Detecting Adaptation in Protein-Coding Genes Using a Bayesian Site-Heterogeneous Mutation-Selection Codon Substitution Model

Nicolas Rodrigue[1] and Nicolas Lartillot[2]

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

[2]Université de Lyon, Laboratoire de Biométrie, Biologie Évolutive, Villeurbanne, France

*Corresponding author: E-mail: nicolas.rodrigue@carleton.ca

Associate editor: Tal Pupko

Abstract

Codon substitution models have traditionally attempted to uncover signatures of adaptation within protein-coding genes by contrasting the rates of synonymous and non-synonymous substitutions. Another modeling approach, known as the mutation–selection framework, attempts to explicitly account for selective patterns at the amino acid level, with some approaches allowing for heterogeneity in these patterns across codon sites. Under such a model, substitutions at a given position occur at the neutral or nearly neutral rate when they are synonymous, or when they correspond to replacements between amino acids of similar fitness; substitutions from high to low (low to high) fitness amino acids have comparatively low (high) rates. Here, we study the use of such a mutation–selection framework as a null model for the detection of adaptation. Following previous works in this direction, we include a deviation parameter that has the effect of capturing the surplus, or deficit, in non-synonymous rates, relative to what would be expected under a mutation–selection modeling framework that includes a Dirichlet process approach to account for across-codon-site variation in amino acid fitness profiles. We use simulations, along with a few real data sets, to study the behavior of the approach, and find it to have good power with a low false-positive rate. Altogether, we emphasize the potential of recent mutation–selection models in the detection of adaptation, calling for further model refinements as well as large-scale applications.

Key words: **nearly neutral evolution, epistasis, fitness landscape, Dirichlet process, Markov chain Monte Carlo.**

Introduction

There has long been a great interest in characterizing the selective regimes involved in the evolution of protein-coding genes. Much of the focus has been on whether traces of molecular adaptation can be detected through the use of phylogenetic statistical methods, including likelihood-based models of codon substitution. The central idea of some of the most well-known codon substitution models is to estimate the ratio of the non-synonymous over the synonymous substitution rates, denoted *dN*/*dS*. Assuming no selection acting on synonymous mutations, a *dN*/*dS* greater than 1 indicates that non-synonymous substitutions accumulate more rapidly than expected in the absence of selection on them. Such a result is considered a typical signature of an adaptive process, and is indeed usually qualified as *positive selection*.

Stemming from the early versions of codon models (Goldman and Yang 1994; Muse and Gaut 1994), the most basic approach would invoke a single global parameter, usually denoted ω, capturing the average *dN*/*dS* over the entire protein and along all branches of the phylogenetic tree. Combined with a set of parameters controlling a point-mutation rate from one codon *i* to another *j*, denoted μ_{ij}, the substitution rate matrix Q has entries of the form:

$$Q_{ij} = \begin{cases} \mu_{ij}, & \text{if } i \text{ and } j \text{ are synonymous}, \\ \mu_{ij}\omega, & \text{if } i \text{ and } j \text{ are nonsynonymous}. \end{cases} \qquad (1)$$

However, positive selection is rarely detected using such simple models. This is because adaptation most often operates on a background of strong purifying selection, against mutations that would disrupt conformational stability or other fundamental biochemical and cellular requirements. As a result, even in the presence of some adaptation, the overall $\omega = dN/dS$ of a protein is typically well below 1.

Hoping to increase statistical power, several variants of these codon models have been developed, most of which rely on the original idea of detecting a value of ω that would exceed 1. These variants, however, operate at a more fine-grained level—typically, at the level of individual branches (Yang and Nielsen 1998), sites (Nielsen and Yang 1998; Yang et al. 2000; Yang and Swanson 2002; Yang et al. 2005), or their combination (Yang and Nielsen 2002; Guindon et al., 2004; Zhang et al. 2005; Pond et al. 2011)—in the hope that positive selection is sufficiently concentrated, either in space or in time, to be detected in this more local fashion.

With this in mind, an important strategy has been to attempt to capture the modulations of ω along the sequence

doi:10.1093/molbev/msw220 Advance Access publication October 15, 2016

(Nielsen and Yang 1998; Yang et al. 2000; Yang and Swanson 2002; Yang et al. 2005). Technically, these *site models* assume that the codon sites along the gene sequence are composed of a mixture of several categories of sites, being either under negative selection ($\omega < 1$), in a neutral regime ($\omega = 1$), or under positive selection ($\omega > 1$). The proportion of sites for which $\omega > 1$ is typically estimated by maximum likelihood (e.g., as with the M8 model in Yang et al. 2000), and the model is tested against a null model without sites under positive selection (e.g., the M7 model in Yang et al. 2000).

When applied to genome-wide empirical data, site models uncover a number of interesting candidate genes under adaptation (Kosiol et al. 2008). On the other hand, these genes represent only a small fraction of the proteome (around 5%), and the fraction of sites inferred to be under positive selection among detected genes is also typically small (around 10%). This is in sharp contrast with alternative methods for measuring selection, comparing patterns of synonymous and non-synonymous polymorphism and divergence (McDonald and Kreitman 1991; Sawyer and Hartl 1992; Keightley and Eyre-Walker 2007; Eyre-Walker and Keightley 2009; Keightley and Eyre-Walker 2010; Galtier 2016), which tend to infer that a large fraction of non-synonymous substitutions might be adaptive (Sawyer et al. 2003; Halligan et al. 2010). A possible cause for the lack of power of site models could be that adaptive processes, rather than being intensely concentrated on a small number of sites, are more diluted over a larger number of positions across a functional region of a protein. In addition, even at the level of a single site, only a small subset of possible amino acid-changing mutations are likely to be adaptive at any given instant, whereas all other possible non-synonymous mutations at that site might still be highly deleterious, and thus subject to strong purifying selection.

Branch-site models represent a second strategy (Yang and Nielsen 2002; Zhang et al. 2005; Pond et al. 2011). Their motivation is to test for the presence of a short episode of adaptive evolution along a particular branch of a phylogenetic tree. These models are based on contrasting the patterns of synonymous and non-synonymous substitution rates along the branch of interest with those of the background provided by the remaining branches of the tree.

Branch-site models have uncovered many cases of putative episodes of adaptive evolution on some branches (Clark et al. 2003; Sawyer et al. 2004, 2005; Kosiol et al. 2008). By design, however, they have to assume that no adaptive regime operates along most of the tree (which can thus be taken as a meaningful neutral background) and are therefore inherently driven toward the detection of rare and isolated spikes of adaptive evolution. They are much less helpful in the case of proteins that are constantly under adaptation over very long evolutionary periods. In such situations, the putative episodes detected by branch-site models would merely be the emerging part of the iceberg—essentially, the branches over which adaptation has been strongest—revealing but a small fraction of the true extent of adaptation undergone by these proteins.

As an alternative to the classical codon models, an emerging trend in recent developments is the recognition of the potential to devise mechanistic codon substitution models that are rooted in first principles (Rodrigue et al. 2010a; Thorne et al. 2012; Echave et al. 2016). In particular, a modeling strategy based on an underlying population-genetics rationale has been studied with increased interest in recent years (Rodrigue et al. 2010b; Bloom 2014; Tamuri et al. 2014; Spielman and Wilke 2015). These *mutation–selection models* rely on an explicit account of the underlying fitness landscape, such that the selection coefficient associated with any particular mutation at any given instant is specified. Thus far, the models have essentially assumed a fixed fitness landscape, with multiplicative fitnesses across sites. However, the importance of accounting for heterogeneity of amino acid fitness profiles across codon sites is now well recognized (Halpern and Bruno 1998; Rodrigue et al. 2010b; Bloom 2014; Tamuri et al. 2014), which has led to models in which the fitness landscape is entirely characterized by an array of site-specific fitness vectors; for instance, the scaled fitness of the amino-acid encoded by codon *i* at position *n*, would be denoted $F_i^{(n)}$. The Markov process operating at codon site *n* is then given as (Halpern and Bruno 1998; Yang and Nielsen 2008):

$$Q_{ij}^{(n)} = \begin{cases} \mu_{ij}, & \text{if } i \text{ and } j \text{ are synonymous}, \\ \mu_{ij} \dfrac{S_{ij}^{(n)}}{1 - e^{-S_{ij}^{(n)}}}, & \text{if } i \text{ and } j \text{ are nonsynonymous}, \end{cases} \tag{2}$$

where $S_{ij}^{(n)} = F_j^{(n)} - F_i^{(n)} = 4N_e s_{ij}$ is the scaled selection coefficient (scaled by the effective population size N_e and a ploidy-dependent constant, here 4) associated with a mutant protein with the amino acid encoded by *j*, in a wild-type population where the amino-acid encoded by *i* is fixed at that position. Site-specific fitness profiles have been estimated either by maximum likelihood (Holder et al. 2008; Tamuri et al. 2014; Bloom 2016), by experimental means (Bloom 2014, 2016), or, have been considered as random-effects integrated over a distribution (Rodrigue 2013), sometimes itself non-parametrically estimated using Dirichlet process priors (Rodrigue et al. 2010b; Rodrigue and Lartillot 2014).

Under these mutation–selection models with constant fitness landscapes, the long-term evolutionary process is such that the protein-coding gene essentially fluctuates at the mutation–selection balance, maintaining itself around the optimum defined by the fitness landscape. We refer to this as a *nearly neutral* regime, since most mutations are deleterious (having a negative $S_{ij}^{(n)}$) and are purified away by selection, whereas the mutations that reach fixation, the substitutions, are typically either mildly deleterious or mildly advantageous, being neutral on average (having $S_{ij}^{(n)} \approx 0$). As a result, the *dN*/*dS induced* by the modeling of purifying selection, which we refer to herein as ω_0, is predicted to be in the 0–1 range (Halpern and Bruno 1998; Spielman and Wilke 2015). The extreme cases are also noteworthy: if, for some reason, all 20 amino acid have a very similar fitness, then $\omega_0 \approx 1$, whereas if the fitness landscape is greatly dominated by a single amino acid, then $\omega_0 \approx 0$.

Importantly, however, even if they capture only purifying selection, mutation–selection models with a constant fitness landscape nevertheless imply that, at mutation–selection equilibrium, some of the non-synonymous mutations (and half of the non-synonymous substitutions, owing to the time-reversibility of the Markov process; Yang and Nielsen 2008) are still characterized by a positive selection coefficient ($S_{ij}^{(n)} > 0$). This connects to a more general point: that one should not identify purifying and adaptive evolutionary regimes with $S_{ij}^{(n)} < 0$ and $S_{ij}^{(n)} > 0$ (Mustonen and Lässig 2009). The distinction between purifying selection versus ongoing adaptation should primarily be seen as a more global question, not linked to the selection coefficients attached to specific mutations or substitutions, but instead, relating to the evolutionary regime followed by the protein under study: fundamentally, whether the protein of interest is at equilibrium under a fixed fitness landscape, or whether it is constantly challenged by changing ecological conditions or ongoing evolutionary Red-Queens, such that it is evolving under a constantly fluctuating fitness landscape. If the fitness landscape changes at a very high rate, and with sufficient amplitude, this may lead to situations where $\omega = dN/dS > 1$. However, one can readily imagine less extreme Red-Queen regimes that would remain unapparent to current *dN*/*dS* models. For instance, building on a mutation–selection framework with a site-heterogeneous amino acid fitness landscape assuming multiplicative fitness across codon sites, suppose a process where, at an average rate of 1 per Myr, a position of the protein is taken at random, and the amino acid fitness profile at that position is mildly changed. Each round of such a Red-Queen process mildly changes a profile at a different position, and each time for a potentially specific amino acid target. If the rate of the Red-Queen were to be reduced to 0, the protein-coding gene would be evolving in the nearly-neutral regime specified by Equation (2), typically inducing a *dN*/*dS* well below 1 (Spielman and Wilke 2015), as described above. In contrast, with the Red-Queen active, the protein sequence is tracking a constantly moving fitness optimum. Even in this regime, however, most non-synonymous mutations still tend to move the sequence away from the target and are therefore deleterious. However, since the protein sequence is always lagging behind the moving target defined by the amino acid profiles, while accepting substitutions preferentially in the direction of this target, substitutions are on average adaptive. This results in a net increase in the rate of non-synonymous substitutions, and thus a higher overall *dN*/*dS*, even if it is still well below 1. Only when the Red-Queen operates at an extremely high rate, drastically perturbating the fitness landscape as fast or faster than substitutions occur, would we observe a *dN*/*dS* greater than 1.

In their current form (Halpern and Bruno 1998; Rodrigue et al. 2010b; Tamuri et al. 2012, 2014), site-heterogeneous mutation–selection models have been constructed under the assumption of multiplicative fitness across sites. In reality, it is suspected that epistasis represents an important feature of protein evolution (Lunzer et al. 2010; Ashenberg et al. 2013; McCandlish et al. 2013; Weinreich and Knies 2013; Gong and Bloom 2014). One consequence of epistatic interactions is to change the fitness landscape experienced by each site, as the sequence at other interacting sites changes over time. Thus, like Red-Queen adaptive regimes, epistasis also results in a fluctuating fitness landscape at each site. However, and unlike in the case of most ecological or intra-genomic Red-Queens, which unfold at a relatively high rate, these fluctuations are slow (reviewed by Bazykin 2015). Furthermore, under epistasis, fluctuations at a given site are such that the fitness landscape at that site tends to change so as to stabilize the fitness of the current state, a phenomenon referred to as *entrenchment* (Shah et al. 2015), or *evolutionary Stokes shift* (Pollock et al. 2012). Also, whereas it has been pointed out that when a deleterious mutation becomes fixed in the protein, it may be subsequently compensated by clusters of mutations becoming fixed at interacting sites, so as to restore the fitness (Cutler 2000), at mutation–selection balance, such a strongly deleterious mutation would unlikely to be fixed in the first place, precisely because it is strongly deleterious; thus, such clusters would be rare. For these reasons, and in contrast to Red-Queen situations, epistatic interactions will tend to result in a decrease of the *dN*/*dS*, compared with the expectation under the nearly neutral null model with plain multiplicative fitness across sites. This reasoning has been put to use for estimating the contribution of epistasis in protein evolution (McCandlish et al. 2013), although without relying on explicit mutation–selection models.

New Approaches

An interesting idea suggested by all these observations is that, in order to study molecular evolutionary regimes at protein-coding genes, one could try to detect *deviations* in *dN*/*dS*, relative to that induced by the nearly neutral regime. Doing so represents a fundamental change in perspective, which places the emphasis on the adoption a more adequate null model of neutrality against which to conduct statistical tests. Using a more realistic null model of what happens in the absence of on-going adaptation should result in a higher sensitivity of the test.

Detecting deviations from the null hypothesis implied by the mutation–selection model under a constant and multiplicative fitness landscape can be done by introducing a parameter, denoted ω_*, absorbing any deviation from the mechanistic mutation–selection formulation. This idea can be traced to several previous works (Robinson et al. 2003; Yang and Nielsen 2008). In the present context, rate matrix entries are given as:

$$Q_{ij}^{(n)} = \begin{cases} \mu_{ij}, & \text{if } i \text{ and } j \text{ are synonymous,} \\ \mu_{ij}\omega_* \dfrac{S_{ij}^{(n)}}{1 - e^{-S_{ij}^{(n)}}}, & \text{if } i \text{ and } j \text{ are non-synonymous.} \end{cases} \tag{3}$$

Significant upward deviation of ω_* from 1 then signals that there are too many non-synonymous substitutions, compared with the null expectation under the nearly-neutral regime, potentially due to the presence of ongoing Red-Queen-like adaptation. Significant downward deviation of ω_* signals a deficit in non-synonymous substitutions, perhaps due to the presence of epistatic effects.

It is noteworthy that the overall *dN*/*dS* of the model given in 3 will be the *dN*/*dS* induced by the modeling of purifying selection (which we have denoted ω_0) multiplied by the deviation parameter (denoted ω_*), written symbolically as $\omega = \omega_* \times \omega_0$. Re-writing this into $\omega_* = \omega/\omega_0$ highlights the interpretation of ω_* as measure of deviation in the overall $\omega = dN/dS$ from what would be expected under the nearly neutral regime of the mutation–selection model, ω_0. The standard application of codon substitution models is in fact a special case of this viewpoint, where $\omega_0 = 1$; in other words, with standard codon models, the (nearly-)neutral regime is one that assigns the same fitness to each amino acid, for all sites. However, Kimura emphasized that this is not the intended meaning of neutrality, stating (Kimura 1983, p. 53): "The neutral theory does not assert that *all* [original emphasis] amino acids are equivalent at a certain site, only that the majority of evolutionary changes concern those mutant that are equivalent." The misconstrued concept of neutrality of classical codon models may explain their lack of sensitivity, which would be only partially compensated for by focusing the detection of adaption on specific sites and/or specific branches. In contrast, by explicitly accounting for the background of strong purifying selection, such as in Equation (3), ω_* should capture very modest shifts in fitness landscapes, even if these shifts are distributed across several different positions, and are evenly distributed over the branches of the phylogeny. Specifically, a signature of molecular adaptation would not need to be pronounced to the point of producing $\omega > 1$ to be detected, but instead only requiring $\omega_* > 1$.

Bloom (2016) has very recently introduced an analogous framework in a maximum likelihood context. In his main approach, $S_{ij}^{(n)} = F_j^{(n)} - F_i^{(n)}$ values are built from experimentally derived site-specific amino acid profiles; site-specific ML inference of ω_* is performed, along with single-observation likelihood ratio tests at each of the sites. However, most of currently available protein-coding DNA data are not amenable to experimental construction of amino acid profiles, and there remains an interest for random-effects models that include a deviation parameter ω_* in the context of a joint probabilistic inference.

We perform simulations under the mutation–selection construct of the nearly-neutral regime, along with simulations of Red-Queen adaptive regimes, and protein-structure-based epistatic regimes, to study the above models. We show that, when combined with the Dirichlet process capturing across-site amino acid fitness heterogeneity, ω_* is able to detect the evolutionary regimes of the simulations in accordance with the above predictions. Applying the approach to a few empirical cases suggests that, for the purpose of detecting adaptation in protein coding sequences, it may provide a promising alternative to current codon models.

Results and Discussion

Simulated Data

We simulated protein-coding sequence data under a mutation–selection-based approach at the scale of placental mammals, as explained in the "Materials and Methods" section. Among other values, these simulations require amino acid profiles to be defined. Three different sets of site-specific amino acid fitness profiles were explored, as inferred from three arbitrarily chosen real protein-coding sequence alignments under a model based on Equation (2) (Rodrigue et al. 2010b), again at the mammalian scale, for the following genes: SAMHD1, TRIM5α, and BRCA1. We refer to these simulations as those of the *nearly neutral regime*. We also simulated data taking the nearly-neutral regime parameter values as a starting point, but making small changes to the amino acid profiles along the branches of the phylogenetic tree; these are the Red-Queen *adaptive regime* simulations. And, finally, we simulated a contact map and statistical potential to further modulate non-synonymous rates of a nearly neutral setting, with a model similar to that in Robinson et al. (2003); these are referred to as the *epistatic regime* simulations (see "Materials and Methods" for details on simulations).

Figure 1 shows examples of results for inferences on three simulated data sets, produced from the three types of evolutionary regimes, and utilizing different sets of site-specific amino acid profiles in each column of panels. The top three panels (A, B, and C) show the posterior distributions of ω under the plain MG model of Equation (1). All three top panels show distributions with $\omega < 1$, and in all three cases the central distribution (green) is the one from the simulation under the nearly neutral regime. The epistatic regime (blue) thus shifts the distributions to the left, whereas the adaptive regime (red) shifts it the right. Such shifts match well with the predicted behavior of the models applied to these simulated data, with epistatic effects having a tendency to reduce non-synonymous rates, and Red-Queen-like adaptation increasing non-synonymous rates. Based on the simple MG model, however, the three regimes would still be qualified as being under purifying selection.

Among the top three panels of figure 1, only in panel C does ω approach 1; this panel shows the simulations based on the posterior mean amino acid profiles inferred from the BRCA1 data set. Indeed, the posterior mean ω obtained on the real BRCA1 data set with the MG model is about 0.8 (table 1, first column), whereas the posterior mean ω obtained from simulations using profiles "trained" from that data set are close, but mildly below, in this case at about 0.7 (fig. 1C, green histogram). When using the BRCA1 profiles as a starting point, and activating the Red-Queen adaptive process, the distribution of ω (fig. 1C, red histogram) shifts to the right to the point of slightly over-stepping 1 in this instance. On the other hand, when simulating using the other two sets of amino acid profiles, obtained from data sets that lead to comparatively lower ω values under the plain MG model (with posterior means of about 0.3 and 0.45, respectively, for SAMHD1 and TRIM5α, see table 1, first column), the distributions of ω are always well below 1 (fig. 1A and B).

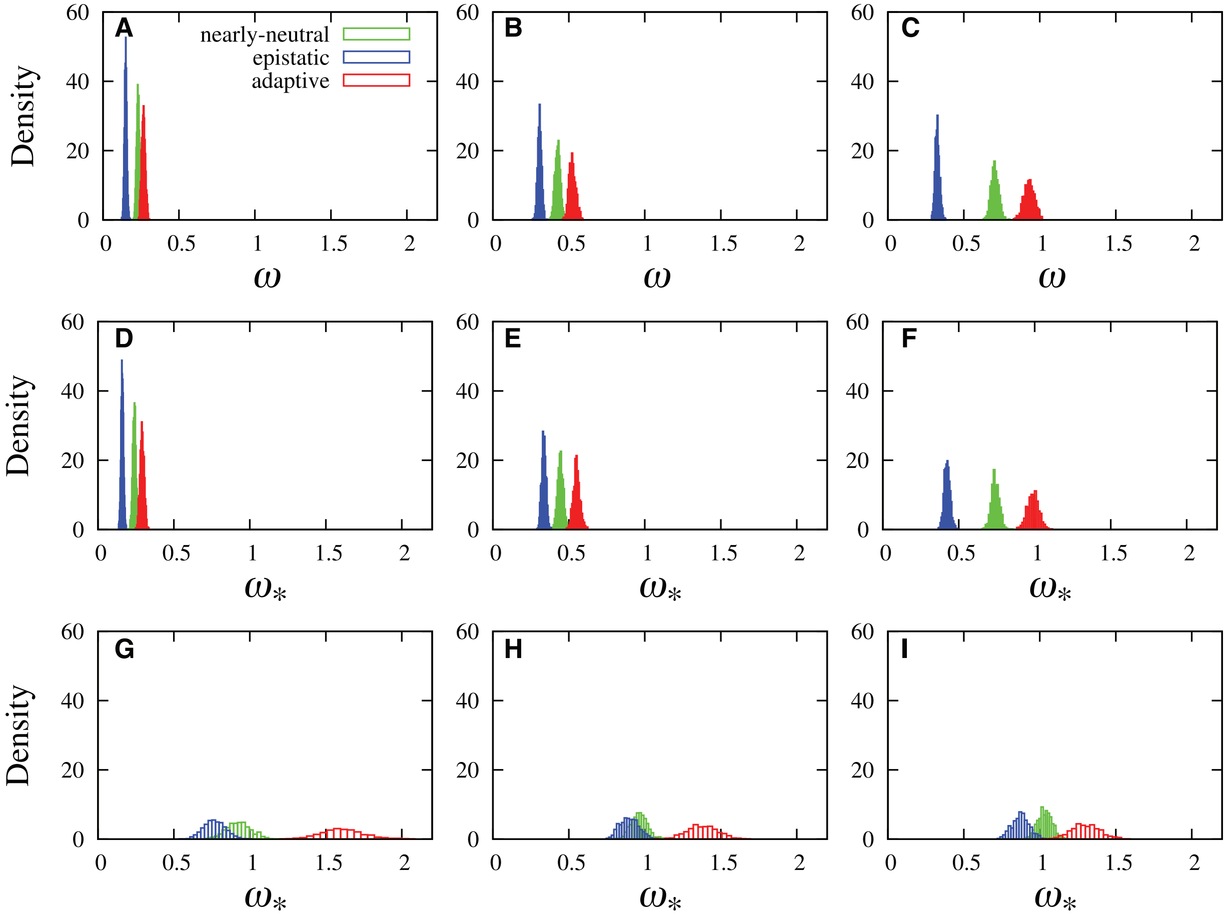

Fig. 1. Posterior distributions of ω (for MG model, panels *A*, *B*, and *C*) and ω_* (for MutSelYN model in *D*, *E*, and *F*, and MutSelDP in *G*, *H*, and *I*) using three different sets of amino acid profiles (those obtained from SAMHD1 in left panels, those obtained from TRIM5α in middle panels, and those obtained from BRCA1 in right panels). Three simulation conditions were used: the nearly neutral regime (green), an adaptive regime (red) and an epistatic regime (blue).

Table 1. Posterior Means and 95% Credibility Intervals, in Parentheses, of ω (with the MG model) and ω_* (with MutSelYN and MutSelDP models) on Six Mammalian Genes is shown in square brackets.

Data	MG	MutSelYN	MutSelDP
S1pr1-67-325	**0.049 (0.042, 0.055)**	**0.058 (0.051, 0.065)**	**0.681 (0.538, 0.857) [0.001]**
Rbp3-54-412	**0.190 (0.177, 0.203)**	**0.193 (0.181, 0.206)**	**0.726 (0.654, 0.806) [0.000]**
Vwf-62-392	**0.205 (0.188, 0.220)**	**0.212 (0.199, 0.226)**	**0.960 (0.865, 1.063) [0.220]**
Samhd1-67-543	**0.309 (0.288, 0.332)**	**0.324 (0.300, 0.348)**	**1.731 (1.542, 1.935) [1.000]**
Trim5α-68-363	**0.454 (0.426, 0.484)**	**0.468 (0.439, 0.498)**	**1.240 (1.128, 1.355) [1.000]**
Brca1-64-941	**0.783 (0.750, 0.817)**	**0.802 (0.770, 0.837)**	**1.188 (1.123, 1.257) [1.000]**

Note.—With the MutSelDP model, the posterior probability $p(\omega_* > 1|D)$ is shown in square brackets.

With the nearly neutral regime simulations, the values of ω are again found to be close (with posterior means at 0.23 and 0.43, respectively, for these example SAMHD1 and TRIM5α simulations) to the values obtained on the true data (at 0.31 and 0.45, respectively, table 1, first column).

The three panels of the middle row in figure 1*D–F* show the posterior distributions of ω_* under the site-homogeneous formulation of Yang and Nielsen (2008), referred to herein as MutSelYN; this is essentially the model written in Equation (3), but with all sites having the same amino acid fitness profile. As had been previously noted (Yang and Nielsen 2008), these distributions hardly differ from those of the panels above. We believe these and other previous attempts at combining a modeling of purifying selection with a scalar parameter on nonsynonymous rates were simply not suitably capturing the overall form of the sequence fitness landscape. In this particular case, the explanation is straightforward: amino acid profiles are not global quantities, but highly variable across different coding positions, such that a global approach is likely very far from reflecting a realistic fitness landscape.

Finally, the bottom three panels of figure 1G–I show the posterior distributions of ω_* when invoked with the Dirichlet

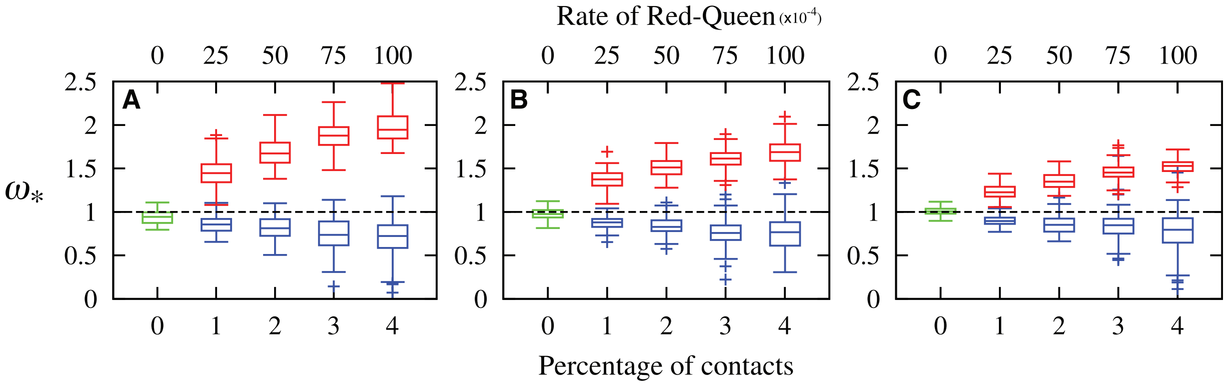

Fig. 2. Summary of posterior means of 100 replicates for each boxplot. Results for simulations under the nearly-neutral regime are in green, whereas results for four different degrees of epistatic regimes (with increasing percentage of possible pairwise amino acids being in contact) are in blue, and results of four different rates of the Red-Queen are in red. Simulations were based on three initial sets of amino acid profiles, taken from SAMHD1 in *A*, TRIM5α in *B*, and BRCA1 in *C*.

process controlling across-site heterogeneity in amino acid profiles, referred to as MutSelDP. These distributions have greater variance than those of the upper panels, since they are based on a much richer underlying model, and their locations are clearly very different than those of the panels above. In the three cases, the nearly neutral regime simulations lead to distributions that clearly straddle 1 (in green), indicating that the approach recovers well the actual simulation conditions. The adaptive Red-Queen simulations lead to distributions that clearly surpass 1 (red), indicating that the approach clearly recognizes the presence of an adaptive regime. The epistatic simulations show a tendency to lead to distributions of ω_* below 1 (blue), again suggesting that the approach can detect the presence of such an effect.

Each histogram in figure 1 is obtained from a single synthetic 300 codon alignment, from one of the three types of regimes (and with three sets of amino acid profiles), under one of three inference models. The histograms serve to illustrate the general behavior of the models under the simulation conditions. However, we performed several instances of our simulations. Moreover, we incremented the rate of the Red-Queen process of the adaptive regimes over four values. We did likewise for the degree of epistatic interaction across sites, by progressively increasing the number of interacting sites. Under each simulation condition, we generated 100 replicates. For each replicate, with ran a Bayesian MCMC with the model combining the Dirichlet process prior on amino acid profiles, along with ω_* and all other parameters, and computed the posterior mean, and 95% credibility interval of ω_*. The posterior means of ω_* are displayed across the various simulation settings for each batch of 100 replicates as boxplots in figure 2, with simulations conducted with the three sets of amino acid profiles from SAMHD1, TRIM5α, and BRCA1 in panels A, B, and C, respectively.

Most simulations conducted under nearly-neutral conditions have posterior means close to 1 (fig. 2, green boxplots). Looking more closely at each of the distributions, we find that 87 of the 100 nearly neutral replicates using the SAMHD1 amino acid profiles included 1 in their posterior 95% credibility interval of ω_*. All 13 cases that did not include 1 were below it. For nearly neutral simulations with TRIM5α amino acid profiles, 95 of the 100 replicates included 1 in their posterior 95% credibility interval, with only 1 of the 5 error cases being above 1. For the nearly neutral simulations with the BRCA1 profiles, 97 of 100 replicates included 1 in their posterior 95% credibility interval of ω_*; of the three replicates that did not, one had a 95% credibility interval below 1, and two are above 1. Overall, these results generally indicate a well-behaved system, although there appears to be a bias toward under-estimation of the value of ω_*. More work is required to investigate this mild bias.

We next note from figure 2 that the posterior mean value of ω_* tends to increase when the rate of the Red-Queen increases (red boxplots), i.e., when amino acid profiles change more rapidly along the branches of the tree, the detected deviation is indeed greater. Only 16 replicates of the 1,200 simulations under the adaptive regimes had 95% credibility intervals of ω_* that included 1, and these were all amongst the slowest setting of the Red-Queen. In all other cases, the credible intervals were above 1. This is in spite of the model's under-estimation of ω_*, revealed by the results from the nearly neutral simulations, which is likely still the case in these Red-Queen simulations. Altogether, these results suggests that uncovering genes with high posterior probability support for $\omega_* > 1$ provides a powerful test for the presence of ongoing, Red-Queen-like adaptation in protein coding sequences, while having a tendency to be conservative.

We also ran CODEML from the PAML package (Yang 2007) with several of the classic site models on these adaptive regimes simulations. Applied to the SAMHD1-based simulations, even with the most intense Red-Queen rate, none the site models detected any traces of adaptation (i.e., the likelihood was not significantly improved by including a class with $\omega > 1$). Applied to the BRCA1-based simulations, on the other hand, site models in CODEML can detect on average 5% of sites at the lowest rate of the Red-Queen, and this climbs to 11% of sites at the highest rate of the Red-Queen. The TRIM5α-based simulations exhibit an intermediate behavior, going from 2% to 9% across the range of Red-Queen rates explored. That site models can readily detect adaptive regimes in some cases and not others can be understood from the fact that the BRCA1-based simulations start from a model that already induces a relatively high non-synonymous rate; it does not take very much for the Red-Queen to push ω beyond 1 at certain sites, whereas the SAMHD1-based simulations start out from a point that induces low non-synonymous rates, such that there is a long way to go for the Red-Queen to attain $\omega > 1$. Still, detecting 11% of sites as being under an adaptive regime is very inaccurate, since 100% of sites are in fact under a mild (on average) adaptive regime.

Figure 2 also shows that epistatic effects tend to lead to $\omega_* < 1$ (blue boxplots), although in some cases (7 out of 1,200 replicates) it does surpass 1 with a posterior probability greater than 0.99. We are still unsure as to why a few epistatic simulations result in $\omega_* > 1$, but we suspect that the greater variance in the results across replicates as the extent of epistasis increases comes from the fact that a higher degree of epistasis requires one to simulate more contacts within the contact map protein structure representation; there are many more possible contact maps when 4% of each possible pair of amino acid positions are in contact than when only one 1% of pairs are. Given the random simulation of contact maps and statistical potentials, these simulations are not necessarily representative of epistasis at large. Rather, they show that, in principle, epistasis can occasionally lead to an $\omega_* > 1$, but that it's overall tendency will be to produce $\omega_* < 1$.

Empirical Data

As example applications, six empirical data sets, corresponding to 6 genes sampled in placental mammals (S1Pr1, RBP3, VWF, TRIM5α, SAMHD1, and BRCA1) were analyzed with the models of interest herein (table 1). The MG model leads to posterior distributions of ω that are well below 1, with $p(\omega > 1|D) \sim 0$ in all cases. As noted on the simulations, the MutSelYN model has little impact, and leads to ω_* values that hardly differ from the ω of the MG model. The MutSelDP model with ω_* is able to detect a signal of adaptive evolution in the TRIM5α, SAMHD1, and BRCA1 genes, which all lead to posterior distributions of ω_* beyond 1; as shown in table 1, their posterior probability of ω_* being greater than 1 is essentially 1 in these three cases. Interestingly, TRIM5α and SAMHD1 are known to be involved in the immune response against retroviruses in primates (Lee and KewalRamani 2004; Laguette et al. 2012; Zheng et al. 2012). In the case of BRCA1, a Red-Queen-like evolutionary regime is also suspected, possibly due to antagonistic selection between mother and offspring (Crespi and Summers 2004). The other three data sets (VWF, RPB, and S1PR1) either have ω_* below 1, or include 1 in its 95% credibility interval.

When analyzed with the site models, only the TRIM5α, SAMHD1, and BRCA1 data sets show evidence of adaptive evolution. Thus, in the present case, site-models would detect the same protein-coding genes as being under adaptation as would MutSelDP. On the other and, for the three proteins thus selected, the M8 model estimates that 13%, 8%, and 11% of sites, respectively, for TRIM5α, SAMHD1, and BRCA1, are under adaptive evolution. Given that these percentages are around the same as those found with site models in some of the most extreme Red-Queen adaptive simulations applied to 100% of sites, it seems plausible that a large proportion of adaptive evolution may remain undetected by the classical approaches.

Table 2. Number of Replicates among sets of 100 Where $p(\omega_* > 1 |D) \geq 0.95$ (and ≥ 0.99)

	Samhd1-67-543	Trim5α-68-363	Brca1-64-941
Nearly-neutral	0 (0)	1 (0)	2 (0)
Epistatic			
1	0 (0)	0 (0)	0 (0)
2	0 (0)	0 (0)	0 (0)
3	0 (0)	0 (0)	3 (3)
4	0 (0)	3 (3)	1 (1)
Adaptive			
25	95 (89)	97 (95)	96 (85)
50	100 (100)	100 (100)	100 (100)
75	100 (100)	100 (100)	100 (100)
100	100 (100)	100 (100)	100 (100)

Conclusions and Future Work

Here, we introduce a codon model framework relying on a fine-grained mechanistic model of the nearly neutral regime acting on protein coding sequences. A deviation parameter, ω_*, is introduced, such that upward deviations of ω_* away from 1 of are suggestive of the presence of an ongoing Red-Queen-like adaptive regime (diversifying selection). Technically, this codon framework can be used as a test for the presence of adaptation, by uncovering genes such that the posterior probability that $\omega_* > 1$ is sufficiently high. Being based on a more realistic null model of neutrality, our approach has the potential to be more powerful than classic codon models. This is particularly apparent in simulation conditions where adaptation occurs on a background of strong purifying selection (table 2).

More generally, the simulations performed here suggest that, if conducted at a 95% posterior probability threshold, the resulting test has a good power (with only 16 out of 1,200 true positive cases, i.e., about 1.3%, missed by the test), while having a good control of the rate of false positives (<1% on genes evolving under a pure nearly-neutral regime or when epistatic interactions are explicitly accounted for in the simulation model). Indeed, results under the pure nearly neutral simulations indicate a tendency to under-estimate ω_*. The Dirichlet process may not always capture the distribution of amino acid profiles across sites sufficiently well, which would most likely lead it to an under-estimate of the intensity of purifying selection, or, in other words, an over-estimate of ω_0. One modeling avenue to explore this issue would be to extend the base distribution of the Dirichlet process, perhaps so as to be a mixture itself.

It will also be important to investigate many more simulation conditions, such as varying a relative effective population size over the branches of the tree, introducing codon usage bias, or context-dependent mutation rates, in order to better understand the possible reasons for results where $\omega_* \neq 1$. Moreover, combining these simulation complexities together, or even just combining the different simulation conditions explored herein, would be revealing; given our results, one can readily imagine that a sort of tug-of-war between adaptive Red-Queen-like regimes across certain sites, and epistatic interactions across certain sites, such that, for instance, $\omega_* \sim 1$. One would thus be mislead into thinking that the evolutionary regime is a nearly neutral one.

Another extension from this point would be to combine the heterogeneous modeling ideas of site models to the new ω_* parameter within the MutSelDP. Indeed, Rodrigue (2008) discussed the idea of a model that combined two independent Dirichlet processes: one to account for heterogeneous amino acid profiles across sites, and another to account for heterogeneous ω_* parameters across sites. Note that such a modeling framework differs from the site-specific ML ω_* of Bloom (2016). In any case, approaches to heterogeneous modeling of ω_* within a rich mutation–selection framework seem much more accessible than explicitly accounting for the features we introduced within the simulation conditions (Robinson et al. 2003; Blanquart and Lartillot 2008; Rodrigue et al. 2009), and would enable a powerful analysis. The present work could act as a stepping-stone toward a re-appraisal of evolutionary regimes discernible from interspecific molecular data.

Materials and Methods

Data

We used six protein-coding genes at the scale of placental mammals: Rbp3-54-412: retinol-binding protein 3 (former IRBP gene), 54 taxa; S1pr1-67-325: sphingosine-1-phosphate receptor 1, 67 taxa; Vwf-62-392: von Willebrand factor, 62 taxa; Brca1-64-941: breast cancer 1, 64 taxa; Trim5α-68-363: tripartite motif-containing protein 5, 68 taxa; a virus restriction factor; Samhd1-67-543: SAM domain and HD domain-containing protein, 67 taxa; cellular enzyme responsible for blocking retroviral replication.

For Samhd1, Trim5α, placental mammalian sequences were retrieved from GenBank, translated and aligned using Muscle (Edgar 2004). The protein alignment was used as a template to align the nucleotide sequences whereas respecting the coding structure. Finally, the sequences were filtered using Gblocks (Castresana 2000), with the default options. For all other datasets, the alignments were obtained from Lartillot and Delsuc (2012).

Simulations

We simulated the evolution of sequences of length $N = 300$ codons, along a pre-specified tree, with 38 tips. The simulation model is parameterized by a nucleotide mutation process and a fitness landscape defined at the protein level (the model assumes no selection on synonymous variants). Three alternative models of the fitness landscape were considered:

Nearly neutral and multiplicative. In the absence of epistasis, and if constant through time, the fitness landscape is entirely characterized by the fitness of each amino-acid at each position. Each site is thus endowed with a 20-dimensional vector, w, specifying a fitness profile. Since fitnesses are relative, w profiles are by convention normalized to sum to 1. We used three different sets of profiles, which we obtained by running the model of Equation (2) on the Samhd1- *67-543*, Trim5α- *68-363*, and Brca1- *64-941* data sets. In each case the site-specific posterior mean amino acid profiles were then computed. With all of these values at hand, each coding site of the simulated sequence is attributed an amino-acid profile selected at random (with replacement) from one of the three sets of possible profiles. Let us call $F^{(n)}(a) = \ln w_n(a)$, the fitness of amino-acid a at position n induced by this random set of profiles, and let $s = (s_n)_{n=1..N}$ be an amino-acid sequence (with $s_n = 1 \ldots 20$ for each position n). The fitness of the entire sequence s is then given by:

$$F(s) = \ln W(s) = \sum_{n=1}^{N} \ln w_n(s_n).$$

Nearly neutral with epistatic interactions. Epistatic interactions are introduced on the top of the site-specific fitness profiles introduced above, as follows: a pre-specified proportion of pairwise contacts is chosen uniformly at random across all possible contacts. For each contact belonging to this subset, the following is simulation is applied: a series of 210 normal variates of mean 0 and standard deviation σ_{epi} are drawn, so as to define a contact potential, say between positions m and n, denoted $\epsilon_{mn}(a, b)$, with $a, b = 1..20$ running over all possible pairs of amino-acids. For all other pairs of positions, we set $\epsilon_{mn}(a, b) = 0$ for all a and b. Once this pairwise contact potential is defined, the fitness of an amino-acid sequence s of length N is given by:

$$F(s) = \ln W(s) = \sum_{n=1}^{N} \ln w_n(s_n) + \sum_{1 \le m < n \le N} \epsilon_{mn}(s_m, s_n).$$

Fluctuating, adaptive Red-Queen fitness landscape. In this regime, the fitness landscape is allowed to fluctuate through time, according to a Markov-modulated process. This process is characterized by its rate of change ρ, a standard deviation for the fluctuations σ_{RQ}, and a number of hidden states K per fluctuating site; here, we set $K = 2$. For a given fluctuating site, say n, each hidden state $k = 1..K$ defines a modulating profile $h_{nk}(a)$, where $a = 1 \ldots 20$ is running over the 20 amino-acids. Here, only two entries in the two modulating profiles have non-zero values. Specifically, the entries corresponding to the two highest fitness amino acids in the starting empirical profile, say a and b, are set as $h_{n1}(a) = \delta$, $h_{n1}(b) = -\delta$, $h_{n2}(a) = -\delta$, and $h_{n2}(b) = \delta$, where δ is a random variable drawn from of normal distribution of mean 0 and standard deviation $\sigma_{RQ} = 6$. At each fluctuating site, the hidden state $k_n(t)$ is time-dependent and evolves according to a simple Jukes–Cantor model of rate ρ; we refer to ρ as the *rate of the Red Queen*, for which we explored four values (as displayed in figure 2). Then, the fitness of a sequence s at time t is given by:

$$F(s) = \ln W(s) = \sum_{n=1}^{N} \ln w_n(s_n) + h_{nk_n(t)}(s_n).$$

The mutation process is assumed to be strand-symmetric. Accordingly, it is characterized by six relative mutation rates (given below). An absolute mutation rate is also defined here, scaling the rates below by 2×10^{-4}.

$$\mu = \left(\begin{array}{ccccc} & A & C & G & T \\ A & - & \mu_{A:T>C:G} = 3.16 & \mu_{A:T>G:C} = 8.01 & \mu_{A:T>T:A} = 3.26 \\ C & \mu_{C:G>A:T} = 2.29 & - & \mu_{C:G>G:C} = 2.18 & \mu_{C:G>T:A} = 5.75 \\ G & \mu_{C:G>T:A} = 5.75 & \mu_{C:G>G:C} = 2.18 & - & \mu_{C:G>A:T} = 2.29 \\ T & \mu_{A:T>T:A} = 3.26 & \mu_{A:T>G:C} = 8.01 & \mu_{A:T>C:G} = 3.16 & - \end{array}\right).$$

The overall simulation process is Markovian. The events can be either a point substitution (to any one of all possible single-nucleotide mutants away from the current sequence, except those resulting in a premature stop codon) or, in the case of the fluctuating fitness regime, a modulation of the fitness parameters at any of the fluctuating sites (i.e., a transition undergone by the hidden state at that site). Given the current sequence at time *t*, and given the hidden states at all fluctuating sites, the rate associated to each possible event is calculated. In the case of substitution events, these rates are given by the mutation rate multiplied by the fixation factor, which itself depends on the fitness of the final and the initial sequence variant (Halpern and Bruno 1998). Note that, in the presence of epistasis, the fitness of each single-mutant depends on the sequence at all other positions. The total rate R_{tot}, summed over all possible events at time *t*, is calculated, and the time until the next event is randomly drawn from an exponential distribution of rate R_{tot}. Then, the exact nature of the next event is chosen with a probability equal to the relative rate of this event (from the mutation rate, for synonymous mutations, or from the product of the mutation rate and the fixation factor, for non-synonymous cases). Whenever the waiting time to the next event exceeds the amount of time remaining until the next branching event along the phylogenetic tree, the simulation process is started with the current state, independently along each of the two daughter branches. The procedure is started at a time $T = 100$ time units before the root (so as to ensure that the process has reached stationarity before starting from the root) and is propagated forward in time down to all tips of the tree.

Priors and Implementation

We used the same priors as in previous works (Lartillot et al. 2013; Rodrigue and Lartillot 2014):

Branch lengths are i.i.d. exponential of rate λ, itself exponential of rate 0.1.

We use a Dirichlet process over amino-acid fitness profiles, with base distribution a Dirichlet(α_i), where the α_i are i.i.d. exponential of rate 1.

The granularity parameter of the Dirichlet process is exponential of rate 0.1 (mean 10).

Nucleotide exchangeability parameters and nucleotide frequency parameters are each flat Dirichlets.

Non-synonymous rate factors ω and ω_* are ratios of two exponential random variables (Huelsenbeck et al. 2006).

The use of PhyloBayes-MPI with the mutation–selection model is explained within the online manual, and activating the ω_* parameter is done by adding the option -freeomega to the command. To obtain the plain MG model, the options -freeomega and -catfix uniform are applied, whereas to obtain the MutSelYN model, the options -freeomega, -rigidbaseprior, and -ncat 1 are applied. For simulated data, inferences based on MCMC calculations were conducted under fixed tree topology, as originally used for the simulations, and were run for 1,100 cycles, discarding the first 100 as burn-in. Note that each cycle itself includes hundreds of Gibbs and Metropolis-Hastings updates within PhyloBayes-MPI. Real data analyses were run with 5,500 cycles (500 as burn-in), treating the topology (with uniform priors) as a nuisance variable of the inference. Source code is freely available within the PhyloBayes-MPI package, distributed at www.phylobayes.org.

Acknowledgments

We thank Jesse Bloom and an anonymous reviewer for constructive comments on the manuscript. This work was supported by the Natural Sciences and Engineering Research Council of Canada (NR), by a start-up fund from Carleton University (NR), and by the French National Research Agency, Grant ANR-10-BINF-01-01 “Ancestrome” (NL).

References

Ashenberg O, Gong LI, Bloom JD. 2013. Mutational effects on stability are largely conserved during protein evolution. *Proc Natl Acad Sci U S A.* 110:21071–21076.

Bazykin GA. 2015. Changing preferences: deformation of single position amino acid fitness landscapes and evolution of proteins. *Biol Lett.* 11: 20150315.

Blanquart S, Lartillot N. 2008. A site- and time-heterogeneous model of amino acid replacement. *Mol Biol Evol.* 25:842–858.

Bloom J. 2016. Identification of positive selection in genes is greatly improved by using experimentally informed site-specific models. *bioRxiv* URL http://www.biorxiv.org/content/early/2016/01/22/037689.

Bloom JD. 2014. An experimentally informed evolutionary model improves phylogenetic fit to divergent lactamase homologs. *Mol Biol Evol.* 31:2753–2769.

Castresana J. 2000. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Mol Biol Evol.* 17:540–552.

Clark AG, Glanowski S, Nielsen R, Thomas PD, Kejariwal A, Todd MA, Tanenbaum DM, Civello D, Lu F, Murphy B, et al. 2003. Inferring nonneutral evolution from human-chimp-mouse orthologous gene trios. *Science* 302:1960–1963.

Crespi B, Summers K. 2004. In defense of the cell: TRIM5alpha interception of mammalian retroviruses. *Proc Natl Acad Sci U S A.* 101:10496–10497.

Cutler DJ. 2000. Understanding the overdispersed molecular clock. *Genetics* 154:1403–1417.

Echave J, Spielman SJ, Wilke CO. 2016. Causes of evolutionary rate variation among protein sites. *Nat Rev Genet.* 17:109–121.

Edgar, RC. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res*. 32:1792–1797.

Eyre-Walker A, Keightley PD. 2009. Estimating the rate of adaptive molecular evolution in the presence of slightly deleterious mutations and population size change. *Mol Biol Evol*. 26:2097–2108.

Galtier N. 2016. Adaptive protein evolution in animals and the effective population size hypothesis. *PLoS Genet*. 12:e1005774.

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

Gong LI, Bloom JD. 2014. Epistatically interacting substitutions are enriched during adaptive protein evolution. *PLoS Genet*. 10:e1004328.

Guindon S, Rodrigo AG, Dyer KA, Huelsenbeck JP. 2004. Modeling the site-specific variation of selection patterns along lineages. *Proc Natl Acad Sci U S A*. 101:12957–12962.

Halligan DL, Oliver F, Eyre-Walker A, Harr B, Keightley PD. 2010. Evidence for pervasive adaptive protein evolution in wild mice. *PLoS Genet*. 6:e1000825.

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

Holder MT, Zwickl DJ, Dessimoz C. 2008. Evaluating the robustness of phylogenetic methods to among-site variability in substitution processes. *Phil Trans R Soc B* 363:4013–4021.

Huelsenbeck JP, Jain S, Frost SWD, Pond SLK. 2006. A Dirichlet process model for detecting positive selection in protein-coding DNA sequences. *Proc Natl Acad Sci U S A*. 103:6263–6268.

Keightley PD, Eyre-Walker A. 2007. Joint inference of the distribution of fitness effects of deleterious mutations and population demography based on nucleotide polymorphism frequencies. *Genetics* 177:2251–2261.

Keightley PD, Eyre-Walker A. 2010. What can we learn about the distribution of fitness effects of new mutations from DNA sequence data? *Philos Trans R Soc Lond B Biol Sci*. 365:1187–1193.

Kimura M. 1983. *The neutral theory of molecular evolution*. Cambridge: Cambridge University Press.

Kosiol C, Vinar T, da Fonseca RR, Hubisz MJ, Bustamante CD, Nielsen R, Siepel A. 2008. Patterns of positive selection in six Mammalian genomes. *PLoS Genet*. 4:e1000144.

Laguette N, Rahm N, Sobhian B, Chable-Bessia J, Münch C, Snoeck J, Sauter D, Switzer WM, Heneine W, Kirchhoff F, et al. 2012. Evolutionary and functional analyses of the interaction between the myeloid restriction factor SAMHD1 and the lentiviral Vpx protein. *Cell Host Microbe* 11:205–217.

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

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

Lee K, KewalRamani VN. 2004. In defense of the cell: TRIM5alpha interception of mammalian retroviruses. *Proc Natl Acad Sci U S A*. 101:10496–10497.

Lunzer M, Golding GB, Dean AM. 2010. Pervasive cryptic epistasis in molecular evolution. *PLoS Genet*. 6:e1001162.

McCandlish DM, Rajon E, Shah P, Ding Y, Plotkin JB. 2013. The role of epistasis in protein evolution. *Nature* 497:E1–2, discussion E2–3.

McDonald JH, Kreitman M. 1991. Adaptive protein evolution at the Adh locus in Drosophila. *Nature* 351:652–654.

Muse SV, Gaut BS. 1994. A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates, with application to the chloroplast genome. *Mol Biol Evol*. 11:715–724.

Mustonen V, Lässig M. 2009. From fitness landscapes to seascapes: non-equilibrium dynamics of selection and adaptation. *Trends Genet*. 25:111–119.

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

Pollock DD, Thiltgen G, Goldstein RA. 2012. Amino acid coevolution induces an evolutionary Stokes shift. *Proc Natl Acad Sci U S A*. 109:E1352–E1359.

Pond SLK, Murrell B, Fourment M, Frost SDW, Delport W, Scheffler K. 2011. A random effects branch-site model for detecting episodic diversifying selection. *Mol Biol Evol*. 28:3033–3043.

Robinson DM, Jones DT, Kishino H, Goldman N, Thorne JL. 2003. Protein evolution with dependence among codons due to tertiary structure. *Mol Biol Evol*. 18:1692–1704.

Rodrigue N. 2008. Phylogenetic structural modeling of molecular evolution. Doctoral dissertation, Université de Montréal, Canada.

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

Rodrigue N, Kleinman CL, Philippe H, 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, Lartillot N. 2014. Site-heterogeneous mutation-selection models within the phylobayes-mpi package. *Bioinformatics* 30:1020–1021.

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

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

Sawyer SA, Hartl DL. 1992. Population genetics of polymorphism and divergence. *Genetics* 132:1161–1176.

Sawyer SA, Kulathinal RJ, Bustamante CD, Hartl DL. 2003. Bayesian analysis suggests that most amino acid replacements in Drosophila are driven by positive selection. *J Mol Evol*. 57 Suppl 1:S154–S164.

Sawyer SL, Emerman M, Malik HS. 2004. Ancient adaptive evolution of the primate antiviral DNA-editing enzyme APOBEC3G. *PLoS Biol*. 2:E275.

Sawyer SL, Wu LI, Emerman M, Malik HS. 2005. Positive selection of primate TRIM5alpha identifies a critical species-specific retroviral restriction domain. *Proc Natl Acad Sci U S A*. 102: 2832–2837.

Shah P, McCandlish DM, Plotkin JB. 2015. Contingency and entrenchment in protein evolution under purifying selection. *Proc Natl Acad Sci U S A*. 112:E3226–E3235.

Spielman SJ, Wilke CO. 2015. The relationship between dN/dS and scaled selection coefficients. *Mol Biol Evol*. 32:1097–1108.

Tamuri AU, Goldman N, dos Reis M. 2014. A penalized likelihood method for estimating the distribution of selection coefficients from phylogenetic data. *Genetics* 197:257–271.

Tamuri AU, dos Reis M, Goldstein RA. 2012. Estimating the distribution of selection coefficients from phylogenetic data using sitewise mutation-selection models. *Genetics* 190:1101–1115.

Thorne JL, Lartillot, N, Rodrigue N, Choi SC. 2012. Codon models as a vehicle for reconciling population genetics with inter-specific sequence data. In: Cannarozzi GM, Schneider A, editors. Codon evolution. Oxford: Oxford University Press. p. 97–110.

Weinreich DM, Knies JL. 2013. Fisher's geometric model of adaptation meets the functional synthesis: data on pairwise epistasis for fitness yields insights into the shape and size of phenotype space. *Evolution* 67:2957–2972.

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

Yang Z, Nielsen R. 1998. Synonymous and nonsynonymous rate variation in nuclear genes of mammals. *J Mol Evol*. 46:409–418.

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

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

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

Yang Z, Swanson WJ. 2002. Codon-substitution models to detect adaptive evolution that account for heterogeneous selective pressures among site classes. *Mol Biol Evol*. 19:49–57.

Yang Z, Wong WSW, Nielsen R. 2005. Bayes empirical Bayes inference of amino acid sites under positive selection. *Mol Biol Evol*. 22:1107–1118.

Zhang J, Nielsen R, Yang Z. 2005. Evaluation of an improved branch-site likelihood method for detecting positive selection at the molecular level. *Mol Biol Evol*. 22:2472–2479.

Zheng YH, Jeang KT, Tokunaga K. 2012. Host restriction factors in retroviral infection: promises in virus-host interaction. *Retrovirology* 9:112.