somehow reframe sig statement and abstract. more story, more umph.

Significance Statement

Models of protein evolution represent some of the most rigorous and sophisticated tools in comparative sequence analysis. Various frameworks which discern the strength of natural selection in phylogenetic data exist, yet how these distinct models relate to one another remains entirely unknown. Therefore, researchers have no concrete rationale informing their methodological choices. Here, we derive a precise mathematical relationship beween two competing models: dN/dS and selection coefficient models. Using this relationship, we uncover previously unrecognized properities inherent to both frameworks. In particular, we find that the most widely-used dN/dS model parameterization produces systematically biased parameter estimates. Moreover, model fit metrics cannot necessarily identify which model parameterization produces the most accurate inferences, and can therefore be counterproductive in model selection.

Abstract

The evolutionary rate ratio dN/dS, representing the ratio nonsynonymous to synonymous substitution rates, is widely used to infer the strength of natural selection in protein-coding sequences. Recently, mutation-selection-balance (MutSel) models have become a popular alternative to the dN/dS framework. MutSel models estimate scaled selection coefficients, indicating the selective response to particular codon and/or amino-acid changes, from phylogenetic data. However, it remains unknown how these two modeling frameworks relate to one another. Do dN/dS estimates contain similar or distinct information from scaled selection coefficients? To answer this question, we have derived a formal mathematical relationship to calculate dN/dS from scaled selection coefficients. We prove that, when synonymous changes are neutral, MutSel models strictly produce $dN/dS \leq 1$. Alternatively, if selection acts on synonymous changes, dN/dS can readily be greater than 1, even though positive selection is not occurring. In addition, we used the relationship between models to examine the extent to which dN/dS estimated using maximum likelihood (ML) agree with dN/dS computed from selection coefficients. This analysis served as a novel and robust strategy for assessing the performance of ML dN/dS inference frameworks. ML methods yielded biased dN/dS estimates in the presence of mutational asymmetry, but this bias was ameliorated by fitting models which more closely matched the mechanistic process which generated the data. Strikingly, we found that the best-fitting model, based on AIC scores, was not necessarily the model with the least bias and highest precision for parameter estimates of interest. Therefore, we conclude that selecting models based solely on fit can be counterproductive and positively misleading.

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 ($dN/dS > 1$) [1, 2, 3, 4]. In phylogenetic contexts, dN/dS is typically calculated using a maximum likelihood (ML) approach [5, 6, 1, 7]. 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 [8] 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 dN/dS-based models. While dN/dS-based models describe how quickly a protein's constituent amino acids change, MutSel models assess the strength of natural selection operating on specific amino-acid or codon changes. 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 or amino acid changes [9, 10, 11, 12]. Although first introduced over 15 years ago [9], MutSel models have seen little use due to their high computational expense. Recently, however, several computationally tractable model implementations have emerged [13, 14], 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 [6, 15] and episodic [16, 17] and/or lineage-specific selection [18, 19, 20], and they can also incorporate information regarding protein structure and epistatic interactions [21, 22, 23, 24, 25]. This flexibility, along with accessible software implementations [26, 27, 28], makes dN/dS-based models an attractive modeling choice. On the other hand, some have argued that MutSel models, given their explicit consideration of population genetics theory and attention to site-specific amino-acid fitness differences, offer a more fine-grained approach to studying coding-sequence evolution [9, 11, 12, 29]. 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 [30, 31, 32].

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 properties and/or limitations.

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 derived 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 have proven that, when synonymous mutations are neutral, dN/dS calculated from selection coefficients is necessarily less than 1. This proof demonstrates that MutSel models are inherently only able to model purifying selection, and therefore would be an inappropriate model choice if positive selection is expected. However, when synonymous codons have different fitnesses, it is possible to recover dN/dS values above 1, even though no positive selection is occurring.

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 (with a notable exception of ref. [33]). 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

inherently cannot assess model performance when the data arises from a different mechanism than that of the inference framework, as is obviously the case in real sequence analysis. A more sensitive test of model performance would examine how a given method performs when data is simulated under entirely different conditions. Unfortunately, such an approach is typically infeasible as true parameter values would be unknown, and thus model performance would remain untested.

The relationship we have established 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 to 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 that, in the absence of mutational bias, dN/dS values inferred in an ML framework agreed precisely with those calculated from scaled selection coefficients. However, as mutational bias increases, dN/dS MLEs become increasingly biased away from their true values, even under a variety of ML model parameterizations. Strikingly, the best-performing ML model parameterizations are not those which exhibited the best fit to the data (measured by AIC), 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 rederivation of results presented in refs. [9, 10], reproduced here 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 [9, 10, 12, 29]. This continuous-time reversible Markov process is governed by the 61×61 transition matrix $M(t) = e^{Qt}$, 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 $s_{ij} = f_j^{\text{codon}} - f_i^{\text{codon}}$ [34, 10]. The fixation probability for this mutation is [35, 9, 10]

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

. We further define $S_{ij} = 2N_e s_{ij}$ as the scaled selection coefficient for this change [10]. 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 [9, 34]. If we assume that m_{ij} only describes single-nucleotide changes, it can be written as a nucleotide 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.

We now show how S_{ij} can be written in terms of mutation rates and stationary (equilibrium) codon frequencies P_i. As this system statisfies detailed balance (reversibility) [9], 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. [9], 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 $S_{ij} = \ln\left(P_j/P_i\right)$ [34].

Mathematical relationship between scaled selection coefficients and *dN/dS*.

Using the theory laid out in the previous subsection, we 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. 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 [5, 7] 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

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

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 $u_{ij|j\in\mathcal{S}_i} = 1/N_e$ [36], 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 $\mathcal{N}(0, \sigma^2)$, where a unique σ^2 for each fitness distribution was drawn 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, higher σ^2 values indicate strong

purifying selection, low values indicate weaker purifying selection, and finally $\sigma^2 = 0$ 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 [13, 14].

We then converted each amino-acid fitnesses 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. Using equations (6) - (9), we computed dN/dS for each distribution of codon fitnesses. For these calculations, we assumed the symmetric mutation model HKY85 [37], which is specified by the parameters μ, the nucleotide mutation rate, and κ, the ratio of transitiitions to tranversions. Thus, transitiitions 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 κ value for each simulation from $\mathcal{U}(1, 6)$.

COME BACK TO PARAGRAPH: Under neutral evolution, we expect that $dN/dS = 1$, and moreover that dN/dS will decline as fitness differences among amino acids increase. 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, negatively correlation between these quantities. When fitness differences among amino acids were very high, dN/dS took on lower values, properly reflecting stronger purifying selection (Figure 1). This correlation was much stronger for fitness distribution 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 $dN/dS = 1$. Furthermore, we never recovered $dN/dS > 1$ when synonymous changes were neutral. This finding reveals an important property of MutSel models: they inherently cannot describe positive, diversifying selection. Indeed, in Appendix 1, we have proven that scaled selection coefficients strictly yield $dN/dS \leq 1$, under the assumptions that synonymous changes are neutral and nucleotide mutation is unbiased. This proof formalizes the MutSel model's underlying assumption that selection pressure is constant over the phylogeny, and thus the protein evolves under equilibrium conditions. Although this proof assume symmetric nucleotide mutation rates, we do not expect that deviations from this assumption will have dramatic effects on dN/dS estimates.

However, the restriction $dN/dS \leq 1$ does not hold when synonymous changes are not neutral, as seen in Figure 1B. Even though the underlying evolutionary model explicitly assumes that the system is at equilibrium, dN/dS can readily be greater than 1. Indeed, 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 $K_{\text{S}} = 0$, and thus the value for dN/dS approaches infinity. 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, in theory, what is classically termed positive selection can result simply from strong synonymous fitness differences.

That sequences 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 [38], from viruses [39, 40] to plants [41] to Metazoa [42, 43, 44, 45, 46]. Particular genomic regions with strong synonymous selection include exonic splicing enhancers [47, 48, 49], RNA secondary structures [50, 47, 39, 40], and sites near translation initiation [38]. In addition, both selection against protein misfolding and for translation efficiency tend to induce synonymous selection in a gene-specific manner [51, 52], most notably in highly expressed genes [53, 46]. While synonymous selection may not dominate genomes in organisms with relatively small effective population sizes [43, 45], it certainly acts strongly at specific sites and/or small, local regions. Thus, as dN/dS ratios are typically measured on a per-site basis, we expect that such sites $dN/dS > 1$ 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. Therefore, we applied a novel benchmarking approach which used the theoretical relationship among modeling frameworks to assess the accuracy and specific utility of those models. This approach, outlined in Figure 2A, entailed comparing dN/dS values calculated from selection coefficients to those inferred by an dN/dS-based model. Importantly, as MutSel models are based explicitly on population genetics theory, these simulated alignments are likely far more similar to real sequence data than alignments simulated under a dN/dS-based model.

Using the selection coefficients and symmtric mutation rates from the previous subsection, we simulated alignments using standard methods [7] according the Halpern-Bruno MutSel model [9]. We then inferred dN/dS for each alignment using the M0 model [5, 2], as implemented in the HyPhy batch language [26]. Throughout the remaining text, we refer to dN/dS inferred using ML as ω, and to dN/dS computed using equations (1) - (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, represented by simulated alignment length, increased. These results unequivocally showed that the dN/dS quantity is fully contained within MutSel model parameters, and importantly that dN/dS-based model-inference methods behave exactly as expected (when nucleotide mutation is symmetric), yielding precise 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),

given by ref. [54]. We combined each of these fitness distributions with three sets of experimentally-determined mutation rates, either for NP [54], yeast [55], or polio virus [56], to determine $498 \times 3 = 1494$ distinct distributions of steady-state codon frequencies (see *Methods* section for details). While all three mutation matrices were asymmetric, each featured a differing degree of mutational bias; in the absence of amino-acid level selection, the GC contents that the NP, yeast, and polio mutation rates would generate are 0.518, 0.336, and 0.192, 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.

dN/dS-based models account for nucleotide mutational bias by incorporating either target codon [5] or target nucleotide [6] frequencies; these frameworks are known, respectively, as GY-style and MG-style models [57]. 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. Although the GY-style and MG-style matrices follow different forms, it is possible to write MG-style models such that they conform to the reversible GY-style framework, as we show in Appendix 2.

Previous works have suggested that MG-style and GY-style models yield different ω estimates [15, 58], so we inferred ω according to both GY- and MG-style frameworks. For GY-style models, we used the frequency estimators F61 [5], F3x4 [5], CF3x4 [57], and F1x4 [6]. For MG-style models, we considered both a parameterization in which four global nucleotide frequency parameters were used [6] and a parameterization which employed twelve nucleotide frequency parameters to allow for different frequencies at each codon position [15]. We term the former framework MG1, and the latter MG3. Note that our MG1 corresponds to the original MG-style model [6], whereas our MG3 corresponds to the so-called MG94×HKY84 model [15].

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, the average discrepancy between the true dN/dS value and the ω MLEs, of ω MLEs. Figure 3B displays 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 dN/dS vs. ω 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. [58] had previously suggested that GY-style models produce negatively biased ω estimates. By contrast, our results show that this bias is pervasive and remains approximately constant through a wide range of dN/dS values as mutational asymmetry increases (Figure 3A, Table S1, Figure S1). Furthermore, 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.

Strikingly, when we examined model fit using AIC scores [59, 60] 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 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).

Therefore, evaluating model performance based strictly on model fit can be highly misleading, as model fit is clearly at odds with model performance. Going forward, we highly recommend that researchers employ MG-style matrix frameworks in their dN/dS inferences to minimize bias.

That a standard measure of model fit (AIC) could not identify the model which produced the most accurate estimates for the parameter of interest has broad implications for model selection, in general. In particular, model fit cannot necessarily identify the best-performing model if the data arose from a process distinct from the inference model. Therefore, model selection based on AIC scores may not guard effectively against spurious inferences. Instead, the process by which the data arose should be carefully considered, and an appropriate inference method which best approximates this process should then be selected. Otherwise, model fit criteria may lend unfounded support to models which yield biased parameter estimates.

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 rates and neutral synonymous changes) 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 [9, 22, 11, 29]. Thus, if the aim is to identify positive selection, of the two frameworks examined here, only dN/dS-based models are appropriate.

However, we also found that dN/dS can readily be greater than 1 when selection acts on synonymous changes, even though the protein sequence is strictly evolving under purifying selection. This seemingly paradoxical finding actually reflected 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 selection acts on synonymous changes. 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 theoretically be greater than 1, even under strictly purifying selection, if sequences considered contain segregating polymorphisms rather than strictly fixed differences [30, 31, 32]. 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, and it is therefore critical to ensure that the data adhere to model assumptions before conclusions from dN/dS are drawn.

We suggest several strategies to limit such false positives under synonymous selection. For one, certain formulations (particularly those implemented in HyPhy package [26]) of dN/dS-based methods model dN and dS rate variation separately [6, 15, 61], rather than using a single parameter representing dN/dS. These sorts of methods may be able to identify when $dN/dS > 1$ not because of positive selection, but instead because dS is vanishingly small. In addition, other dN/dS-based models have proposed corrections for dS to account for synonymous selection, an approach which may prove fruitful in the future [62]. Finally, our benchmarking approach, in which we simulate sequences according to MutSel models and infer dN/dS using both analytically and using ML, will help to identify circumstances under which synonymous selection confounds dN/dS interpretations.

textbfnext paragraph is repetitive and not strong enough for point it is trying to make. Finally, we emphasize the utility of establishing relationships among distinct modeling frameworks in order to examine model behavior and assess model performance. Assessing the extent to which distinct modeling frameworks overlap represents a promising future avenue of research for model develop-

ment. 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 going forward.

Methods

Simulation of scaled selection coefficients.

We examined the relationship between dN/dS and scaled selection coefficients by simulating 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 $F_i^{\text{codon}} = F_a^{\text{aa}}$, such that all synonymous codons had the same fitness. For the other 100 fitness distributions, we assigned synonymous codons different fitnesses by randomly selected 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 $F_i^{\text{codon}} = F_a^{\text{aa}} - \lambda$. We drew a unique λ for each fitness distribution from $\mathcal{U}(0, 2)$. We then computed stationary codon frequencies as

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

where the sum in the denominator runs over all 61 sense codons [34]. Equation (10) gives the analytically precise stationary frequencies for a MutSel model, under the assumption of symmetric mutation rates [34]. We used equations (6) - (9) to compute dN/dS for each resulting set of stationary codon frequencies. For these calculations, we assumed the HKY85 [37] nucleotide mutation model, and accordingly we set the mutation rate for transitions as $\mu\kappa$, and the rate for all transversions 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 [7] according to the Halpern-Bruno MutSel model [9]. In simplified form, 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 [7]. 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. Code used to generate all simulated alignments is available at `https://github.com/sjspielman/MutSel`.

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. [54], in combination with experimental nucleotide mutation rates for either NP [54], yeast [55], or polio virus [56], to derive realistic distributions of stationary codon frequencies. Ref. [54] reported 498 distinct site-wise amino acid preference distributions for NP [54]. We combined these 498 amino acid preference sets with each of the three mutation rate matrices sets to construct a total of $498 \times 3 = 1494$ unique experimental evolutionary Markov models, using the approach in refs. [54, 63]. The instantaneous rate matrix Q for each experimental model was populated by

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

for a substitution from codon i to codon j, where F_i^{codon} is the fitness of codon i [54, 63]. We calculated F_i^{codon} values by simply assigning a given amino acid's experimental fitness F_a^{aa} to each of its consistituent 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 [2], as implemented in the HyPhy batch language [26]. The M0 model uses the GY94 instantaneous rate matrix, which is populated by elements

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

for a substitution from codon i to codon j, where κ is the transition-transversion bias, π_j is the equilibrium frequency of the target codon j, and ω represents dN/dS [5, 1]. Importantly, the π parameters are intended to represent those codon frequencies which would exist in absence of selection pressure generated by mutation alone [5, 6, 64, 7]. Thus, when inferring ω on datasets which used symmetric mutation rates, we assigned the value 1/61 to all parameters π, 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 [5] (target codon frequency) and MG-style [6] (target nucleotide frequency) parameterizations. Codon frequency parameterizations considered include F61 [5], F3x4 [5], CF3x4 [57], and F1x4 [6]. We additionally implemented two varieties of MG-style models; the first, MG1, employs four parameters for nucleotide frequencies (one per nucleotide) and the second, MG3, employs twelve nucleotide frequency parameters, with four nucleotide frequency parameters for each of the three codon positions. All models included the parameters κ and ω.

Availability

All code is freely available from `https://github.com/clauswilke/Omega_MutSel`, and simulated alignments are available from **dryad???**. Code used to simulate all alignments is available from `https://github.com/sjspielman/pyvolve`.

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 of this paper, 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 term in L_N, and hence $dN/dS \leq 1$.

For this proof, we consider the pair of nonsynonymous codons i and j, whose respective stationary frequencies P_i and P_j satisfy $P_i \leq P_j$ and $P_j > 0$. 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 the K_N (numerator) calculation for dN. 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 (the denominator) 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

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

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

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

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 decomposd into a 61×61 symmetric substitution rate matrix and a 61-dimensional vector containing the equilibrium codon frequencies, which correspond to the stationary distribution of the Markov chain. On the other hand, MG-style rate matrices explicitly consider nucleotide, not codon, frequencies, and thus do not clearly follow this paradigm. We now describe how the MG-style matrix can be rewritten in a form that follows 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. We note that the MG-style matrix yields the stationary codon frequency $P_i = \pi_{i_1} \pi_{i_2} \pi_{i_3} C$ for a given codon i, where $C = 1 - \Pi_{\text{stop}}$ and $\Pi_{\text{stop}} = \pi_{\text{T}} \pi_{\text{A}} \pi_{\text{G}} + \pi_{\text{T}} \pi_{\text{G}} \pi_{\text{A}} + \pi_{\text{T}} \pi_{\text{A}} \pi_{\text{A}}$ [6]. Therefore, we can simply rewrite the term $\theta_{s_i t_j} \pi_{t_j}$ as $[P_j \theta_{s_i t_j} / (\pi_m \pi_n)] \times C$, 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

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

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

References

[1] 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.

[2] Yang ZH, Nielsen R, Goldman N, Pedersen AMK (2000) Codon-substitution models for heterogeneous selection pressure at amino acid sites. *Genetics* 155:431–449.

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

[4] Huelsenbeck JP, Jain S, Frost SWD, Kosakovsky Pond SL (2006) A Dirichlet process model for detecting positive selection in protein-coding DNA sequences. *Proc Natl Acad Sci USA* 103:6263–6268.

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

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

[7] Yang Z (2006) *Computational Molecular Evolution* (Oxford University Press).

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

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

[10] 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.

[11] 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.

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

[13] Rodrigue N, Lartillot N (2014) Site-heterogeneous mutation-selection models within the PhyloBayes-MPI package. *Bioinformatics* :1020–1021.

[14] Tamuri AU, Goldman N, dos Reis M (2014) A penalized-likelihood method to estimate the distribution of selection coefficients from phylogenetic data. *Genetics* 197:257–271.

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

[16] Kosakovsky Pond S, *et al.* (2011) A random effects branch-site model for detecting episodic diversifying selection. *Mol Biol Evol* 28:3033–3043.

[17] Murrell B, *et al.* (2012) Detecting individual sites subject to episodic diversifying selection. *PloS Genet* 8:e1002764.

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

[19] 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.

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

[21] Robinson DM, Jones DT, Kishino H, Goldman N, Thorne JL (2003) Protein evolution with dependence among codons due to tertiary structure. *Mol Biol Evol* 20:1692–1704.

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

[23] 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* 26:1663–1676.

[24] Scherrer MP, Meyer AG, Wilke CO (2012) Modeling coding-sequence evolution within the context of residue solvent accessibility. *BMC Evol Biol* 12:179.

[25] Meyer AG, Wilke CO (2012) Integrating sequence variation and protein structure to identify sites under selection. *Mol Biol Evol* 30:36–44.

[26] Kosakovsky Pond SL, Frost SDW, Muse SV (2005) HyPhy: hypothesis testing using phylogenetics. *Bioinformatics* 21:676–679.

[27] Yang Z (2007) PAML 4: Phylogenetic analysis by maximum likelihood. *Mol Biol Evol* 24:1586–1591.

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

[29] Thorne JL, Lartillot N, Rodrigue N, Choi SC (2012) Codon models as vehicles for reconciling population genetics with inter-specific data. In Cannarozzi G, Schneider A, eds., *Codon evolution: mechanisms and models* (Oxford University Press, New York).

[30] Rocha E, *et al.* (2006) Comparisons of *dN/dS* are time dependent for closely related bacterial genomes. *J Theor Biol* 239:226 – 235.

[31] Kryazhimskiy S, Plotkin JB (2008) The population genetics of *dN/dS*. *PLoS Genet* 4:e1000304.

[32] Mugal CF, Wolf JBW, Kaj I (2014) Why time matters: Codon evolution and the temporal dynamics of *dN/dS*. *Mol Biol Evol* 31:212–231.

[33] 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.

[34] Sella G, Hirsh AE (2005) The application of statistical physics to evolutionary biology. *Proc Natl Acad Sci USA* 102:9541–9546.

[35] Kimura M (1962) On the probability of fixation of mutant genes in a population. *Genetics* 4:713–719.

[36] Crow JF, Kimura M (1970) *An Introduction to Population Genetics Theory* (Burgess Pub. Co., California).

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

[38] Gu W, Zhou T, Wilke CO (2010) A universal trend of reduced mRNA stability near the translation-initiation site in prokaryotes and eukaryotes. *PLoS Computational Biology* 6:e1000664.

[39] Cuevas JM, Domingo-Calap P, Sanjuan R (2011) The fitness effects of synonymous mutations in DNA and RNA viruses. *Mol Biol Evol* 29:17–20.

[40] Zanini F, Neher RA (2013) Quantifying Selection against Synonymous Mutations in HIV-1 env Evolution. *J Virol* 87:11843–11850.

[41] 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* 29:3037–3044.

[42] Duret L (2002) Evolution of synonymous codon usage in metazoans. *Curr Opin Genet Dev* 12:640–649.

[43] Chamary JV, Parmley JL, Hurst LD (2006) Hearing silence: non-neutral evolution at synonymous sites in mammals. *Nature Rev Genet* 7:98–108.

[44] Hershberg R, Petrov D (2008) Selection on codon bias. *Annu Rev Genet* 42.

[45] Plotkin JB, Kudla G (2011) Synonymous but not the same: the causes and consequences of codon bias. *Nature Rev Genet* 12:32–42.

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

[47] Schattner P, Diekhans M (2006) Regions of extreme synonymous codon selection in mammalian genes. *Nucleic Acids Research* 34:1700–1710.

[48] Parmley JL, Chamary JV, Hurst LD (2006) Evidence for purifying selection against synonymous mutations in mammalian exonic splicing enhancers. *Mol Biol Evol* 23:301–309.

[49] Parmley JL, Hurst LD (2007) Exonic splicing regulatory elements skew synonymous codon usage near intron-exon boundaries in mammals. *Mol Biol Evol* 24:1600–1603.

[50] Chamary JV, Hurst LD (2005) Evidence for selection on synonymous mutations affecting stability of mRNA secondary structure in mammals. *Genome Biol* 6:R75.

[51] Williford A, Demuth JP (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* 29:3755–3766.

[52] Agashe D, Martinez-Gomez NC, Drummond DA, Marx CJ (2013) Good codons, bad transcript: Large reductions in gene expression and fitness arising from synonymous mutations in a key enzyme. *Mol Biol Evol* 30:549–560.

[53] Drummond DA, Wilke CO (2008) Mistranslation-Induced Protein Misfolding as a Dominant Constraint on Coding-Sequence Evolution. *Cell* 134:341–352.

[54] Bloom JD (2014) An experimentally determined evolutionary model dramatically improves phylogenetic fit. *Mol Biol Evol* :To appear.

[55] Zhu YO, Siegal ML, Hall DW, Petrov DA (2014) Precise estimates of mutation rate and spectrum in yeast. *Proc Natl Acad Sci USA* :FORTHCOMING.

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

[57] 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.

[58] Yap LHES V B, Huttley G (2010) Estimates of the effect of natural selection on protein-coding content. *Mol Biol Evol* 27:726 – 734.

[59] Akaike H (1974) A new look at the statistical model identification. *IEEE Transactions on Automatic Control* 19:6:716–723.

[60] Burnham KP, Anderson DR (2004) Multimodel inference: understanding AIC and BIC in model selection. *Sociological Methods & Research* 33:261–304.

[61] Murrell B, *et al.* (2013) FUBAR: A Fast, Unconstrained Bayesian AppRoximation for inferring selection. *Mol Biol and Evol* 30:1196–1205.

[62] Zhou T, Gu W, Wilke CO (2010) Detecting positive and purifying selection at synonymous sites in yeast and worm. *Mol Biol Evol* 27:1912–1922.

[63] Bloom JD (2014) An experimentally informed evolutionary model improves phylogenetic fit to divergent lactamase homologs. *Mol Biol Evol* 31:1956–1978.

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

Figures and Tables

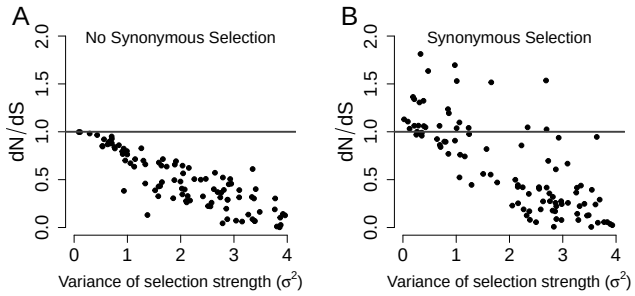

Figure 1: dN/dS decreass in proportion to amino-acid level selection strength. dN/dS is plotted against the variance (σ^2) of the simulated distribution of amino-acid scaled selection coefficients. Higher variances indicate larger fitness differences among amino acids, whereas the limiting value of $\sigma^2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

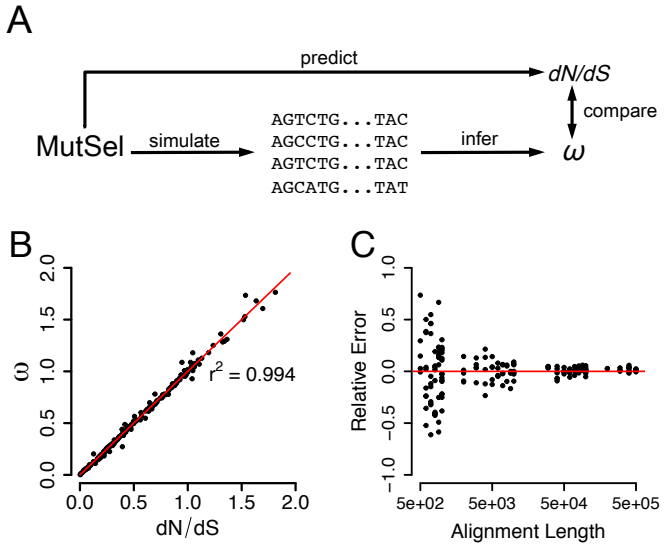

Figure 2: Combined modeling approach to assess performance of dN/dS inference frameworks. (A) Protein-coding alignments are simulated in the MutSel modeling framework. dN/dS can then be calculated from scaled selection coefficients as well as through a ML inference framework. Comparing resulting quantities reveals the accuracy of the chosen dN/dS inference framework. (B) Regression between dN/dS values as calculated from scaled selection coefficients and as inferred using the M0 model [5, 1, 2]. Each point corresponds to a single simulated alignment, and the red 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 maximum likelihood dN/dS estimate, 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, the maximum likelihood estimates converges to the dN/dS values calculated using equations (6)-(9). The red line is the $y = 0$ line, indicating no error.

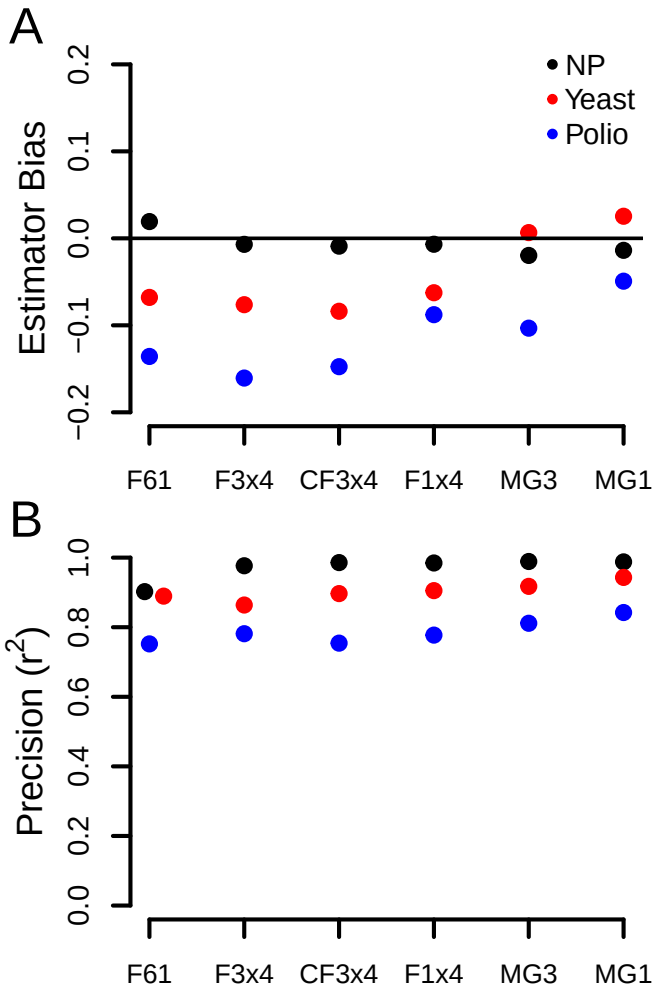

Figure 3: Estimator bias and precision of ω estimates for various model frequency parameterizations. (A) Estimator bias and (B) r^2 values between dN/dS and ω MLEs across model frequency parameterizations, for each set of nucleotide mutation rates. To calculate bias, we fit a linear model, with ω as 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 estimates that are, on average, lower than dN/dS. For all frequency parameterizations, bias generally increases as mutation rates became increasingly asymmetric. Even so, MG-style models tend to yield far less biased ω estimates than do GY-style models.

Table 1: Mean ΔAIC for datasets simulated with NP, Yeast, or Polio mutation rates.

Frequencies	NP	Yeast	Polio
F61	0	0	0
CF3x4	-9627.5	-7951.8	-7975.9
MG1	-13325.5	-10042.0	-5147.6
F1x4	-13524.5	-13658.5	-15468.3
MG3	-14401.3	-12851.6	-8624.9
F3x4	-14807.2	-17385.3	-19384.6

Based on AIC scores, the F61 parameterization strongly outperforms all other model parameterizations for all mutation rates, even though the F61 framework yields neither the most accurate nor the most precise parameter estimate. Note that 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 [59, 60]. As codon and/or nucleotide frequency parameters were directly computed from the data, all models have 3 free parameters (ω, κ, and a global branch-length scaling parameter).

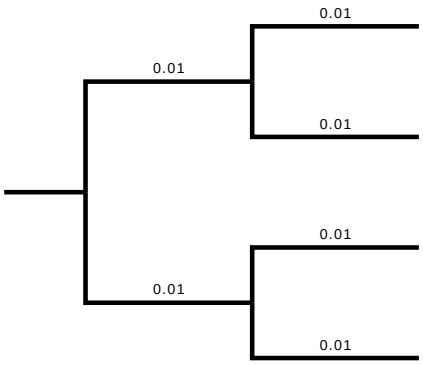

Figure 4: Phylogeny used for all simulated alignments.

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. Negative bias values indicate that ω MLEs are, on average, lower than dN/dS. 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

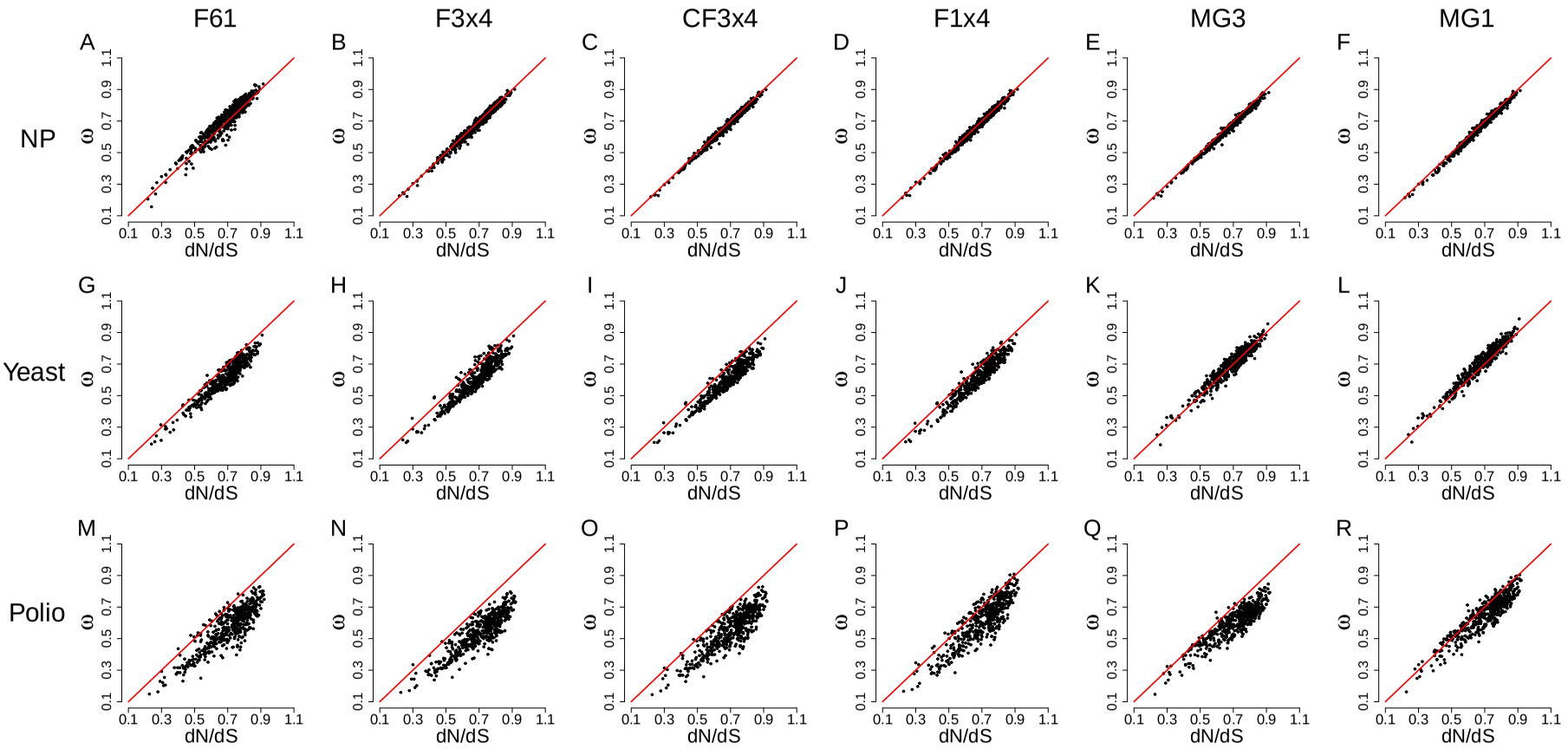

Figure S1. Regressions for inferred ω estimates and dN/dS values, as calculated from scaled selection coefficients, for datasets simulated using experimental fitnesses and mutation rates. Each point represents an alignment, and each red line is that $x = y$ line.