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:

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

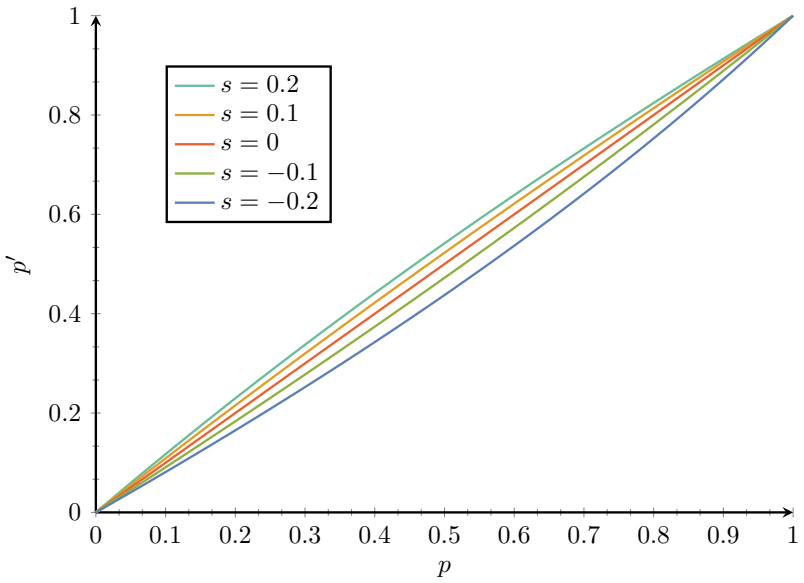

Figure 2.1: *Frequency of derived allele p' after a generation in the vertical axis a function of the frequency in the previous generation p in the horizontal axis, shown for several selection coefficients in coloured solid lines. Positive selection coefficients ($s > 0$) result in increased derived 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:

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

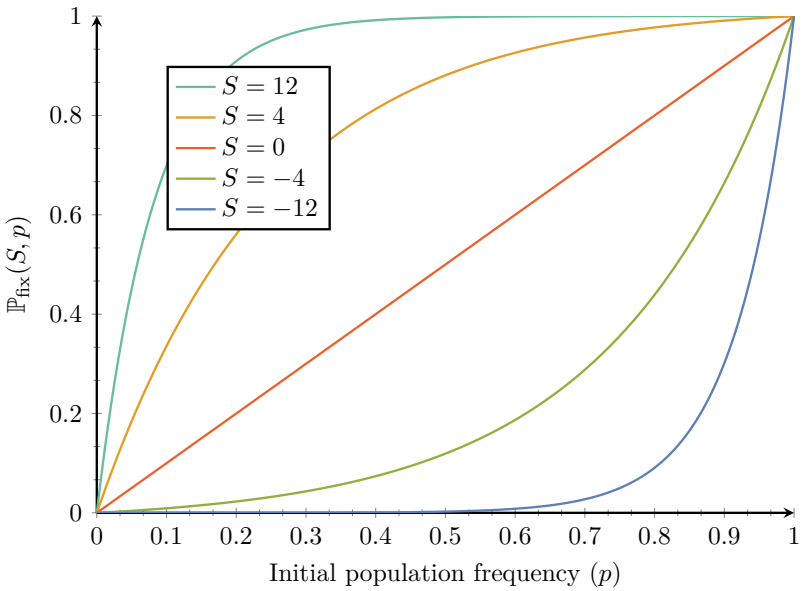

Figure 2.2: *Probability of fixation $\mathbb{P}_{fix}(S, p)$ in the vertical axis as a function of the initial frequency p in the horizontal axis, shown for different scaled effective population size $S = 4N_e s$. In contrast to changes of frequency during a generation, the probability of fixation is 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.

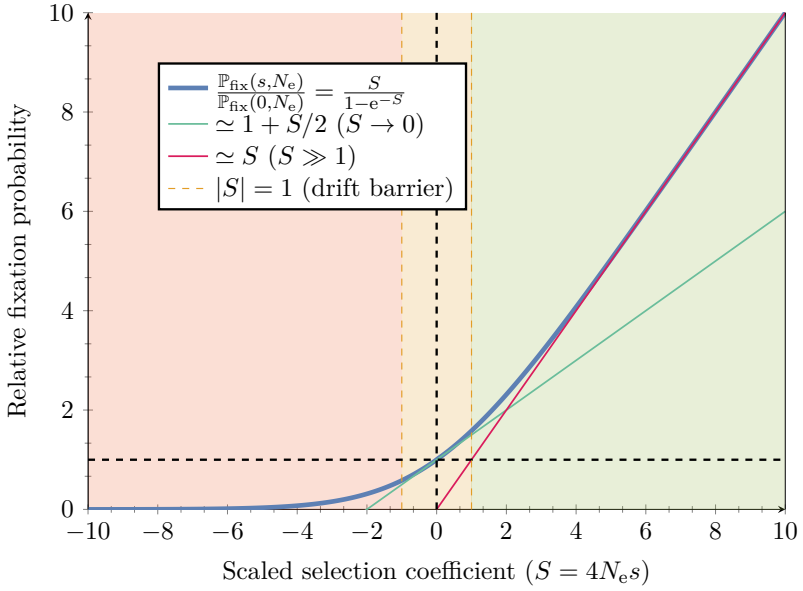

Figure 2.3: *Fixation probability of a selected allele relative to a neutral allele, shown in the vertical axis, as function of the scaled selection coefficient $S = 4N_e s$ in the horizontal axis. For a substantial negative scaled selection coefficient ($s \leq -1/N_e$, red-filled area), the probability 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:

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

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

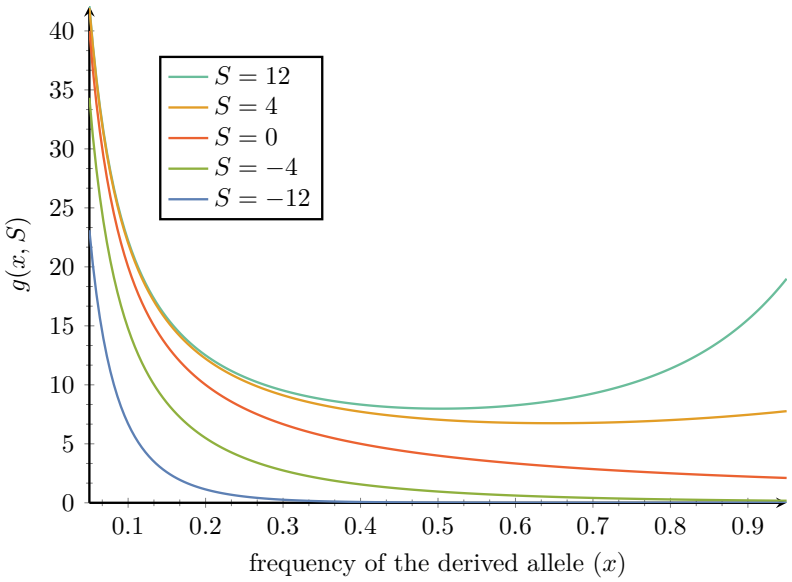

Figure 2.4: *Expected time at a derived frequency $g(x, S)$ in the vertical axis as a function of the frequency x, shown for different scaled selection coefficient. Alleles with a positive selection coefficient can be observed at high frequency, while alleles with negative selection coefficients are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

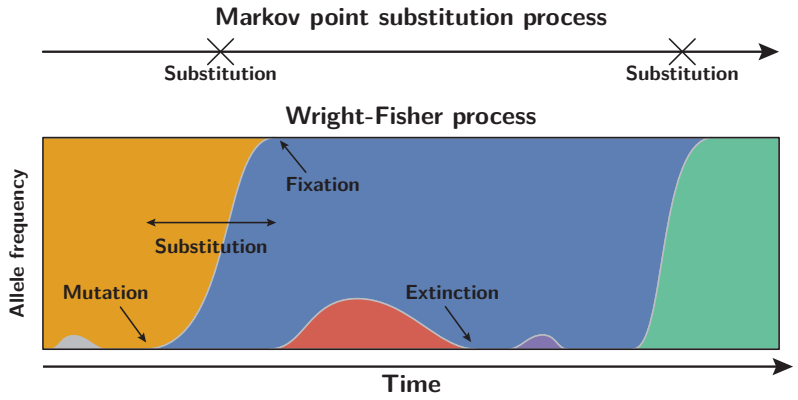

Figure 2.5: *Mutation-selection substitutions models. The trajectory of alleles inside a population is collapsed into a single point substitution process. This approximation is valid under low mutation rates such that a mutation originates uniquely whenever the gene is monomorphic (with a single allele).*

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-

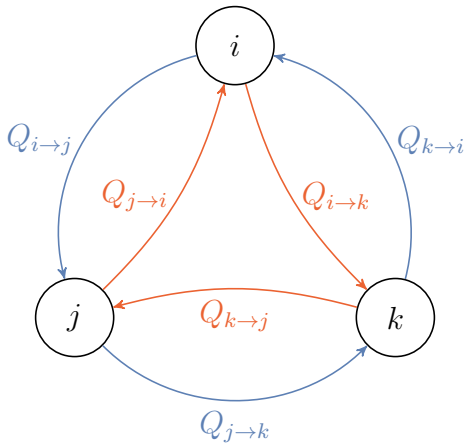

Figure 2.6: *The continuous-time Markov chain is reversible if the process fulfils Kolmogorov's criterion. Namely, the product of the transition rates for a closed loop is equal whether traversed in one sense (blue arrows) or the other (red arrows).*

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.

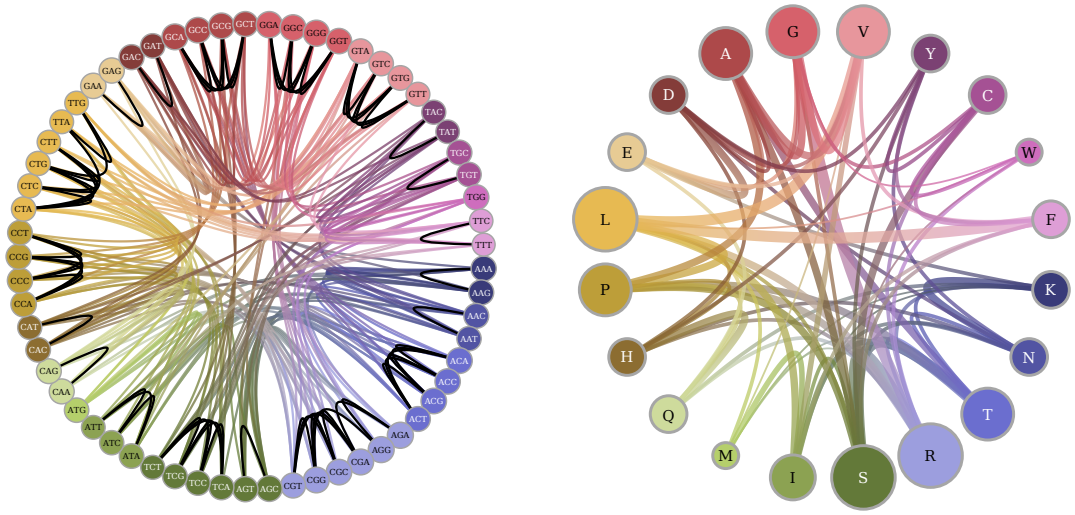

Figure 3.1: *Graphs of possible one nucleotide transitions between codons (left panel) and between amino acids (right panel). Nodes correspond to codons (left panel) and amino acids (right panel), and their colour represents the encoded amino acid. Additionally, for amino acids, the size of nodes represents 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).

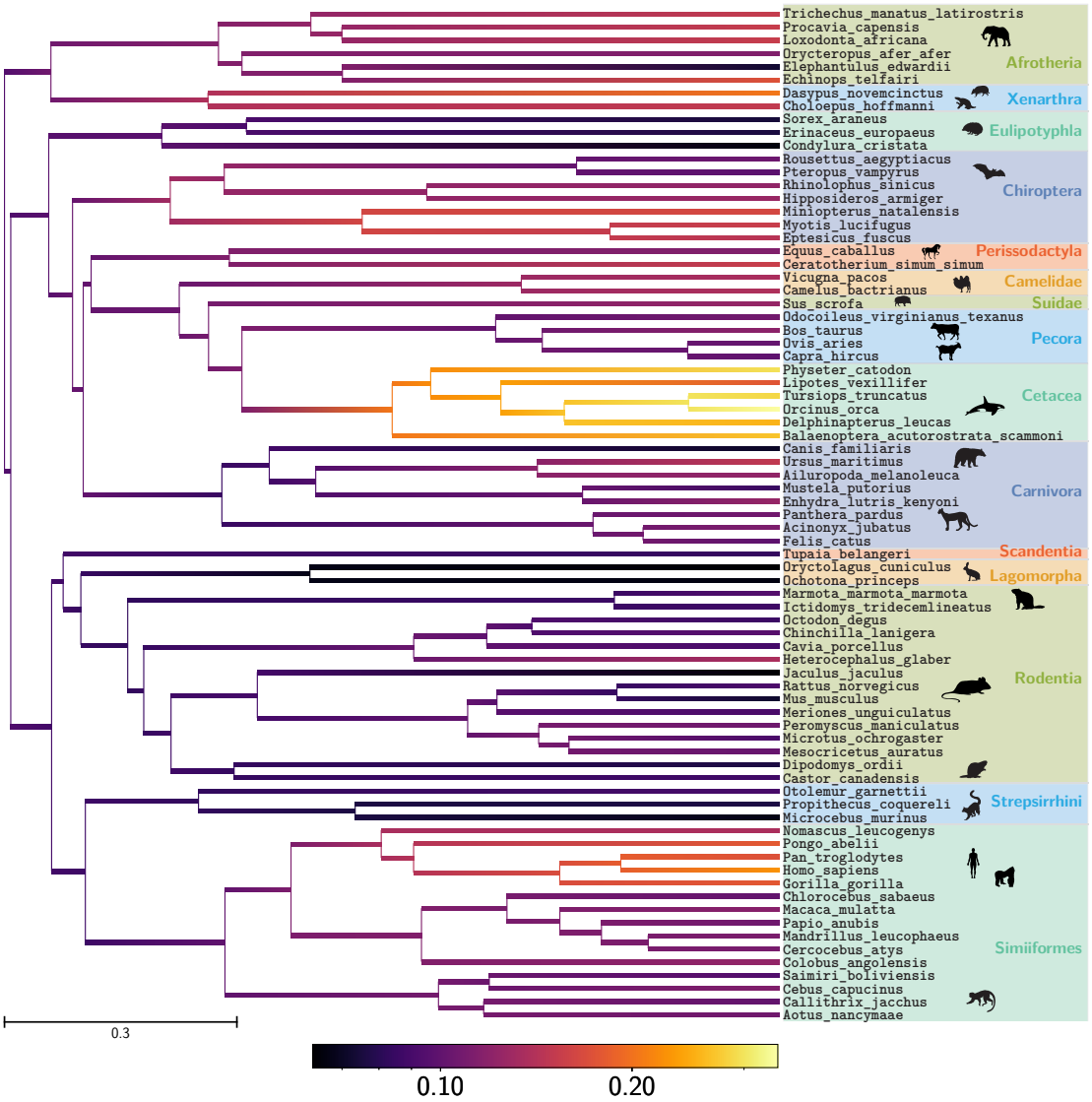

Figure 3.2: *d_N/d_S variations across branches in mammals. The Brownian process (i.e. logarithm of d_N/d_S) starts at the root of the dated tree, runs along branches and splits at each node of the tree into two independent children processes until reaching the extant species. Along each branch, the 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.