September 6, 2020

A Bayesian mutation-selection framework for detecting site-specific adaptive evolution in protein-coding genes

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

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

[2]Université de Lyon, Université Lyon 1, CNRS; UMR 5558, Laboratoire de Biométrie et Biologie Évolutive, F-69622 Villeurbanne, France

Running head: Detecting site-specific adaptation with mutation-selection models

Keywords: Nearly-neutral evolution; fitness landscape; Dirichlet process; Markov chain Monte Carlo.

Correspondence: Nicolas Rodrigue

209 Nesbitt Biology Building,

1125 Colonel By Drive Ottawa, Ontario, CANADA

K1A 0C6

nicolas.rodrigue@carleton.ca

tel: +1 613 520 2600 x 4194

Abstract

In recent years, codon substitution models based on the mutation-selection principle have been extended for the purpose of detecting signatures of adaptive evolution in protein-coding genes. However, the approaches used to date have either focused on detecting global signals of adaptive regimes—across the entire gene—or on contexts where experimentally derived site-specific amino acid fitness profiles are available. Here, we present a Bayesian site-heterogeneous mutation-selection framework for site-specific detection of adaptive substitution regimes given a protein-coding DNA alignment. We offer implementations, briefly present simulation results, and apply the approach on a few real data sets. Our analyses suggest that the new approach shows greater sensitivity than traditional methods. However, more study is required to assess the impact of potential model violations on the method, and gain a greater empirical sense its behaviour on a broader range of real data sets. We propose an outline of such a research program.

Introduction

Codon substitution models (Goldman and Yang, 1994; Muse and Gaut, 1994) are among the important modern tools used for uncovering potential signals of molecular adaptation from protein-coding gene alignments. One set of broadly used models focuses on estimating the ratio of rates of nonsynonymous (dN) and synonymous (dS) substitutions. These models introduce a multiplicative parameter, denoted ω, to entries in a codon substitution matrix corresponding to nonsynonymous events. Because ω is the only distinction between the rate specification of nonsynonymous and synonymous events, it directly corresponds to $\omega = dN/dS$.

Fitting a model with a single (global) nonsynonymous multiplicative parameter almost always leads to $\omega < 1$ (Yang, 2006), given the pervasive purifying selection that operates at most codon sites over most of evolutionary history. Many efforts were thus made to develop codon substitution models with distributions of ω values across sites and/or across the branches of a phylogeny (reviewed in Yang, 2019). A common objective of such developments is to uncover specific sites having evolved under an adaptive regime (e.g., with $\omega > 1$), perhaps along a particular branch of the phylogeny.

Meanwhile, another set of codon substitution models was proposed, with a focus on accounting for purifying selection at the amino acid level in a site-heterogeneous manner (Halpern and Bruno, 1998). Having nucleotide-level parameters controlling a mutational process, and amino acid fitness profiles controlling selection, they have come to be known as *mutation-selection* models (e.g., Yang and Nielsen, 2008; Rodrigue et al., 2010). In these models, the dN/dS ratio is not explicitly parameterized. Instead, it is an emerging quantity, induced by the interplay between mutation, selection, and drift. Spielman and Wilke (2015) have shown how to calculate the dN/dS induced by the mutation-selection framework—which we denote ω_0 (Rodrigue and Lartillot, 2017)—and found that, under specific conditions (i.e., a substitution process at equilibrium, without selection on synonymous variants), it is always true that $\omega_0 \leq 1$, as expected from a model focused on purifying selection.

In the last few years, the mutation-selection framework has been extended for the purpose of detecting genes having evolved under an adaptive regime, in either a global (Rodrigue and Lartillot, 2017) or site-specific (Bloom, 2017) manner. Like their traditional predecessors, these recent mutation-selection models introduce a multiplicative parameter on nonsynonymous rates. However, because amino acid profiles are also involved in modulating nonsynonymous rates, such a multiplicative parameter—which we denote as ω_* (Rodrigue and Lartillot, 2017)—cannot be interpreted as the dN/dS ratio; we chose to emphasize this distinction with an asterix in the notation. Given that the mutation-selection formulation itself induces a certain dN/dS ratio, ω_0, the net overall dN/dS ratio, ω, can be thought of as $\omega = \omega_0 \times \omega_*$, which can be rearranged to $\omega_* = \omega/\omega_0$. The latter equation helps clarify the interpretation of ω_* as a measure of the deviation in nonsynonymous rates from the expectation under the pure mutation-selection equilibrium; in particular, $\omega_* > 1$ indicates that nonsynonymous rates are higher than expected, even though they might not be so high as to lead to $\omega > 1$.

New Approaches

Here, we conduct a first exploration of a Bayesian mutation-selection model with site-heterogeneous amino acid fitness profiles and site-heterogeneous ω_* values. The Bayesian nature of the model qualifies it as a *random-effects* approach, in contrast to the *fixed-effects* approach utilized to date in maximum-likelihood versions of mutation-selection models (Halpern and Bruno, 1998; Holder et al., 2008; Tamuri et al., 2014; Bloom, 2017).

Results and Discussion

Models With Global ω or ω_*

We first contrasted the difference in behaviour between a traditional codon substitution model inspired from Muse and Gaut (1994), with a global ω parameter (a traditional model we denote MG-M0, described in detail in the Materials and Methods section), and a mutation-selection model with a Dirichlet process prior on amino acid profiles across sites and a global ω_* parameter (a model presented in Rodrigue and Lartillot, 2017, which we denote here as MutSel-M0$_*$, and also described in the Materials and Methods section).

Simulations

Figure 1 shows results of the two models on data generated through a simulation approach explicitly allowing for fluctuating selection at some sites; for these sites, amino acid fitness profiles change along the branches of the phylogeny, as described in Rodrigue and Lartillot (2017, and in the Materials and Methods section). The simulation system is an attempt at mimicking an adaptive substitution process, where the simulated substitution history tracks a changing amino acid fitness optimum along the branches of the tree, and thus accrues more nonsynonymous substitutions than expected under a pure nearly-neutral regime (i.e., mutation-selection balance). An important distinction with Rodrigue and Lartillot (2017) is that here the simulated data set contains only 10% of codon sites generated under adaptive regimes, and 90% of codon sites generated under a pure nearly-neutral mutation-selection formulation (Rodrigue et al., 2010). We produced alignments of 300 codons in length, repeating the simulation thrice, with different sets of empirically inferred amino acid profiles (see Lowe and Rodrigue, 2020, and the Materials and Methods section).

Results under the traditional MG-M0 model (red) reflect the overall purifying selection governing most of the data-generating processes, with posterior mean ω values at 0.14, 0.15, 0.13 in three replicates displayed in panels 1A, 1B, and 1C respectively. The fact that 10% of sites where produced under an adaptive regime is underwhelming to the MG-M0 model, and indeed little is generally expected of it in practice. Results under the MutSel-M0$_*$ model (blue) show a posterior distribution for ω_* situated above 1, with $p(\omega_* > 1 \mid D) \geq 0.99$ (where D refers to the data set) for the first two replicates (fig. 1A and 1B); indeed, the second replicate has a posterior mean that surpasses 2. For the third replicate, we find a slightly lower probability, at $p(\omega_* > 1 \mid D) \sim 0.93$, still highly suggestive of a signal for adaptive evolution.

Previous studies (Rodrigue and Lartillot, 2017; Lowe and Rodrigue, 2020) have shown that a simulation conducted with 100% of sites under the pure nearly-neutral mutation-selection formulation leads to a posterior distribution of ω_* situated around 1 (while the ω parameter inferred under the MG-M0 model on such simulated data tends to be closer to 0 than to 1, as shown in Rodrigue and Lartillot, 2017). Here, however, 10% of sites have evolved with higher than expected nonsynonymous rates, which pulls the distribution of ω_* to the right. Already with the use of single additional parameter, ω_*, the mutation-selection framework allows us to detect adaptation where the traditional framework with a single ω parameter would not. Note that these results are under ideal conditions, however, free of the numerous potential model violations present in real data that could sway inferences of ω_*.

Real Data

Figure 2 shows results of these models on a hand-full of real alignments. Panel 2A displays the results on the well-known β-Globin alignment sampled across 17 vertebrates (Yang et al., 2000). As in the simulation, the MG-M0 model indicates that $\omega < 1$. In contrast, under MutSel-M0$_*$, the posterior mean of ω_* is around 1.8, with a high posterior probability in favor of a value greater than 1, $p(\omega_* > 1 \mid D) > 0.99$, suggesting the presence of adaptive evolution in this gene. As described with the simulation experiment presented above, and assuming negligible effects of potential model violations, adaptive evolution on even a relatively small fraction of the sites of the gene could be sufficient to induce such a rightward shift in the posterior distribution of ω_*.

Panel 2B displays results on an alignment of the alcohol dehydrogenase (ADH) gene sampled across 23 species of *Drosophila*. Here again, the MG-M0 model indicates that $\omega < 1$, with a posterior mean $\sim$0.13. In contrast, with the MutSel-M0$_*$ model we find a posterior mean ω_* $\sim$1.2, and $p(\omega_* > 1 \mid D) > 0.95$. As for the β-GLOBIN alignment, and again assuming no major effects from potential model violations, this result could be explained by a fraction of sites evolving under adaptive evolution regimes. No previous phylogenetic approach has found signals of adaptive evolution in this gene, in spite of the fact that population-genetic approaches have long suggested adaptation in many instances (e.g., McDonald and Kreitman, 1991; Matzkin and Eanes, 2003; Matzkin, 2004). While a specific scenario of ADH adaptation in specific species has been refuted by Siddiq and Thornton (2019), their study nonetheless provides strong experimental evidence of major fitness effects of some mutations, suggesting adaptive opportunities across *Drosophila*.

The four lower panels of figure 2 (2C, 2D, 2E, and 2F) show results on four genes sampled across placental mammals (Lartillot and Delsuc, 2012). Again, $\omega < 1$ in all four genes, whereas ω_* is either around 1, or slightly below, which does not suggest adaptive evolution in these genes across placental mammals. This does not rule out the possibility that some of these genes have some sites under adaptive evolution, but perhaps these sites are too few and/or too mildly adaptive to raise ω_* beyond 1. Previous simulations studies have pointed to epistasis (Rodrigue and Lartillot, 2017) or weak evolutionary signal (Lowe and Rodrigue, 2020) as potential reasons for $\omega_* < 1$. In the absence of major effects from model violations, these are conditions that tend to make the model conservative in the detection of adaptive regimes.

Models With Heterogeneous ω or ω_*

In spite of the potential of the MutSel-M0$_*$ model—able to capture relatively subtle signals of adaptive evolution—it still does not directly allow us to pinpoint which sites are most responsible for such signals. This is one of the motivations of *site-models*. Classical site-models (Nielsen and Yang, 1998; Yang et al., 2000; Yang and Swanson, 2002) consider alignment sites as having been produced from a distribution of possible ω values. They are typically used in the context of an empirical Bayes approach for identifying sites with a strong statistical support for a $\omega > 1$; and they are more efficient at detecting positive selection than the simple MG-M0 model with a single ω for all sites. For instance, they do find sites under positive selection in the case of the β-Globin gene (detailed below, but also see Yang et al., 2000). On the other hand, site-models might still miss those sites under weaker positive selection. In particular, an adaptive regime at a site could be sufficiently strong to increase the dN/dS ratio, but not to the point of driving it well above 1. In other words, at least in their current version, these models might present the same limitation as the classical MG-M0 model, as compared with MutSel-M0$_*$ model, although now at the level of the single site. This in turn suggests that the rationale of estimating ω_* in the context of a mutation-selection model should be explored not just globally over the whole gene (Rodrigue and Lartillot, 2017), but as a distribution across sites of the gene (Bloom, 2017).

To illustrate this point, and for simplicity here, we work with the classical MG-M3 model, inspired from Muse and Gaut (1994) and Yang et al. (2000), which invokes a finite mixture of three ω values—with their respective weights—jointly estimated with all other parameters given the data. We also study a new model referred to as MutSel-M3$_*$, which is built from a finite mixture of three ω_* values, and respective weights, combined with the Dirichlet process prior on amino acid profiles across sites, and global mutational parameters. The two forms of across-site heterogeneity are independent in the model construction, in that each site draws its amino acid profile and its ω_*

independently from the two corresponding mixtures.

Simulations

As a verification, figure 3 shows the results under the MG-M3 (red) and MutSel-M3$_*$ (blue) models on three simulated data sets, this time generated entirely under the pure mutation-selection framework (i.e., no adaptive regimes within the data-generating processes). In accordance with the simulation, no sites have high probabilities of having $\omega_* > 1$ (or $\omega > 1$). Most sites have posterior probabilities of $\omega_* > 1$ ranging from 0 to 0.5, or not much more, suggesting that the MutSel-M3$_*$ model tends to mildly under-estimate some site-specific ω_* values. One possible reason for such under-estimates is the fact that, in its current form, the mutation-selection apparatus utilized tends to over-estimate ω_0 (the nonsynonymous to synonymous rate ratio *induced* by the amino acid fitness profiles), as shown by Spielman and Wilke (2015). Overall, however, if the data-generating process doesn't depart too drastically from the model's assumptions, this behaviour tends to make MutSel-M3$_*$ conservative vis-à-vis inferences of adaptive evolution.

These simulations also highlight an inherent risk built into the MutSelM3$_*$ model's construction, in comparison with MG-M3: the threshold for a site to considered of interest—in terms of potential adaptive evolution—is much closer to the value expected under the null (of no adaptive regime) under MutSelM3$_*$ than under MG-M3, with the latter reporting site-specific probabilities of having $\omega > 1$ that are close to 0; for the second replicate in particular (fig. 3B), $p(\omega > 1 \mid D)$ never surpasses 0.007. In other words, finding a site with $p(\omega > 1 \mid D) > 0.95$ under the MG-M3 model represents a dramatic increase in nonsynonyomous rate, compared to finding one with $p(\omega_* > 1 \mid D) > 0.95$ under the MutSelM3$_*$ model, which could make MutSelM3$_*$ more vulnerable to false positives from stochastic effects, or from the effects of model violations.

Figure 4 shows the results on the three simulated data sets studied in figure 1 (i.e., with 10% of sites simulated with an adaptive regime). The panels include vertical marks at the top, showing the 30 codon sites simulated under adaptive regimes. Sites evolving under an adaptive regime tend to accrue more nonsynonymous substitutions than under a nearly-neutral regime, which would shift ω_* to the right of the unit. With a threshold posterior probability of 0.95 for $p(\omega_* > 1 \mid D)$, the MutSel-M3$_*$ model correctly identifies 23/30 sites (76%), calls 1 false positive, and misses 7 sites for the first and second replicates, whereas for the third replicate it correctly identifies 20/30, with no false positives. Of note, a single false positive out of 24 discoveries, using a threshold of 0.95, corresponds to an accuracy of ∼96%, thus suggesting that the posterior probabilities are reasonably well-calibrated, reflecting our actual rate of true discovery. The MG-M3 models identifies no sites at this threshold, although the plot suggests that it nonetheless faintly detects some adaptive signal. Interestingly, the sites leading to false positives under the MutSel-M3$_*$ model also tempt the MG-M3 model; the simulations are stochastic processes, and can, from time to time, accumulate a disproportionately high number of nonsynonymous substitutions, even when the configuration of the simulating model is one of pure mutation-selection balance. In other words, false positives may not come about solely as a result of a problem with MutSel-M3$_*$ model itself, but rather, at least partly, from a chance occurrence in the simulation. Still, this demonstrates the increased risk of the MutSelM3$_*$ model over MG-M3. However, the MG-M3 model also clearly lacks sensitivity; the sure way of having no false positives is to have no positives at all. It is particularly noteworthy that some of the sites correctly identified by MutSel-M3$_*$ show virtually no signal under MG-M3 (e.g., sites 52, 103, 285 in the first replicate, panel 4A). In contrast, all of the sites simulated with an adaptive regime but missing the 0.95 threshold under MutSel-M3$_*$ nonetheless have relatively high probabilities of having $\omega_* > 1$. Overall, under ideal conditions, the MutSel-M3$_*$ model seems to have considerably greater sensitivity than the traditional-style MG-M3, at the cost of a mildly increased risk of false positives.

Real Data

Figures 5 and 6 display the results obtained from analyzing the six real data sets mentioned above with the MG-M3 and MutSel-M3$_*$ models. For the β-Globin alignment (fig. 5A), our Bayesian version of the classic MG-M3 model leads to the same set of sites identified with these traditional models in the maximum likelihood context (Yang et al., 2000): at the 95% threshold, the sites are 7, 11, 42, 48, 50, 54, 67, 85, and 123. Under the MutSel-M3$_*$ model, these same sites are also found, and the following three are added: 10, 74, and 84. (The complete lists of sites identified at different thresholds are reported in table 1.) It is interesting to note that the MG-M3 model found $p(\omega > 1 \mid D) = 0.381$ for site 10, $p(\omega > 1 \mid D) = 0.244$ for site 74, and $p(\omega > 1 \mid D) = 0.074$ for site 84. These last three sites, and site 84 in particular, yield results compatible with the interpretation of having evolved under a mild adaptive regime, of changing amino acid fitness profiles over time, leading to an increase in nonsynonymous rate; the increase is not to the point where $\omega > 1$ at a site in question, although it is enough for $\omega_* > 1$. Sites 10 and 74 are known to be involved in oxygen affinity, which could indeed make them a target for adaptive evolution.

The sites uncovered by MutSel-M3$_*$ on the β-Globin data set are conditional on the overall construction of the model, which makes many over-simplified assumptions. As such, the list of sites should be considered provisional, in need of more thorough investigation by external means, and in the context of a larger scale application of the model. Some of the model violations potentially at play here, and that have mislead other types of approaches to detecting adaptive evolution, include variable effective population size (Rousselle et al., 2018), biased gene-conversion (Ratnakumar et al., 2010), multi-nucleotide mutations (Venkat et al., 2018), and non-homogeneous/non-neutral synonymous substitution rates (Wisotsky et al., in press). Richer simulation studies will be needed to better understand how the MutSel-M3$_*$ model reacts to such violations, and the extent to which they could be responsible for false positives.

Results of the analysis of ADH (fig. 5B, table 1) suggest several sites under adaptive evolution under the MutSel-M3$_*$ model, whereas the MG-M3 yields posterior probabilities of $\omega > 1$ at all sites that are numerically indistinguishable from 0. Given that most studies suggesting adaptation in this gene have relied on population-genetic methodologies, which pool the statistics across all sites, a comparison of sites uncovered by the MutSel-M3$_*$ model with previous results is not possible.

As with the analyses of the β-GLOBIN data set, much more work will be required to determine the plausibility of these new results on the ADH data set. In addition to the aforementioned potential model violations, with a sampling across *Drosophila*, which have high effective population sizes, features such as uneven codon usage can become highly pronounced (Powell and Moriyama, 1997), potentially mis-leading inferences of site-specific adaptation as well. As a hypothetical example, suppose that the codon TTG is used almost exclusively for encoding leucine, and that GTG is similarly strongly favored for encoding valine. Also suppose that leucine and valine are of equivalent fitness at a given site. In such a context, nonsynonymous substitutions between TTG and GTG accumulate more readily than synonymous substitutions. If this feature were to be present to a high extent, it could mislead the MutSel-M3$_*$ model into inferring $\omega_* > 1$, thus suggesting adaptive evolution where the regime is in fact one of strict purifying selection on codon usage. Simulations should eventually be used to study effects relevant to high effective population sets of taxa—such as codon usage—on the inferences of MutSel-M3$_*$.

Our analysis of the mammalian-level alignment of the gene VWF also suggests several sites with adaptive signatures under the MutSel-M3$_*$ model, and none under the MG-M3 model (fig. 5C, table 1). A previous study, utilizing branch-heterogeneous models, has suggested adaptive evolution conferring venom resistance to opposoms that prey on pitvipers (Jansa and Voss, 2011). Moreover, variants of this gene have been found to have dramatic effects on its own expression levels in mice (Lemmerhirt et al., 2006), and hence with high potential for strong fitness effects.

While these latter studies are precedents to finding sites with signatures of adaptive evolution under the MutSel-M3$_*$ model, many of the model violations mentioned above could apply here as well. At the mammalian scale of this Vwf data set, a mutation-selection-based test of selection on codon usage has been shown to be misled by the effect of CpG hypermutability (Laurin-Lemay et al., 2018). This context-dependent mutational feature could have the effect of inflating ω_* values beyond 1 at sites where there is in fact no adaptive evolution (Suzuki et al., 2009). Again, however, more simulation work is required to better understand how such issues play out with the MutSel-M3$_*$.

Of the remaining mammalian gene alignments studied with the MutSel-M3$_*$ model, two suggest very few sites having evolved under adaptive regimes (ADORA3 and S1PR1, in fig. 6A and 6C respectively), and one (RBP3, fig. 6B) with none. The traditional MG-M3 model suggests no sites under adaptive evolution for these data sets. These three genes may be typical of results under the MutSel-M3$_*$ model at the mammalian scale (i.e., few, if any sites with high $p(\omega_* > 1 \mid D)$), but broader empirical studies evaluating the relative proportion of genes with several sites having high probabilities of $\omega_* > 1$ are pressing.

Future directions

The traditional codon models based on ω have become increasingly well understood thanks to decades of empirical applications and simulation studies. A similar project should be considered within the mutation-selection framework. We have already suggested several lines of research meriting further attention, and we expand on these themes below.

Simulation studies

A flurry of recent research has shown how a variety of approaches are highly susceptible to model violations, with many instances of purported signals of molecular adaptation being the result of unaccounted features of the evolutionary process (e.g., Ratnakumar et al., 2010; Rousselle et al., 2018; Venkat et al., 2018; Laurin-Lemay et al., 2018; Wisotsky et al., in press). From the codon substitution modeling perspective, this raises important questions regarding the mutation-selection-based approach we propose here: whereas the biological expectation under traditional models is for ω values closer to 0 than to 1, such that $\omega > 1$ is a drastic threshold, representing a very pronounced increase in nonsynonymous rates, the biological expectation under the new approach is for ω_* values closer to 1, and thus naturally approaching threshold of $\omega_* > 1$. This could make the mutation-selection-based methods highly susceptible to model violations that mildly increase non-synonymous rates for reasons other than adaptive evolution. We plan to use richer simulations to study how the new approach reacts to such model violations, and if expanding the model to recognize features such as variable effective population size, CpG hypermutability, codon usage and gene conversion biases, could introduce greater robustness to inferences of adaptive evolution.

Empirical studies

A more detailed examination, ideally combined with experimental corroborations, of the sites uncovered by the model is pressing, and hopefully based on far more than the hand-full of data sets of the present study. This would help build our empirical understanding how the model behaves in a variety of different contexts (Moutinho et al., 2019; Slodkowicz and Goldman, 2020). We hope to apply the model on a few thousand genes from the OrthoMamm database (Scornavacca et al., 2019) in a first step, before engaging broader applications across varied taxanomic contexts.

Model forms

While we have outlined the modeling strategy with a three-component finite mixture of ω_* values, in combination with a Dirichlet process prior on amino acid profiles, many other possibilities could be considered: various parametric families on ω_* (as did Yang et al., 2000, with ω), non-parametric approaches on ω_* (as proposed for ω by Huelsenbeck et al., 2006), grids of predetermined ω_* values (in the spirit of Murrell et al., 2013), along with similar choices on modeling amino acid fitness heterogeneity (e.g., Rodrigue et al., 2010; Rodrigue, 2013; Rodrigue and Lartillot, 2014). The potentially complex interactions between the numerous combinations also entail a large study.

Applications

We propose these modeling ideas in two independent software packages (see below). One of our Markov chain Monte Carlo implementations can run under fixed topology as well as sample over trees, and thus enable studies of the impact of phylogenetic uncertainty in inferences of adaptive evolution, utilizing both traditional and mutation-selection codon substitution models; this also suggests more extensive studies on the potential of such models for phylogenetic inference *per se*. Another implementation we offer lends itself to integrative modeling objectives, with a wide suite of potential research avenues utilizing the mutation-selection-based approaches. Foreseeable directions with the latter implementation include capturing the evolution of effective population size over the phylogeny, along with joint inferences of continuous-trait evolution, as formalized by Lartillot and Poujol (2011).

Materials and methods

Data

For convenience, all data sets (simulated and real) studied herein, and described below, are included in a supplementary information file.

Simulated Data

We used the simulation system described in Rodrigue and Lartillot (2017) to generate artificial datasets using a mutation-selection framework with global mutation parameters and site-specific amino acid fitness profiles. The mutation-level parameters (which assume no selection on synonymous variants) are as given in Rodrigue and Lartillot (2017), as is the phylogenetic tree (with 38 tips). With nearly-neutral simulations (i.e., with the pure mutation-selection formulation, such as detailed below), the amino acid fitness profile used to simulate a codon site is chosen at random from a set of empirically derived profiles. We obtained such profiles by running the pure Dirichlet process-based mutation-selection model (Rodrigue et al., 2010) on a multi-gene data set at the scale of placental mammals (Lartillot and Delsuc, 2012), and calculating the posterior mean amino acid profile at each site. The simulation draws at random (with replacement) one such site-specific posterior mean profile to run the evolutionary process along the tree at one codon site, repeating to produce alignments of 300 codons. For simulations with adaptive evolutionary regimes, the starting profiles are altered along the branches of the phylogeny as detailed in Rodrigue and Lartillot (2017), with the *Red Queen* parameter set to 0.01. In contrast to the simulations in Rodrigue and Lartillot (2017), however, the adaptive simulations herein are applied to only 10% of sites of the alignment; these 30 sites were chosen at random, i.e., they were spread out randomly across the alignment. The remaining 270 codon sites are simulated with the Red Queen parameter set to 0, thus constituting pure mutation-selection regimes.

Real Data

We used previously studied alignments of protein-coding genes provided by the authors of earlier works:

- β-Globin: 17 vertebrate sequences of β-globin gene, 144 codons in length, taken from Yang et al. (2000);
- Adh: 23 *Drosophila* sequences of the alcohol dehydrogenase gene, 254 codons in length, taken from Yang et al. (2000);
- Vwf: 62 sequences, at the scale of placental mammals, of the von Willbrand factor gene, 392 codons in length, taken from Lartillot and Delsuc (2012), as were the next three alignments;
- Adora3: 67 sequences of the adenosine receptor A3 gene, 107 codons in length;
- Rbp3: 54 sequences of the retinol binding protein 3, 412 codons in length;
- S1pr1: 67 sequences of the sphingosine-1-phosphate receptor 1 gene, 325 codons in length.

Substitution models

The MG-M0 codon substitution model, inspired from Muse and Gaut (1994), but with a single ω parameter distinguishing nonsynonymous events, has entries given as:

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

Here, μ_{ij} is the mutational parameterization, which we set as a *general-time reversible* nucleotide-level model (Lanave et al., 1984), with six exchangeability parameters (five degrees of freedom) and four frequency parameters (three degrees of freedom). The MG-M3 model has the same form, but rather than a single ω parameter, it invokes three different values (with their respective weights), and has a likelihood function consisting of the a weighted average of likelihood scores under each of the three ω values (Yang et al., 2000).

The MutSel-M0$_*$ model, presented in Rodrigue and Lartillot (2017), is given as:

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

where $S_{ij}^{(n)} = F_j^{(n)} - F_i^{(n)} = 4N_e s_{ij} = 4N_e f_j^{(n)} - f_i^{(n)}$ is the *scaled selection coefficient* (scaled by the effective population size N_e and a ploidy-dependent constant, in this example set at 4 Yang and Nielsen, 2008), calculated from the difference in fitness associated with a mutant protein with the amino acid encoded by codon j at site n, denoted $F_j^{(n)}$, with that of the wild-type population where the amino-acid encoded by i is fixed at that position, $F_i^{(n)}$. Site-specific fitness profiles are treated as random effects within a Dirichlet process system (Rodrigue et al., 2010; Rodrigue and Lartillot, 2014). As with the MG-M3 model, the MutSel-M3$_*$ model invokes three distinct ω_* values, with their respective weights, as a finite mixture model of heterogeneity across sites.

Priors

Branch lengths are endowed with an exponential prior of mean controled by a hyperprior, itself endowed with an exponential prior of mean 1. Nucleotide exchangeabilities and frequencies are each endowed with flat Dirichlet priors, whereas ω and ω_* have priors following a gamma law, controlled by two hyperparameters, each endowned with exponential priors of mean 1. Weights of finite mixture on ω or ω_* follow with flat Dirichlet prior. Amino acid fitness profiles follow a Dirichlet process prior (Rodrigue et al., 2010), implemented under a stick-breaking representation (Lartillot et al., 2013; Rodrigue and Lartillot, 2014).

Implementations

The models presented have been implemented in an experimental version (2) of PhyloBayes-MPI (`https://github.com/bayesiancook/pbmpi2`), allowing for a joint sampling across parameter space, auxiliary variables, and tree topology space. We have also implemented the models in a new software called BayesCode, which is focused on integrative comparative methods under fixed topology (`https://github.com/bayesiancook/bayescode`). Example scripts demonstrating the use of the software are provided in the supplementary file.

Acknowledgements

The work was funded by the Natural Sciences and Engineering Research Council of Canada (NR), Carleton University (NR), and ... ask Nico and Thibault if they'd like to mention anything here.

References

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

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

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

Holder, M. T., D. J. Zwickl, and C. Dessimoz. 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., S. Jain, S. W. D. Frost, and S. L. K. Pond. 2006. A Dirichlet process model for detecting positive selection in protein-coding DNA sequences. *Proc. Natl. Acad. Sci. USA* 103:6263–6268.

Jansa, S. A., and R. S. Voss. 2011. Adaptive evolution of the venom-targeted vwf protein in opossums that eat pitvipers. *PLoS One* 6:e20997.

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

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

Lartillot, N., and R. Poujol. 2011. A phylogenetic model for investigating correlated evolution of substitution rates and continuous phenotypic characters. *Mol. Biol. Evol.* 28:729–744.

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

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

Lemmerhirt, H. L., J. A. Shavit, G. G. Levy, S. M. Cole, J. C. Long, and D. Ginsburg. 2006. Enhanced VWF biosynthesis and elevated plasma VWF due to a natural variant in the murine Vwf gene. *Blood* 108:3061–3067.

Lowe, C., and N. Rodrigue. 2020. Detecting adaptation from multi-species protein-coding dna sequence alignments alignments. *Phylogenetics in the Genomic Era* 4–5.

Matzkin, Luciano M. 2004. Population Genetics and Geographic Variation of Alcohol Dehydrogenase (Adh) Paralogs and Glucose-6-Phosphate Dehydrogenase (G6pd) in Drosophila mojavensis. *Mol. Biol. Evol.* 21:276–285.

Matzkin, Luciano M., and Walter F. Eanes. 2003. Sequence variation of alcohol dehydrogenase (adh) paralogs in cactophilic *drosophila*. *Genetics* 163:181–194.

McDonald, J. H., and M. Kreitman. 1991. Adaptive protein evolution at the *adh* locus in *drosophila*. *Nature* 351:652–654.

Moutinho, A. F., F. F. Trancoso, and J. Y. Dutheil. 2019. The impact of protein architecture on adaptive evolution. *Mol. Biol. Evol.* 36:2013–2028.

Murrell, Ben, Sasha Moola, Amandla Mabona, Thomas Weighill, Daniel Sheward, Sergei L Kosakovsky Pond, and Konrad Scheffler. 2013. Fubar: a fast, unconstrained Bayesian approximation for inferring selection. *Mol. Biol. Evol.* 30:1196–1205.

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

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

Powell, J. R., and E. N. Moriyama. 1997. Evolution of codon usage bias in *drosophila*. *Proc. Natl. Acad. Sci. USA* 94:7784–7790.

Ratnakumar, A., S. Mousset, S. Glémin, J. Berglund, N. Galtier, L. Duret, and M. T. Webster. 2010. Detecting positive selection within genomes: the problem of biased gene conversion. *Phil. Trans. R. Soc.iety B* 365:2571–2580.

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

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

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

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

Rousselle, M., M. Mollion, B. Nabholz, T. Bataillon, and N. Galtier. 2018. Overestimation of the adaptive substitution rate in fluctuating populations. *Biology letters* 14:20180055.

Scornavacca, C., K. Belkhir, J. Lopez, R. Dernat, F. Delsuc, E. J. P. Douzery, and V. Ranwez. 2019. OrthoMaM v10: Scaling-Up Orthologous Coding Sequence and Exon Alignments with More than One Hundred Mammalian Genomes. *Mol. Biol. Evol.* 36:861–862.

Siddiq, M. A., and J. W Thornton. 2019. Fitness effects but no temperature-mediated balancing selection at the polymorphic *adh* gene of *drosophila melanogaster*. *Proc. Natl. Acad. Sci. USA* 116:21634–21640.

Slodkowicz, G., and N. Goldman. 2020. Integrated structural and evolutionary analysis reveals common mechanisms underlying adaptive evolution in mammals. *Proc. Natl. Acad. Sci. USA* 117:5977–5986.

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

Suzuki, Y., T. Gojobori, and S. Kumar. 2009. Methods for Incorporating the Hypermutability of CpG Dinucleotides in Detecting Natural Selection Operating at the Amino Acid Sequence Level. *Mol. Biol. Evol.* 26:2275–2284.

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

Venkat, A., M. W. Hahn, and J. W. Thornton. 2018. Multinucleotide mutations cause false inferences of lineage-specific positive selection. *Nat. Ecol. Evol.* 2:1280–1288.

Wisotsky, S. R., S. L. Kosakovsky Pond, S. D. Shank, and S. V. Muse. in press. Synonymous site-to-site substitution rate variation dramatically inflates false positive rates of selection analyses: ignore at your own peril. *Mol. Biol. Evol.* .

Yang, Z. 2006. *Computational Molecular Evolution*. Oxford Series in Ecology and Evolution.

Yang, Z. 2019. Adaptive molecular evolution. In *Handbook of statistical genomics, vol. i*, ed. D. J. Balding, I. Moltke, and J. Marioni. John Wiley & Sons, Ltd.

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

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

Table 1. Amino acid sites under positive selection.

Data	Model	Sites
$\beta-$Globin	MG-M3	**7**, 11, **42**, **48**, **50 54**, **67**, **85**, **123**
	MutSel-M3$_*$	**7**, 10, 11, *14*, **42**, **48**, **50**, **54**, **67**, **74**, 84, **85**, *110*, *113*, **123**
Adh	MG-M3	-
	MutSel-M3$_*$	*9*, 39, 49, **57**, **68**, 69, *72*, *81*, 85, 98, 133, 163, 165, *170*, **185**, *187*, *197*, 201, *205*, 208, **216**, 229, 253
Vwf	MG-M3	-
	MutSel-M3$_*$	5, 9, 26, **41**, *82*, 85, *103*, **108**, 125, *147*, 148, *158*, *177*, 182, *197*, 226, 227, **235**, **239**, 241, **242**, 247, 288, *291*, 307, **313**, 318, *324*, *339*, 371, *379*, 390
Adora3	MG-M3	-
	MutSel-M3$_*$	2, *4*, *93*, **96**
Rbp3	MG-M3	-
	MutSel-M3$_*$	-
S1pr1	MG-M3	-
	MutSel-M3$_*$	*1*, *58*, *144*, *145*, *146*, *148*

Note.—Numbers in *italic* font are at the 0.9 level, in plain font at the 0.95 level, and in **bold** font at 0.99 level.

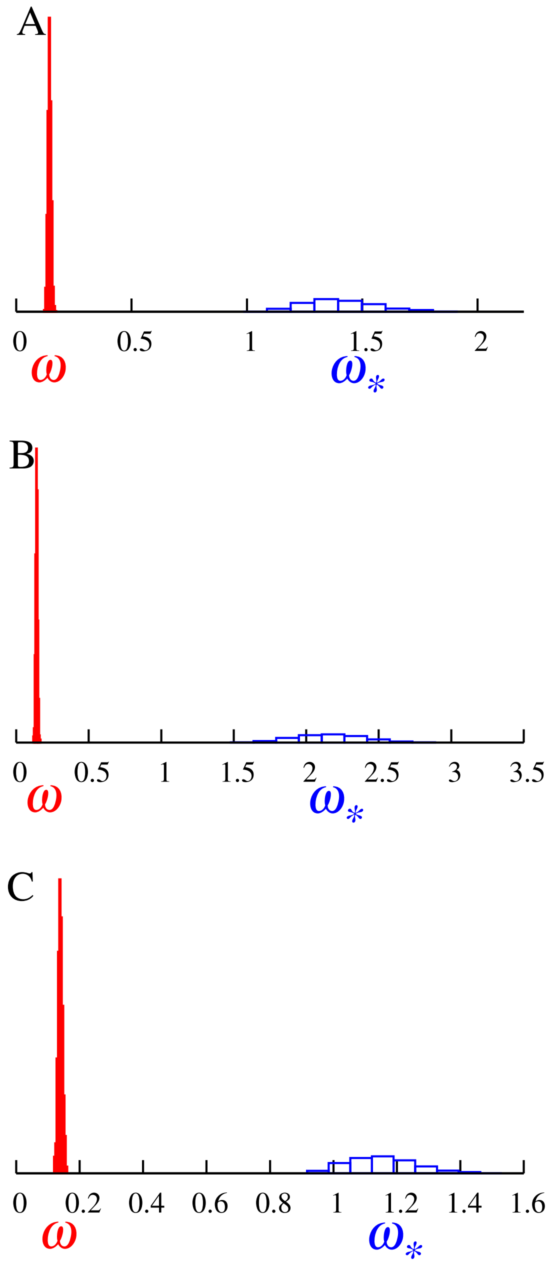

Figure 1. Posterior distributions of ω (red, under MG-M0) and ω_* (blue, under MutSel-M0$_*$) on simulated data sets with 10% of sites evolved under adaptive evolution (see methods).

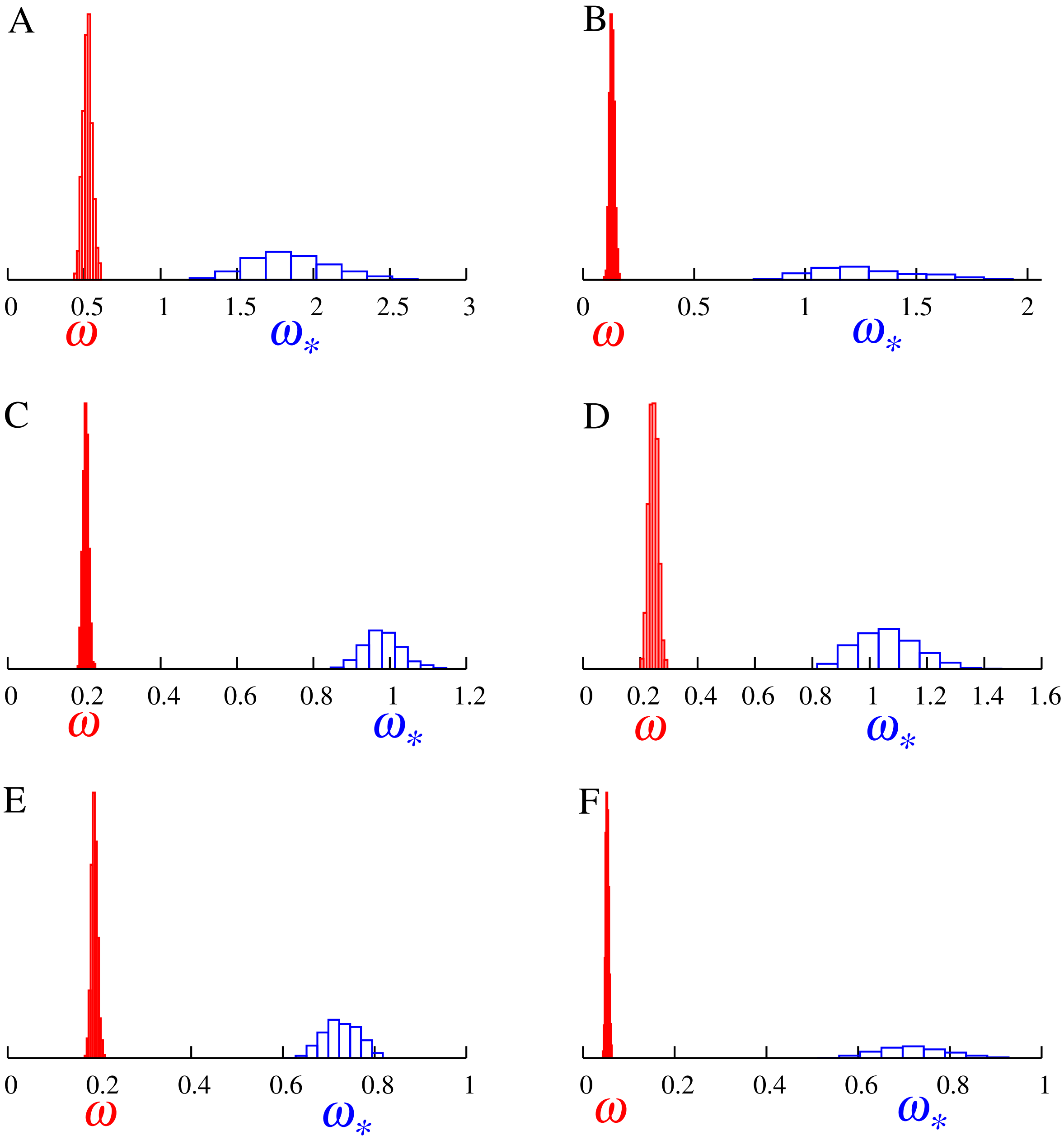

Figure 2. Posterior distributions of ω (red, under MG-M0) and ω_* (blue, under MutSel-M0$_*$) on β-globin17-144, adh, vwf, adora3, rbp3, s1pr1 data sets (see methods).

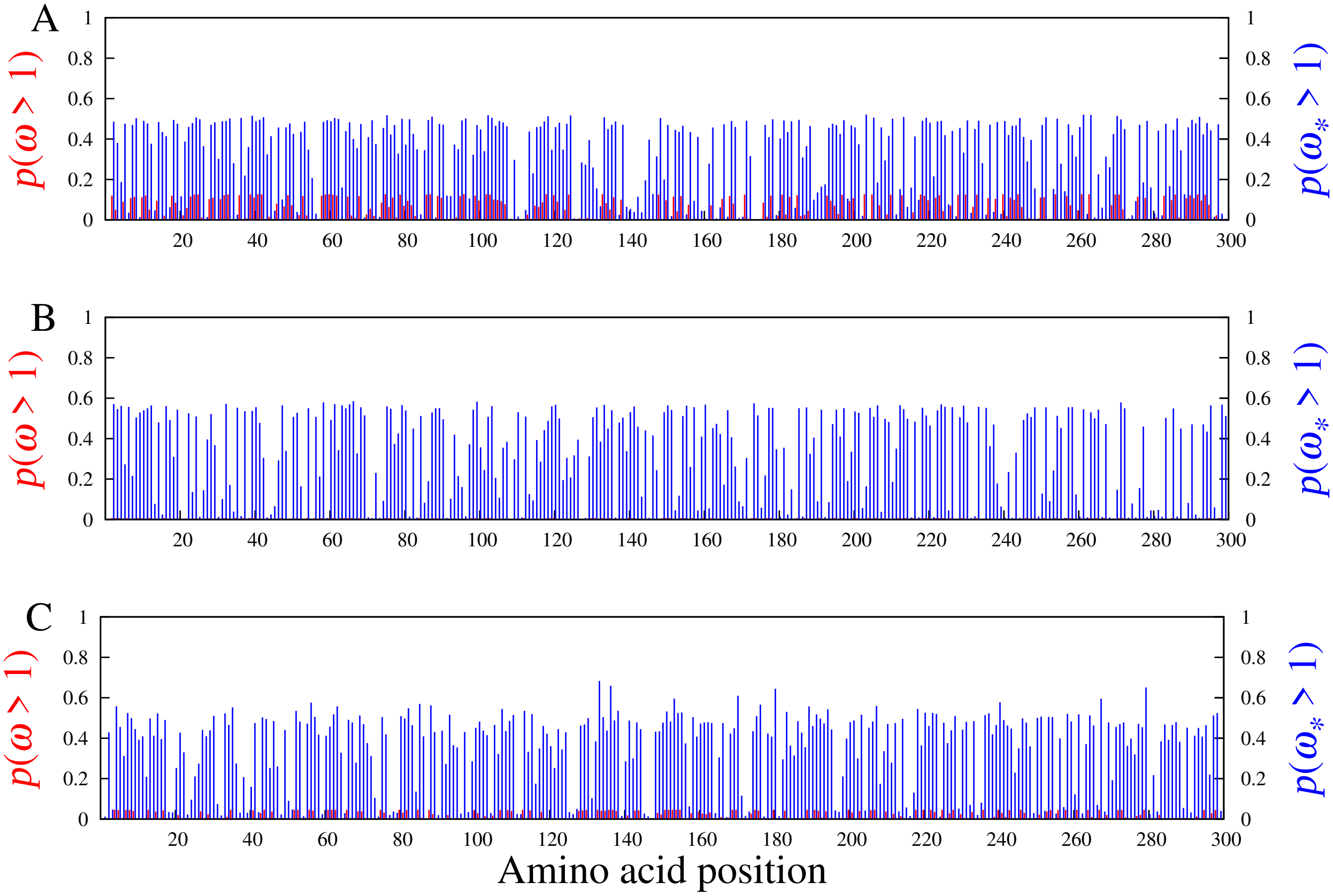

Figure 3. Site-specific posterior probabilities of ω (red, under MG-M3) and ω_* (blue, under MutSel-M3$_*$) being greater than 1 on data sets simulated under the pure mutation-selection framework.

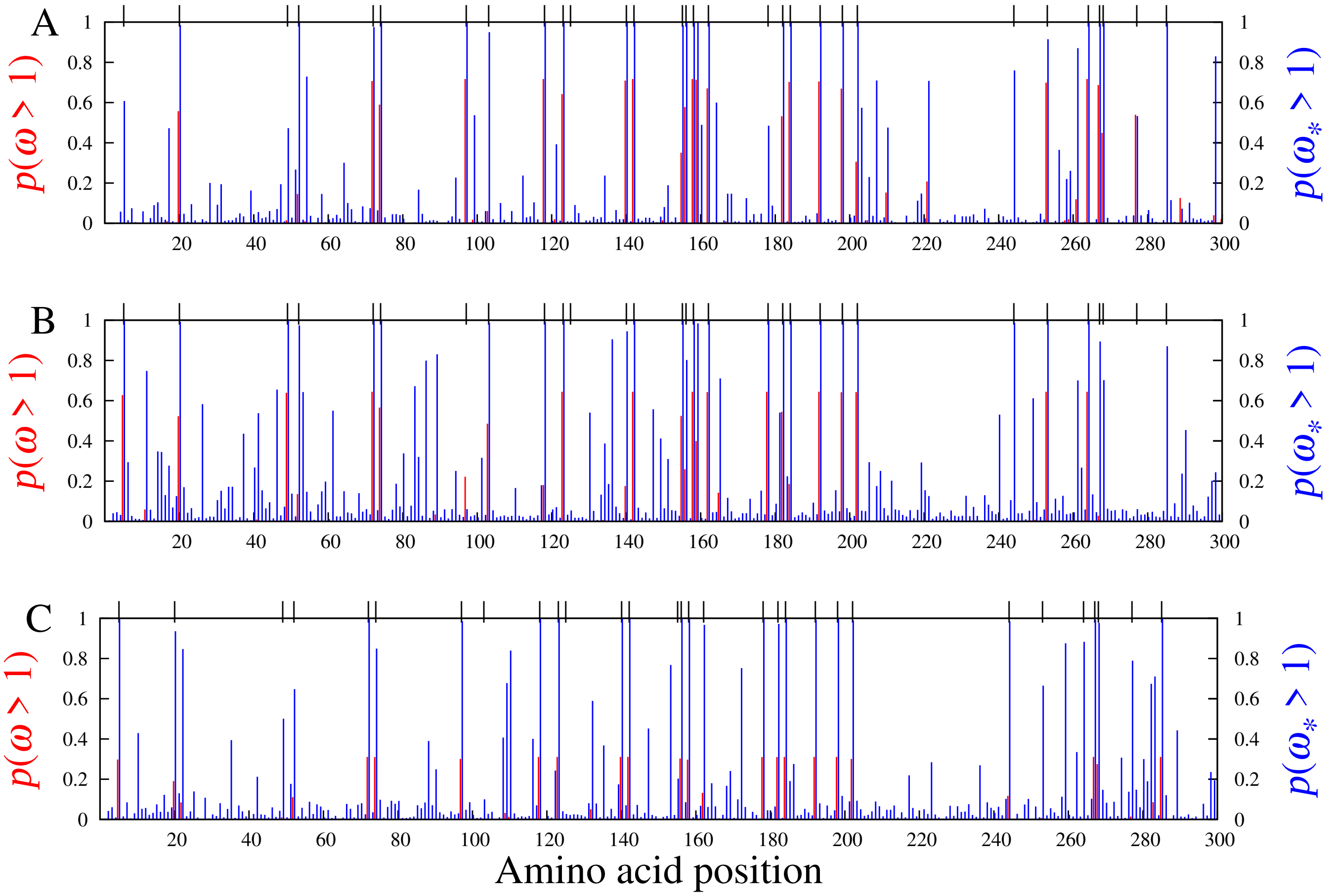

Figure 4. Site-specific posterior probabilities of ω (red, under MG-M3) and ω_* (blue, under MutSel-M3$_*$) being greater than 1 on data sets simulated with 30 sites (marked with at top of panels) under an adaptive regime, and the remaining 270 site under the pure mutation-selection framework.

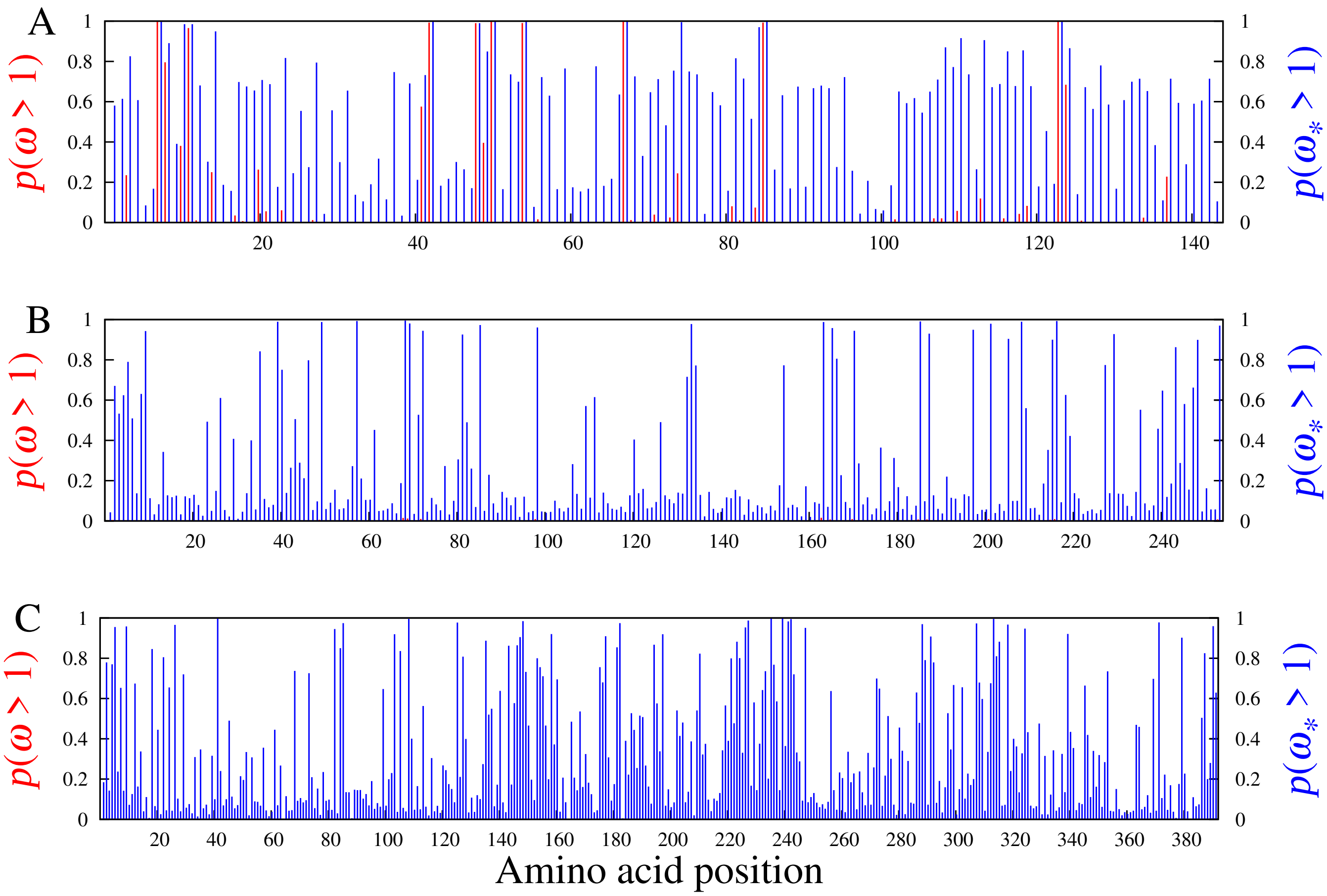

Figure 5. Site-specific posterior probabilities of ω (red, under MG-M3) and ω_* (blue, under MutSel-M3$_*$) being greater than 1 on β-globin, adh, and vwf.

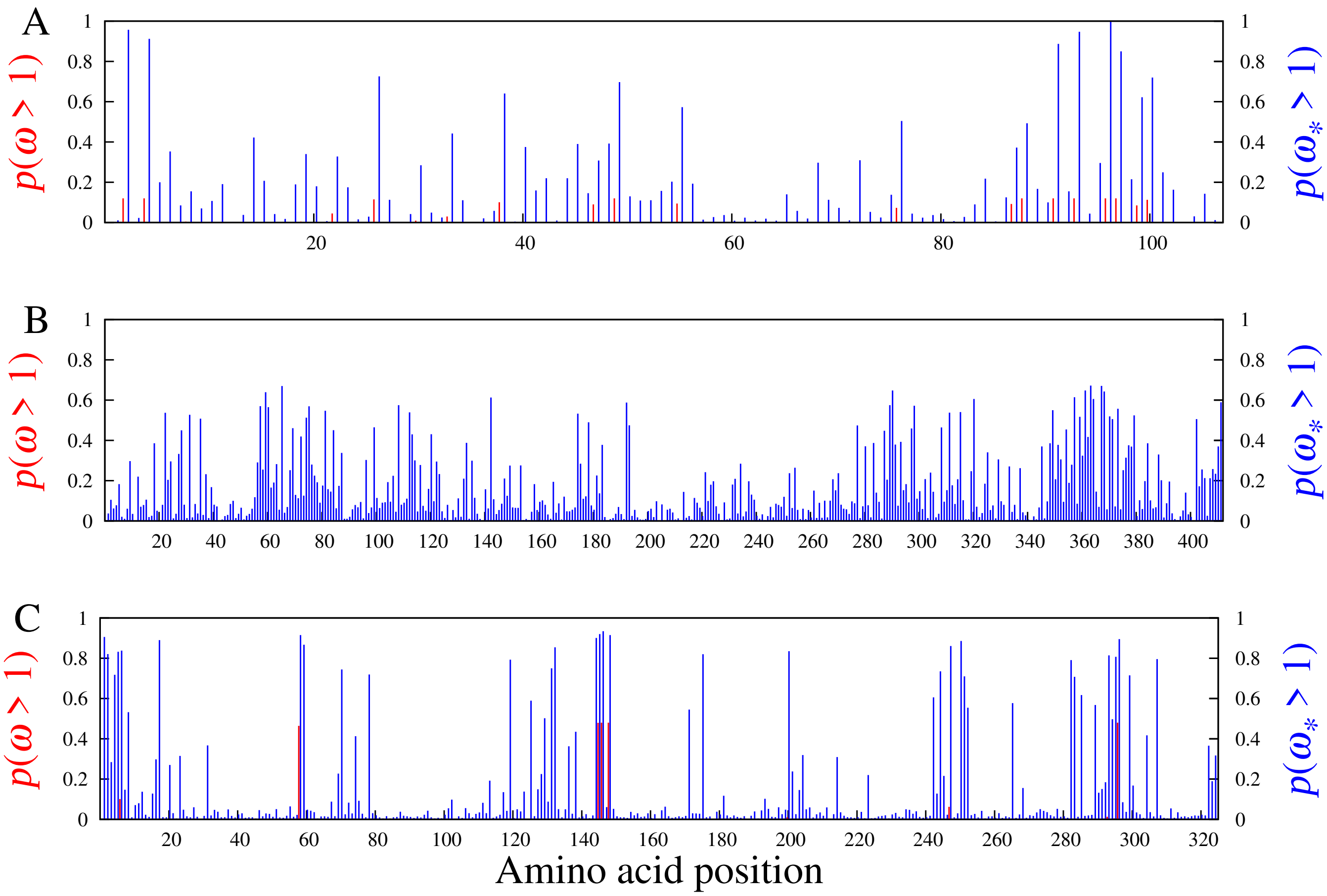

Figure 6. Site-specific posterior probabilities of ω (red, under MG-M3) and ω_* (blue, under MutSel-M3$_*$) being greater than 1 on adora3, rbp3, s1pr1.