Detecting Adaptation from Multi-species Protein-coding DNA Sequence Alignments

Christine Lowe, Nicolas Rodrigue

Chapter 4.5 Detecting Adaptation from Multi-species Protein-coding DNA Sequence Alignments

Christine Lowe
Agriculture and Agri-Food Canada
Biological Informatics Centre of Excellence, Ottawa, ON, Canada
christine.lowe@canada.ca

Nicolas Rodrigue
Carleton University
Department of Biology, Institute of Biochemistry, and School of Mathematics and Statistics, Ottawa, ON, Canada
nicolas.rodrigue@carleton.ca

Abstract

Modern methods to detecting adaptive evolution from interspecific protein-coding gene alignments rely on statistical models of sequence evolution formulated at the level of codons. By performing model comparisons, one measures the evidence for signals of adaptive substitution processes, relative to a null model that disallows any adaptive regime. In this chapter, we present the detailed form of these models of sequence evolution, and how they are applied to real data sets. The classical codon substitution models are based on evaluating the relative nonsynonymous to synonymous substitution rates, and the main focus has traditionally been placed on devising models allowing for increasingly more subtle manifestations of adaptive substitution processes. We also overview a contrasting modeling direction that has emerged in the last decade—although with roots two decades back—in which the emphasis is placed on devising a richer modeling of purifying selection. Using simulations, we expand the characterization of this latter approach, followed by a contrasting of its conclusions on real data with those of classical codon models. Finally, we discuss the numerous model violations that can lead to erroneous inferences on various tests, and potential future directions meriting attention.

How to cite: Christine Lowe and Nicolas Rodrigue (2020). Detecting Adaptation from Multi-species Protein-coding DNA Sequence Alignments. In Scornavacca, C., Delsuc, F., and Galtier, N., editors, *Phylogenetics in the Genomic Era*, chapter No. 4.5, pp. 4.5:1–4.5:18. No commercial publisher | Authors open access book. The book is freely available at https://hal.inria.fr/PGE.

1 Introduction

When alignments of protein-coding DNA sequences from different species became available in the second half of the 20th century, evolutionary biologists quickly sought to contrast the rate of substitutions that do not alter the encoded amino acid sequence (the *synonymous* substitutions) to those that imply an amino acid replacement (the *nonsynonymous* substitutions). Early methods were based on simple counting schemes (Miyata and Yasunaga, 1980; Perler et al., 1980; Gojobori, 1983; Li et al., 1985; Nei and Gojobori, 1986). One of their objectives was to account for the fact that not all codon states have the same potential for synonymous and nonsynonymous substitutions; for instance, a codon encoding tryptophan has no synonymous opportunity, given that it is alone in encoding this amino acid, whereas leucine is encoded by six codons, and therefore has high synonymous opportunity. These

Phylogenetics in the genomic era.
Editors: Celine Scornavacca, Frédéric Delsuc and Nicolas Galtier; chapter No. 4.5; pp. 4.5:1–4.5:18

A book completely handled by researchers.
No publisher has been paid.

early methods also relied on multiple pairwise sequence comparisons, for all or most of the possible pairs from the multiple sequence alignment. By the 1990s, however, a number of statistical models were proposed, working within a full phylogenetic framework (Goldman and Yang, 1994; Muse and Gaut, 1994; Halpern and Bruno, 1998). Rather than utilizing counting schemes on pairwise comparisons, the models were based on the idea of fitting parameters (e.g., by maximum likelihood) representing features of a codon substitution process running over the branches of a phylogenetic tree relating all sequences of an alignment.

In this chapter, we follow two main threads in the development of codon substitution models aimed at detecting adaption in protein-coding genes. The first consists of models based on a parameter denoted ω, which corresponds to the rate ratio of nonsynonymous (dN) and synonymous (dS) substitutions ($\omega = dN/dS$). The traditional interpretation of ω is that instances where model fitting leads to $\omega \sim 1$ correspond to (nearly) neutral evolution, whereas $\omega < 1$ indicates purifying, or negative selection, and $\omega > 1$ corresponds to an adaptive regime, or positive selection. We present the detailed form of such models, as well as the commonly used extensions allowing for variation in ω across codon sites of an alignment. The second modeling thread we present is focused on better capturing the subtleties of purifying selection, in what has come to be known as the *mutation-selection* framework. These approaches attempt to account for the heterogeneity of amino acid fitness profiles across sites, and form a null model against which to test for nonsynonymous rates greater than expected under the mutation-selection balance. We describe the mutation-selection rationale in detail, and present a simulation study to further characterize the approach. We also contrast its results on several thousands of real data sets with those of the classical codon substitution models. Finally, we discuss a number of model violations that can influence inferences under codon substitution models, and outline future research directions that merit greater attention.

2 The classic codon substitution models

Appearing back-to-back in the 1994 September issue of Molecular Biology and Evolution, the papers by Muse and Gaut (1994, "MG") and Goldman and Yang (1994, "GY") took the ideas of likelihood-based phylogenetic analysis in the nucleotide state space, and proposed to expand the state space to in-frame nucleotide triplets: rather than specifying a 4 by 4 matrix of nucleotide substitution rates, a 61 by 61 (assuming a universal genetic code) matrix of codon substitution rates is specified; the lethality of stop-codons is a built-in assumption of the model, in being disallowed in the state space. Another built-in assumption is that of a point-mutation process, where the substitution rate between codons that differ by two or three nucleotides is set to zero. However, this latter assumption is not new to the codon-level context, but inherent to the nucleotide-level context as well, since the probability of two nucleotide sites undergoing a substitution within a given time interval vanishes as the interval approaches zero (see Chapter 1.1 [Pupko and Mayrose 2020]).

2.1 MG-style models

Thanks to the point-mutation assumption, one can re-formulate a nucleotide-level model, such as the general-time-reversible (GTR) model (Lanave et al., 1984), into a nucleotide triplet state space. Let $\rho = (\rho_{lm})_{1 \leq l,m \leq 4}$ be a set of (symmetrical) nucleotide relative exchangeability parameters, with the constraint $\sum_{1 \leq l < m \leq 4} \rho_{lm} = 1$. Also let $\varphi = (\varphi_m)_{1 \leq m \leq 4}$, with $\sum_{m=1}^{4} \varphi_m = 1$, be a set of nucleotide equilibrium frequency parameters. The GTR model specifies the entries of a 4 by 4 rate matrix as $Q_{lm} = \rho_{lm}\varphi_m$. The exact same model

can be written into a 64 by 64 matrix, specifying rates from one codon i to another j as:

$$Q_{ij} = \begin{cases} \rho_{i_c j_c} \varphi_{j_c}, & \text{if } i \text{ and } j \text{ differ only at } c^{\text{th}} \text{ codon position,} \\ 0, & \text{otherwise,} \end{cases} \quad (1)$$

where i_c corresponds to an index of the nucleotide at the c^{th} codon position ($c = 1$, 2, or 3) of codon i ($i_c = 1$, 2, 3, or 4, as indices for A, C, G, and T). The distinction between Equation 1 and the GTR model is only one of a game of indices, but the formulation given in Equation 1 suggests that we could further recognize different types of codon substitutions. For instance, if we suppress stop codons from the process (reducing it to a 61 by 61 rate matrix, as stated above), we could recognize the distinction between synonymous and nonsynonymous substitution rates, specifying the entries in the matrix as:

$$Q_{ij} = \begin{cases} \rho_{i_c j_c} \varphi_{j_c}, & \text{if } i \text{ and } j \text{ are synonymous} \\ \rho_{i_c j_c} \varphi_{j_c} \omega, & \text{if } i \text{ and } j \text{ are nonsynonymous} \\ 0, & \text{otherwise.} \end{cases} \quad (2)$$

The model given in Equation 2 resembles closely the one proposed by Muse and Gaut (1994). The most notable difference is that in their seminal paper, Muse and Gaut invoked a new multiplicative parameter for nonsynonymous and synonymous rates, whereas in Equation 2 we effectively set the synonymous rate multiplier to 1, and invoke a single parameter, ω, on nonsynonymous rates, and hence $\omega = dN/dS$. The other difference with the original model is that Equation 2 includes parameters controlling nucleotide exchangeabilities (as in a GTR nucleotide-level model), whereas Muse and Gaut originally did not, and only used the frequency parameter of the target nucleotide (as in a F81 nucleotide-level model, Felsenstein, 1981). The stationary probability of codon i, denoted π_i, which can be thought of as the proportion of time spent in codon state i when running the substitution process for a very long time, is given by:

$$\pi_i = \frac{\varphi_{i_1} \varphi_{i_2} \varphi_{i_3}}{\sum_j \varphi_{j_1} \varphi_{j_2} \varphi_{j_3}}, \quad (3)$$

where the summation in the denominator is over all 61 (non-stop) codon states. In other words, the stationarity of the model given by Equation 2 is nearly identical to what it would be under the GTR model over three nucleotide positions, but with a slight re-normalization for the absence of the stop codons from the state space.

One of the widely used extensions to this formulation, often denoted as F3x4, is to invoke three distinct nucleotide frequency vectors, $\varphi^{(1)}$, $\varphi^{(2)}$, and $\varphi^{(3)}$, for the three within-codon positions (Yang, 2006), and hence with a rate matrix given by:

$$Q_{ij} = \begin{cases} \rho_{i_c j_c} \varphi^{(c)}_{j_c}, & \text{if } i \text{ and } j \text{ are synonymous} \\ \rho_{i_c j_c} \varphi^{(c)}_{j_c} \omega, & \text{if } i \text{ and } j \text{ are nonsynonymous} \\ 0, & \text{otherwise,} \end{cases} \quad (4)$$

and stationary probability given by:

$$\pi_i = \frac{\varphi^{(1)}_{i_1} \varphi^{(2)}_{i_2} \varphi^{(3)}_{i_3}}{\sum_j \varphi^{(1)}_{j_1} \varphi^{(2)}_{j_2} \varphi^{(3)}_{j_3}}. \quad (5)$$

The idea behind F3x4 formulation—so called because it involves three sets of four-dimensional frequency vectors—is to account for the uneven frequencies observed across the three within-codon positions that result from the structure of the genetic code; for instance, if there is

a highly skewed compositional bias in the data, it is most reflected in the third codon positions, whose state typically has no impact on the encoded amino acid, whereas first and second codon positions, whose states generally dictate the amino acid, would have different nucleotide frequencies. In other words, the differences in nucleotide frequencies across the three positions are features of different selective pressures operating at the amino acid level. While it may seem artificial to have an enriched parameterization at the nucleotide-level to account for features of selection at the amino acid level, the justification is phenomenological: regardless of the mechanistic details at the amino acid level, this modeling approach attempts to capture the net effect of these mechanisms, at least partially.

Pond et al. (2010) proposed a correction to the common practice of setting the values of these parameters to the nucleotide frequencies observed at the three codon positions of the data set at hand. The parameters can also be estimated by maximum likelihood, or become part of a Bayesian inference (Rodrigue et al., 2008a).

2.2 GY-style models

Models inspired by Goldman and Yang (1994) differ from those inspired by Muse and Gaut (1994) in a few ways. Perhaps the most significant difference, however, is the fact that with GY-style models the entries in the substitution rate matrix are proportional to the frequency (or stationary probability) of the target *codon*:

$$Q_{ij} = \begin{cases} \rho_{i_c j_c} \pi_j, & \text{if } i \text{ and } j \text{ are synonymous} \\ \rho_{i_c j_c} \pi_j \omega, & \text{if } i \text{ and } j \text{ are nonsynonymous} \\ 0, & \text{otherwise.} \end{cases} \tag{6}$$

This contrasts with the MG-style models given in Equation 2, which have entries proportional to the frequency of the target *nucleotide* (φ_{j_c}).

Most practitioners set the values of the 61-dimensional π-vector based on nucleotide-level specifications. For instance, it is common to work with a single vector of nucleotide frequencies, denoted F1x4, and setting $\pi_i \propto \varphi_{i_1}\varphi_{i_2}\varphi_{i_3}$. This practice, however, leads to peculiar effects in the specification of codon substitution rates, that could sometimes contradict the modeling intention. Rodrigue et al. (2008a) point out an example: suppose a context that is highly sucesptible to events leading to A or to T (in other words, a context where φ_A and φ_T are at higher values than φ_C and φ_G); all else being equal, the substitution rate from CGC to CTC (i.e., toward a high-frequency state T) will be lower than the rate from ATA to AGA (i.e., toward a low-frequency state G). Stated in reciprocally, the rate will be higher for the event that goes against the compositional bias at the nucleotide level (to a final state G), simply because of the context of the event. It is unclear if practitioners fully realize these sorts of peculiarities of GY-style F1x4 models, or if they consider them to be negligible to the inference of interest, usually ω (or its distribution).

Naturally, the F3x4 configuration, setting $\pi_i \propto \varphi^{(1)}_{i_1}\varphi^{(2)}_{i_2}\varphi^{(3)}_{i_3}$ is also utilized extensively in GY-style models. However, Huelsenbeck and Dyer (2004), as well as Rodrigue et al. (2008b), have shown that such approximations of π are often well outside the 95% credibility intervals of full Bayesian inferences of the 61-dimensional vector (a setting often denoted F61). On the other hand, the main drawback of the GY-F61 model is the confounded account of nucleotide, amino acid, and codon propensities. As discussed in a later section, the MG-style models offer opportunities to account for these propensities separately, within the mutation-selection framework.

3 Distributions of ω

Beyond issues relating to MG versus GY, or F1x4, F3x4, and F61, the main objective of codon substitution models is to characterize the ratio of nonsynonymous rates to synonymous rates, or ω, as denoted above, with a particular interest in cases where $\omega > 1$. In the previous section, we presented the classic codon substitution models in their homogeneous versions; each codon column of the alignment is considered to be the realization of a strictly identical Markov process, remaining unchanged across the branches of the phylogeny, or across the positions of the alignment. When fitting such global models to real data, one virtually never encounters cases where $\omega > 1$, simply because adaptive evolution is unlikely to be operating across all states, sites, and branches. In this section, we present the ideas behind the most widely known models of codon substitutions that account for variation in ω, with a focus on across-site heterogeneity.

3.1 Variable ω across sites

The first models to account for variation in ω values across the codon alignment were inspired from the random effects approaches of Yang (1993, 1994) to model rates across sites in nucleotide-level models. They consider each codon column of the alignment to have been produced from a model with ω drawn from a parametric statistical law (Nielsen and Yang, 1998; Yang et al., 2000). A problem remains, however, in that there is no inherent reason to choose one statistical law over another. The approach taken in the seminal works of Yang, Nielsen and collaborators was empirical: explore many different statistical laws, and perform likelihood-based model comparisons to identify the most appropriate one.

The simplest strategy to capturing an unknown distribution is to discretize it, into a finite mixture model. Suppose that we allow for K different ω parameters operating across codon alignment sites, denoted ω_1, ω_2,..., ω_K. The probability of the nth codon alignment column, denoted D_n, given the parameters of the model, denoted collectively as θ, is given as a weighted average over the K components of the mixture:

$$p(D_n \mid \theta) = \sum_{k=1}^{K} w_k p(D_n \mid \theta, \omega_k), \tag{7}$$

where $w = (w_k)_{1 \leq k \leq K}$, with $\sum_{k=1}^{K} w_k = 1$ is a set of weights associated with each component; these weights can be thought of as the prior probabilities that a particular alignment codon column D_n was generated by each of the K components.

In what they refer to as their *neutral* model, Nielsen and Yang (1998) set $K = 2$, $\omega_1 = 0$, and $\omega_2 = 1$, and infer the remaining parameters by maximum likelihood. In other words, their neutral model assumes a mixture of two codon sites, one in which the nonsynonymous substitution rate matches the synonymous substitution rate, and one in which nonsynonymous events are disallowed. Their *positive selection* model adds a third class ($K = 3$), with $\omega_3 > 1$. A likelihood ratio test can be performed between these two models, to establish if there is evidence of positive selection. Such a test is an example of the general approach to detecting adaptive evolution in the maximum likelihood framework, often followed by Empirical Bayes methods for identifying sites with high probability of having $\omega > 1$ (reviewed in Anisimova, 2012).

Capturing the unknown distribution of ω-values across sites can also be explored using continuous parametric distributions. For instance, rather than a two-component neutral model where sites either belong to a component with $\omega_1 = 0$ or $\omega_2 = 1$, one could invoke a

continuous distribution in the [0, 1] range, such as the beta distribution (Yang et al., 2000). Under such a model, a site likelihood takes the following form:

$$p(D_n \mid \theta) = \int_{\omega_n} p(\omega_n \mid \alpha, \beta) p(D_n \mid \theta, \omega_n) d\omega_n, \tag{8}$$

where $p(\omega_n \mid \alpha, \beta)$ is the density under the beta distribution (analogous to w_k in Equation 7), parameterized by α and β (which become part of the ML inference). In practice, the integral in Equation 8 is approximated through a discretization technique, reducing it to a weighted sum much like in Equation 7, but the overall approach still allows for a compact parameterization accounting for the heterogeneity across sites.

Of course, one could also envisage models built from a mixture of a continuous distribution and a discrete category $\omega_p > 1$ to allow for sites with positive selection, leading to a site likelihood with the form:

$$p(D_n \mid \theta) = w_1 \left(\int_{\omega_n} p(\omega_n \mid \alpha, \beta) p(D_n \mid \theta, \omega_n) d\omega_n \right) + w_2 p(D_n \mid \theta, \omega_p). \tag{9}$$

As before, this latter model, along with the previous one based on the beta distribution alone, can form the basis of a likelihood ratio test for the presence of sites with signatures of adaptive evolution.

Models based on mixtures of discrete and continuous distributions, or several continuous distributions, can be built in a similar fashion, which led Yang et al. (2000) to propose the suite of M-class models, with associated likelihood-ratio tests for adaptive substitution regimes. Indeed, the latter test based on comparing a model invoking a beta distribution and a additional component $\omega_p > 1$ against a model with only the beta distribution is known as the M8 versus M7 test.

These types of models have also been studied in the Bayesian context (Huelsenbeck and Dyer, 2004), within which they were also extended into non-parametric versions based on the Dirichlet process (Huelsenbeck et al., 2006; Rodrigue et al., 2008b,a).

3.2 Increasingly subtle modeling of variation in ω

Historically, the development of codon models has mainly progressed in the manner described above: the focus has been on proposing models that would allow for increasingly subtle manifestations of adaptive evolution, such as adaptive regimes operating at particular sites, and/or along particular branches (Yang and Nielsen, 2002; Yang et al., 2005; Yang and Dos Reis, 2010; Guindon et al., 2004), typically through the use of a variety of statistical devices controlling the values of ω across sites and branches. This focus may have turned attention away from a richer modeling of mutational features, as well as impeded a more general questioning of the use of ω as an appropriate means of detecting adaptation.

Indeed, the interpretation of ω values has been a point of contention (Nielsen and Yang, 2003; Seo and Kishino, 2008). Nielsen and Yang (2003) made indirect connections between this parameter and basic population genetics theory. They related the value of ω to the *scaled selection coefficient*, denoted S, which quantifies the change in fitness associated to a particular amino acid replacement, through the expression $\omega = S/(1 - e^{-S})$. This relation raises some odd scenarios. For instance, $\omega > 1$ implies that all nonsynonymous substitutions (at a given site and/or branch) have $S > 0$; an amino acid replacement from 'L' to 'I' would have a positive selection coefficient, and so would one from 'I' to 'L'. Conversely, cases where $\omega < 1$ imply that every nonsynonymous substitution decreases fitness ($S < 0$), including, say, 'D' to 'E' *and* 'E' to 'D'.

Meanwhile, another modeling rationale had emerged with an entirely different focus: devising a better representation of the pervasive underlying purifying selection operating on protein-coding DNA sequences.

4 The mutation-selection framework

In 1998, Halpern and Bruno (1998, "HB') introduced a model formulation with a more straightforward interpretation directly rooted in population genetics concepts. Their basic idea was to explicitly recognize both the initial and final states of a codon substitution, rather than simply distinguishing between synonymous and nonsynonymous events, and the identity of the final nucleotide or codon state. Yang and Nielsen (2008) offered a clear presentation of the idea of the model, introducing a fitness parameter f_i for codon i. A change from a wild-type state i to a mutant state j then implies a selection coefficient $s_{ij} = f_j - f_i$. The fixation probability associated to the mutant is given (approximated) by $2s_{ij}/(1 - e^{-2N_e s_{ij}})$, where N_e is the effective chromosomal population size. In a context involving haploids, N_e directly corresponds to the effective population size, whereas with diploids, our notation implies that N_e is twice the effective population size; in other words, the N_e we refer to here includes a ploidy-dependent multiplicative factor (see Yang and Nielsen, 2008, for details). A mutational process can be specified, for instance as a nucleotide-level GTR model, where the mutation rate from codon i to j is given by $\mu_{ij} = \rho_{i_c j_c} \varphi_{j_c}$, and the chromosomal population-level mutation rate is thus $N_e \mu_{ij}$. These two concepts are combined multiplicatively to specify the substitution rate:

$$Q_{ij} = \begin{cases} N_e \mu_{ij} \frac{2s_{ij}}{1-e^{(-2N_e s_{ij})}}, & \text{if } i \text{ and } j \text{ differ by one nucleotide,} \\ 0 & \text{otherwise.} \end{cases} \tag{10}$$

By multiplying the leftmost N_e factor through the fixation probability, and replacing $2N_e s_{ij}$ with S_{ij}, the scaled selection coefficient (scaled by twice the effective chromosomal population size), the model given by Equation 10 can be re-written as:

$$Q_{ij} = \begin{cases} \mu_{ij} \frac{S_{ij}}{1-e^{-S_{ij}}}, & \text{if } i \text{ and } j \text{ differ by one nucleotide,} \\ 0 & \text{otherwise,} \end{cases} \tag{11}$$

where $S_{ij} = F_j - F_i$, and where $F_i = 2N_e f_i$ is the scaled fitness of codon i. One can estimate scaled fitness parameters by anchoring one of them at $F_i = 0$ (for instance), and estimating the remaining fitness parameters around this constraint; what matters is the *relative* scaled fitness. An alternative, however is to set $F_i = \ln \psi_i$, where $\psi = (\psi_i)_{1 \leq i \leq 61}$, with $\sum_{i=1}^{61} \psi_i = 1$, is a codon profile. With this mutation-selection framework, the interpretations of $S_{ij} < 0$, $S_{ij} = 0$, and $S_{ij} > 0$ as negative, neutral, and positive selection coefficients apply to specific events, as opposed to being the coefficients of a long-standing regime implied from Nielsen and Yang (2003)'s interpretation.

The stationary probability of codon i under such a model is given by

$$\pi_i = \frac{\varphi_{i_1} \varphi_{i_2} \varphi_{i_3} e^{F_i}}{\sum_j \varphi_{j_1} \varphi_{j_2} \varphi_{j_3} e^{F_j}}, \tag{12}$$

or equivalently

$$\pi_i = \frac{\varphi_{i_1} \varphi_{i_2} \varphi_{i_3} \psi_i}{\sum_j \varphi_{j_1} \varphi_{j_2} \varphi_{j_3} \psi_j}. \tag{13}$$

In Equations 12 and 13, we have defined parameters controlling nucleotide propensities, φ, and parameters controlling codon fitness, ψ. The stationary probability, π, is calculated on the basis of φ and ψ, but the presentation of the model suggests that it is φ and ψ (or F) that are being estimated. In the presentation of the model by Halpern and Bruno (1998), however, it is π and φ that are estimated, with ψ (or F) being implicit. Specifically, they construct the substitution matrix as:

$$Q_{ij} = \begin{cases} \mu_{ij} \dfrac{\ln\left(\frac{\pi_j \mu_{ji}}{\pi_i \mu_{ij}}\right)}{1-\left(\frac{\pi_i \mu_{ij}}{\pi_j \mu_{ji}}\right)}, & \text{if } i \text{ and } j \text{ differ by one nucleotide,} \\ 0 & \text{otherwise.} \end{cases} \tag{14}$$

The equivalence of these approaches is made plain by noting that substituting Equation 13 into Equation 14 yields Equation 11.

4.1 HB-style models

Above, the form of the mutation-selection framework is presented as a global model. A core idea of Halpern and Bruno HB model, however, is to have a unique set of codon profiles for each site. Moreover, in practice, they reduce the model to having only amino acid profiles; this yields a model for each site n given by:

$$Q_{ij}^{(n)} = \begin{cases} \mu_{ij} \dfrac{\ln \phi_{f(j)}^{(n)} - \ln \phi_{f(i)}^{(n)}}{1-\left(\phi_{f(i)}^{(n)} / \phi_{f(j)}^{(n)}\right)}, & \text{if } i \text{ and } j \text{ are nonsyn. and differ by one nucleotide,} \\ \mu_{ij}, & \text{if } i \text{ and } j \text{ are syn. and differ by one nucleotide,} \\ 0 & \text{otherwise,} \end{cases} \tag{15}$$

where $\phi^{(n)} = (\phi_a^{(n)})_{1 \le a \le 20}$ is the amino acid profile operating at site n, and $f(i)$ returns an index for the amino acid encoded by codon i. Note that with $S_{ij}^{(n)} = \ln \phi_{f(j)}^{(n)} - \ln \phi_{f(i)}^{(n)}$, we can re-write the model in manner similar to Equation 11:

$$Q_{ij}^{(n)} = \begin{cases} \mu_{ij} \dfrac{S_{ij}^{(n)}}{1-e^{-S_{ij}^{(n)}}}, & \text{if } i \text{ and } j \text{ are nonsyn. and differ by one nucleotide,} \\ \mu_{ij}, & \text{if } i \text{ and } j \text{ are syn. and differ by one nucleotide,} \\ 0 & \text{otherwise.} \end{cases} \tag{16}$$

As such, the HB model is one that specifies as many codon substitution matrices as there are codon columns in the alignment, with each sharing a single nucleotide-level parameterization, but having a unique set of amino acid fitness parameters to it alone.

The adoption of the approach was hindered by its very high dimensionality, with a decade passing before such site-specific parameters were used again (Holder et al., 2008; Tamuri et al., 2012, 2014). There are also concerns with the treatment of site-specific profiles as *bona fide* parameters to be estimated by ML, given that such conditions do not conform with those of asymptotic theory of likelihood inference (Rodrigue, 2013): site-specific approaches do not have asymptotic conditions, since applying them to data sets with increasingly greater number of sites implies introducing new amino acid profiles, and thus changing the model; applying them to data sets with more sequences also changes the model, by requiring additional branch lengths, over a different tree. An alternative modeling strategy, routinely applied in most phylogenetic analyses, is the random variable approach.

4.2 Random-variable approaches for mutation-selection models

As described earlier in the context of classical codon models focused on ω, the random variable approach has also been invoked for amino acid fitness profiles, with both parametric (Rodrigue, 2013) and non-parametric (Rodrigue et al., 2010b; Rodrigue and Lartillot, 2014) methods. As before, the amino acid fitness profiles are considered random variables, integrated over a statistical law (Lartillot 2006; Chapter 1.4 [Lartillot 2020]). Along the parametric versions, the statistical laws utilized in previous studies have included a plain flat Dirichlet on the 20 amino acid states (Rodrigue and Aris-Brosou, 2011; Rodrigue, 2013), a free Dirichlet, itself with parameters controlling its center and concentration (Lartillot, 2006; Rodrigue, 2013), or finite mixture models with empirically derived values (Rodrigue and Aris-Brosou, 2011; Kazmi and Rodrigue, 2019).

Along the non-parametric versions, the Dirichlet process on amino acid fitness profiles has been utilized, implemented via both "Chinese restaurant" (Rodrigue et al., 2010b) and "stick-breaking" (Rodrigue and Lartillot, 2014) representations. Both representations utilize an auxiliary variable $z = (z_n)_{1 \leq n \leq N}$, specifying for each codon site the current allocation to one of K sets of "active"[1] amino acid fitness profiles, and the substitution model at site n is often presented as:

$$Q_{ij}^{(n)} = \begin{cases} \mu_{ij} \frac{S_{ij}^{(z_n)}}{1 - e^{-S_{ij}^{(z_n)}}}, & \text{if } i \text{ and } j \text{ are nonsyn. and differ by one nucleotide,} \\ \mu_{ij}, & \text{if } i \text{ and } j \text{ are syn. and differ by one nucleotide,} \\ 0 & \text{otherwise,} \end{cases} \tag{17}$$

where $S_{ij}^{(z_n)} = \ln \phi_{f(j)}^{(z_n)} - \ln \phi_{f(i)}^{(z_n)}$ is the scaled selection coefficient based on the amino acid profile allocated to site n, denoted $\phi^{(z_n)}$. The model given in Equation 17 is sometimes denoted MutSelDP.

In spite of being referred to as *non-parametric*, the Dirichlet process indeed involves parameters (often referred to as *hyper-parameters*), specifying a *base distribution*, analogous to a mean, and a *granularity* parameter, controlling the coarseness of the estimation of the unknown distribution. The likelihood function can be thought of as an integral over the infinite set of mixture models, conditional on these hyperparameters; the integral is effectively approximated using Monte Carlo methods. To date, relatively little work has been done to study the Dirichlet process with alternative hyperparameters; previous studies have endowed them with their own simple statistical laws (hyperpriors), and treated them as free elements of the inference. Richer hyperpriors, such as a base distribution itself consisting of a mixture of Dirichlets, should be studied in future work.

4.3 The mutation-selection framework as a null model for detecting adaptation

The basic motivation behind the mutation-selection framework set out by Halpern & Bruno is to define a better null model as a starting point to understanding features of the evolution of protein-coding genes (Rodrigue et al., 2010b). Specifically, the framework is focused on capturing purifying selection in a site-heterogeneous manner. Spielman and Wilke (2015) clearly lay out out how these models have the effect of inducing a dN/dS ratio less than 1

[1] The Monte Carlo devices invoke large sets of amino acid profiles, some of which are not actually allocated to any sites in the alignment (see Lartillot et al., 2013, for details).

at stationarity. We refer to the dN/dS ratio induced by the mutation-selection formulation as ω_0, and it is intuitively straightforward to see why it is constrained in the 0 to 1 range, by examining two extreme cases. First, suppose that, for some reason, all amino acids have nearly the same fitness; then, all values of S_{ij} would be close to 0, and thus $S_{ij}/(1-e^{-S}) \sim 1$, such that, when accounting for mutational opportunity (see Spielman and Wilke, 2015, for details), the nonsynonymous rate and the synonymous rate closely match (i.e., ω_0 is close to 1). At the other extreme, suppose that the amino acid fitness profile is strongly dominated by a single amino acid; then, at stationarity, it would be rare to be in a state other than this dominating amino acid, with all possible mutations away from it leading to very negative values of S_{ij}, and thus $S_{ij}/(1-e^{-S}) \sim 0$, such that the induced nonsynonymous rate is close to 0 (i.e., ω_0 is close to 0). Other configurations on amino acid profiles, between these two extreme scenarios, lead to ω_0 in the 0 to 1 range.

These ideas suggest an alternative approach to detecting adaptive regimes: rather than aiming to detect cases where $\omega > 1$, we could aim to detect cases where the overall dN/dS is greater than what would be expected under a pure mutation-selection formulation. One approach to this is to introduce a multiplicative parameter to nonsynonymous events, which we denote ω_* (the asterisk is used to clearly distinguish this parameter from ω), leading to the following model by Rodrigue and Lartillot (2017):

$$Q_{ij}^{(n)} = \begin{cases} \mu_{ij}\omega_* \frac{S_{ij}^{(zn)}}{1-e^{-S_{ij}^{(zn)}}}, & \text{if } i \text{ and } j \text{ are nonsyn. and differ by one nucleotide,} \\ \mu_{ij}, & \text{if } i \text{ and } j \text{ and syn. and differ by one nucleotide,} \\ 0 & \text{otherwise.} \end{cases} \tag{18}$$

The model given in Equation 18 is referred to as MutSelDP-ω_*. With such a formulation, the overall $\omega = dN/dS$ of the model at stationarity can be thought of as $\omega = \omega_*\omega_0$. Or, written alternatively as, $\omega_* = \omega/\omega_0$, this new parameter can be thought of as a measure of the deviation in the overall dN/dS (ω) with respect to what is expected from the pure mutation-selection formulation (ω_0). A value of $\omega_* > 1$ signals that the overall nonsynonymous rate is greater than expected under the mutation-selection balance, indicating a potential adaptive regime. The approach is much less demanding than classical codon models requiring that the nonsynonymous rate actually surpasses the synonymous rate, and could therefore have the potential to detect manifestations of adaptation that would otherwise be overlooked. More fundamentally, the approach demonstrates a different modeling perspective: that of formulating a better null framework, in order to uncover more subtle deviations from this new null. Such ideas were also studied by Bloom (2017).

5 Simulation study

Rodrigue and Lartillot (2017) conducted a brief simulation study to highlight the general behaviour of the MutSelDP-ω_* model given in Equation 18 when encountering data generated under an adaptive regime. Without going into the details of the simulation methods (described in full in Rodrigue and Lartillot, 2017), the idea is to change the amino acid fitness parameters (used to evolve sequences) over a phylogenetic tree; at a certain *rate* (ρ in Rodrigue and Lartillot, 2017) over the branches of the phylogeny, the amino acid profiles are altered (referred to as a *Red Queen* regime in Rodrigue and Lartillot, 2017). Thus, a sequence evolving along the branches with such changes in amino acid profiles is never quite at equilibrium, since it is tracking a repeatedly changing fitness optimum. In order to achieve a state of higher fitness, the sequence will accumulate a greater number of nonsynonymous substitutions than it would have in the absence of changes in fitness over time.

Note, however, that the increased nonsynonymous rate is generally still far from surpassing the synonymous rate.

In addition to controlling the rate of the Red Queen—the rate at which amino acid profiles are changed over the phylogeny—we can control the *magnitude* (σ_{RQ} in Rodrigue and Lartillot, 2017) of the change brought about to entries in the amino acid profile being altered. We can also control the *degree* (K in Rodrigue and Lartillot, 2017) of the change in profile, which corresponds to the number of pairs of entries in a profile that are being changed. We also alter the *proportion of sites* that are evolved under a Red Queen, with the remaining sites evolved under a pure mutation-selection process. Finally, we have an absolute *mutation rate*, which acts as a multiplier in the data-generating substitution matrices, and controls the overall amount of evolutionary signal in the simulations (set at 2×10^{-4} in Rodrigue and Lartillot, 2017). The simulations originally presented by Rodrigue and Lartillot (2017) only altered the rate of the Red Queen, leaving all other parameters of the simulation set to arbitrary values. Here, we further explore how the model reacts to a range of different values for other simulation parameters, with the results displayed as box-plots of the posterior mean values of ω_* over 20 replicate simulations in Figure 1. As done by Rodrigue and Lartillot (2017), we varied the rate of the Red-Queen (Figure 1A,B), however, rather than applying the Red-Queen regime to all sites, we explore results with the proportion of sites in an adaptive regime ranging from 0 to 1. We also adjust the number of pairs altered, or degreE,From 1 to 10 (Figure 1E,F), and study different magnitudes of changes brought to each pair of profiles (Figure 1C,D). All simulation scenarios were repeated on the basis of two starting sets of amino acid fitness profiles: those obtained from running the plain MutSelDP model (without ω_*) as given in Equation 17 on the BRCA1 alignment described by Rodrigue and Lartillot (2017), or on the concatenated alignment by Lartillot and Delsuc (2012).

In examining all of the performed simulations, increasing the proportion of sites in the alignment evolving under a Red-Queen evolutionary regime led to progressively higher ω_* values (Figure 1). This is easy to understand, as ω_* is a global parameter, estimated on the basis of the joint information found in the alignment. Therefore, if only a few sites experience an adaptive process, the model will accommodate the majority of the positions. As more sites are under a Red-Queen regime, the value of ω_* increases, and if the Red-Queen is sufficiently pronounced, the probability that ω_* is greater than 1 given the data, written symbolically as $p(\omega_* > 1 \mid D)$, approaches 1. When the Red-Queen regime is applied to 10% of sites, only 8% of simulations were found to be under positive selection using the BRCA1 amino acid profiles and 6% using the nuclear concatenation-based profiles ($p(\omega_* > 1 \mid D) \geq 0.95$). When all sites are subject to the Red-Queen regime, 89% of simulations were found to be under positive selection based on the BRCA1-derived profiles and 93% based on the concatenation-derived profiles.

Simulations based on concatenation- and BRCA1-derived profiles exhibited similar overall trends. We note, however, that in simulations based on the concatenation-derived profiles without any sites evolving under a Red-Queen, ω_* appears to be slightly below 1 (Figure 1B,D,F). Under null conditions, 2% of concatenation-based simulations where above 1, similar to the 1% observed with the BRCA1-based simulations. More noteworthy, under null conditions, 10% of BRCA1-based simulations were below 1 and 18% of concatenation-based simulations ($p(\omega_* < 1 \mid D) \geq 0.95$). The tendency of these simulations to lead to ω_* values below 1 supports the conclusions of Spielman and Wilke (2015) that the current MutSelDP model overestimates ω_0. However, null-generated data leading to ω_* values less than 1 suggests that detecting adaptive evolution using $\omega_* > 1$ will be conservative.

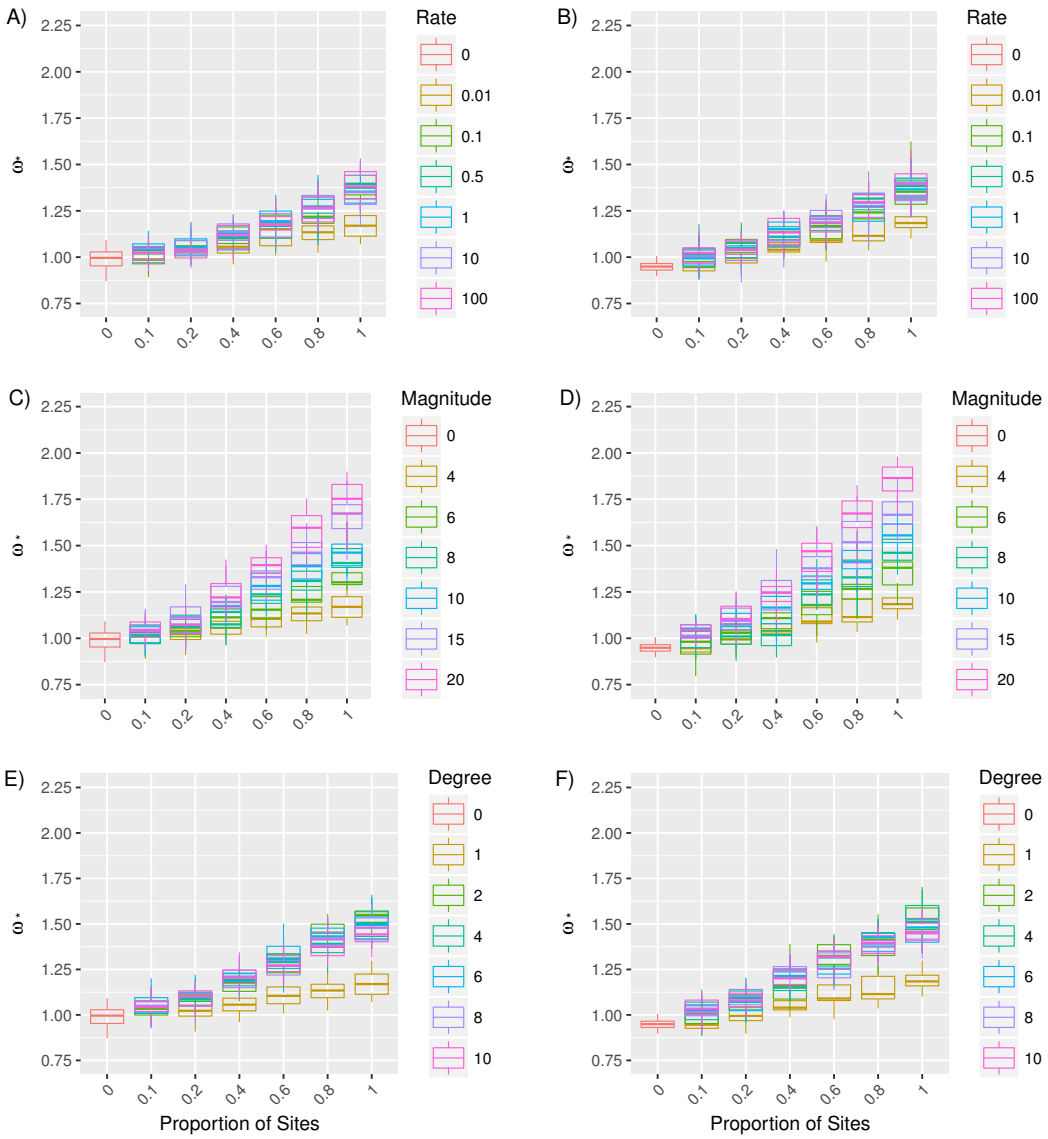

Figure 1 Simulations of an adaptive fitness landscape were performed using amino acid derived from the BRCA1 gene (Rodrigue and Lartillot, 2017) (A, C, E) and the nuclear concatenation by Lartillot and Delsuc (2012) (B, D, F). Varying the proportion of sites in combination with the magnitude of change had the greatest impact on ω_* under both sets of amino acid profiles (C,D). Increasing the rate of the Red Queen and the degree (the number of amino acid fitness values changed) eventually leads to a plateau effect (A,B, E,F).

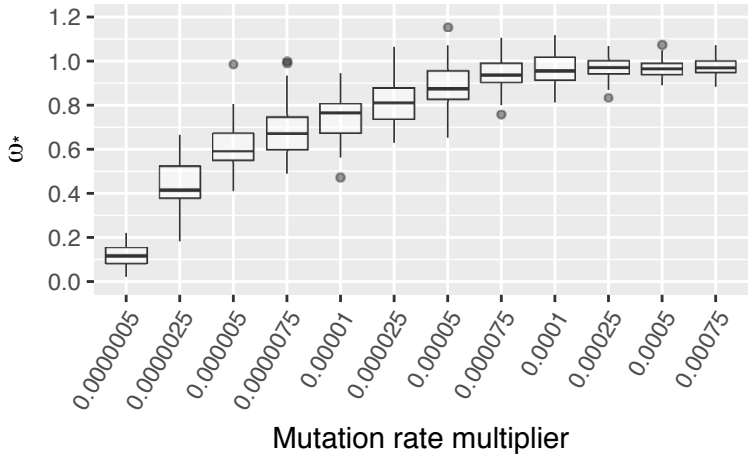

Figure 2 The effect of overall mutation rate on inferred values of ω_*.

Reproducing the results of Rodrigue and Lartillot (2017), we varied the *rate* of the Red-Queen (Figure 1A,B). When simulating alignments with 10% to 20% of sites under increasing rates of evolution, simulations were detected to be under adaptive evolution ($p(\omega_* > 1 \mid D) \geq 0.95$) with a frequency less than or equal to 30% of replicates. At least 40% of sites were required to be under the Red-Queen regime in conjunction with a moderate rate of change, to find adaptive evolution in more than half of the simulations. When increasing the rate of change for the evolutionary regime there appears to be a plateau effect, where subsequent increases to the rate have no further impact on ω_* (Figure 1A,B). In other words, increasing the rate of the Red Queen eventually leads to a saturation effect: if the fitness profiles are altered at a rate greater than the substitution rate, there is not sufficient time between Red Queen changes for the sequence to evolve toward the intermediate fitness optima, and the Red Queen starts "spinning its wheels".

The highest values of ω_* were reached by increasing the magnitude of changes in amino acid profiles, when the proportion of Red-Queen sites was high (Figure 1C,D). Simply stated, more drastic changes to amino acid profiles will increase the nonsynonymous flux in sequence evolution, leading to higher values of ω_*.

Altering the number of pairs of amino acid fitness values subject to change at a site led to a similar pattern observed with rate changes (Figure 1E,F). Examining simulations where only 10% or 20% of the sites in the alignment are considered, the proportion of simulations resulting in inferences with $\omega_* > 1$ shows little variation. However, beyond 40% of Red-Queen sites in the alignment, there is a rapid jump to greater than 90% of simulations being detected as being under adaptive evolution ($p(\omega_* > 1 \mid D) \geq 0.95$). However, with more than 40% of Red-Queen sites, after the increase in degree from one pair of amino acids to two, subsequent increases beyond two pairs have minimal apparent effect on on the value of ω_*. This can be understood from the fact that many changes to amino acid profiles will have little impact on the evolving sequence, if those changes are made to amino acid states that are not accessible via point mutation, which will tend to happen increasingly when changing multiple pairs of values.

The simulations described above were done with the objective of recreating adaptive evolution, through changes over time to the parameters controlling the amino acid fitness landscape. The next set of simulations was aimed at studying the model's behaviour with

pure mutation-selection simulations, where the overall amount of evolutionary signal is controlled by modulating the underlying mutation rate over time. We varied a multiplicative parameter to the mutation matrix given by Rodrigue and Lartillot (2017) from 0.000025 to 0.00075, again with 20 replicates for each setting. The results of the posterior mean ω_* across replicates are displayed with box-plots at each set of simulation conditions in Figure 2.

We see from Figure 2 that when the mutation rate is sufficiently high, the simulations tend to lead to ω_* around 1. When the mutation rate is low, however, ω_* is below one, and can be very low when the mutation rate is sufficiently low. Under these simulations, many sites have only a few substitutions (or none at all), such that when an analysis is conducted on the resulting artificial data sets, the Dirichlet process has very little evolutionary signal from which to infer the distribution of amino acid profiles across sites; loosely speaking, the model "chooses" to dispense with capturing purifying selection with the Dirichlet process apparatus (by adopting a low-dimensional, nearly flat configuration). In other words, when the overall evolutionary signal is very weak, the model given in Equation 18 reverts back to the simpler MG model given in (2), with ω_* effectively playing the role of ω.

6 Comparison of mutation-selection and classical frameworks on real data

The MutSelDP-ω_* model has only been applied to a handful of real data sets. To gain a broader empirical view of how the model reacts to real data sets, we analyzed a random sub-set of 4464 alignments taken from the OrthoMaM database (v8, Douzery et al., 2014). To get a general sense of the values of ω obtained across these alignments, we sorted the posterior mean values obtained and plotted them in Figure 3. Of the genes examined, 8.7% (388 genes) had 95% credibility intervals with ω_* greater than 1 (note that this implies that $p(\omega_* > 1 \mid D) > 0.975$). This is not entirely unexpected, as the majority of genes are not likely to be subjected to an ongoing Red-Queen over the mammalian phylogeny. What is somewhat surprising, however, is the extent to which most genes lead to ω_* well below 1. While is has been shown that epistasis can lead to ω_* values below 1 (Rodrigue and Lartillot, 2017), we suspect that many of these data sets are analogous to those simulated with very low mutation rates: they are highly conserved, and generally imply very few nonsynonymous substitutions. In other words, most alignments do not have sufficient evolutionary signal to reliably infer the Dirichlet process on amino acid profiles, such that the model favors a more compact means of fitting low nonsynonymous rates with low ω_* values, rather than several highly peaked amino acid profiles.

The same datasets from OrthoMaM were examined with CODEML (Yang, 2007), under the M7 and M8 models. Based on CODEML, 1301 genes reject the null M7 model in an LRT against M8 ($p < 0.05$). Among these, 497 genes had no sites within the gene detected to be under adaptive evolution as defined by $p(\omega > 1 \mid D) \geq 0.95$ with the Bayes Empirical Bayes approach. Here, we make the distinction between two classes of genes detected to be under adaptive evolution with M8: those that reject M7 in the LRT against M8, and those that also have at least one significant site ($p(\omega > 1 \mid D) \geq 0.95$). We compared genes in the latter class with the genes uncovered with the MutSelDP-ω_* model. There is a 70.1% overlap in the genes identified using the two methods, and this overlap climes to 82.5% if considering only genes with a length greater than 500 codons. Thus, it appears that the two methods are at least partially capturing the same features, but doing so in very different ways. It is particularly noteworthy that the MutSelDP-ω_* model is detecting adaptive regimes globally over the gene, whereas the M8 model has the potential for site-heterogeneous detection.

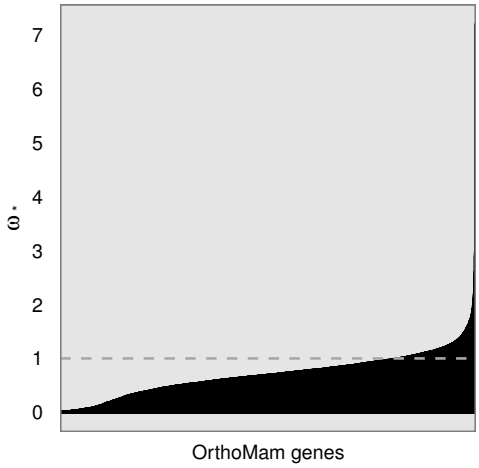

Figure 3 Analysis of a random subset of 4464 genes from the OrthoMam v8 database with the MutSelDP-ω_* model.

7 Conclusion

Although MutSelDP-ω_* is one of the richest codon substitution models proposed to date, its parameterization for detecting adaptive evolution is a simple univariate multiplier (ω_*) on nonsynonymous rates. Yet, it is already capable of detecting subtle instances of adaptive regimes in simulations, and has the potential to detect similar signals of adaptation in real data as the classical models with distributions of nonsynonymous rate multipliers.

It would of course be of great interest to expand the MutSelDP-ω_* model to allow for a distribution of ω_* values across sites, and/or across branches, in the same spirit as has been explored over the last few decades with classical codon models. Such models with multiple independent types of heterogeneity (e.g., distributions of amino acid profiles and distributions of ω_*) pose significant challenges, not the least of which will be their computational burden. Some short-cuts, such as utilizing empirically derived mixtures of amino acid profiles, along with a preset grid of ω_* values, could be worth considering for more speedy first-pass analyses of large data sets.

It would also be of interest to explore more simulations, incorporating more known features of the evolutionary process into the data-generating model. One glaring model violation in real data, laid bare in equation (10), is the assumption of a time-homogeneous effective population size (N_e). Understanding how the MutSelDP-ω_* model, or its eventual extensions, reacts to data simulated with changing effective population size over the tree would be an important first step. Other recent simulations (Laurin-Lemay et al., 2018a) have shown the CpG hypermutability can mislead some codon substitution models into detecting selection on synonymous codon usage. More generally, applying simulations to assessing the effects of these model violations, as well as others (Venkat et al., 2018; Jones et al., 2018), on mutation-selection-based models are in order to more carefully calibrate the reliability of the inferences to which they lead.

Finally, long-term modeling objectives should probably seek to integrate recent innovations into a single model, that could accommodate features such as uneven codon usage (e.g., as in Pouyet et al., 2016), variable effective populations size across the phylogeny, context-

dependent mutation rates, and epistatic effects (both within and across genes). Preliminary works exist that lay out the technical means of pursuing such a project, including ideas from Approximate Bayesian Computation (Laurin-Lemay et al., 2018b), and nested-MCMC systems (Robinson et al., 2003; Rodrigue et al., 2009; Kleinman et al., 2010; Rodrigue et al., 2010a). Bringing these ideas together could enable a framework for building models that are progressively more faithful to modern biological understanding of molecular evolution.

References

Anisimova, M. (2012). Parametric models of codon substitution. In Cannarozzi, G. M. and Schneider, A., editors, *Codon Evolution*, pages 12–33. Oxford University Press.

Bloom, J. (2017). Identification of positive selection in genes is greatly improved by using experimentally informed site-specific models. *Biol. Direct*, 12:1.

Douzery, E. J., Scornavacca, C., Romiguier, J., Belkhir, K., Galtier, N., Delsuc, F., and Ranwez, V. (2014). Orthomam v8: a database of orthologous exons and coding sequences for comparative genomics in mammals. *Mol. Biol. Evol.*, 31(7):1923–1928.

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

Gojobori, T. (1983). Codon substitution in evolution and the saturation of synonymous changes. *Genetics*, 105(4):1011–1027.

Goldman, N. and Yang, Z. (1994). A codon-based model of nucleotide substitution for protein-coding DNA sequences. *Mol. Biol. Evol.*, 11(5):725–736.

Guindon, S., Rodrigo, A. G., Dyer, K. A., and Huelsenbeck, J. P. (2004). Modeling the site-specific variation of selection patterns along lineages. *Proc. Natl. Acad. Sci. USA*, 101(35):12957–12962.

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

Holder, M. T., Zwickl, D. J., and Dessimoz, C. (2008). Evaluating the robustness of phylogenetic methods to among-site variability in substitution processes. *Phil. Tran. R. Soc. B*, 363:4013–4021.

Huelsenbeck, J. P. and Dyer, K. A. (2004). Bayesian estimation of positively selected sites. *J. Mol. Evol.*, 58:661–672.

Huelsenbeck, J. P., Jain, S., Frost, S. W. D., and Pond, S. L. K. (2006). A Dirichlet process model for detecting positive selection in protein-coding DNA sequences. *Proc. Natl. Acad. Sci. USA*, 103:6263–6268.

Jones, C. T., Youssef, N., Susko, E., and Bielawski, J. P. (2018). Phenomenological load on model parameters can lead to false biological conclusions. *Molecular biology and evolution*, 35(6):1473–1488.

Kazmi, S. O. and Rodrigue, N. (2019). Detecting amino acid preference shifts with codon-level mutation-selection mixture models. *BMC Evol. Biol.*, 19:62.

Kleinman, C. L., Rodrigue, N., Lartillot, N., and Philippe, H. (2010). Statistical potentials for improved structurally constrained evolutionary models. *Mol. Biol. Evol.*, 27(7):1546–1560.

Lanave, C., Preparata, G., Saccone, C., and Serio, G. (1984). A new method for calculating evolutionary substitution rates. *J. Mol. Evol.*, 20:86–93.

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

Lartillot, N. (2020). The bayesian approach to molecular phylogeny. In Scornavacca, C., Delsuc, F., and Galtier, N., editors, *Phylogenetics in the Genomic Era*, chapter 1.4, pages 1.4:1–1.4:17. No commercial publisher | Authors open access book.

Lartillot, N. and Delsuc, F. (2012). Joint reconstruction of divergence times and life-history evolution in placental mammals using a phylogenetic covariance model. *Evolution*, 66(6):1773–1787.

Lartillot, N., Rodrigue, N., Stubbs, D., and Richer, J. (2013). PhyloBayes-MPI: Phylogenetic reconstruction with infinite mixtures of profiles in a parallel environment. *Syst. Biol.*, 62:611–615.

Laurin-Lemay, S., Philippe, H., and Rodrigue, N. (2018a). Multiple factors confounding phylogenetic detection of selection on codon usage. *Mol. Biol. Evol.*, 35(6):1463–1472.

Laurin-Lemay, S., Rodrigue, N., Lartillot, N., and Philippe, H. (2018b). Conditional approximate bayesian computation, a new approach for across-site dependency in high-dimensional mutation-selection models. *Mol. Biol. Evol.*, in press.

Li, W.-H., Wu, C.-I., and Luo, C.-C. (1985). A new method for estimating synonymous and nonsynonymous rates of nucleotide substitution considering the relative likelihood of nucleotide and codon changes. *Mol. Biol. Evol.*, 2(2):150–174.

Miyata, T. and Yasunaga, T. (1980). Molecular evolution of mrna: a method for estimating evolutionary rates of synonymous and amino acid substitutions from homologous nucleotide sequences and its applications. *J. Mol. Evol.*, 16:23–36.

Muse, S. V. and Gaut, B. S. (1994). A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates, with application to the chloroplast genome. *Mol. Biol. Evol.*, 11(5):715–724.

Nei, M. and Gojobori, T. (1986). Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Mol. Biol. Evol.*, 3(5):418–426.

Nielsen, R. and Yang, Z. (1998). Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. *Genetics*, 148:929–936.

Nielsen, R. and Yang, Z. (2003). Estimating the distribution of selection coefficients from phylogenetic data with applications to mtDNA. *Mol. Biol. Evol.*, 20:1231–1239.

Perler, F., Efstratiadis, A., Lomedico, P., Gilbert, W., Kolodner, R., and Dodgson, J. (1980). The evolution of genes: the chicken preproinsulin gene. *Cell*, 20(2):555–566.

Pond, S. K., Delport, W., Muse, S. V., and Scheffler, K. (2010). Correcting the bias of empirical frequency parameter estimators in codon models. *PLoS One*, 5(7):e11230.

Pouyet, F., Bailly-Bechet, M., Mouchiroud, D., and Guéguen, L. (2016). SENCA: A Multilayered Codon Model to Study the Origins and Dynamics of Codon Usage. *Gen. Biol. Evol.*, 8:2427–2441.

Pupko, T. and Mayrose, I. (2020). A gentle introduction to probabilistic evolutionary models. In Scornavacca, C., Delsuc, F., and Galtier, N., editors, *Phylogenetics in the Genomic Era*, chapter 1.1, pages 1.1:1–1.1:21. No commercial publisher | Authors open access book.

Robinson, D. M., Jones, D. T., Kishino, H., Goldman, N., and Thorne, J. L. (2003). Protein evolution with dependence among codons due to tertiary structure. *Mol. Biol. Evol.*, 18:1692–1704.

Rodrigue, N. (2013). On the statistical interpretation of site-specific variables in phylogeny-based substitution models. *Genetics*, 193:557–564.

Rodrigue, N. and Aris-Brosou, S. (2011). Fast Baysian choice of phylogenetic models: prospecting data augmentation-based thermodynamic integration. *Syst. Biol.*, 60:881–887.

Rodrigue, N., Kleinman, C., Philippe, H., and Lartillot, N. (2009). Computational methods for evaluating phylogenetic models of coding sequence evolution with dependence between codon. *Mol. Biol. Evol.*, 26:1663–1676.

Rodrigue, N. and Lartillot, N. (2014). Site-heterogeneous mutation-selection models within the phylobayes-mpi package. *Bioinformatics*, 30(7):1020–1021.

Rodrigue, N. and Lartillot, N. (2017). Detecting adaptation in protein-coding genes using a Bayesian site-heterogeneous mutation-selection codon substitution model. *Mol. Biol. Evol.*, 34(1):204–214.

Rodrigue, N., Philippe, H., and Lartillot, N. (2008a). Bayesian comparisons of codon substitution models. *Genetics*, 180:1579–1591.

Rodrigue, N., Philippe, H., and Lartillot, N. (2008b). Uniformization for sampling realizations of markov processes: applications to bayesian implementations of codon substitution models. *Bioinformatics*, 24(1):56–62.

Rodrigue, N., Philippe, H., and Lartillot, N. (2010a). Mechanistic revisions of phenomenological modeling strategies in molecular evolution. *Trends Genet.*, 26:248–252.

Rodrigue, N., Philippe, H., and Lartillot, N. (2010b). Mutation-selection models of coding sequence evolution with site-heterogeneous amino acid fitness profiles. *Proc. Natl. Acad. Sci. USA*, 107(10):4629–4634.

Seo, T. K. and Kishino, H. (2008). Synonymous Substitutions Substantially Improve Evolutionary Inference from Highly Diverged Proteins. *Syst. Biol.*, 57:367–377.

Spielman, S. J. and Wilke, C. O. (2015). The relationship between dN/dS and scaled selection coefficients. *Mol. Biol. Evol.*, 32(4):1097–1108.

Tamuri, A. U., dos Reis, M., and Goldstein, R. A. (2012). Estimating the distribution of selection coefficients from phylogenetic data using sitewise mutation-selection models. *Genetics*, 190:1101–115.

Tamuri, A. U., Goldman, N., and dos Reis, M. (2014). A Penalized Likelihood Method for Estimating the Distributionof Selection Coefficients from Phylogenetic Data. *Genetics*, 197:257–271.

Venkat, A., Hahn, M. W., and Thornton, J. W. (2018). Multinucleotide mutations cause false inferences of lineage-specific positive selection. *Nature ecology & evolution*, 2(8):1280.

Yang, Z. (1993). Maximum-likelihood estimation of phylogeny from DNA sequences when substitution rates differ over sites. *Mol. Biol. Evol.*, 10:1396–1401.

Yang, Z. (1994). Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. *J. Mol. Evol.*, 39:306–14.

Yang, Z. (2006). *Computational Molecular Evolution*. Oxfors Series in Ecology and Evolution.

Yang, Z. (2007). Paml 4: phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.*, 24(8):1586–1591.

Yang, Z. and Dos Reis, M. (2010). Statistical properties of the branch-site test of positive selection. *Molecular biology and evolution*, 28(3):1217–1228.

Yang, Z. and Nielsen, R. (2002). Codon-substitution models for detecting molecular adaptation at individual sites along specific lineages. *Mol. Biol. Evol.*, 19(6):908–917.

Yang, Z. and Nielsen, R. (2008). Mutation-selection models of codon substitution and their use to estimate selective strengths on codon usage. *Mol. Biol. Evol.*, 25:568–579.

Yang, Z., Nielsen, R., Goldman, N., and Pedersen, A.-M. K. (2000). Codon-substitution models for heterogeneous selection pressure at amino acid sites. *Genetics*, 155:431–449.

Yang, Z., Wong, W. S., and Nielsen, R. (2005). Bayes empirical bayes inference of amino acid sites under positive selection. *Mol. Biol. Evol.*, 22(4):1107–1118.