The relationship between *dN/dS* and scaled selection coefficients

Stephanie J. Spielman[1*] and Claus O. Wilke[1]

Address:

[1]Department of Integrative Biology, Center for Computational Biology and Bioinformatics, and Institute of Cellular and Molecular Biology. The University of Texas at Austin, Austin, TX 78712, USA.

[*]Corresponding author

Email: stephanie.spielman@gmail.com

Manuscript type: Article

Keywords: dN/dS, mutation-selection-balance models, scaled selection coefficients, Markov models of sequence evolution

Abstract

Numerous computational methods exist to assess the mode and strength of natural selection in protein-coding sequences, yet how distinct methods relate to one another remains entirely unknown. Here, we elucidate the relationship between two widely-used phylogenetic modeling frameworks: dN/dS models and mutation-selection-balance (MutSel) models. We derive a mathematical relationship between dN/dS and scaled selection coefficients, the focal parameters of MutSel models, and use this relationship to gain unprecedented insight into the behaviors, limitations, and applicabilities of these two modeling frameworks. We prove that, if all synonymous changes are neutral, MutSel models correspond only to $dN/dS \leq 1$. However, if synonymous codons differ in fitness, dN/dS can take on arbitrarily high values even if all selection is purifying. Thus, MutSel models cannot accommodate positive, diversifying selection, while dN/dS cannot distinguish between purifying selection on synonymous codons and positive selection on amino acids. We further propose a new benchmarking strategy of dN/dS inferences against MutSel simulations and demonstrate that the widely-used Goldman-Yang-style dN/dS models yield substantially biased dN/dS estimates on realistic sequence data. By contrast, the less frequently used Muse-Gaut-style models display much less bias. Strikingly, the least-biased and most-precise dN/dS estimates are never found in the models with the best AIC scores. Thus, selecting models based on goodness-of-fit criteria can yield poor parameter estimates if the models considered do not precisely correspond to the underlying mechanism that generated the data. In conclusion, establishing mathematical links among modeling frameworks represents a novel, powerful strategy to pinpoint previously unrecognized model limitations and strengths.

Introduction

The oldest and most-widely used method to infer selection pressure in protein-coding genes calculates the evolutionary rate ratio dN/dS, which represents the ratio of non-synonymous to synonymous substitution rates. This metric indicates how quickly a protein's constituent amino acids change, relative to synonymous changes, and it is commonly used to identify protein sites that experience purifying selection ($dN/dS < 1$), evolve neutrally ($dN/dS \approx 1$), or experience positive, diversifying selection ($dN/dS > 1$) (Nielsen and Yang 1998; Yang et al. 2000; Kosakovsky Pond and Frost 2005b; Huelsenbeck et al. 2006). In phylogenetic contexts, dN/dS is typically calculated using a maximum likelihood (ML) approach (Goldman and Yang 1994; Muse and Gaut 1994; Nielsen and Yang 1998; Yang 2006). ML methods assume a continuous time Markov model of sequence evolution and have become a staple of comparative sequence analysis since their introduction in the

1990s (see ref (Anisimova and Kosiol 2009) for a comprehensive review). Throughout this paper, we will refer to these models as dN/dS-based models.

A second class of Markov models, known as mutation-selection-balance (MutSel) models, are increasingly being viewed as a viable alternative to the dN/dS framework. While dN/dS-based models describe how quickly a protein's constituent amino acids change, MutSel models assess the strength of natural selection acting on specific mutations. Couched firmly in population-genetic theory, the MutSel framework estimates site-specific scaled selection coefficients $S = 2N_e s$, which indicate the extent to which natural selection favors, or disfavors, particular codon and/or amino acid changes (Halpern and Bruno 1998; Yang and Nielsen 2008; Rodrigue et al. 2010; Tamuri et al. 2012). Although first introduced over 15 years ago (Halpern and Bruno 1998), MutSel models have seen little use due to their high computational expense. Recently, however, several computationally tractable model implementations have emerged (Rodrigue and Lartillot 2014; Tamuri et al. 2014), allowing for the first time the potential for widespread adoption.

Over the course of twenty years developement, dN/dS-based models have advanced to a high level of sophistication. These models can accommodate a variety of evolutionary scenarios, including synonymous rate variation (Muse and Gaut 1994; Kosakovsky Pond and Muse 2005) and episodic (Kosakovsky Pond et al. 2011; Murrell et al. 2012) and/or lineage-specific selection (Yang and Nielsen 2002; Zhang et al. 2005; Kosakovsky Pond and Frost 2005a), and they can also incorporate information regarding protein structure and epistatic interactions (Robinson et al. 2003; Thorne et al. 2007; Rodrigue et al. 2000; Scherrer et al. 2012; Meyer and Wilke 2013). This flexibility, along with accessible software implementations (Kosakovsky Pond et al. 2005; Yang 2007; Delport et al. 2010), makes dN/dS-based models an attractive analysis choice. On the other hand, some have argued that MutSel models, given their explicit basis in population-genetics theory and attention to site-specific amino-acid fitness differences, offer a more mechanistically realistic approach to studying coding-sequence evolution (Halpern and Bruno 1998; Rodrigue et al. 2010; Tamuri et al. 2012; Thorne et al. 2012). Moreover, a growing body of literature has demonstrated that dN/dS estimates are particularly sensitive to violations in model assumptions, calling into question the general utility of dN/dS-based models (Rocha et al. 2006; Kryazhimskiy and Plotkin 2008; Mugal et al. 2014).

Although both MutSel and dN/dS-based models describe the same fundamental process of coding-sequence evolution along a phylogeny, it is unknown how these two modeling frameworks relate to one another. In particular, as these inference methods have been developed independently,

it remains an open question whether or not parameter estimates from one model are comparable to those of the other model. As a consequence, although certain rhetorical arguments may be made in favor of using one method over another, there is currently no formalized, concrete rationale to guide researchers in their methodological choices. Elucidating the relationship between these competing modeling frameworks will more precisely reveal under which circumstances the use of these models is justified and has great potential to reveal previously unrecognized model behaviors, limitations, and capabilities.

Here, we formalize the relationship between these two modeling frameworks by examining the extent to which their respective focal parameters, dN/dS and scaled selection coefficients, yield overlapping information about the evolutionary process. To this end, we derive a mathematical relationship between dN/dS and scaled selection coefficients. We find that dN/dS values can be precisely calculated from scaled selection coefficients, and that dN/dS accurately captures the selective pressures indicated by a given distribution of scaled selection coefficients. Furthermore, we prove that, when synonymous mutations are neutral, it is only possible to recover $dN/dS \leq 1$ from selection coefficients, demonstrating that MutSel models are inherently only able to model purifying selection. Therefore, these models would be an inappropriate analysis method if positive selection is expected. However, we also find that, when synonymous codons have different fitnesses and hence purifying selection acts on synonymous changes, it is possible to recover dN/dS values above 1, even though classical positive, diversifying selection is not occurring. Therefore, the dN/dS framework cannot distinguish between positive, diversifying selection on amino acids and purifying selection on synonymous changes.

Finally, this relationship provides a uniquely rigorous platform to examine the performance of dN/dS-based models. Typically, researchers evaluate performance of a given inference framework through simulations that adhere to the underlying model's assumptions (but see refs. (Schoniger and von Haeseler 1995; Minin et al. 2003; Holder et al. 2008; Yap et al. 2010; Rubinstein et al. 2011)). Indeed, simulated data is usually generated according to the same model as the inference framework, allowing for a direct comparison between the true and estimated parameter values. While this strategy is critical for testing whether a model implementation behaves as expected, it cannot assess model performance when the data are generated under a different process than the one modeled in the inference framework. However, in real-world sequence analysis, the inference framework will never exactly match the data-generation process. Therefore, a more sensitive test of model performance would examine how a given method performs when data are simulated

under different mechanistic processes, and how sensitive the method is to violations of its assumptions. Unfortunately, such an approach is typically infeasible, because the relationships between parameters of interest among distinct model classes are generally not known.

The relationship we establish here between dN/dS and selection coefficients allows us to overcome this limitation, as we can determine the true dN/dS value directly from MutSel model parameters. Thus, we can assess performance of dN/dS-based inference frameworks by simulating data with a MutSel model and then comparing inferred dN/dS ML estimates (MLEs) to dN/dS values computed from selection coefficients. Using this strategy, we find, for sequences evolved under a symmetric mutation model, that dN/dS values inferred in an ML framework agreed precisely with those calculated from scaled selection coefficients. However, as mutational asymmetry increases, dN/dS MLEs become increasingly biased away from their true values, under a variety of ML model parameterizations. Surprisingly, the ML model parameterization which produced the most accurate dN/dS estimates was never the model which exhibited the best fit to the data (measured by AIC and BIC), ultimately revealing that relying on model fit as a litmus-test for model performance can be an ineffective and misleading strategy.

Results and Discussion

Theoretical model.

This section contains a re-derivation of results presented in refs. (Halpern and Bruno 1998; Yang and Nielsen 2008), reproduced here to introduce notation and to place the remainder of our work into context. We model sequence evolution using the Halpern-Bruno MutSel modeling framework under the assumptions of a fixed effective population size N_e and constant selection pressure over time (Halpern and Bruno 1998; Yang and Nielsen 2008; Tamuri et al. 2012; Thorne et al. 2012). This continuous-time reversible Markov process is governed by the 61×61 transition matrix $T(t) = e^{Qt}$, where the matrix $Q = q_{ij}$ gives the instantaneous substitution probabilities between all 61 sense codons, and diagonal elements of Q satisfy $q_{ii} = -\sum_{i \neq j} q_{ij}$. We assume that only single-nucleotide substitutions occur instantaneously.

Let f_i^{codon} be the fitness of codon i, and let the selection coefficient acting on a mutation from codon i to codon j be $s_{ij} = f_j^{\text{codon}} - f_i^{\text{codon}}$ (Sella and Hirsh 2005; Yang and Nielsen 2008). The fixation probability for this mutation is (Kimura 1962; Halpern and Bruno 1998; Yang and Nielsen

2008)

$$u_{ij} \approx \frac{2s_{ij}}{1 - e^{-2N_e s_{ij}}} = \frac{1}{N_e} \frac{2N_e s_{ij}}{1 - e^{-2N_e s_{ij}}}. \tag{1}$$

We further define $S_{ij} = 2N_e s_{ij}$ as the scaled selection coefficient for this change (Yang and Nielsen 2008). The probability of a substitution from codon i to j is therefore

$$q_{ij} = N_e m_{ij} u_{ij} = m_{ij} \frac{S_{ij}}{1 - e^{-S_{ij}}}, \tag{2}$$

where m_{ij} is the codon mutation rate, which represents the rate at which codon i transitions to codon j (Halpern and Bruno 1998; Sella and Hirsh 2005). If we assume that m_{ij} only has non-zero entries for single-nucleotide changes, we can write it as $m_{ij} = \mu_{s_i t_j}$, where μ_{kl} is the per-nucleotide mutation rate, s_i is the source (i.e., before mutation) nucleotide in codon i, and t_j is the target (i.e., after mutation) nucleotide in codon j.

We now show how S_{ij} can be written in terms of mutation rates and stationary (equilibrium) codon frequencies P_i. As this system satisfies detailed balance (reversibility) (Halpern and Bruno 1998), we have

$$q_{ij} P_i = q_{ji} P_j. \tag{3}$$

From equations (2) and (3), we can write the ratio of substitution probabilities as

$$\frac{P_i}{P_j} = \frac{m_{ji} S_{ji} (1 - e^{-S_{ij}})}{m_{ij} S_{ij} (1 - e^{-S_{ji}})}. \tag{4}$$

Using $S_{ij} = -S_{ji}$, we find that

$$S_{ij} = \ln \left(\frac{P_j m_{ji}}{P_i m_{ij}} \right). \tag{5}$$

This equation, previously derived in ref. (Halpern and Bruno 1998), establishes a relationship between scaled selection coefficients and the stationary codon frequencies of the Markov chain. Moreover, in the specific case of symmetric mutation rates $m_{ij} = m_{ji}$, we have $S_{ij} = \ln \left(P_j / P_i \right)$ (Sella and Hirsh 2005).

Predicting *dN/dS* from scaled selection coefficients.

We now derive respective expressions for nonsynonymous and synonymous evolutionary rates, which we can divide to obtain the evolutionary rate ratio dN/dS. We write the nonsynonymous rate K_{N} as

$$K_{\mathrm{N}} = \sum_i \sum_{j \in \mathcal{N}_i} P_i q_{ij} \,, \tag{6}$$

where $\mathcal{N}_i$ is the set of codons that are nonsynonymous to codon i and differ from it by one nucleotide, and the substitution probability q_{ij} is defined in equation (2). To normalize K_{N}, we divide it by the number of nonsynonymous sites, which we calculate according to the mutational opportunity definition of a site (Goldman and Yang 1994; Yang 2006) as

$$L_{\mathrm{N}} = \sum_i \sum_{j \in \mathcal{N}_i} P_i m_{ij} \,. \tag{7}$$

Thus, we find that

$$dN = \frac{K_{\mathrm{N}}}{L_{\mathrm{N}}} = \frac{\sum_i \sum_{j \in \mathcal{N}_i} P_i q_{ij}}{\sum_i \sum_{j \in \mathcal{N}_i} P_i m_{ij}} \,. \tag{8}$$

Similarly, for dS, the synonymous evolutionary rate K_{S} per synonymous site L_{S}, we find

$$dS = \frac{K_{\mathrm{S}}}{L_{\mathrm{S}}} = \frac{\sum_i \sum_{j \in \mathcal{S}_i} P_i q_{ij}}{\sum_i \sum_{j \in \mathcal{S}_i} P_i m_{ij}} \,, \tag{9}$$

where $\mathcal{S}_i$ is the set of codons that are synonymous to codon i and differ from it by one nucleotide substitution. The quantities K_{S} and L_{S} are defined as in Eqs. (6) and (7) but sum over $j \in \mathcal{S}_i$ instead of $j \in \mathcal{N}_i$. Moreover, if we assume that mutation rates are symmetric and that all synonymous codons have equal fitness (i.e. synonymous mutations are neutral), the synonymous fixation rate satisfies $u_{ij|j \in \mathcal{S}_i} = 1/N_e$ (Crow and Kimura 1970), and hence the substitution probability becomes $q_{ij} = m_{ij}$. In this circumstance, the value for dS reduces to 1.

MutSel models strictly describe purifying selection.

We examined the relationship between dN/dS and scaled selection coefficients by simulating 200 distributions of amino acid scaled fitness values, $F_a^{\mathrm{aa}} = 2N f_a^{\mathrm{aa}}$, from a normal distribution $\mathcal{N}(0, \sigma^2)$. We drew a unique σ^2 for each fitness distribution from a uniform distribution $\mathcal{U}(0, 4)$. Higher values for σ^2 correspond to larger fitness differences among amino acids, causing selection to act more strongly against nonsynonymous changes. Thus, high σ^2 values indicate strong purifying selection, low values indicate weak purifying selection, and $\sigma^2 = 0$ indicates that all amino acids are equally fit. We note that these F_a^{aa} quantities correspond exactly to the amino-acid propensity parameters estimated by currently available site-specific MutSel inference methods (Rodrigue and Lartillot 2014; Tamuri et al. 2014).

We then converted each amino-acid fitness distribution to a corresponding set of codon fitnesses, as described in *Methods*. Briefly, for 100 of the distributions, we assumed that all synonymous codons had the same fitness, but for the other 100 distributions we allowed synonymous codons to have different fitnesses. In other words, the former 100 distributions do not incorporate purifying

selection on synonymous changes whereas the latter 100 distributions do. Using equations (6) - (9), we computed dN/dS for each distribution of codon fitnesses. For these calculations, we assumed the symmetric mutation model HKY85 (Hasegawa et al. 1985), which is specified by the parameters μ, the nucleotide mutation rate, and κ, the ratio of transitions to tranversions. Specifically, transitions occur at a rate $\mu\kappa$ and tranversions at a rate μ. We used $\mu = 10^{-6}$ for all simulations, while we selected a unique value for κ for each simulation from $\mathcal{U}(1, 6)$.

Under neutral evolution, we expect that $dN/dS = 1$, and as purifying selection increases in strength, dN/dS should correspondingly decrease. Therefore, we expect that dN/dS will decline with the variance (σ^2) of the distribution of amino acid fitness values. Indeed, we observed a strong, negative correlation between these quantities (Figure 1). The larger the fitness differences among amino acids (higher σ^2), the lower dN/dS, properly reflecting increased purifying selection. This correlation was much stronger for fitness distributions without synonymous selection (Figure 1A) than for those with synonymous selection (Figure 1B). This difference emerged because fitness differences among synonymous codons obscured underlying amino-acid fitness differences. Even so, selection on synonymous codons did not negate the significant correlation between dN/dS and overall selection strength.

Importantly, Figure 1A demonstrates that, in the limiting case when σ^2 approaches 0, and thus all codons have virtually the same fitness, dN/dS converges to 1. In other words, when the protein-coding sequence evolved neutrally, selection coefficients correctly yielded a $dN/dS \approx 1$. Furthermore, we never recovered $dN/dS > 1$ when synonymous changes were neutral, revealing a key property of MutSel models: they inherently cannot describe positive, diversifying selection. Indeed, in Appendix 1, we prove that scaled selection coefficients strictly yield $dN/dS \leq 1$, under the assumptions that synonymous changes are neutral and mutation is symmetric. This proof formalizes the MutSel model's underlying assumption that selection pressure is constant over the phylogeny and that the protein evolves under equilibrium conditions. Although this proof assumes symmetric mutation rates, we have found numerically that dN/dS remains bounded from above by 1 even when mutations rates are asymmetric (Figure S1).

Purifying selection on synonymous changes can produce $dN/dS > 1$.

The restriction $dN/dS \leq 1$ does not hold when synonymous changes are not neutral, as seen in Figure 1B. In other words, even though the Halpern-Bruno model explicitly assumes that the system is at equilibrium (Halpern and Bruno 1998; Thorne et al. 2012), dN/dS can readily be greater than

1. In fact, it is theoretically possible to achieve arbitrarily high dN/dS values when synonymous codon substitutions carry fitness changes. In the most extreme case of synonymous selection, where only a single codon per amino acid is selectively tolerated, the number of synonymous changes becomes $K_S = 0$, and thus the value for dN/dS approaches infinity. In other words, $dN/dS > 1$ may indicate either positive, diversifying selection on amino acids or strong purifying selection on synonymous codons.

Given that the MutSel model framework assumes an overarching regime of purifying selection, this finding might seem paradoxical. However, the logical argument that $dN/dS > 1$ represents positive, diversifying selection assumes that the rate of synonymous change may be used as a neutral benchmark, an assumption clearly violated when selection acts on synonymous changes. Thus, the traditional signal of positive, diversifying selection, a dN/dS value in excess of one, can result simply from strong synonymous fitness differences.

That sequences evolving under purifying selection can spuriously bear the hallmark of positive, diversifying selection highlights the pitfalls of naively interpreting dN/dS values. Indeed, evolutionary constraints which induce synonymous selection are pervasive and affect virtually all domains of life (Gu et al. 2010), from viruses (Cuevas et al. 2011; Zanini and Neher 2013) to plants (Gu et al. 2012) to Metazoa (Duret 2002; Chamary et al. 2006; Hershberg and Petrov 2008; Plotkin and Kudla 2011; Lawrie et al. 2013). Recent work has shown that synonymous rate variation is common across myriad proteins and contributes to evolutionary rate heterogeneity in up to 42% of known protein families (Dimitrieva and Anisimova 2014). For example, exonic splicing enhancers (Schattner and Diekhans 2006; Parmley et al. 2006; Parmley and Hurst 2007), regions contributing to mRNA secondary structure such as translation-initiation sites (Chamary and Hurst 2005; Schattner and Diekhans 2006; Gu et al. 2010; Cuevas et al. 2011; Zanini and Neher 2013), and DNA- and RNA-binding sites (Parmley et al. 2006) all experience moderate to strong synonymous selection. It has additionally been suggested that up to 18% of mutational fitness effects in RNA viruses, whose genomes frequently feature sites with $dN/dS > 1$ (Bush et al. 1999; Suzuki 2006; Bhatt et al. 2011; Meyer et al. 2013; Meyer and Wilke 2013), are caused by selection acting on synonymous changes (Cuevas et al. 2011). Finally, both selection against protein mis-folding and for translation efficiency tend to induce synonymous selection in a gene-specific manner (Williford and Demuth 2012; Agashe et al. 2013), most notably in highly expressed genes (Drummond and Wilke 2008; Lawrie et al. 2013). Therefore, while synonymous selection may not dominate genomes in organisms with relatively small effective population sizes (Chamary et al. 2006; Plotkin and Kudla

2011), it certainly acts strongly at specific sites and/or small, local regions. As dN/dS ratios are typically measured on a per-site basis, we expect that some sites with $dN/dS > 1$ may in fact be false positives in the detection of positive, diversifying selection. We offer several approaches to ease this concern in *Conclusions*.

Relationship between *dN/dS* and scaled selection coefficients provides a novel benchmarking approach.

The relationship we have established between dN/dS and scaled selection coefficients offers a unique opportunity to assess the robustness of dN/dS-based inference methods. It is conventional practice in model development to benchmark models against data simulated according to the model itself. While crucial for testing whether a given model has been correctly implemented, this strategy inherently cannot discern how the model behaves when data arose from a different mechanistic process. To overcome this limitation, we applied a novel benchmarking approach which used the theoretical relationship among modeling frameworks to assess the accuracy and specific utility of those models. Outlined in Figure 2A, this approach entails comparing dN/dS values calculated from selection coefficients to those inferred by an dN/dS-based model. In addition, because MutSel models are based explicitly on population genetics theory, these simulated alignments are likely far more similar to real sequence data than are alignments simulated under a dN/dS-based model, and therefore provide excellent benchmarking data.

Using the selection coefficients and symmetric mutation rates from the previous two subsections, we simulated alignments using standard methods (Yang 2006) according to the Halpern-Bruno MutSel model (Halpern and Bruno 1998). We then inferred a dN/dS value for each alignment using the GY94 matrix (Goldman and Yang 1994; Nielsen and Yang 1998), which estimates dN/dS with the parameter ω. Throughout the remaining text, we refer to dN/dS inferred using ML as ω or ω maximum likelihood estimate (MLE), and to dN/dS computed using equations (6) - (9) simply as dN/dS.

We found that dN/dS values agree nearly perfectly with ω MLEs (Figure 2B), and indeed this relationship was robust to both synonymous selection and uneven nucleotide composition (simulated alignments featured GC contents ranging from 0.21-0.89). Additionally, Figure 2C demonstrates that ω converged to the true dN/dS value as the size of the data set (i.e., simulated alignment length) increased. These results unequivocally show that, when nucleotide mutation is symmetric, dN/dS-based model-inference methods behave exactly as expected, yielding precisely

accurate dN/dS estimates. This finding has important implications for modeling choices; although the MutSel framework might model the sequence evolution in a way that more mechanistically matches the evolutionary process, dN/dS-based models may suffice to model selective forces in phylogenetic data.

Biased *dN/dS* estimates under asymmetric mutation models.

We next sought to test the accuracy of dN/dS-based models using more realistic parameter values. To this end, we determined codon fitness distributions from 498 unique distributions of experimentally-derived, site-specific amino-acid fitnesses for H3N2 influenza nucleoprotein (NP) (Bloom 2014a). We combined each of these fitness distributions with three sets of experimentally-determined mutation rates, either for NP (Bloom 2014a), yeast (Zhu et al. 2014), or polio virus (Acevedo et al. 2014), to determine $498 \times 3 = 1494$ distinct distributions of steady-state codon frequencies (see *Methods* for details). While all three mutation matrices were asymmetric, each featured a differing degree of mutational bias; specifically, the mean ratios μ_{ij}/μ_{ji} for NP, yeast, and polio mutation rates are 1.03, 1.69, and 5.25, respectively. For each resulting set of stationary codon frequencies, in combination with its respective set of mutation rates, we calculated dN/dS and simulated alignments from which we inferred ω. Note that we assumed no selection on synonymous codons for these calculations.

dN/dS-based model matrices account for nucleotide mutational bias by incorporating either target codon (Goldman and Yang 1994) or target nucleotide (Muse and Gaut 1994) frequencies; these frameworks are known, respectively, as GY-style and MG-style models (Kosakovsky Pond et al. 2010). For example, the instantaneous rate matrix element giving the substitution probability from codon AAA to AAG would contain the target codon frequency P_{AAG} in GY-style models but the target nucleotide frequency π_{G} in MG-style models. Moreover, the GY-style models conform explicitly to a general-time reversible (GTR) form, whereas MG-style matrices do not, at first glace, appear to follow the same framework. However, as we show in Appendix 2, it is indeed posible to write MG-style matrices such that they conform to the GTR framework. This insight explicitly justifies using a time-reversible Markov process to describe these models, and it additionally demonstrates that the $F1x4$ codon frequency estimator (Muse and Gaut 1994) represents the state frequencies of an MG-style model.

Previous works have suggested that MG-style and GY-style models yield different ω estimates (Kosakovsky Pond and Muse 2005; Yap et al. 2010), so we inferred ω according to both GY- and

MG-style frameworks. For GY-style models, we used the frequency estimators F61 (Goldman and Yang 1994), F3x4 (Goldman and Yang 1994), CF3x4 (Kosakovsky Pond et al. 2010), and F1x4 (Muse and Gaut 1994). For MG-style models, we considered both a parameterization with four global nucleotide frequency parameters and a parameterization which employed twelve nucleotide frequency parameters to allow for different frequencies at each codon position. We term the former framework MG1 and the latter MG3. Note that our MG1 corresponds to the original MG-style model (Muse and Gaut 1994), whereas our MG3 corresponds to the so-called MG94×HKY85 model (Kosakovsky Pond and Muse 2005).

Figure 3 shows the resulting relationships between dN/dS and ω MLEs for each set of mutation rates (NP, yeast, and polio), across model frequency parameterizations. Figure 3A displays the estimator bias, defined as the average difference between the true dN/dS value and the ω MLEs. Figure 3B displays the precision in this relationship, measured by the squared correlation coefficient r^2 between dN/dS and ω. The exact bias and r^2 values are given in Tables S1 and S2, respectively, and full regression plots for ω vs. dN/dS are shown in Figure S1.

Two distinct trends emerge from Figure 3. First, asymmetry in the mutational process consistently induced significant bias in ω estimates. Most often, the model underestimated ω relative to the true dN/dS value. Based on simulations without any selection ($dN/dS = 1$), ref. (Yap et al. 2010) had previously suggested that GY-style models produce negatively biased ω estimates. Our results generalize this finding and show that this bias is pervasive, remains approximately constant through a wide range of dN/dS values, and is not limited to the GY-style framework (Figure 3A, Table S1, Figure S1). Furthermore, we show that this bias systematically increased in magnitude as the underlying mutational process became more asymmetric. Indeed, for all frequency parameterizations, ω MLEs were most accurate under NP mutation rates, and both accuracy and precision tended to decrease as mutational bias progressed from yeast to polio mutation rates. Second, frequency parameterizations which more closely matched the mechanistic process that generated the data (MG1 and MG3) generally outperformed all other frequency estimators. In particular, MG1 clearly performed the best of all frequency estimators considered, featuring by far the least amount of estimator bias for the highly asymmetric polio mutation rates. Thus, going forward, we highly recommend that researchers employ MG-style matrices in their dN/dS inferences to minimize bias (we note that this modeling framework is available through HyPhy (Kosakovsky Pond et al. 2005) and/or the Datamonkey server (Delport et al. 2010)).

Strikingly, when we examined model fit using AIC scores (Akaike 1974; Burnham and Anderson

2004) for the different frequency parameterizations, we found that the F61 parameterization was unequivocally the best-performing model, on average, for all datasets (Table 1). This result dramatically juxtaposed the substantial inaccuracy and imprecision in ω that F61 frequently yielded. In particular, F61 had the most estimator bias for NP datasets as well as the least precision for both NP and polio datasets (Figure 3). Thus, we found AIC could not identify the model which produced the most accurate estimates for the parameter of interest.

Although this result may seem counterintuitive, it is important to note that AIC approximates the Kullback-Leibler (KL) distance between a given candidate model and the true model. As the Halpern-Bruno MutSel framework defines selection coefficients in terms of stationary frequencies, it indeed follows that the F61 estimator, which explicitly incorporates empirical codon frequencies into the rate matrix, should be selected as the best-fitting model, in spite of its biased parameter estimates. Therefore, we additionally assessed whether BIC might provide a more accurate indication of model performance. However, BIC scores yielded the same overarching trend as did AIC scores in which F61 dramatically outperformed all other frequency parameterizations (Table S3).

This finding has broad implications for practices in model selection. In particular, it appears that model fit can be confounded with model accuracy, such that the model with better model fit may produce less accurate parameter estimates. We find that, if the data are generated by a process distinct from the inference model, standard model selection quantities cannot necessarily identify which model produces the most precise and least biased parameter estimates. Good model fit, therefore, may not have any bearing on whether using that model is mechanistically justified, and selecting models based solely on fit may not guard effectively against spurious inferences but instead prove misleading. We suggest that the mechanism producing the data should be carefully considered, and an appropriate inference method which best approximates this process should then be selected.

Finally, these results provide a concrete example of previous theoretical suggestions that AIC might fail in phylogenetic model selection (Liberles et al. 2013). Indeed, previous work has suggested that Bayes Factors might serve as a better indication of model performance than AIC, albeit results were obtained in a Bayesian rather than frequentist framework (Rodrigue et al. 2008). Therefore, it is clear further investigation into the performance of various model fit criteria for complex models is strongly warranted.

Conclusions

By elucidating the relationship between dN/dS and scaled selection coefficients, we have shown that dN/dS-based and MutSel models convey consistent information regarding the strength of natural selection. Importantly, our proof that $dN/dS \leq 1$ (assuming symmetric mutation and no synonymous selection) indicates that the use of MutSel models is only justified under conditions of strictly purifying selection. This restriction is in part indicated by the basic MutSel model assumption of constant selection pressures over time, or in other words a static fitness landscape (Halpern and Bruno 1998; Thorne et al. 2007; Rodrigue et al. 2010; Thorne et al. 2012). Thus, if the aim is to identify positive, diversifying selection, of the two frameworks examined here, only dN/dS-based models are appropriate.

However, we have also found that dN/dS values can readily be greater than 1 when selection acts on synonymous mutations, even though the protein sequence is evolving solely under purifying selection. This seemingly paradoxical finding actually reflects an assumption violation; the assertion that $dN/dS > 1$ necessarily corresponds to positive, diversifying selection requires that synonymous changes are neutral, which clearly does not hold if there are fitness differences among synonymous codons. This result contributes to a growing body of literature which has found that purifying selection can yield $dN/dS > 1$ if model assumptions are not met. For instance, dN/dS can be greater than 1, even under purifying selection, if sequences contain segregating polymorphisms (Rocha et al. 2006; Kryazhimskiy and Plotkin 2008; Mugal et al. 2014). Thus, it is becoming increasingly clear that the $dN/dS = 1$ neutral threshold typically used to distinguish purifying and positive selection is highly sensitive to violations in model assumptions. We emphasize that it is crucial to ensure that data adhere to model assumptions before conclusions from dN/dS are drawn.

We suggest several strategies to limit such false positive results under synonymous selection. For one, certain formulations of dN/dS-based methods (particularly those implemented in the HyPhy package (Kosakovsky Pond et al. 2005)) consider dN and dS rate variation separately (Muse and Gaut 1994; Kosakovsky Pond and Muse 2005; Murrell et al. 2013) rather than using a single parameter to represent dN/dS. These kinds of methods, and others which explicitly model nucleotide-level selection in conjunction with codon-level selection (Rubinstein et al. 2011), may be able to distinguish situations in which $dN/dS > 1$ because dN is unusually large (positive selection) or dS unusually small (purifying selection). In addition, dN/dS-based models which correct dS for synonymous selection may prove fruitful in the future for teasing out the source of elevated dN/dS

estimates (Zhou et al. 2010; Rubinstein et al. 2011). Finally, our benchmarking approach, in which we simulate sequences according to MutSel models and infer dN/dS both from MutSel parameters directly and using ML, may be used to benchmark these kinds of models and may help to identify circumstances under which synonymous selection confounds dN/dS interpretations.

Finally, we emphasize the utility of establishing relationships among distinct modeling frameworks to probe model behavior and evaluate model performance. Such an approach is uniquely able to reveal unrecognized behaviors and/or limitations of different modeling frameworks and can precisely reveal the circumstances in which different models are best suited. We hope that further studies in this spirit will ensure robust model development in future studies.

Methods

Simulation of scaled selection coefficients.

To examine the relationship between dN/dS and scaled selection coefficients, we simulated 200 distributions of amino-acid scaled fitness values, $F_a^{\text{aa}} = 2Nf_a^{\text{aa}}$, from a normal distribution $\mathcal{N}(0, \sigma^2)$, where a unique σ^2 for each fitness distribution was drawn from a uniform distribution $\mathcal{U}(0, 4)$. We converted these amino-acid fitnesses to codon fitnesses as follows. For 100 of the fitness distributions, we directly assigned all codons within a given amino acid family the fitness $F_i^{\text{codon}} = F_a^{\text{aa}}$, so that all synonymous codons had the same fitness. For the other 100 fitness distributions, we assigned synonymous codons different fitnesses by randomly drawing a preferred codon for each amino acid. This preferred codon was assigned the fitness of $F_i^{\text{codon}} = F_a^{\text{aa}} + \lambda$, and all non-preferred codons were given the fitness $F_j^{\text{codon}} = F_a^{\text{aa}} - \lambda$. We drew a unique λ for each fitness distribution from $\mathcal{U}(0, 2)$. We then computed stationary codon frequencies as

$$P_i = \frac{e^{F_i^{\text{codon}}}}{\sum_k e^{F_k^{\text{codon}}}}, \tag{10}$$

where the sum in the denominator runs over all 61 sense codons (Sella and Hirsh 2005). Equation (10) gives the analytically precise stationary frequencies for a MutSel model, under the assumption of symmetric mutation rates (Sella and Hirsh 2005). We used equations (6) - (9) to compute dN/dS for each resulting set of stationary codon frequencies. For these calculations, we assumed the HKY85 (Hasegawa et al. 1985) nucleotide mutation model, and accordingly we set the transition mutation rate as $\mu\kappa$ and the transversion rate as μ. We used the value $\mu = 10^{-6}$ for all dN/dS calculations, and we drew a unique value for κ from $\mathcal{U}(1, 6)$ for each set of codon frequencies.

Alignment simulations.

We simulated protein-coding sequences as a continuous-time Markov process using standard methods (Yang 2006) according to the Halpern-Bruno MutSel model (Halpern and Bruno 1998). In brief, this model's instantaneous rate matrix $Q = q_{ij}$ is populated by elements

$$q_{ij} = \begin{cases} m_{ij} \frac{S_{ij}}{1-1/S_{ij}} & \text{single nucleotide change} \\ \\ 0 & \text{multiple nucleotide changes} \end{cases}, \tag{11}$$

for a mutation from codon i to j, where m_{ij} is the mutation rate, and the scaled selection coefficient S_{ij} is defined in equation (5). All alignments presented here were simulated along a 4-taxon phylogeny (Figure 4), beginning with a root sequence generated in proportion to stationary codon frequencies (Yang 2006). Unless otherwise stated, all simulated alignments contained 500,000 codon positions. A single evolutionary model was applied to all positions in the simulated sequences. While this lack of site-wise heterogeneity is unrealistic for real sequence evolution, it allowed us to verify our derived relationship between scaled selection coefficients and dN/dS with a sufficiently sized data set.

Computation of stationary frequencies for experimental data sets.

We used experimentally-determined site-specific amino-acid fitness parameters F_a^{aa} for influenza nucleoprotein (NP), from ref. (Bloom 2014a), in combination with experimental nucleotide mutation rates for either NP (Bloom 2014a), yeast (Zhu et al. 2014), or polio virus (Acevedo et al. 2014), to derive realistic distributions of stationary codon frequencies. We combined each of the 498 site-wise amino-acid preference sets reported by ref. (Bloom 2014a) with each of the three mutation-rate matrices to construct a total of $498 \times 3 = 1494$ unique experimental evolutionary Markov models, using the approach in refs. (Bloom 2014a,b). The instantaneous rate matrix Q for each experimental model is populated by elements

$$q_{ij} = \begin{cases} F_j^{\text{codon}} / F_i^{\text{codon}} m_{ij} & \text{if } F_j^{\text{codon}} \geq F_i^{\text{codon}} \\ m_{ij} & \text{if } F_j^{\text{codon}} < F_i^{\text{codon}} \\ 0 & \text{multiple nucleotide changes} \end{cases} \tag{12}$$

for a substitution from codon i to codon j, where F_i^{codon} is the fitness of codon i (Bloom 2014a,b). We calculated F_i^{codon} values by simply assigning a given amino acid's experimental fitness F_a^{aa} to

each of its constituent codons; thus, all synonymous changes were neutral. We determined the stationary codon frequencies for each resulting experimental model from the matrix's eigenvector corresponding to the eigenvalue 0. Finally, we simulated alignments for each set of stationary frequencies and corresponding mutation rates according to the Halpern-Bruno model (equation (11)).

Maximum likelihood inference of *dN/dS*.

For the 200 alignments simulated with symmetric mutation rates, we inferred dN/dS using the M0 model (Yang et al. 2000), as implemented in the HyPhy batch language (Kosakovsky Pond et al. 2005). The M0 model uses the GY94 instantaneous rate matrix, which is populated by elements

$$q_{ij} = \begin{cases} P_j & \text{synonymous transversion} \\ \kappa P_j & \text{synonymous transition} \\ \omega P_j & \text{nonsynonymous transversion} \\ \omega\kappa P_j & \text{nonsynonymous transition} \\ 0 & \text{multiple nucleotide changes} \end{cases}, \tag{13}$$

for a substitution from codon i to codon j, where κ is the transition-transversion bias, P_j is the equilibrium frequency of the target codon j, and ω represents dN/dS (Goldman and Yang 1994; Nielsen and Yang 1998). The P_i parameters are intended to represent those codon frequencies which would exist in absence of selection pressure generated by mutation alone (Goldman and Yang 1994; Muse and Gaut 1994; Yang and Nielsen 2000; Yang 2006). Thus, when inferring ω on datasets which used symmetric mutation rates, we assigned the value $1/61$ to all parameters P_i, as all codons are equally probable in the absence of mutational bias.

Alternatively, when inferring ω for alignments simulated with experimental fitness and mutation rates, we used several different model parameterizations, including GY-style (Goldman and Yang 1994) (target codon frequency) and MG-style (Muse and Gaut 1994) (target nucleotide frequency) parameterizations. We considered the GY-style parameterizations F61 (Goldman and Yang 1994), F3x4 (Goldman and Yang 1994), CF3x4 (Kosakovsky Pond et al. 2010), and F1x4 (Muse and Gaut 1994). We implemented two varieties of MG-style models; the first, MG1, employs four parameters for nucleotide frequencies (one per nucleotide) (Muse and Gaut 1994), and the second, MG3, employs twelve nucleotide frequency parameters, with four nucleotide frequency parameters for each of the three codon positions (Kosakovsky Pond and Muse 2005). All models included the

parameters κ and ω. Note that we used the state frequencies of F1x4 for the MG1 framework and F3x4 for the MG3 framework.

Availability.

All code is freely available from
`https://github.com/clauswilke/Omega_MutSel`. Alignments were simulated using pyvolve, available from
`https://github.com/sjspielman/pyvolve`. Simulated alignments are available from Dryad at doi:XXX (data will be submitted upon acceptance of this article).

Appendix 1

We prove that $dN/dS \leq 1$ when calculated from scaled selection coefficients. We assume that mutation rates are symmetric ($m_{ij} = m_{ji}$) and that synonymous codons have the same fitness (synonymous changes are neutral). As described in the main text, these assumptions yield $dS = 1$, and hence we have to show that $dN = K_\mathrm{N}/L_\mathrm{N} \leq 1$. To this end, we note that the sums in K_N and L_N can be reordered such that the substitution probability from codon i to j is always added to the substitution probability from codon j to i. We can then show that the sum of each of these pairs in the expression for K_N is smaller than the corresponding term in L_N, and hence $dN/dS \leq 1$.

Without loss of generality, we consider a pair of nonsynonymous codons i and j whose respective stationary frequencies P_i and P_j satisfy $P_i \leq P_j$ and $P_i, P_j \geq 0$. As follows from equations (2) and (5), the sum of the probability weights of evolving from codon i to j and from codon j to i is

$$N_e m_{ij} u_{ij} + N_e m_{ji} u_{ji} = \frac{2P_iP_j[\log(P_i) - \log(P_j)]}{P_i - P_j}\,. \tag{14}$$

This quantity represents K_N in the dN calculation. To prove $dN \leq 1$, we must show that this quantity is less than or equal to $P_i + P_j$, which represents L_N in the dN calculation. To this end, we introduce the function

$$F(x,y) = x + y - \frac{2xy[\log(x) - \log(y)]}{x - y}\,, \tag{15}$$

and we will now show that $F(x,y) \geq 0$ for $x \leq y$ and $y \geq 0$. It is straightforward to show, using l'Hôpital's rule, that this condition holds for $x = y$. For $x < y$, we show that the first derivative of equation (15) is negative throughout $x \in (0,y)$, which proves that the function monotonically

decreases, and thus $F(x,y) > 0$, in this interval. We calculate the first derivative as

$$\frac{\partial F(x,y)}{\partial x} = \frac{\big[(x-3y)(x-y) - 2y^2(\log x - \log y)\big]}{(x-y)^2}\,. \tag{16}$$

We now replace the expression $\log x - \log y$ by its Taylor expansion, yielding

$$\frac{\partial F(x,y)}{\partial x} = \frac{\left[(x-3y)(x-y) - 2y^2\left(\sum\limits_{n=1}^{\infty} \frac{1}{n}(1-x/y)^n\right)\right]}{(x-y)^2}\,. \tag{17}$$

We note that the first two terms of the Taylor series equal $(x-3y)(x-y)$, and thus expression (17) simplifies to

$$\frac{\partial F(x,y)}{\partial x} = \frac{-2y^2 \sum\limits_{n=3}^{\infty} \frac{1}{n}\big(1-\frac{x}{y}\big)^n}{(x-y)^2}, \tag{18}$$

which is clearly negative. This concludes the proof.

Appendix 2

GY-style matrices may be expressed in the framework of the general-time reversible (GTR) model, in which the instantaneous matrix Q can be decomposed into a 61×61 symmetric substitution rate matrix and a 61-dimensional vector containing the equilibrium codon frequencies. The latter corresponds to the stationary distribution of the Markov chain. By contrast, MG-style rate matrices are written in terms of nucleotide frequencies rather than codon frequencies. Therefore, whether these models fit into the GTR framework is unclear *a priori*. We now describe how the MG-style matrix can be rewritten in terms of a symmetric matrix and a vector of equilibrium codon frequencies, thus demonstrating that these matrices also fit into the GTR framework.

MG-style matrix elements, for a the substitution from codon i to j, are generally given by

$$q_{ij} = \begin{cases} \theta_{s_i t_j}\pi_{t_j} & \text{synonymous change} \\ \omega\theta_{s_i t_j}\pi_{t_j} & \text{nonsynonymous change} \\ 0 & \text{multiple nucleotide changes} \end{cases}, \tag{19}$$

where ω is the ratio of nonsynonymous to synonymous substitution rates and the product $\theta_{s_i t_j}\pi_{t_j}$ corresponds to a nucleotide-level mutation rate $\mu_{s_i t_j}$, where s_i is the source nucleotide in codon i, and t_j is the target nucleotide in codon j. Note that the matrix $\theta_{s_i t_j}$ is symmetric in s_i and t_j.

For a given codon i, the matrix of Eq. (19) yields the stationary frequency $P_i = \pi_{i_1}\pi_{i_2}\pi_{i_3}C$, where $C = 1 - \Pi_{\text{stop}}$ and $\Pi_{\text{stop}} = \pi_{\text{T}}\pi_{\text{A}}\pi_{\text{G}} + \pi_{\text{T}}\pi_{\text{G}}\pi_{\text{A}} + \pi_{\text{T}}\pi_{\text{A}}\pi_{\text{A}}$ (Muse and Gaut 1994). Therefore,

we can rewrite the term $\theta_{s_i t_j}\pi_{t_j}$ as $\theta_{s_i t_j} P_j C/(\pi_m \pi_n)$, where m and n are the nucleotides which do not change in a given instantaneous codon substitution. This allows us to rewrite the rate instantaneous matrix as

$$q_{ij} = \begin{cases} \frac{C\theta_{s_i t_j}}{\pi_m \pi_n} P_j & \text{synonymous change from } i \text{ to } j \\ \omega \frac{C\theta_{s_i t_j}}{\pi_m \pi_n} P_j & \text{nonsynonymous change from } i \text{ to } j \\ 0 & \text{multiple nucleotide changes} \end{cases} \tag{20}$$

for a substitution from codon i to codon j, and this matrix clearly conforms to the GTR framework.

Acknowledgements

This work was supported in part by NIH grant R01 GM088344, ARO grant W911NF-12-1-0390, DTRA grant HDTRA1-12-C-0007, and NSF Cooperative Agreement No. DBI-0939454 (BEACON Center). Computational resources were provided by the University of Texas at Austin's Center for Computational Biology and Bioinformatics (CCBB).

References

Acevedo A, Brodsky L, Andino R. 2014. Mutational and fitness landscapes of an RNA virus revealed through population sequencing. Nature 505:686 – 690.

Agashe D, Martinez-Gomez N C, Drummond D A, Marx C J. 2013. Good codons, bad transcript: Large reductions in gene expression and fitness arising from synonymous mutations in a key enzyme. Mol Biol Evol 30(3):549–560.

Akaike H. 1974. A new look at the statistical model identification. IEEE Transactions on Automatic Control 19:6:716–723.

Anisimova M, Kosiol C. 2009. Investigating protein-coding sequence evolution with probabilistic codon substitution models. Mol Biol Evol 26:255–271.

Bhatt S, Holmes E C, Pybus O G. 2011. The genomic rate of molecular adaptation of the human influenza A virus. Mol Biol Evol 28:2443–2451.

Bloom J D. 2014a. An experimentally determined evolutionary model dramatically improves phylogenetic fit. Mol Biol Evol 31:1956 – 1978.

Bloom J D. 2014b. An experimentally informed evolutionary model improves phylogenetic fit to divergent lactamase homologs. Mol Biol Evol 31:1956–1978.

Burnham K P, Anderson D R. 2004. Multimodel inference: understanding AIC and BIC in model selection. Sociol Method Res 33:261–304.

Bush R M, Bender C A, Subbarao K, Cox N J, Fitch W M. 1999. Predicting the evolution of human influenza A. Science 286:1921–1925.

Chamary J V, Hurst L D. 2005. Evidence for selection on synonymous mutations affecting stability of mRNA secondary structure in mammals. Genome Biol 6(9):R75.

Chamary J V, Parmley J L, Hurst L D. 2006. Hearing silence: non-neutral evolution at synonymous sites in mammals. Nature Rev Genet 7:98–108.

Crow J F, Kimura M. 1970. An Introduction to Population Genetics Theory. California: Burgess Pub. Co.

Cuevas J M, Domingo-Calap P, Sanjuan R. 2011. The fitness effects of synonymous mutations in DNA and RNA viruses. Mol Biol Evol 29(1):17–20.

Delport W, Poon A, Frost S, Pond S. 2010. Datamonkey 2010: a suite of phylogenetic analysis tools for evolutionary biology. Bioinformatics 26(19):2455–2457.

Dimitrieva S, Anisimova M. 2014. Unraveling patterns of site-to-site synonymous rates variation and associated gene properties of protein domains and families. PLoS ONE 9(6):e95034.

Drummond D A, Wilke C O. 2008. Mistranslation-Induced Protein Misfolding as a Dominant Constraint on Coding-Sequence Evolution. Cell 134(2):341–352.

Duret L. 2002. Evolution of synonymous codon usage in metazoans. Curr. Opin. Genet. Dev. 12:640–649.

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

Gu W, Wang X, Zhai C, Xie X, Zhou T. 2012. Selection on synonymous sites for increased accessibility around miRNA binding sites in plants. Mol Biol Evol 29(10):3037–3044.

Gu W, Zhou T, Wilke C O. 2010. A universal trend of reduced mRNA stability near the translation-initiation site in prokaryotes and eukaryotes. PLoS Comput Biol 6(2):e1000664.

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

Hasegawa M, Kishino H, Yano T. 1985. Dating the humanape splitting by a molecular clock of mitochondrial DNA. J Mol Evol 22:160 – 174.

Hershberg R, Petrov D. 2008. Selection on codon bias. Annu Rev Genet 42.

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

Kimura M. 1962. On the probability of fixation of mutant genes in a population. Genetics 4:713–719.

Kosakovsky Pond S, Delport W, Muse S, Scheffler K. 2010. Correcting the bias of empirical frequency parameter estimators in codon models. PLoS One 5:e11230.

Kosakovsky Pond S, Frost S. 2005a. A genetic algorithm approach to detecting lineage-specific variation in selection pressure. Mol Biol Evol 22:478–485.

Kosakovsky Pond S, Frost S. 2005b. Not so different after all: A comparison of methods for detecting amino acid sites under selection. Mol Biol Evol 22:1208–1222.

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

Kosakovsky Pond S, Muse S. 2005. Site-to-site variation of synonymous substitution rates. Mol Biol Evol 22:2375–2385.

Kosakovsky Pond S L, Frost S D W, Muse S V. 2005. HyPhy: hypothesis testing using phylogenetics. Bioinformatics 21:676–679.

Kryazhimskiy S, Plotkin J B. 2008. The population genetics of *dN/dS*. PLoS Genet 4:e1000304.

Lawrie D, Messer P, Hershberg R, Petrov D. 2013. Strong purifying selection at synonymous sites in *D. melanogaster*. PLoS Genet 9(5):e1003527.

Liberles D, Teufel L AI Liu, Stadler T. 2013. On the need for mechanistic models in computational genomics and metagenomics. Genome Biol Evol 5:2008 – 2018.

Meyer A G, Dawson E T, Wilke C O. 2013. Cross-species comparison of site-specific evolutionary-rate variation in influenza hemagglutinin. Phil Trans R Soc B 368:1614.

Meyer A G, Wilke C O. 2013. Integrating sequence variation and protein structure to identify sites under selection. Mol Biol Evol 30:36–44.

Minin V, Abdo Z, Joyce P, Sullivan J. 2003. Performance-based selection of likelihood models for phylogeny estimation. Syst Biol 52:674 – 683.

Mugal C F, Wolf J B W, Kaj I. 2014. Why time matters: Codon evolution and the temporal dynamics of *dN/dS*. Mol Biol Evol 31:212–231.

Murrell B, Moola S, Mabona A, Weighill T, Scheward D, Kosakovsky Pond S L, Scheffler K. 2013. FUBAR: A Fast, Unconstrained Bayesian AppRoximation for inferring selection. Mol Biol Evol 30:1196–1205.

Murrell B, Wertheim J O, Moola S, Weighill T, Scheffler K, Pond S L K. 2012. Detecting individual sites subject to episodic diversifying selection. PloS Genet 8:e1002764.

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

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

Parmley J L, Chamary J V, Hurst L D. 2006. Evidence for purifying selection against synonymous mutations in mammalian exonic splicing enhancers. Mol Biol Evol 23:301–309.

Parmley J L, Hurst L D. 2007. Exonic splicing regulatory elements skew synonymous codon usage near intron-exon boundaries in mammals. Mol Biol Evol 24(8):1600–1603.

Plotkin J B, Kudla G. 2011. Synonymous but not the same: the causes and consequences of codon bias. Nature Rev Genet 12:32–42.

Robinson D M, Jones D T, Kishino H, Goldman N, Thorne J L. 2003. Protein evolution with dependence among codons due to tertiary structure. Mol Biol Evol 20:1692–1704.

Rocha E, Maynard Smith J, Hurstd L, Holden M, Cooper J, Smith N, Feil E. 2006. Comparisons of *dN/dS* are time dependent for closely related bacterial genomes. J Theor Biol 239:226 – 235.

Rodrigue N, Kleinman C, Phillipe H, Lartillot N. 2000. Computational methods for evaluating phylogenetic models of codong sequence evolution with dependence between codons. Mol Biol Evol 26(7):1663–1676.

Rodrigue N, Lartillot N. 2014. Site-heterogeneous mutation-selection models within the PhyloBayes-MPI package. Bioinformatics 30:1020–1021.

Rodrigue N, Lartillot N, Phillipe H. 2008. Bayesian comparisons of codon substitution models. Genetics 180:1579 – 1591.

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

Rubinstein N, Faigenboim-Doron A, Mayrose I, Pupko T. 2011. Evolutionary models accounting for layers of selection in protein-coding genes and their impact on the inference of positive selection. Mol Biol Evol 28:3297 – 3308.

Schattner P, Diekhans M. 2006. Regions of extreme synonymous codon selection in mammalian genes. Nucleic Acids Res 34(6):1700–1710.

Scherrer M P, Meyer A G, Wilke C O. 2012. Modeling coding-sequence evolution within the context of residue solvent accessibility. BMC Evol Biol 12:179.

Schoniger M, von Haeseler A. 1995. Performance of the maximum likelihood, neighbor joining, and maximum parsimony methods when sequence sites are not independent. Syst Biol 44:533 – 547.

Sella G, Hirsh A E. 2005. The application of statistical physics to evolutionary biology. Proc Natl Acad Sci USA 102:9541–9546.

Suzuki Y. 2006. Natural selection on the influenza virus genome. Mol Biol Evol 23:1902–1911.

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

Tamuri A U, Goldman N, dos Reis M. 2014. A penalized-likelihood method to estimate the distribution of selection coefficients from phylogenetic data. Genetics 197:257–271.

Thorne J, Choi S, Yu J, Higgs P, Kishino H. 2007. Population genetics without intraspecific data. Mol Biol Evol 24(8):1667–1677.

Thorne J L, Lartillot N, Rodrigue N, Choi S C. 2012. Codon models as vehicles for reconciling population genetics with inter-specific data. In G Cannarozzi, A Schneider, editors, Codon evolution: mechanisms and models, New York: Oxford University Press.

Williford A, Demuth J P. 2012. Gene expression levels are correlated with synonymous codon usage, amino acid composition, and gene architecture in the red flour beetle, Tribolium castaneum. Mol Biol Evol 29(12):3755–3766.

Yang Z. 2006. Computational Molecular Evolution. Oxford University Press.

Yang Z. 2007. PAML 4: Phylogenetic analysis by maximum likelihood. Mol Biol Evol 24:1586–1591.

Yang Z, Nielsen R. 2000. Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. Mol Biol Evol 17:32–42.

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

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

Yap V B, Lindsay H, Easteal S, Huttley G. 2010. Estimates of the effect of natural selection on protein-coding content. Mol Biol Evol 27:726 – 734.

Zanini F, Neher R A. 2013. Quantifying Selection against Synonymous Mutations in HIV-1 env Evolution. J Virol 87(21):11843–11850.

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.

Zhou T, Gu W, Wilke C O. 2010. Detecting positive and purifying selection at synonymous sites in yeast and worm. Mol Biol Evol 27(8):1912–1922.

Zhu Y O, Siegal M L, Hall D W, Petrov D A. 2014. Precise estimates of mutation rate and spectrum in yeast. Proc Natl Acad Sci USA 111(22):E2310 – E2318.

Figures and Tables

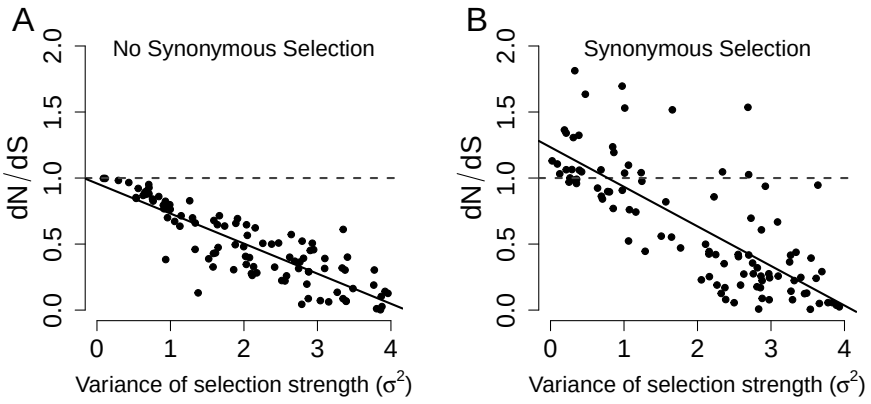

Figure 1: dN/dS decreases in proportion to amino-acid level selection strength. dN/dS is plotted against the variance (σ^2) of the simulated distribution of amino-acid scaled fitness values. Higher variances indicate larger fitness differences among amino acids, whereas the limiting value of $\sigma^2 = 0$ indicates that all amino acids have the same fitness. (A) Synonymous codons have equal fitness values ($r^2 = 0.83, P < 2^{-16}$). (B) Synonymous codons have different fitness values ($r^2 = 0.45, , P < 2^{-16}$). Note that panel B, but not A, shows dN/dS values greater than 1, in spite of the steady-state evolutionary process. In each panel, the dashed line indicates the $y = 1$ line, and the solid line indicates the regression line.

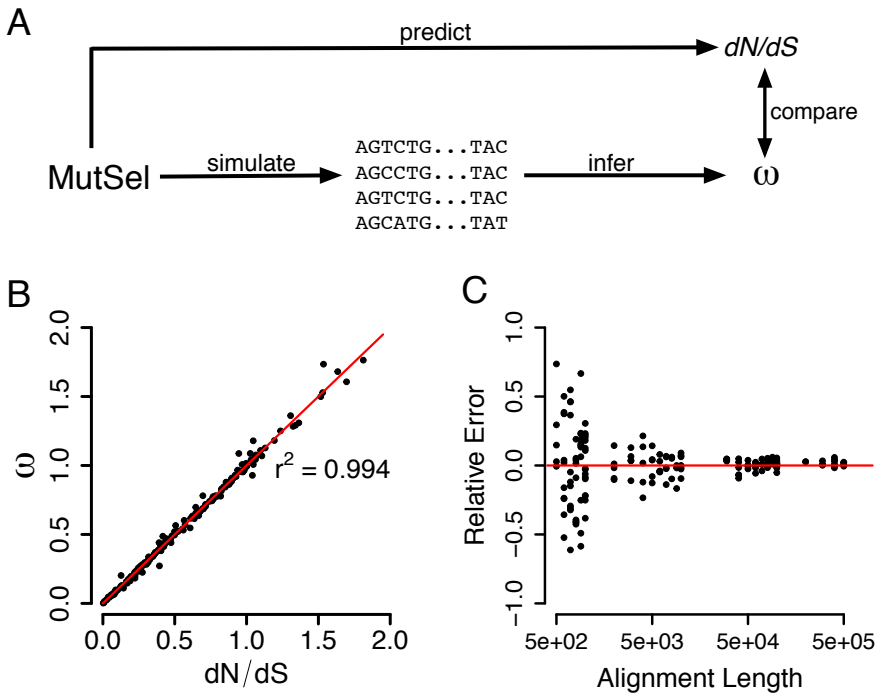

Figure 2: Combined modeling approach to assess performance of dN/dS inference frameworks. (A) Protein-coding alignments are simulated in the MutSel modeling framework. dN/dS can then be calculated ("predict") from scaled selection coefficients as well as through an ML inference framework ("infer"). Comparing resulting quantities reveals the accuracy of the chosen inference framework. (B) Regression between predicted dN/dS values and inferred ω MLEs. Each point corresponds to a single simulated alignment, and the solid line is the $x = y$ line. (C) Convergence of ω MLEs to the true dN/dS value. The y-axis indicates the relative error of the ω MLE, and the x-axis indicates the number of positions in the simulated alignment. As the number of positions and hence the size of the data set increases, ω converges to the predicted dN/dS value. The solid line is the $y = 0$ line, indicating no error.

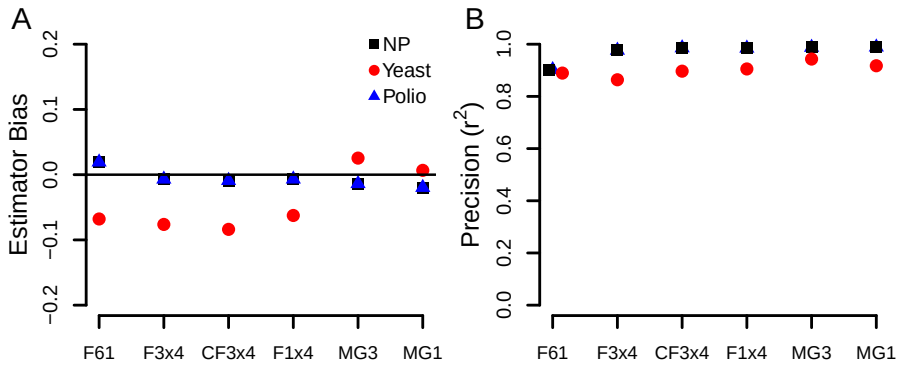

Figure 3: Estimator bias and precision of ω estimates for various model frequency parameterizations. (A) Estimator bias and (B) Precision (r^2) values between dN/dS and ω MLEs across model frequency parameterizations, for each set of nucleotide mutation rates. To calculate bias, we fit a linear model with ω as the response and dN/dS as the predictor, with a fixed slope of 1, and the resulting intercept value represents the bias. Negative biases indicate ω MLEs that are, on average, lower than dN/dS. Note that all standard errors for bias are smaller than the symbol size.

Table 1: Mean ΔAIC for datasets simulated with NP, yeast, or polio virus mutation rates.

Frequencies	NP	Yeast	Polio
F61	0	0	0
CF3x4	-9519.53	-7843.77	-9519.53
MG1	-13207.5	-9924.05	-13207.5
F1x4	-13410.54	-13544.47	-13410.54
MG3	-14287.28	-12737.57	-14287.28
F3x4	-14699.22	-17277.3	-14699.22

Note: The order of frequency models shown in the table corresponds to the model ranking for NP, and the ranking differs somewhat for yeast and polio datasets. AIC is computed as $AIC = 2(k - \ln L)$, where k is the number of free parameters of the model, and $\ln L$ is the log-likelihood (Akaike 1974; Burnham and Anderson 2004). Number of free parameters for each model are F61, 63; CF3x4, 12; MG1, 6; F1x4, 6; MG3, 12; and F3x4, 12. Note that, for each model, 3 of the parameters are ω, κ, and a global branch-length scaling parameter, and the remaining parameters are either empirical codon or nucleotide frequencies.

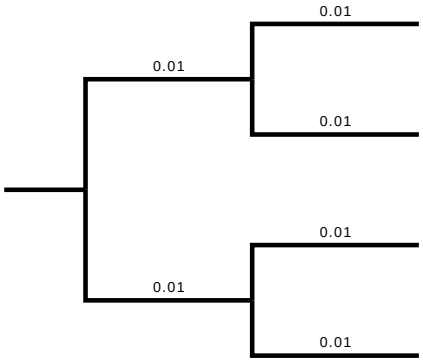

Figure 4: Phylogeny used for all simulated alignments.

Supplementary Information

Table S1. Estimator bias of ω MLEs and the true dN/dS values, for all nucleotide mutation rates and model frequency parameterizations examined. All biases are statistically significant (different from 0), with all $P < 2 \times 10^{-16}$ except for the estimator bias associated with yeast mutation rates for MG3, where $P = 5.4 \times 10^{-5}$.

Mutation rate	MG1	F1x4	MG3	CF3x4	F3x4	F61
NP	-0.014	-0.02	-0.007	-0.009	-0.007	0.019
Yeast	0.025	0.007	-0.063	-0.084	-0.076	-0.068
Polio	-0.049	-0.103	-0.088	-0.148	-0.161	-0.136

Table S2. Precision, measured as the squared correlation coefficient r^2, of ω MLEs relative to the true dN/dS values, for all nucleotide mutation rates and model frequency parameterizations examined. All values shown are statistically significant, with all $P < 2 \times 10^{-16}$.

Mutation rate	MG1	F1x4	MG3	CF3x4	F3x4	F61
NP	0.988	0.989	0.985	0.986	0.977	0.902
Yeast	0.943	0.917	0.905	0.897	0.864	0.889
Polio	0.842	0.811	0.777	0.754	0.781	0.752

Table S3. The order of frequency models shown in the table corresponds to the model ranking for NP, and the ranking differs somewhat for yeast and polio datasets. BIC is computed as $BIC =$

$-2 \ln L + k \ln n$, where k is the number of free parameters of the model, $\ln L$ is the log-likelihood, and n is the sample size (Burnham and Anderson 2004). For all models, $n = 500000$, which corresponds to the number of alignment columns. The number of free parameters for each model are F61, 63; CF3x4, 12; MG1, 6; F1x4, 6; MG3, 12; and F3x4, 12. Note that, for each model, 3 of the parameters are ω, κ, and a global branch-length scaling parameter, and the remaining parameters are either empirical codon or nucleotide frequencies.

Frequencies	NP	Yeast	Polio
F61	0	0	0
CF3x4	-8918.92	-7243.16	-8918.92
MG1	-12551.28	-9267.83	-12551.28
F1x4	-12776.56	-12910.5	-12776.56
MG3	-13653.31	-12103.59	-13653.31
F3x4	-14098.61	-16676.69	-14098.61

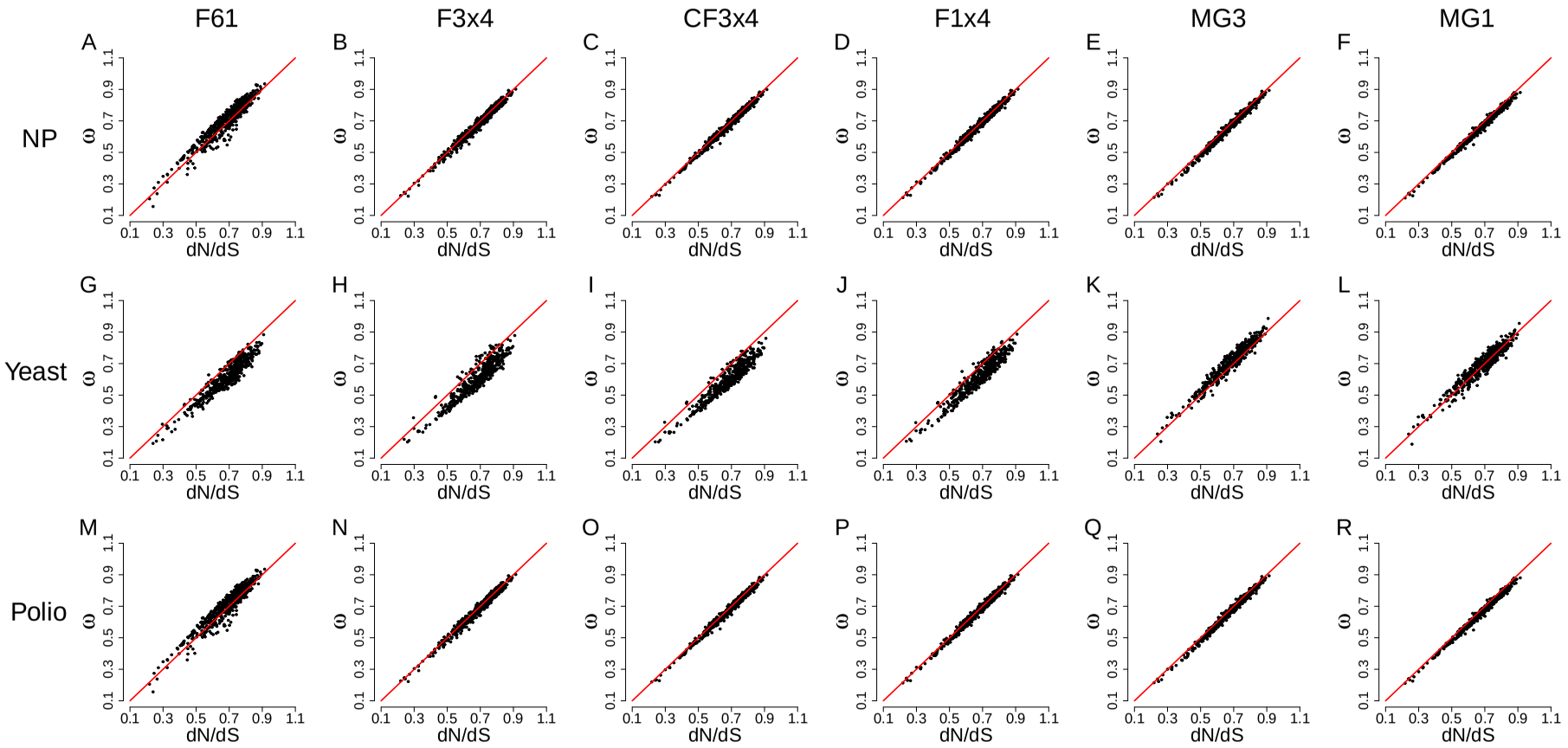

Figure S1. Regressions of ω MLEs on the true dN/dS values, as calculated from scaled selection coefficients, for datasets simulated using experimental fitnesses and mutation rates. Each point represents an alignment, and each red line is the $x = y$ line.