Arnold, F. H. 2005. Why highly expressed proteins evolve slowly. *Proceedings of the National Academy of Sciences of the United States of America*, 102(40): 14338–14343. *Cited at pages 68, 122, 129, 199*

Dunn, P. M. 2003. Gregor Mendel, OSA (1822–1884), founder of scientific genetics. *Archives of Disease in Childhood-Fetal and Neonatal Edition*, 88(6): F537–F539. *Cited at page 4*

Duret, L. and Galtier, N. 2009. Biased Gene Conversion and the Evolution of Mammalian Genomic Landscapes. *Annual Review of Genomics and Human Genetics*, 10(1): 285–311. *Cited at page 12*

Duret, L. and Mouchiroud, D. 1999. Expression pattern and, surprisingly, gene length shape codon usage in Caenorhabditis, Drosophila, and Arabidopsis. *Proceedings of the National Academy of Sciences of the United States of America*, 96(8): 4482–4487. *Cited at page 111*

Duret, L. and Mouchiroud, D. 2000. Determinants of Substitution Rates in Mammalian Genes: Expression Pattern Affects Selection Intensity but Not Mutation Rate. *Molecular Biology and Evolution*, 17(1): 68–070. *Cited at pages 67, 122*

Dutheil, J. 2008. Detecting site-specific biochemical constraints through substitution mapping. *Journal of Molecular Evolution*, 67(3): 257–265. *Cited at page 41*

Dutheil, J., Gaillard, S., Bazin, E., Glémin, S., Ranwez, V., Galtier, N., and Belkhir, K. 2006. Bio++: A set of C++ libraries for sequence analysis, phylogenetics, molecular evolution and population genetics. *BMC Bioinformatics*, 7(1): 1–6. *Cited at page 56*

Dutheil, J. Y., Galtier, N., Romiguier, J., Douzery, E. J., Ranwez, V., and Boussau, B. 2012. Efficient Selection of Branch-Specific Models of Sequence Evolution. *Molecular Biology and Evolution*, 29(7): 1861–1874. *Cited at pages 42, 100*

Echave, J. and Wilke, C. O. 2017. Biophysical Models of Protein Evolution: Understanding the Patterns of Evolutionary Sequence Divergence. *Annual Review of Biophysics*, 46(1): 85–103. *Cited at pages 63, 68*

Echave, J., Spielman, S. J., and Wilke, C. O. 2016. Causes of evolutionary rate variation among protein sites. *Nature Reviews Genetics*, 17(2): 109–121. *Cited at pages 68, 99, 100*

Ellegren, H., Smith, N. G., and Webster, M. T. 2003. Mutation rate variation in the mammalian genome. *Cited at page 109*

Elyashiv, E., Bullaughey, K., Sattath, S., Rinott, Y., Przeworski, M., and Sella, G. 2010. Shifts in the intensity of purifying selection: An analysis of genome-wide polymorphism data from two closely related yeast species. *Genome Research*, 20(11): 1558–1573. *Cited at page 10*

Eme, D., Malard, F., Konecny-Dupré, L., Lefébure, T., and Douady, C. J. 2013. Bayesian phylogeographic inferences reveal contrasting colonization dynamics among European groundwater isopods. *Molecular Ecology*, 22(22): 5685–5699. *Cited at page 107*

Enard, D., Messer, P. W., and Petrov, D. A. 2014. Genome-wide signals of positive selection in human evolution. *Genome Research*, 24(6): 885–895. *Cited at page 10*

Enard, D., Cai, L., Gwennap, C., and Petrov, D. A. 2016. Viruses are a dominant driver of protein adaptation in mammals. *eLife*, 5: e12469. *Cited at pages 10, 42*

Eyre-Walker, A. 2002. Changing Effective Population Size and the McDonald-Kreitman Test. *Genetics*, 162(4): 2017–2024. *Cited at page 145*

Eyre-Walker, A. and Eyre-Walker, Y. C. 2014. How much of the variation in the mutation rate along the human genome can be explained? *G3: Genes, Genomes, Genetics*, 4(9): 1667–1670. *Cited at page 109*

Eyre-walker, A. and Keightley, P. D. 2007. The distribution of fitness effects of new mutations. *Nature*, 8(August). *Cited at pages 108, 213*

Eyre-Walker, A. and Keightley, P. D. 2009. Estimating the Rate of Adaptive Molecular Evolution in the Presence of Slightly Deleterious Mutations and Population Size Change. *Molecular Biology and Evolution*, 26(9): 2097–2108. *Cited at pages 10, 22, 145*

Eyre-Walker, A., Woolfit, M., and Phelps, T. 2006. The distribution of fitness effects of new deleterious amino acid mutations in humans. *Genetics*, 173(2): 891–900. *Cited at page 22*

Felsenstein, J. 1981. Evolutionary trees from DNA sequences: A maximum likelihood approach. *Journal of Molecular Evolution*, 17(6): 368–376. *Cited at pages 56, 99*

Felsenstein, J. 1985. Phylogenies and the Comparative Method. *The American Naturalist*, 125(1): 1–15. *Cited at pages 42, 99*

Felsenstein, J. and Churchill, G. A. 1996. A Hidden Markov Model approach to variation among sites in rate of evolution. *Molecular Biology and Evolution*, 13(1): 93–104. *Cited at page 71*

Figuet, E., Ballenghien, M., Romiguier, J., and Galtier, N. 2014. Biased gene conversion and GC-content evolution in the coding sequences of reptiles and vertebrates. *Genome Biology and Evolution*, 7(1): 240–250. *Cited at pages 12, 81*

Figuet, E., Nabholz, B., Bonneau, M., Mas Carrio, E., Nadachowska-Brzyska, K., Ellegren, H., and Galtier, N. 2016. Life History Traits, Protein Evolution, and the Nearly Neutral Theory in Amniotes. *Molecular Biology and Evolution*, 33(6): 1517–1527. *Cited at pages 44, 100, 122*

Figuet, E., Ballenghien, M., Lartillot, N., and Galtier, N. 2017. Reconstruction of body mass evolution in the Cetartiodactyla and mammals using phylogenomic data. *bioRxiv*, pages 139147, ver. 3 peer–reviewed and recommended by PC. *Cited at pages 44, 100, 106, 109, 122*

Fisher, R. A. 1919. Xv. the correlation between relatives on the supposition of mendelian inheritance. *Earth and Environmental Science Transactions of The Royal Society of Edinburgh*, 52(2): 399–433. *Cited at page 4*

Fisher, R. A. 1930. *The Genetical Theory of Natural Selection*. The Clarendon Press. *Cited at page 4*

Franklin, R. E. and Gosling, R. G. 1953. Molecular configuration in sodium thymonucleate. *Nature*, 171(4356): 740. *Cited at page 5*

Fürnkranz, J., Scheffer, T., Spiliopoulou, M., Kocsis, L., and Szepesvári, C. 2006. Monte-carlo UCT Search. *Proceedings of the Eleventh conference on Uncertainty in artificial intelligence*, 4212(June): 282–293–293. *Cited at page 30*

Galtier, N. 2016. Adaptive protein evolution in animals and the effective population size hypothesis. *PLoS Genetics*, pages 1–23. *Cited at pages 10, 100, 101, 108, 109, 145, 213*

Galtier, N. and Duret, L. 2007. Adaptation or biased gene conversion? Extending the null hypothesis of molecular evolution. *Trends in Genetics*, 23(6): 273–277. *Cited at page 12*

Galtier, N. and Rousselle, M. 2020. How much does Ne vary among species? *bioRxiv*, pages 861849, ver. 3 peer–reviewed and recommended by PC. *Cited at page 109*

Galtier, N., Duret, L., Glé min, S., and Ranwez, V. 2009. GC-biased gene conversion promotes the fixation of deleterious amino acid changes in primates. *Trends in Genetics*. *Cited at pages 12, 81*

Gao, Z., Wyman, M. J., Sella, G., and Przeworski, M. 2016. Interpreting the Dependence of Mutation Rates on Age and Time. *PLOS Biology*, 14(1): e1002355. *Cited at pages 11, 106*

Gaut, B. S., Muse, S. V., Clark, W. D., and Clegg, M. T. 1992. Relative rates of nucleotide substitution at the rbcl locus of monocotyledonous plants. *Journal of Molecular Evolution*, 35(4): 292–303. *Cited at page 11*

Geman, S. and Geman, D. 1984. Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-6(6): 721–741. *Cited at page 60*

Geraldes, A., Basset, P., Gibson, B., Smith, K. L., Harr, B., YU, H., Bulatova, N., Ziv, Y., and Nachman, M. W. 2008. Inferring the history of speciation in house mice from autosomal, X-linked, Y-linked and mitochondrial genes. *Molecular Ecology*, 17(24): 5349–5363. *Cited at page 109*

Gillespie, D. T. 1977. Exact stochastic simulation of coupled chemical reactions. *The Journal of Physical Chemistry*, 81(25): 2340–2361. *Cited at pages 93, 118, 136, 156, 160, 161*

Goldman, N. and Yang, Z. 1994. A codon-based model of nucleotide substitution for protein-coding DNA sequences. *Molecular biology and evolution*, 11(5): 725–736. *Cited at pages 10, 37, 40, 80, 81, 99*

Goldstein, R. A. 2011. The evolution and evolutionary consequences of marginal thermostability in proteins. *Proteins: Structure, Function and Bioinformatics*, 79(5): 1396–1407. *Cited at pages 65, 100, 104, 124, 125, 131*

Goldstein, R. A. 2013. Population Size Dependence of Fitness Effect Distribution and Substitution Rate Probed by Biophysical Model of Protein Thermostability. *Genome Biology and Evolution*, 5(9): 1584–1593. *Cited at pages 10, 69, 104, 110, 121, 127, 131, 133, 197, 205*

Goldstein, R. A. and Pollock, D. D. 2016. The tangled bank of amino acids. *Protein Science*, 25(7): 1354–1362. *Cited at pages 51, 71, 100, 118, 135*

Goldstein, R. A. and Pollock, D. D. 2017. Sequence entropy of folding and the absolute rate of amino acid substitutions. *Nature Ecology & Evolution*, 1(12): 1923–1930. *Cited at pages 51, 83, 100, 119, 137, 142, 160, 204*

Gong, L. I. and Bloom, J. D. 2014. Epistatically Interacting Substitutions Are Enriched during Adaptive Protein Evolution. *PLoS Genetics*, 10(5). *Cited at page 51*

Goodenbour, J. M. and Pan, T. 2006. Diversity of tRNA genes in eukaryotes. *Nucleic Acids Research*, 34(21): 6137–6146. *Cited at page 34*

Gossmann, T. I., Woolfit, M., and Eyre-Walker, A. 2011. Quantifying the variation in the effective population size within a genome. *Genetics*, 189(4): 1389–1402. *Cited at page 109*

Gould, N. E.-S. J. and Eldredge, N. 1972. Punctuated equilibria: an alternative to phyletic gradualism. *Essential readings in evolutionary biology*, pages 82–115. *Cited at page 5*

Gout, J.-F., Kahn, D., and Duret, L. 2010. The Relationship among Gene Expression, the Evolution of Gene Dosage, and the Rate of Protein Evolution. *PLoS Genetics*, 6(5): e1000944. *Cited at page 130*

Grandaubert, J., Dutheil, J. Y., and Stukenbrock, E. H. 2019. The genomic determinants of adaptive evolution in a fungal pathogen. *Evolution Letters*, 3(3): 299–312. *Cited at page 10*

Grantham, R. 1974. Amino acid difference formula to help explain protein evolution. *Science*, 185(4154): 862–864. *Cited at pages 41, 131, 137*

Gromiha, M. M., Anoosha, P., and Huang, L. T. 2016. Applications of protein thermodynamic database for understanding protein mutant stability and designing stable mutants. In *Methods in Molecular Biology*, volume 1415, pages 71–89. Humana Press Inc. *Cited at page 65*

Guéguen, L. and Duret, L. 2018. Unbiased estimate of synonymous and nonsynonymous substitution rates with nonstationary base composition. *Molecular Biology and Evolution*, 35(3): 734–742. *Cited at page 61*

Guéguen, L., Gaillard, S., Boussau, B., Gouy, M., Groussin, M., Rochette, N. C., Bigot, T., Fournier, D., Pouyet, F., Cahais, V., Bernard, A., Scornavacca, C., Nabholz, B., Haudry, A., Dachary, L., Galtier, N., Belkhir, K., and Dutheil, J. Y. 2013. Bio++: Efficient extensible libraries and tools for computational molecular evolution. *Molecular Biology and Evolution*, 30(8): 1745–1750. *Cited at page 56*

Haldane, J. B. S. 1932. *The Causes of Evolution.* Longsman, Green and Co. *Cited at page 4*

Halligan, D. L., Oliver, F., Eyre-Walker, A., Harr, B., and Keightley, P. D. 2010. Evidence for Pervasive Adaptive Protein Evolution in Wild Mice. *PLoS Genetics*, 6(1): e1000825. *Cited at page 10*

Halpern, A. L. and Bruno, W. J. 1998. Evolutionary distances for protein-coding sequences: modeling site-specific residue frequencies. *Molecular biology and evolution*, 15(7): 910–917. *Cited at pages 23, 45, 46, 48, 80, 100, 102, 121, 132*

Harris, H. 1966. C. Genetics of Man Enzyme polymorphisms in man. *Proceedings of the Royal Society of London. Series B. Biological Sciences*, 164(995): 298–310. *Cited at page 6*

Hartl, D. L., Dykhuizen, D. E., and Dean, A. M. 1985. Limits of adaptation: the evolution of selective neutrality. *Genetics*, 111(3): 655 LP – 674. *Cited at page 9*

Hastings, W. K. 1970. Monte Carlo Sampling Methods Using Markov Chains and Their Applications. *Biometrika*, 57(1): 97–109. *Cited at page 29*

Hawkins, J. A., Kaczmarek, M. E., Müller, M. A., Drosten, C., Press, W. H., and Sawyer, S. L. 2019. A metaanalysis of bat phylogenetics and positive selection based on genomes and transcriptomes from 18 species. *Proceedings of the National Academy of Sciences of the United States of America*, 166(23): 11351–11360. *Cited at page 144*

He, Z., Chen, Q., Yang, H., Chen, Q., Shi, S., and Wu, C.-I. 2020. Two decades of suspect evidence for adaptive DNA-sequence evolution - Failure in consistent detection of positive selection. *bioRxiv*, pages 1–18. *Cited at page 146*

Hilton, S. K., Doud, M. B., and Bloom, J. D. 2017. Phydms: Software for phylogenetic analyses informed by deep mutational scanning. *PeerJ*, 2017(7): e3657. *Cited at page 70*

Horvilleur, B. and Lartillot, N. 2014. Monte Carlo algorithms for Brownian phylogenetic models. *Bioinformatics*, 30(21): 3020–3028. *Cited at page 116*

Hubby, J. L. and Lewontin, R. C. 1966. A molecular approach to the study of genic heterozygosity in natural populations. I. The number of alleles at different loci in Drosophila pseudoobscura. *Genetics*, 54(2): 577. *Cited at page 6*

Huelsenbeck, J. P. and Dyer, K. A. 2004. Bayesian estimation of positively selected sites. *Journal of Molecular Evolution*, 58(6): 661–672. *Cited at page 41*

Huelsenbeck, J. P. and Rannala, B. 2003. Detecting correlation between characters in a comparative analysis with uncertain phylogeny. *Evolution*, 57(6): 1237–1247. *Cited at page 42*

Huelsenbeck, J. P. and Ronquist, F. 2001. MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics*, 17(8): 754–755. *Cited at pages 42, 56*

Huelsenbeck, J. P., Rannala, B., and Masly, J. P. 2000. Accommodating phylogenetic uncertainty in evolutionary studies. *Science*, 288(5475): 2349–2350. *Cited at page 57*

Huelsenbeck, J. P., Jain, S., Frost, S. W. D., and Kosakovsky Pond, S. L. 2006. A Dirichlet process model for detecting positive selection in protein-coding DNA sequences. *Proceedings of the National Academy of Sciences*, 103(16): 6263–6268. *Cited at page 41*

Hughes, A. L. 2005. Evidence for abundant slightly deleterious polymorphisms in bacterial populations. *Genetics*, 169(2): 533–538. *Cited at page 9*

Huxley, J. 1942. *Evolution. The Modern Synthesis.* George Allen & Unwin. *Cited at page 5*

Irvahn, J. and Minin, V. N. 2014. Phylogenetic Stochastic Mapping Without Matrix Exponentiation. *Journal of Computational Biology*, 21(9): 676–690. *Cited at pages 61, 117*

Jacquier, H., Birgy, A., Le Nagard, H., Mechulam, Y., Schmitt, E., Glodt, J., Bercot, B., Petit, E., Poulain, J., Barnaud, G., Gros, P. A., and Tenaillon, O. 2013. Capturing the mutational landscape of the beta-lactamase TEM-1. *Proceedings of the National Academy of Sciences of the United States of America*, 110(32): 13067–13072. *Cited at page 65*

James, J., Castellano, D., and Eyre-Walker, A. 2017. DNA sequence diversity and the efficiency of natural selection in animal mitochondrial DNA. *Heredity*, 118(1): 88–95. *Cited at page 10*

Janin, J. 1995. Protein-protein recognition. *Cited at page 135*

Jiang, Q., Teufel, A. I., Jackson, E. L., and Wilke, C. O. 2018. Beyond thermodynamic constraints: Evolutionary sampling generates realistic protein sequence variation. *Genetics*, 208(4): 1387–1395. *Cited at page 135*

Jimenez, M. J., Arenas, M., and Bastolla, U. 2018. Substitution rates predicted by stability-constrained models of protein evolution are not consistent with empirical data. *Molecular Biology and Evolution*, 35(3): 743–755. *Cited at page 135*

Jones, C. T., Youssef, N., Susko, E., and Bielawski, J. P. 2016. Shifting Balance on a Static Mutation–Selection Landscape: A Novel Scenario of Positive Selection. *Molecular Biology and Evolution*, 34(2): msw237. *Cited at pages 48, 84, 100, 123, 132, 133, 137*

Jühling, F., Mörl, M., Hartmann, R. K., Sprinzl, M., Stadler, P. F., and Pütz, J. 2008. tRNAdb 2009: compilation of tRNA sequences and tRNA genes. *Nucleic Acids Research*, 37(suppl_1): D159–D162. *Cited at page 35*

Kimura, M. 1962. On the probability of fixation of mutant genes in a population. *Genetics*, 47(6): 713–719. *Cited at page 18*

Kimura, M. 1968. Evolutionary rate at the molecular level. *Nature*, 217(5129): 624–626. *Cited at pages 6, 37*

Kimura, M. 1979. Model of effectively neutral mutations in which selective constraint is incorporated. *Proceedings of the National Academy of Sciences of the United States of America*, 76(7): 3440–3444. *Cited at pages 99, 121*

Kimura, M. 1983. *The Neutral Theory of Molecular Evolution*. Cambridge University Press. *Cited at pages 9, 10, 80*

Kimura, M. 1991. The neutral theory of molecular evolution: A review of recent evidence. *The Japanese Journal of Genetics*, 66(4): 367–386. *Cited at page 6*

Kimura, M. and Ohta, T. 1969. The average number of generations until fixation of a mutant gene in a finite population. *Genetics*, 61(3): 763. *Cited at page 23*

Kimura, M., Clarke, B. C., Robertson, A., and Jeffreys, A. J. 1986. Dna and the neutral theory. *Philosophical Transactions of the Royal Society of London. B, Biological Sciences*, 312(1154): 343–354. *Cited at page 6*

King, J. L. and Jukes, T. H. 1969. Non-darwinian evolution. *Science*, 164(3881): 788–798. *Cited at pages 6, 7, 9, 37*

Kishino, H., Thorne, J. L., and Bruno, W. J. 2001. Performance of a Divergence Time Estimation Method under a Probabilistic Model of Rate Evolution. *Molecular Biology and Evolution*, 18(3): 352–361. *Cited at page 11*

Kleinman, C. L., Rodrigue, N., Lartillot, N., and Philippe, H. 2010. Statistical Potentials for Improved Structurally Constrained Evolutionary Models. *Molecular Biology and Evolution*, 27(7): 1546–1560. *Cited at page 71*

Kleppe, K., Ohtsuka, E., Kleppe, R., Molineux, I., and Khorana, H. G. 1971. Studies on polynucleotides. XCVI. Repair replication of short synthetic DNA's as catalyzed by DNA polymerases. *Journal of Molecular Biology*, 56(2): 341–361. *Cited at page 5*

Kocsis, L. and Szepesvári, C. 2006. Bandit based Monte-Carlo planning. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 4212 LNAI, pages 282–293. Springer Verlag. *Cited at page 30*

Kosakovsky Pond, S. L. and Muse, S. V. 2005a. HyPhy: Hypothesis Testing Using Phylogenies. In *Statistical Methods in Molecular Evolution*, pages 125–181. Springer-Verlag. *Cited at pages 56, 97*

Kosakovsky Pond, S. L. and Muse, S. V. 2005b. Site-to-site variation of synonymous substitution rates. *Molecular Biology and Evolution*, 22(12): 2375–2385. *Cited at pages 40, 42, 81*

Kosakovsky Pond, S. L., Murrell, B., Fourment, M., Frost, S. D. W., Delport, W., and Scheffler, K. 2011. A random effects branch-site model for detecting episodic diversifying selection. *Molecular biology and evolution*, 28(11): 3033–3043. *Cited at page 44*

Kosakovsky Pond, S. L., Poon, A. F., Velazquez, R., Weaver, S., Hepler, N. L., Murrell, B., Shank, S. D., Magalis, B. R., Bouvier, D., Nekrutenko, A., Wisotsky, S., Spielman, S. J., Frost, S. D., and Muse, S. V. 2020. HyPhy 2.5 - A Customizable Platform for Evolutionary Hypothesis Testing Using Phylogenies. *Molecular Biology and Evolution*, 37(1): 295–299. *Cited at pages 56, 92*

Kosiol, C. and Anisimova, M. 2019. Selection acting on genomes. In *Methods in Molecular Biology*, volume 1910, pages 373–397. Humana Press Inc. *Cited at page 80*

Kosiol, C., Holmes, I., and Goldman, N. 2007. An Empirical Codon Model for Protein Sequence Evolution. *Molecular Biology and Evolution*, 24(7): 1464–1479. *Cited at page 80*

Kosiol, C., Vinař, T., da Fonseca, R. R., Hubisz, M. J., Bustamante, C. D., Nielsen, R., and Siepel, A. 2008. Patterns of Positive Selection in Six Mammalian Genomes. *PLOS Genetics*, 4(8): e1000144. *Cited at pages 42, 99*

Kumar, M. D., Bava, K. A., Gromiha, M. M., Prabakaran, P., Kitajima, K., Uedaira, H., and Sarai, A. 2006. ProTherm and ProNIT: thermodynamic databases for proteins and protein-nucleic acid interactions. *Nucleic acids research*, 34(Database issue): D204–D206. *Cited at page 65*

Kumar, S., Stecher, G., Suleski, M., and Hedges, S. B. 2017. TimeTree: A Resource for Timelines, Timetrees, and Divergence Times. *Molecular Biology and Evolution*, 34(7): 1812–1819. *Cited at page 106*

Lande, R. 1976. Natural Selection and Random Genetic Drift in Phenotypic Evolution. *Evolution*, 30(2): 314. *Cited at page 6*

Lande, R. 1980. Sexual dimorphism, sexual selection, and adaptation in polygenic characters. *Evolution*, pages 292–305. *Cited at page 6*

Lande, R. and Arnold, S. J. 1983. The Measurement of Selection on Correlated Characters. *Evolution*, 37(6): 1210. *Cited at page 6*

Lanfear, R., Ho, S. Y., Love, D., and Bromham, L. 2010a. Mutation rate is linked to diversification in birds. *Proceedings of the National Academy of Sciences of the United States of America*, 107(47): 20423–20428. *Cited at pages 44, 100, 106, 122*

Lanfear, R., Welch, J. J., and Bromham, L. 2010b. Watching the clock: Studying variation in rates of molecular evolution between species. *Trends in Ecology and Evolution*, 25(9): 495–503. *Cited at page 11*

Lanfear, R., Kokko, H., and Eyre-Walker, A. 2014. Population size and the rate of evolution. *Cited at pages 44, 100, 122, 134*

Larget, B. and Simon, D. L. 1999. Markov Chasin Monte Carlo Algorithms for the Bayesian Analysis of Phylogenetic Trees. *Molecular Biology and Evolution*, 16(6): 750–759. *Cited at page 56*

Lartillot, N. 2006. Conjugate Gibbs sampling for Bayesian phylogenetic models. *Cited at page 61*

Lartillot, N. 2020. The Bayesian Approach to Molecular Phylogeny. In C. Scornavacca, F. Delsuc, and N. Galtier, editors, *Phylogenetics in the Genomic Era*, pages 1.4:1—-1.4:17. A book completely handled by researchers. No publisher has been paid. *Cited at page 57*

Lartillot, N. and Delsuc, F. 2012. Joint reconstruction of divergence times and life-history evolution in placental mammals using a phylogenetic covariance model. *Evolution*, 66(6): 1773–1787. *Cited at pages 11, 100, 106, 109, 122*

Lartillot, N. and Poujol, R. 2011. A Phylogenetic Model for Investigating Correlated Evolution of Substitution Rates and Continuous Phenotypic Characters. *Molecular Biology and Evolution*, 28(1): 729–744. *Cited at pages 11, 42, 44, 80, 99, 100, 102, 116, 122, 173, 188*

Latrille, T., Duret, L., and Lartillot, N. 2017. The Red Queen model of recombination hot-spot evolution: a theoretical investigation. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, 372(1736): 20160463. *Cited at page 10*

Lepage, T., Bryant, D., Philippe, H., and Lartillot, N. 2007. A general comparison of relaxed molecular clock models. *Molecular Biology and Evolution*, 24(12): 2669–2680. *Cited at page 11*

Levy, S. E. and Myers, R. M. 2016. Advancements in Next-Generation Sequencing. *Cited at page 5*

Lewontin, R. C. and Hubby, J. L. 1966. A molecular approach to the study of genic heterozygosity in natural populations. II. Amount of variation and degree of heterozygosity in natural populations of Drosophila pseudoobscura. *Genetics*, 54(2): 595. *Cited at page 6*

Li, H. and Durbin, R. 2011. Inference of human population history from individual whole-genome sequences. *Nature*, 475(7357): 493–496. *Cited at page 109*

Li, S., Pearl, D. K., and Doss, H. 2000. Phylogenetic tree construction using markov chain monte carlo. *Journal of the American Statistical Association*, 95(450): 493–508. *Cited at page 56*

Li, W. H., Tanimura, M., and Sharp, P. M. 1987. An evaluation of the molecular clock hypothesis using mammalian DNA sequences. *Journal of Molecular Evolution*, 25(4): 330–342. *Cited at page 11*

Liberles, D. A. 2007. *Ancestral sequence reconstruction*. Oxford University Press on Demand. *Cited at page 99*

Liberles, D. A., Teichmann, S. A., Bahar, I., Bastolla, U., Bloom, J., Bornberg-Bauer, E., Colwell, L. J., de Koning, A. P. J., Dokholyan, N. V., Echave, J., Elofsson, A., Gerloff, D. L., Goldstein, R. A., Grahnen, J. A., Holder, M. T., Lakner, C., Lartillot, N., Lovell, S. C., Naylor, G., Perica, T., Pollock, D. D., Pupko, T., Regan, L., Roger, A., Rubinstein, N., Shakhnovich, E., Sjölander, K., Sunyaev, S., Teufel, A. I., Thorne, J. L., Thornton, J. W., Weinreich, D. M., and Whelan, S. 2012. The interface of protein structure, protein biophysics, and molecular evolution. *Protein Science*, 21(6): 769–785. *Cited at page 63*

Lin, B. Y., Chan, P. P., and Lowe, T. M. 2019. tRNAviz: explore and visualize tRNA sequence features. *Nucleic Acids Research*, 47(W1): W542–W547. *Cited at page 35*

Lowe, T. M. and Eddy, S. R. 1997. tRNAscan-SE: A Program for Improved Detection of Transfer RNA Genes in Genomic Sequence. *Nucleic Acids Research*, 25(5): 955–964. *Cited at page 35*

Lunzer, M., Golding, G. B., and Dean, A. M. 2010. Pervasive cryptic epistasis in molecular evolution. *PLoS Genetics*, 6(10). *Cited at pages 51, 67*

Lynch, M. and Walsh, B. 2007. *The origins of genome architecture*, volume 98. Sinauer Associates Sunderland, MA. *Cited at page 8*

Manhart, M. and Morozov, A. V. 2015a. Protein folding and binding can emerge as evolutionary spandrels through structural coupling. *Proceedings of the National Academy of Sciences of the United States of America*, 112(6): 1797–1802. *Cited at pages 67, 135*

Manhart, M. and Morozov, A. V. 2015b. Scaling properties of evolutionary paths in a biophysical model of protein adaptation. *Physical Biology*, 12(4). *Cited at page 71*

Marais, G. 2003. Biased gene conversion: Implications for genome and sex evolution. *Cited at page 12*

March, J. G. 1991. Exploration and Exploitation in Organizational Learning. *Organization Science*, 2(1): 71–87. *Cited at page 30*

Mardis, E. R. 2008. The impact of next-generation sequencing technology on genetics. *Cited at page 5*

Mau, B., Newton, M. A., and Larget, B. 1999. Bayesian Phylogenetic Inference via Markov Chain Monte Carlo Methods. *Biometrics*, 55(1): 1–12. *Cited at page 56*

Mayr, E. 1959. Where are we? *Cold Spring Harbor Symposia on Quantitative Biology*, 24: 1–14. *Cited at page 5*

McCandlish, D. M. and Stoltzfus, A. 2014. Modeling Evolution Using the Probability of Fixation: History and Implications. *The Quarterly Review of Biology*, 89(3): 225–252. *Cited at pages 22, 82, 136*

Mccandlish, D. M., Rajon, E., Shah, P., Ding, Y., and Plotkin, J. B. 2013. The role of epistasis in protein evolution. *Nature*, 497(7451): E1—-E2. *Cited at pages 51, 67*

McDonald, J. H. and Kreitman, M. 1991. Adaptative protein evolution at Adh locus in Drosophila. *Nature*. *Cited at pages 10, 145*

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. 1953. Equation of state calculations by fast computing machines. *The Journal of Chemical Physics*, 21(6): 1087–1092. *Cited at page 59*

Miller, C. R., Van Leuven, J. T., Wichman, H. A., and Joyce, P. 2018. Selecting among three basic fitness landscape models: Additive, multiplicative and stickbreaking. *Theoretical Population Biology*, 122: 97–109. *Cited at page 142*

Miyazawa, S. and Jernigan, R. L. 1985. Estimation of effective interresidue contact energies from protein crystal structures: quasi-chemical approximation. *Macromolecules*, 18(3): 534–552. *Cited at pages 65, 131, 161, 204*

Moran, N. A. 1996. Accelerated evolution and Muller's rachet in endosymbiotic bacteria. *Proceedings of the National Academy of Sciences of the United States of America*, 93(7): 2873–2878. *Cited at page 10*

Moutinho, A. F., Trancoso, F. F., and Dutheil, J. Y. 2019a. The Impact of Protein Architecture on Adaptive Evolution. *Molecular Biology and Evolution*, 36(9): 2013–2028. *Cited at page 146*

Moutinho, A. F., Bataillon, T., and Dutheil, J. Y. 2019b. Variation of the adaptive substitution rate between species and within genomes. *Evolutionary Ecology*, 34(3): 315–338. *Cited at page 145*

Mugal, C. F., Wolf, J. B., and Kaj, I. 2014. Why time matters: Codon evolution and the temporal dynamics of dN/dS. *Molecular Biology and Evolution*, 31(1): 212–231. *Cited at page 147*

Murrell, B., Wertheim, J. O., Moola, S., Weighill, T., Scheffler, K., and Kosakovsky Pond, S. L. 2012. Detecting individual sites subject to episodic diversifying selection. *PLoS Genetics*, 8(7): 1002764. *Cited at page 44*

Murrell, B., Moola, S., Mabona, A., Weighill, T., Sheward, D., Kosakovsky Pond, S. L., and Scheffler, K. 2013. FUBAR: A fast, unconstrained bayesian AppRoximation for inferring selection. *Molecular Biology and Evolution*, 30(5): 1196–1205. *Cited at page 44*

Muse, S. V. and Gaut, B. S. 1994. A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates, with application to the chloroplast genome. *Molecular biology and evolution*, 1(5): 715–724. *Cited at pages 10, 37, 39, 40, 48, 80, 86, 87, 99*

Mustonen, V. and Lässig, M. 2005. Evolutionary population genetics of promoters: Predicting binding sites and functional phylogenies. *Proceedings of the National Academy of Sciences of the United States of America*, 102(44): 15936–15941. *Cited at page 27*

Mustonen, V. and Lässig, M. 2009. From fitness landscapes to seascapes: non-equilibrium dynamics of selection and adaptation. *Trends in genetics*, 25(3): 111–119. *Cited at pages 22, 50, 135, 143*

Nabholz, B., Uwimana, N., and Lartillot, N. 2013. Reconstructing the Phylogenetic History of Long-Term Effective Population Size and Life-History Traits Using Patterns of Amino Acid Replacement in Mitochondrial Genomes of Mammals and Birds. *Genome Biology and Evolution*, 5(7): 1273–1290. *Cited at pages 44, 100, 106, 109, 122*

Naumenko, S. A., Kondrashov, A. S., and Bazykin, G. A. 2012. Fitness conferred by replaced amino acids declines with time. *Biology Letters*, 8(5): 825–828. *Cited at pages 46, 67*

Nielsen, R. 2002. Mapping Mutations on Phylogenies. *Systematic Biology*, 51(5): 729–739. *Cited at pages 61, 117*

Nielsen, R. and Yang, Z. 1998. Likelihood Models for Detecting Positively Selected Amino Acid Sites and Applications to the HIV-1 Envelope Gene. *Genetics*, 148(3): 929 LP – 936. *Cited at pages 37, 41, 80*

Nikolaev, S. I., Montoya-Burgos, J. I., Popadin, K., Parand, L., Margulies, E. H., Antonarakis, S. E., Bouffard, G. G., Idol, J. R., Maduro, V. V., Blakesley, R. W., Guan, X., Hansen, N. F., Maskeri, B., McDowell, J. C., Park, M., Thomas, P. J., and Young, A. C. 2007. Life-history traits drive the evolutionary rates of mammalian coding and noncoding genomic elements. *Proceedings of the National Academy of Sciences of the United States of America*, 104(51): 20443–20448. *Cited at page 122*

Noivirt-Brik, O., Horovitz, A., and Unger, R. 2009. Trade-off between Positive and Negative Design of Protein Stability: From Lattice Models to Real Proteins. *PLoS Computational Biology*, 5(12): e1000592. *Cited at page 65*

Nowak, M. A. 2006. *Evolutionary dynamics: exploring the equations of life*. Harvard University Press. *Cited at page 31*

Ohta, T. 1972. Population size and rate of evolution. *Journal of Molecular Evolution*, 1(4): 305–314. *Cited at page 121*

Ohta, T. 1973. Slightly deleterious mutant substitutions in evolution. *Nature*, 246(5428): 96–98. *Cited at pages 7, 49*

Ohta, T. 1992. The nearly neutral theory of molecular evolution. *Annual Review of Ecology and Systematics*, 23(1992): 263–286. *Cited at pages 8, 49, 99, 121*

Ohta, T. 1993. Amino acid substitution at the Adh locus of Drosophila is facilitated by small population size. *Proceedings of the National Academy of Sciences of the United States of America*, 90(10): 4548–4551. *Cited at page 10*

Ohta, T. 1995. Synonymous and nonsynonymous substitutions in mammalian genes and the nearly neutral theory. *Journal of Molecular Evolution*, 40(1): 56–63. *Cited at pages 10, 80*

Ohta, T. and Kimura, M. 1971. On the constancy of the evolutionary rate of cistrons. *Journal of Molecular Evolution*, 1(1): 18–25. *Cited at page 7*

Oliver, P. L., Goodstadt, L., Bayes, J. J., Birtle, Z., Roach, K. C., Phadnis, N., Beatson, S. A., Lunter, G., Malik, H. S., and Ponting, C. P. 2009. Accelerated Evolution of the Prdm9 Speciation Gene across Diverse Metazoan Taxa. *PLoS Genetics*, 5(12): e1000753. *Cited at page 10*

Parto, S. and Lartillot, N. 2017. Detecting consistent patterns of directional adaptation using differential selection codon models. *BMC Evolutionary Biology*, 17(1): 1–17. *Cited at pages 47, 143*

Parto, S. and Lartillot, N. 2018. Molecular adaptation in Rubisco: Discriminating between convergent evolution and positive selection using mechanistic and classical codon models. *PLOS ONE*, 13(2): e0192697. *Cited at pages 47, 143*

Pavlovich, S. S., Lovett, S. P., Koroleva, G., Guito, J. C., Arnold, C. E., Nagle, E. R., Kulcsar, K., Lee, A., Thibaud-Nissen, F., Hume, A. J., Mühlberger, E., Uebelhoer, L. S., Towner, J. S., Rabadan, R., Sanchez-Lockhart, M., Kepler, T. B., and Palacios, G. 2018. The Egyptian Rousette Genome Reveals Unexpected Features of Bat Antiviral Immunity. *Cell*, 173(5): 1098–1110.e18. *Cited at page 144*

Perelman, P., Johnson, W. E., Roos, C., Seuánez, H. N., Horvath, J. E., Moreira, M. A. M., Kessing, B., Pontius, J., Roelke, M., Rumpler, Y., Schneider, M. P. C., Silva, A., O'Brien, S. J., and Pecon-Slattery, J. 2011. A Molecular Phylogeny of Living Primates. *PLoS Genetics*, 7(3): e1001342. *Cited at page 108*

Piganeau, G. and Eyre-Walker, A. 2009. Evidence for Variation in the Effective Population Size of Animal Mitochondrial DNA. *PLoS ONE*, 4(2): e4396. *Cited at page 10*

Plata, G. and Vitkup, D. 2017. Protein Stability and Avoidance of Toxic Misfolding Do Not Explain the Sequence Constraints of Highly Expressed Proteins. *Molecular Biology and Evolution*, 35(3): 700–703. *Cited at page 135*

Platt, A., Weber, C. C., and Liberles, D. A. 2018. Protein evolution depends on multiple distinct population size parameters. *BMC Evolutionary Biology*, 18(1): 17. *Cited at page 112*

Plotkin, J. B. and Kudla, G. 2011. Synonymous but not the same: The causes and consequences of codon bias. *Cited at pages 35, 111*

Pollock, D. D. and Goldstein, R. A. 2014. Strong evidence for protein epistasis, weak evidence against it. *Cited at pages 104, 110*

Pollock, D. D., Thiltgen, G., and Goldstein, R. A. 2012. Amino acid coevolution induces an evolutionary Stokes shift. *Proceedings of the National Academy of Sciences of the United States of America*, 109(21): E1352–E1359. *Cited at pages 51, 67, 104, 124, 131*

Ponting, C. P. 2011. What are the genomic drivers of the rapid evolution of PRDM9? *Cited at page 10*

Popadin, K., Polishchuk, L. V., Mamirova, L., Knorre, D., and Gunbin, K. 2007. Accumulation of slightly deleterious mutations in mitochondrial protein-coding genes of large versus small mammals. *Proceedings of the National Academy of Sciences of the United States of America*, 104(33): 13390–13395. *Cited at pages 10, 42, 44, 100, 109, 122*

Pritchard, J. K. and Cox, N. J. 2002. The allelic architecture of human disease genes: Common disease - Common variant... or not? *Cited at page 11*

Provine, W. B. 2001. *The Origins of Theoretical Population Genetics: With a New Afterword*. University of Chicago Press. *Cited at page 4*

Rak, R., Dahan, O., and Pilpel, Y. 2018. Repertoires of tRNAs: The Couplers of Genomics and Proteomics. *Annual Review of Cell and Developmental Biology*, 34(1): 239–264. *Cited at page 35*

Ramsey, D. C., Scherrer, M. P., Zhou, T., and Wilke, C. O. 2011. The relationship between relative solvent accessibility and evolutionary rate in protein evolution. *Genetics*, 188(2): 479–88. *Cited at page 68*

Ranwez, V., Delsuc, F., Ranwez, S., Belkhir, K., Tilak, M.-K., and Douzery, E. J. 2007. OrthoMaM: A database of orthologous genomic markers for placental mammal phylogenetics. *BMC Evolutionary Biology*, 7(1): 241. *Cited at pages 43, 119*

Ratnakumar, A., Mousset, S., Glemin, S., Berglund, J., Galtier, N., Duret, L., and Webster, M. T. 2010. Detecting positive selection within genomes: the problem of biased gene conversion. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 365(1552): 2571–2580. *Cited at pages 12, 81*

Razban, R. M. 2019. Protein Melting Temperature Cannot Fully Assess Whether Protein Folding Free Energy Underlies the Universal Abundance–Evolutionary Rate Correlation Seen in Proteins. *Molecular Biology and Evolution*, 36(9): 1955–1963. *Cited at page 135*

Rich, A. and RajBhandary, U. L. 1976. Transfer RNA: Molecular Structure, Sequence, and Properties. *Annual Review of Biochemistry*, 45(1): 805–860. *Cited at page 34*

Robinson, D. M., Jones, D. T., Kishino, H., and Thorne, J. L. 2003. Protein evolution with dependence among codons due to tertiary structure. *Molecular biology and evolution*, 20(10): 1692–1704. *Cited at page 71*

Rocha, E. P. C. and Danchin, A. 2004. An Analysis of Determinants of Amino Acids Substitution Rates in Bacterial Proteins. *Molecular Biology and Evolution*, 21(1): 108–116. *Cited at page 122*

Rocklin, G. J., Chidyausiku, T. M., Goreshnik, I., Ford, A., Houliston, S., Lemak, A., Carter, L., Ravichandran, R., Mulligan, V. K., Chevalier, A., Arrowsmith, C. H., and Baker, D. 2017. Global analysis of protein folding using massively parallel design, synthesis, and testing. *Science*, 357(6347): 168–175. *Cited at page 65*

Rodrigue, N. and Lartillot, N. 2014. Site-heterogeneous mutation-selection models within the PhyloBayes-MPI package. *Bioinformatics*, 30(7): 1020–1021. *Cited at pages 46, 100, 101, 102*

Rodrigue, N. and Lartillot, N. 2016. Detecting adaptation in protein-coding genes using a Bayesian site- heterogeneous mutation-selection codon substitution model. *Molecular biology and evolution*, 34(1): 204–214. *Cited at pages 10, 48, 49, 50, 51, 80, 101, 111, 144*

Rodrigue, N. and Philippe, H. 2010. Mechanistic revisions of phenomenological modeling strategies in molecular evolution. *Trends in Genetics*, 26(6): 248–252. *Cited at pages 80, 148*

Rodrigue, N., Lartillot, N., Bryant, D., and Philippe, H. 2005. Site interdependence attributed to tertiary structure in amino acid sequence evolution. *Gene*, 347(2 SPEC. ISS.): 207–217. *Cited at page 71*

Rodrigue, N., Lartillot, N., and Philippe, H. 2008a. Bayesian comparisons of codon substitution models. *Genetics*, 180(3): 1579–1591. *Cited at pages 40, 81, 92*

Rodrigue, N., Philippe, H., and Lartillot, N. 2008b. Uniformization for sampling realizations of Markov processes: applications to Bayesian implementations of codon substitution models. *Bioinformatics*, 24(1): 56–62. *Cited at pages 61, 117*

Rodrigue, N., Kleinman, C. L., and Lartillot, N. 2009. Computational methods for evaluating phylogenetic models of coding sequence evolution with dependence between codons. *Molecular biology and evolution*, 26(7): 1663–1676. *Cited at page 71*

Rodrigue, N., Philippe, H., and Lartillot, N. 2010. Mutation-selection models of coding sequence evolution with site-heterogeneous amino acid fitness profiles. *Proceedings of the National Academy of Sciences of the United States of America*, 107(10): 4629–34. *Cited at pages 45, 46, 80, 100, 108, 121, 132*

Romiguier, J., Figuet, E., Galtier, N., Douzery, E. J. P., Boussau, B., Dutheil, J. Y., and Ranwez, V. 2012. Fast and Robust Characterization of Time-Heterogeneous Sequence Evolutionary Processes Using Substitution Mapping. *PLoS ONE*, 7(3): e33852. *Cited at page 61*

Romiguier, J., Gayral, P., Ballenghien, M., Bernard, A., Cahais, V., Chenuil, A., Chiari, Y., Dernat, R., Duret, L., Faivre, N., Loire, E., Lourenco, J. M., Nabholz, B., Roux, C., Tsagkogeorga, G., Weber, A. A.-T., Weinert, L. A., Belkhir, K., Bierne,

N., Glémin, S., and Galtier, N. 2014. Comparative population genomics in animals uncovers the determinants of genetic diversity. *Nature*, 515(7526): 261–263. *Cited at pages 10, 44, 100, 106, 122*

Ronquist, F., Teslenko, M., Van Der Mark, P., Ayres, D. L., Darling, A., Höhna, S., Larget, B., Liu, L., Suchard, M. A., and Huelsenbeck, J. P. 2012. Mrbayes 3.2: Efficient bayesian phylogenetic inference and model choice across a large model space. *Systematic Biology*, 61(3): 539–542. *Cited at page 42*

Rousselle, M., Mollion, M., Nabholz, B., Bataillon, T., and Galtier, N. 2018. Overestimation of the adaptive substitution rate in fluctuating populations. *Biology Letters*, 14(5): 20180055. *Cited at page 148*

Russell, S. and Norvig, P. 2010. *Artificial Intelligence A Modern Approach Third Edition*. Pearson. *Cited at page 30*

Saclier, N., François, C. M., Konecny-Dupre, L., Lartillot, N., Guéguen, L., Duret, L., Malard, F., Douady, C. J., and Lefébure, T. 2018. Life history traits impact the nuclear rate of substitution but not the mitochondrial rate in isopods. *Molecular Biology and Evolution*, 35(12): 2900–2912. *Cited at page 107*

Salser, W., Bowen, S., Browne, D., El-Adli, F., Fedoroff, N., Fry, K., Heindell, H., Paddock, G., Poon, R., Wallace, B., and Whitcome, P. 1976. Investigation of the organization of mammalian chromosomes at the DNA sequence level. In *Federation Proceedings*, volume 35, pages 23–35. *Cited at page 6*

Sanderson, M. J. 1997. A Nonparametric Approach to Estimating Divergence Times in the Absence of Rate Constancy. *Molecular Biology and Evolution*, 14(12): 1218–1231. *Cited at page 11*

Sanger, F. and Coulson, A. R. 1975. A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. *Journal of Molecular Biology*, 94(3): 441–448. *Cited at page 5*

Sanger, F., Nicklen, S., and Coulson, A. R. 1977. DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences of the United States of America*, 74(12): 5463–5467. *Cited at page 5*

Sawyer, S. A. and Hartl, D. L. 1992. Population genetics of polymorphism and divergence. *Genetics*, 132(4). *Cited at pages 21, 215*

Schrempf, D., Minh, B. Q., De Maio, N., von Haeseler, A., and Kosiol, C. 2016. Reversible polymorphism-aware phylogenetic models and their application to tree inference. *Journal of Theoretical Biology*, 407: 362–370. *Cited at page 147*

Schrempf, D., Minh, B. Q., von Haeseler, A., and Kosiol, C. 2019. Polymorphism-Aware Species Trees with Advanced Mutation Models, Bootstrap, and Rate Heterogeneity. *Molecular Biology and Evolution*, 36(6): 1294–1301. *Cited at page 147*

Scornavacca, C., Belkhir, K., Lopez, J., Dernat, R., Delsuc, F., Douzery, E. J. P., and Ranwez, V. 2019. 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. *Cited at pages 43, 119*

Sella, G. and Barton, N. H. 2019. Thinking About the Evolution of Complex Traits in the Era of Genome-Wide Association Studies. *Annual Review of Genomics and Human Genetics*, 20(1): 461–493. *Cited at page 11*

Sella, G. and Hirsh, A. E. 2005. The application of statistical physics to evolutionary biology. *Proceedings of the National Academy of Sciences of the United States of America*, 102(27): 9541–9546. *Cited at pages 22, 27, 135*

Seo, T. K., Kishino, H., and Thorne, J. L. 2004. Estimating absolute rates of synonymous and nonsynonymous nucleotide substitution in order to characterize natural selection and date species divergences. *Molecular Biology and Evolution*, 21(7): 1201–1213. *Cited at pages 42, 100*

Serohijos, A. W. and Shakhnovich, E. I. 2014. Merging molecular mechanism and evolution: Theory and computation at the interface of biophysics and evolutionary population genetics. *Cited at page 63*

Serohijos, A. W., Rimas, Z., and Shakhnovich, E. I. 2012. Protein Biophysics Explains Why Highly Abundant Proteins Evolve Slowly. *Cell Reports*, 2(2): 249–256. *Cited at pages 68, 69, 122, 123, 129, 134, 199, 205*

Serohijos, A. W., Lee, S. Y., and Shakhnovich, E. I. 2013. Highly abundant proteins favor more stable 3D structures in yeast. *Biophysical Journal*, 104(3): L1–L3. *Cited at page 123*

Shah, P., Mccandlish, D. M., and Plotkin, J. B. 2015. Contingency and entrenchment in protein evolution under purifying selection. *Proceedings of the National Academy of Sciences*, 112(5): 3226–3235. *Cited at pages 67, 104, 110*

Sikosek, T. and Chan, H. S. 2014. Biophysics of protein evolution and evolutionary protein biophysics. *Cited at pages 63, 65*

Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., Guez, A., Hubert, T., Baker, L., Lai, M., Bolton, A., Chen, Y., Lillicrap, T., Hui, F., Sifre, L., Van Den Driessche, G., Graepel, T., and Hassabis, D. 2017. Mastering the game of Go without human knowledge. *Nature*, 550(7676): 354–359. *Cited at page 30*

Silver, D., Hubert, T., Schrittwieser, J., Antonoglou, I., Lai, M., Guez, A., Lanctot, M., Sifre, L., Kumaran, D., Graepel, T., Lillicrap, T., Simonyan, K., and Hassabis, D. 2018. A general reinforcement learning algorithm that masters chess, shogi, and Go through self-play. *Science*, 362(6419): 1140–1144. *Cited at page 30*

Simons, Y. B., Bullaughey, K., Hudson, R. R., and Sella, G. 2018. A population genetic interpretation of GWAS findings for human quantitative traits. *PLoS Biology*, 16(3): e2002985. *Cited at page 11*

Simpson, G. G. 1944. *Tempo and mode in evolution*. Columbia University Press, New York. *Cited at page 6*

Simpson, G. G. 1953. *The major features of evolution*. Columbia University Press, New York. *Cited at page 6*

Singer, G. A. and Hickey, D. A. 2000. Nucleotide bias causes a genomewide bias in the amino acid composition of proteins. *Molecular Biology and Evolution*, 17(11): 1581–1588. *Cited at pages 9, 40, 81, 91*

Smith, J. M. 1982. *Evolution and the Theory of Games*. Cambridge university press. *Cited at page 31*

Smith, J. M. and Price, G. R. 1973. The logic of animal conflict. *Nature*, 246(5427): 15–18. *Cited at page 31*

Smith, N. G. and Eyre-Walker, A. 2002. Adaptive protein evolution in Drosophila. *Nature*, 415(6875): 1022–1024. *Cited at page 10*

Song, H., Gao, H., Liu, J., Tian, P., and Nan, Z. 2017. Comprehensive analysis of correlations among codon usage bias, gene expression, and substitution rate in Arachis duranensis and Arachis ipaënsis orthologs. *Scientific Reports*, 7(1): 1–12. *Cited at page 122*

Spielman, S. J. and Wilke, C. O. 2015. The relationship between dN/dS and scaled selection coefficients. *Molecular biology and evolution*, 32(4): 1097–1108. *Cited at pages 28, 40, 48, 49, 81, 84, 101, 121, 137*

Spielman, S. J. and Wilke, C. O. 2016. Extensively Parameterized Mutation-Selection Models Reliably Capture Site-Specific Selective Constraint. *Molecular Biology and Evolution*, 33(11): 2990–3001. *Cited at page 46*

Spielman, S. J., Wan, S., and Wilke, C. O. 2016. A comparison of one-rate and two-rate inference frameworks for site-specific dN/dS estimation. *Genetics*, 204(2): 499–511. *Cited at page 42*

Starr, T. N. and Thornton, J. W. 2016. Epistasis in protein evolution. *Cited at pages 124, 125*

Stebbins, G. L. 1966. *Processes of Organic Evolution.* Englewood Cliffs. *Cited at page 5*

Stiffler, M. A., Hekstra, D. R., and Ranganathan, R. 2015. Evolvability as a Function of Purifying Selection in TEM-1 β-Lactamase. *Cell*, 160(5): 882–892. *Cited at page 70*

Tacutu, R., Craig, T., Budovsky, A., Wuttke, D., Lehmann, G., Taranukha, D., Costa, J., Fraifeld, V. E., and de Magalhães, J. P. 2012. Human Ageing Genomic Resources: Integrated databases and tools for the biology and genetics of ageing. *Nucleic Acids Research*, 41(D1): D1027–D1033. *Cited at page 119*

Tamuri, A. U. and Goldstein, R. A. 2012. Estimating the distribution of selection coefficients from phylogenetic data using sitewise mutation-selection models. *Genetics*, 190(March): 1101–1115. *Cited at pages 45, 46, 80, 100, 108, 121, 132*

Tamuri, A. U., dos Reis, M., Hay, A. J., and Goldstein, R. A. 2009. Identifying Changes in Selective Constraints: Host Shifts in Influenza. *PLoS Computational Biology*, 5(11): e1000564. *Cited at pages 47, 143*

Tamuri, A. U., Goldman, N., and dos Reis., M. 2014. A penalized-likelihood method to estimate the distribution of selection coefficients from phylogenetic data. *Genetics*, 197(May): 257–271. *Cited at pages 46, 100, 101*

Tavaré, S. 1986. Some probabilistic and statistical problems in the analysis of DNA sequences. *Lectures on mathematics in the life sciences*, 17(2): 57–86. *Cited at pages 38, 87, 113*

Taverna, D. M. and Goldstein, R. A. 2002. Why are proteins marginally stable? *Proteins: Structure, Function, and Bioinformatics*, 46(1): 105–109. *Cited at pages 9, 65, 66, 127*

Tenaillon, O. 2014. The Utility of Fisher's Geometric Model in Evolutionary Genetics. *Annual Review of Ecology, Evolution, and Systematics*, 45(1): 179–201. *Cited at pages 110, 159*

Thomas, J. H., Emerson, R. O., and Shendure, J. 2009. Extraordinary Molecular Evolution in the PRDM9 Fertility Gene. *PLoS ONE*, 4(12): e8505. *Cited at page 10*

Thorne, J. L. and Kishino, H. 2002. Divergence Time and Evolutionary Rate Estimation with Multilocus Data. *Systematic Biology*, 51(5): 689–702. *Cited at page 99*

Thorne, J. L., Kishino, H., and Painter, I. S. 1998. Estimating the rate of evolution of the rate of molecular evolution. *Molecular Biology and Evolution*, 15(12): 1647–1657. *Cited at pages 11, 42*

Thorne, J. L., Lartillot, N., Rodrigue, N., and Choi, S. C. 2012. Codon models as a vehicle for reconciling population genetics with inter-specific sequence data. In *Codon evolution: mechanisms and models*, pages 97–110. Oxford University Press. *Cited at page 147*

Tokuriki, N. and Tawfik, D. S. 2009a. Protein Dynamism and Evolvability. *Science*, 324(5924): 203 LP – 207. *Cited at page 63*

Tokuriki, N. and Tawfik, D. S. 2009b. Stability effects of mutations and protein evolvability. *Current Opinion in Structural Biology*, 19(5): 596–604. *Cited at page 63*

Vandewege, M. W., Sotero-Caio, C. G., and Phillips, C. D. 2020. Positive selection and gene expression analyses from salivary glands reveal discrete adaptations within the ecologically diverse bat family Phyllostomidae. *Genome Biology and Evolution*, 12(8): 1419–1428. *Cited at page 144*

Vikhar, P. A. 2017. Evolutionary algorithms: A critical review and its future prospects. In *Proceedings - International Conference on Global Trends in Signal Processing, Information Computing and Communication, ICGTSPICC 2016*, pages 261–265. Institute of Electrical and Electronics Engineers Inc. *Cited at page 30*

Von Neumann, J. and Morgenstern, O. 1947. *Theory of games and economic behavior, 2nd rev. ed.* Princeton University Press, Princeton, NJ, US. *Cited at page 31*

Watson, J. D. and Crick, F. H. 1953. Molecular structure of nucleic acids: A structure for deoxyribose nucleic acid. *Nature*, 171(4356): 737–738. *Cited at page 5*

Welch, J. J., Eyre-Walker, A., and Waxman, D. 2008. Divergence and polymorphism under the nearly neutral theory of molecular evolution. *Journal of Molecular Evolution*, 67(4): 418–426. *Cited at pages 10, 99, 100, 121, 132, 206*

Wilke, C. O. and Drummond, D. A. 2006. Population Genetics of Translational Robustness. *Genetics*, 173(1): 473 LP – 481. *Cited at pages 66, 68, 122, 129, 199*

Wilkins, M. H. F., Stokes, A. R., and Wilson, H. R. 1953. Molecular structure of nucleic acids: Molecular structure of deoxypentose nucleic acids. *Nature*, 171(4356): 738. *Cited at page 5*

Williams, P. D., Pollock, D. D., Blackburne, B. P., and Goldstein, R. A. 2006. Assessing the accuracy of ancestral protein reconstruction methods. *PLoS Computational Biology*, 2(6): 0598–0605. *Cited at pages 104, 124, 131*

Wilson, D. J., Hernandez, R. D., Andolfatto, P., and Przeworski, M. 2011. A Population Genetics-Phylogenetics Approach to Inferring Natural Selection in Coding Sequences. *PLoS Genetics*, 7(12): e1002395. *Cited at pages 112, 147*

Wilson, G., Aruliah, D. A., Brown, C. T., Chue Hong, N. P., Davis, M., Guy, R. T., Haddock, S. H. D., Huff, K. D., Mitchell, I. M., Plumbley, M. D., Waugh, B., White, E. P., and Wilson, P. 2014. Best Practices for Scientific Computing. *PLoS Biology*, 12(1): e1001745. *Cited at page 150*

Woolfit, M. and Bromham, L. 2003. Increased rates of sequence evolution in endosymbiotic bacteria and fungi with small effective population sizes. *Molecular Biology and Evolution*, 20(9): 1545–1555. *Cited at page 10*

Woolfit, M. and Bromham, L. 2005. Population size and molecular evolution on islands. *Proceedings of the Royal Society B: Biological Sciences*, 272(1578): 2277–2282. *Cited at page 10*

Wright, S. 1931. Evolution in Mendelian populations. *Genetics*, 16(2): 97–159. *Cited at page 17*

Wright, S. 1932. *The roles of mutation, inbreeding, crossbreeding, and selection in evolution*, volume 1. Proceedings of the Sixth International Congress on Genetics. *Cited at page 4*

Yang, J. R., Liao, B. Y., Zhuang, S. M., and Zhang, J. 2012. Protein misinteraction avoidance causes highly expressed proteins to evolve slowly. *Proceedings of the National Academy of Sciences of the United States of America*, 109(14): E831–E840. *Cited at pages 67, 135*

Yang, Z. 1997. Paml: A program package for phylogenetic analysis by maximum likelihood. *Bioinformatics*, 13(5): 555–556. *Cited at pages 42, 56*

Yang, Z. 2007. PAML 4: Phylogenetic analysis by maximum likelihood. *Molecular Biology and Evolution*, 24(8): 1586–1591. *Cited at pages 42, 56*

Yang, Z. and Nielsen, R. 1998. Synonymous and nonsynonymous rate variation in nuclear genes of mammals. *Journal of Molecular Evolution*, 46(4): 409–418. *Cited at pages 42, 122*

Yang, Z. and Nielsen, R. 2002. Codon-substitution models for detecting molecular adaptation at individual sites along specific lineages. *Molecular Biology and Evolution*, 19(6): 908–917. *Cited at page 44*

Yang, Z. and Nielsen, R. 2008. Mutation-selection models of codon substitution and their use to estimate selective strengths on codon usage. *Molecular biology and evolution*, 25(3): 568–579. *Cited at pages 47, 111*

Yang, Z. and Rannala, B. 1997. Bayesian phylogenetic inference using DNA sequences: a Markov Chain Monte Carlo Method. *Molecular Biology and Evolution*, 14(7): 717–724. *Cited at page 56*

Yang, Z. and Swanson, W. J. 2002. Codon-substitution models to detect adaptive evolution that account for heterogeneous selective pressures among site classes. *Molecular biology and evolution*, 19(1): 49–57. *Cited at page 10*

Yang, Z., Nielsen, R., and Hasegawa, M. 1998. Models of amino acid substitution and applications to mitochondrial protein evolution. *Molecular Biology and Evolution*, 15(12): 1600–1611. *Cited at page 41*

Yang, Z., Nielsen, R., Goldman, N., and Pedersen, A.-m. K. 2000. Codon-substitution models for heterogeneous selection pressure at amino acid sites. *Genetics*, 155(May): 431–449. *Cited at page 41*

Yang, Z., Wong, W. S., and Nielsen, R. 2005. Bayes empirical Bayes inference of amino acid sites under positive selection. *Molecular Biology and Evolution*, 22(4): 1107–1118. *Cited at page 41*

Yap, V. B., Lindsay, H., Easteal, S., and Huttley, G. 2010. Estimates of the Effect of Natural Selection on Protein-Coding Content. *Molecular Biology and Evolution*, 27(3): 726–734. *Cited at page 40*

Yeh, S.-W., Liu, J.-W., Yu, S.-H., Shih, C.-H., Hwang, J.-K., and Echave, J. 2013. Site-Specific Structural Constraints on Protein Sequence Evolutionary Divergence: Local Packing Density versus Solvent Exposure. *Molecular Biology and Evolution*, 31(1): 135–139. *Cited at page 68*

Zeldovich, K. B., Chen, P., and Shakhnovich, E. I. 2007. Protein stability imposes limits on organism complexity and speed of molecular evolution. *Proceedings of the National Academy of Sciences of the United States of America*, 104(41): 16152–16157. *Cited at pages 133, 134*

Zhang, J. and Nielsen, R. 2005. Evaluation of an improved branch-site likelihood method for detecting positive selection at the molecular level. *Molecular biology and evolution*, 22(12): 2472–2479. *Cited at pages 10, 42, 44, 80, 122*

Zhang, J. and Yang, J. R. 2015. Determinants of the rate of protein sequence evolution. *Cited at pages 67, 68, 80, 99, 122, 134, 203*

Zhang, J., Maslov, S., and Shakhnovich, E. I. 2008. Constraints imposed by non-functional protein–protein interactions on gene expression and proteome size. *Molecular Systems Biology*, 4(1): 210. *Cited at pages 135, 203*

Zhang, X., Perica, T., and Teichmann, S. A. 2013. Evolution of protein structures and interactions from the perspective of residue contact networks. *Cited at pages 67, 135*

Zuckerkandl, E. and Pauling, L. 1965. Molecules as documents of evolutionary history. *Journal of theoretical biology*, 8(2): 357–366. *Cited at pages 6, 11, 99*