One-rate models outperform two-rate models in site-specific dN/dS estimation

Stephanie J. Spielman[1*], Suyang Wan[1,2], and Claus O. Wilke[1]

Address:
[1]Department of Integrative Biology, Center for Computational Biology and Bioinformatics, and Institute for Cellular and Molecular Biology. The University of Texas at Austin, Austin, TX 78712, USA.
[2]School of Physics and Astronomy, The University of Minnesota, Minneapolis, MN 55455, USA.

[*]Corresponding author
Email: stephanie.spielman@gmail.com

Manuscript type: Article

Keywords: dN/dS, evolutionary rate, sequence simulation, molecular evolution, mutation-selection models

Abstract

Methods that infer site-specific dN/dS, the ratio of nonsynonymous to synonymous substitution rates, from coding data have been developed primarily to identify positively selected sites ($dN/dS > 1$). As a consequence, it is largely unknown how well different inference methods can infer dN/dS point estimates at individual sites. In particular, dN/dS may be estimated using either a one-rate approach, where dN/dS is parameterized as a single parameter, or a two-rate approach, in which dN and dS are estimated separately. While some have suggested that the two-rate paradigm may be preferred for positive-selection inference, the relative merits of these two paradigms for site-specific dN/dS estimation remain largely untested. Here, we systematically assess how accurately several popular inference frameworks infer site-specific dN/dS values using alignments simulated within a mutation-selection framework rather than within a dN/dS-based framework. As mutation-selection models describe long-term evolutionary constraints, our simulation approach further allows us to study under what conditions inferred dN/dS captures the underlying equilibrium evolutionary process. We find that one-rate inference models universally outperform two-rate models. Surprisingly, we recover this result even for data simulated with codon bias (i.e., dS varies among sites). Therefore, even when extensive dS variation exists, modeling this variation substantially reduces accuracy. We additionally find that high levels of divergence among sequences, rather than the number of sequences in the alignment, are more critical for obtaining precise point estimates. We conclude that inference methods which model dN/dS with a single parameter are the preferred choice for estimating reliable site-specific dN/dS ratios.

Introduction

A variety of computational approaches have been developed to infer selection pressure from protein-coding sequences in a phylogenetically-aware context. Among the most commonly-used methods are those which compute the evolutionary rate ratio dN/dS, which represents the ratio of non-synonymous to synonymous substitution rates. Beginning in the mid-1990s, this value has been calculated using maximum-likelihood (ML) approaches (Goldman and Yang 1994; Muse and Gaut 1994), and since then, a wide variety of inference frameworks have been developed to infer dN/dS at individual sites in protein-coding sequences (Nielsen and Yang 1998; Yang et al. 2000; Yang and Nielsen 2002; Yang and Swanson 2002; Kosakovsky Pond and Frost 2005; Kosakovsky Pond and Muse 2005; Murrell et al. 2012b; Lemey et al. 2012; Murrell et al. 2013).

Most commonly, the goal of dN/dS inference is to identify sites subject to positive and/or diversifying selection, as indicated when $dN/dS > 1$. As a consequence, the performances of dN/dS inference methods have largely been evaluated based on how well they detect if a given site evolves with a dN/dS significantly above or below 1. Indeed, many positive-selection inference methods do not make a concerted attempt to calculate precise dN/dS point estimates, but rather focus only on obtaining "good enough" estimates so that the value of dN/dS relative to 1 can be formally tested (Murrell et al. 2012b,a; Scheffler et al. 2014).

By contrast, how accurately such methods estimate dN/dS at individual sites has not been rigorously studied, and therefore it remains unclear which methods, or indeed model parameterizations, provide the most reliable dN/dS point estimates. This dearth of research has hindered advancements of mechanistic studies which seek to understand the relationship between site-specific coding-sequence evolutionary rate and structural properties, such as solvent accessibility, packing density, or flexibility (Shahmoradi et al. 2014; Meyer and Wilke 2015a,b). If site-specific evolutionary rate inference is unreliable, then naturally it will be difficult to ascertain underlying mechanisms driving evolutionary rate.

We therefore seek to assess how well various dN/dS inference frameworks estimate site-wise evolutionary rates from coding sequences. We adopt a robust simulation strategy through which we simulate alignments using the mutation-selection (MutSel) modeling framework. Unlike dN/dS models, MutSel models use population genetics principles to model the site-specific evolutionary process as a dynamic interplay between mutational and selective forces (Halpern and Bruno 1998; Yang and Nielsen 2008). Therefore, many regard MutSel models as more mechanistically representative of real coding sequence evolution than dN/dS-based models, which are primarily phenomenological in nature (Thorne et al. 2007; Holder et al. 2008; Rodrigue et al. 2010; Thorne et al. 2012; Tamuri et al. 2012; Liberles et al. 2013). Indeed, substitution rate itself is not an evolutionary mechanism, but rather an emergent property of various interacting evolutionary processes.

Recently, we introduced a mathematical framework which allows us to accurately calculate a dN/dS ratio directly from the parameters of a MutSel model (Spielman and Wilke 2015b) [we note that dos Reis (2015) introduced a similar framework shortly after]. This framework gives rise to a robust benchmarking strategy through which we can simulate sequences using a MutSel model, and we can subsequently infer dN/dS using established approaches. Previously, we successfully used such an approach to identify biases in dN/dS inference methods for whole-gene evolutionary rates (Spielman and Wilke 2015b). Here, we leverage the power of the established relationship between dN/dS and MutSel models to evaluate the performance of site-specific dN/dS inference approaches.

Two primary questions motivate the present study: i) How accurate are various inference methods for dN/dS point estimation?, and ii) Under what conditions does dN/dS capture the long-term evolutionary dynamics of site-specific coding-sequence evolution? For the first question, we focus

our efforts on distinguishing performance between two dN/dS inference paradigms: one-rate and two-rate models. One-rate models parameterize dN/dS with a single parameter for dN, effectively fixing $dS = 1$ at all sites, whereas two-rate models use separate parameters for dN and dS at each site. Some studies have suggested that the two-rate paradigm leads to more robust positive-selection inference (Kosakovsky Pond and Muse 2005; Murrell et al. 2013), whereas others have suggested that the extra dS parameter may actually confound positive selection inference (Yang et al. 2005; Wolf et al. 2009). Therefore, it remains unclear how the one-rate vs. two-rate parameterization choice influences positive-selection inferences, and consequently it is an open question how this parameterization affects dN/dS point estimation.

The second question arises naturally from our use of MutSel models, which describe the equilibrium site-specific codon fitness values. In other words, any dN/dS calculated from MutSel model parameters describes, by definition, the steady-state dN/dS. As dN/dS is an inherently time-sensitive measurement (Rocha et al. 2006; Kryazhimskiy and Plotkin 2008; Mugal et al. 2014; Meyer et al. 2015), it is not necessarily true that dN/dS measured from a given dataset will reflect the equilibrium value. Therefore, our approach additionally enables us to identify the conditions under which site-specific dN/dS ratios are expected to reflect the long-term, rather than transient, evolutionary dynamics.

Results

Approach

We simulated fully heterogeneous alignments under the HB98 MutSel model (Halpern and Bruno 1998) using the simulation software Pyvolve (Spielman and Wilke 2015a). To derive site-specific MutSel model parameterizations, we simulated 100 distinct sets of amino-acid frequencies from a Boltzmann distribution (Ramsey et al. 2011), reflecting the shape of empirical amino-acid distributions observed in conserved protein sequences (see *Methods* for details). We ensured that these simulated distributions resulted in a range of selective pressure, from extremely stringent to nearly neutral. From each amino-acid frequency distribution, we derived two distinct stationary codon frequency distributions: one where all synonymous codons had the same fitness (i.e. no codon bias), and one where synonymous codons differed in fitness values (i.e. codon bias).

Using these derived MutSel model parameterizations, we simulated a set of fully heterogeneous alignments, with 100 sites, each for set of codon fitnesses. All simulations were conducted along balanced phylogenies with the number of sequences N set as either 128, 256, 512, 1024, or 2048 and with branch lengths B set as either 0.0025, 0.01, 0.04, 0.16, or 0.64. For each of the 25 possible combination of parameters N and B, we simulated 50 replicate alignments. Importantly, the site-specific evolutionary models were the same within each simulation set, making inferences across conditions directly comparable.

We inferred site-specific dN/dS for each simulated alignment using three approaches: fixed-effects likelihood (FEL) (Kosakovsky Pond and Frost 2005), single-likelihood ancestor counting (SLAC) (Kosakovsky Pond and Frost 2005), and FUBAR (Murrell et al. 2013). Each of these methods employs a somewhat different approach when computing site-specific dN/dS values. FEL fits a unique dN/dS model to each alignment site (Kosakovsky Pond and Frost 2005), SLAC directly counts nonsynonymous and synonymous changes along the phylogeny where ancestral states are inferred with maximum likelihood (Kosakovsky Pond and Frost 2005), and FUBAR employs a Bayesian approach to determine dN/dS ratios according to a pre-specified grid of rates (Murrell et al. 2013).

For each inference method, we inferred dN/dS at each site in both a two-rate context (separate

dN and dS parameters per site) and in a one-rate context (a single dN/dS parameter per site). Although SLAC, as a counting-based method, always enumerates both dN and dS on a per-site basis, one can derive an effectively one-rate SLAC by normalizing each site-wise dN estimate by the mean of all site-wise dS estimates. We refer to one-rate inferences with these methods as FEL1, FUBAR1, and SLAC1, and similarly to two-rate inferences as FEL2, FUBAR2, and SLAC2, respectively. All inferences were conducted using the HyPhy batch language (Kosakovsky Pond et al. 2005), specifying the MG94xHKY85 model with F1x4 state frequencies. Note that we did not consider the popular random-effects likelihood methods introduced by Yang et al. (2000) (e.g. M3, M5, M8) because these methods are used predominantly in a one-rate context. Available two-rate extensions to this framework are computationally burdensome and cannot model the amount of rate heterogeneity required to calculate per-site rates (Kosakovsky Pond and Muse 2005). Finally, we computed true dN/dS values from the MutSel parameters, using the approach described in Spielman and Wilke (2015b).

Modeling synonymous rate variation reduces inference accuracy

After inferring site-wise dN/dS for all simulated alignments, we correlated the resulting estimates with true dN/dS values. In Figure 1, we show resulting Pearson correlation coefficients, averaged across all 50 replicates, between inferred and true dN/dS for each inference method. Importantly, our simulation strategy necessitates a somewhat different interpretation of results than would more traditional simulation approaches. In particular, the true dN/dS ratios calculated from the MutSel parameterizations used during simulation correspond to the dN/dS expected at steady state, which in turn indicates the signature of natural selection at evolutionary equilibrium. We can only expect to recover this true dN/dS value if the simulated data reflect the full steady-state distribution of codons. When either the simulated divergence or number of sequences analyzed is low, then, it not necessarily possible to capture this distribution. Therefore, to determine the relative performance of dN/dS inference methods, we considered the most accurate inference method to be the one with the highest dN/dS correlations within a given choice of N and B.

In the absence of codon bias, dS was equal to 1 at all sites. As such, we expected that one-rate inference methods would outperform two-rate inference methods. We indeed found that one-rate inference models showed the best performance when there was no synonymous selection (Figure 1A), in particular at low-to-intermediate divergence levels (B of 0.01 or 0.04). As the sequences became more diverged, and hence more informative, two-rate models increasingly performed as well as one-rate models did. Even so, two-rate models never outperformed one-rate models.

In the presence of codon bias, both dN and dS varied at each site. As a consequence, there are two approaches for calculating the true site-wise dN/dS ratio: One can either calculate the ratio of each site's dN and dS values, or one can take each site's dN value and divide by the average dS over the entire sequence. The former corresponds to a two-rate model (there are two independent rates at each site), while the latter corresponds to a one-rate model (only dN varies per site, and dS is taken as a gene-wide normalization factor). Here, we refer to these two dN/dS ratios as True2 and True1, respectively.

We correlated, for data simulated with codon bias, inferred dN/dS with both True2 and True1 dN/dS values, as shown in Figures 1B and 1C, respectively. *A priori*, we would expect that two-rate inference models would perform best when benchmarked against True2, and similarly one-rate inference models would perform best when benchmarked against True1. Surprisingly, however, one-rate models outperformed two-rate models across N and B conditions, regardless of whether True2 or True1 was considered. We did, however, find that two-rate models yielded higher correlations with True2 than with True1, and vice versa.

Importantly, although two-rate models appear to have outperformed one-rates models when $B = 0.0025$ (Figure 1), nearly all such inferences were poor; correlations from two-rate model virtually never exceeded an average of 0.4. In other words, at low divergence levels, inferred dN/dS could explain at most only $\sim$16% of the rate variation expected at equilibrium, likely indicating that, at $B = 0.0025$, the data was mostly uninformative. Similarly, all estimates, from both one- and two-rate models, were strongly biased at $B = 0.0025$ (Figure S1). As divergence increased, and hence the data became more informative, estimator bias dropped substantially for both one- and two-rate models. However, at $B = 0.04$, one-rate models had virtually no estimator bias, but two-rate models still strongly overestimated dN/dS, indicating that two-rate models were more biased than were one-rate models.

Together, these results demonstrate, for both data with and without codon bias, that one-rate models inferred more accurate dN/dS ratios, on average, relative to two-rate models. For data simulated with codon bias, this result was robust to whether inferences were benchmarked against True1 or True2. Modeling synonymous selection with its own parameter reduced inference accuracy, especially when the data contained pervasive codon bias. Indeed, the accuracy boost achieved with one-rate inference models was far more pronounced for data with codon bias than for data without codon bias. In addition, correlations between inferred and true dN/dS were, on average, higher for data simulated without codon bias (Figure 1A) compared to data simulated with codon bias (Figures 1B and 1C). Our results therefore revealed that dN/dS inference methods were generally more reliable in the absence of codon bias.

One-rate inference methods have minimal performance differences

We next quantified performance differences among methods more rigorously using linear models. For each simulation set, we built mixed-effects linear models with Pearson correlation as the response, inference method as a fixed effect, and replicate as well as interaction between N and B as random effects. We performed multiple comparisons tests, with corrected P-values, to ascertain the relative performance across methods. In this analysis, we additionally tested the performance of derived one-rate inferences made with FEL2 and FUBAR2, which we called FEL2_1 and FUBAR2_1, respectively. These inferences represent rates calculated by normalizing site-specific dN values by the average inferred site-specific dS value, similar to how SLAC1 values were computed.

Linear model analysis confirmed observations from Figure 1 that each one-rate inference framework outperformed its respective two-rate counterpart (Figure 2). Further, FEL2_1 and FUBAR2_1 did not display significant performances differences from FEL1 and FUBAR1, respectively, indicating that fixing dS to 1 is essentially equivalent to normalizing all dN by an inferred average dS. Across panels in Figure 2, both SLAC1 and FEL1 generally outperformed FUBAR1, with SLAC1 tending to be the most accurate method. Importantly, even when performance differences were statistically significant, the effect magnitudes were exceedingly small; mean correlations never differed by more than 0.025. Thus, whether dN/dS was modeled by one or two parameters mattered more than the specific inference method used (e.g. FEL1, SLAC1, FUBAR1) did for obtaining accurate estimates, although SLAC1 and FEL1 may be somewhat preferable to FUBAR1.

Estimation error is higher for lower dN/dS values

We next examined whether certain dN/dS values were more difficult to estimate. Our simulation setup ensured an evenly-spaced range of true dN/dS values, from 0.03–0.92 for simulations without codon bias and, for True2, 0.05–0.99 for simulations with codon bias. We calculated, for each simulated dN/dS value, the average relative error, across replicates, for each N and B parame-

terization, from SLAC1 inference (Figure 3). As seen in each panel of Figure 3, error declined as dN/dS increases, indicating that it was more difficult to precisely estimate dN/dS at slowly evolving sites. Importantly, we observed this trend across N and B conditions, although the overall error decreased as datasets became more informative. Figures S2, S3, and S4 shows results for all simulation conditions and display broadly the same trends as seen in Figure 3.

We suggest that lower dN/dS values were more difficult to estimate because natural selection tolerates fewer codons at slowly-evolving sites, and there are often relatively large fitness differences among the codons that are tolerated (Spielman and Wilke 2015b). Therefore, it was less likely that slowly evolving sites reflected the full MutSel steady-state distribution of codons, compared to quickly evolving sites, which ultimately incurred higher estimation error.

Divergence is more important than is the number of sequences for identifying long-term evolutionary constraint

We observed that correlations between true and inferred dN/dS values increased both as the number of sequences N and the branch lengths B (divergence) grew (Figure 1), suggesting that large and/or highly informative datasets are necessary for the inferred dN/dS to capture the actions of natural selection at evolutionary equilibrium. However, it was not immediately clear from Figure 1 whether N, B, or some combination of these conditions drove this trend. Therefore, we next assessed the relative importance of N and B.

We calculated the tree length (expected number of substitutions per site across the entire tree) for each N and B parameterization. If N and B served roughly equal roles in terms of providing information, then any combination of N and B corresponding to the same tree length should have produced similar dN/dS correlations. We did not, however, observe this trend; instead, all else being equal, B had a significantly greater influence than did N on resulting correlations. For example, as shown in Figure 4, we compared dN/dS correlations from SLAC1 for three combinations of N and B conditions which all had virtually the same tree lengths (162–164). Simulations with lower N and higher B yielded far more accurate dN/dS estimates, even though all simulations in Figure 4 experienced the same average number of substitutions. This increase was highly significant; for data simulated without codon bias, correlations increased an average $\sim$20% from $B = 0.04$ to $B = 0.64$ ($P < 10^{-15}$). As shown, neither codon bias nor the manner of true dN/dS calculation influenced this overarching trend ($P > 0.17$), although correlations between true and inferred dN/dS were generally lower when codon bias was present.

Discussion

In this study, we have examined the accuracy of different site-specific dN/dS inference approaches in the context of dN/dS point estimation. In particular, we have assessed performance differences between two dN/dS model parameterization paradigms: one-rate, where dN/dS is modeled with a single parameter, and two-rate, where dN and dS are modeled with separate parameters. We have found that one-rate inference models virtually always produce more accurate dN/dS inferences than do two-rate models. Strikingly, the presence of codon bias does not influence this result. In fact, the increased accuracy of one-rate compared to two-rate models is even more pronounced when codon bias is present (Figures 1 and 2). Therefore, our findings suggest that, even in the presence of synonymous selection, site-specific evolutionary rates should be measured using methods which estimate only dN and implicitly fix $dS = 1$ or consider a global dS for the entire sequence. We did not, however, examine how one- and two-rate inference models compare when dS variation is driven by mutational rather than selective processes.

For this study, we simulated fully heterogeneous sequences with each site evolving according a unique MutSel model. While MutSel models have shortcomings (e.g. they assume constant site-specific fitness values across the phylogeny), they take a far more mechanistic approach to coding-sequence evolution than dN/dS-based models do, and they have therefore been regarded as more evolutionarily realistic. A key benefit of simulating with the MutSel framework is that we are able to directly model synonymous rate variation by specifying different fitnesses for synonymous codons, instead of relying on a phenomenological rate parameter dS. We note that this simulation setup, however, cannot test performance accuracy on positively-selected sites ($dN/dS > 1$), as MutSel models can only correspond to sites under either purifying selection or neutral evolution ($dN/dS \leq 1$) (Spielman and Wilke 2015b). As such, we emphasize that our results here apply specifically to the question of site-specific dN/dS point estimation, and not to the question of positive-selection inference. Future work may be needed to fully understand how one-rate vs. two-rate models compare for positive-selection inference.

We demonstrate that, in the context of dN/dS point estimation, two-rate methods do not properly accomplish their intended goal of accounting for the effects of selection pressure on synonymous codons. Logically, one would presume that, when dS differs among sites, estimating dS separately across sites would produce more accurate dN/dS estimates than would fixing dS to a constant value. Indeed, an assumed presence of synonymous substitution rate variation is the very justification for using a two-rate dN/dS model (Kosakovsky Pond and Muse 2005). However, including this additional parameter hindered accuracy under virtually all simulation conditions, and we therefore conclude that including a dS parameter is not an effective way to model the presence of synonymous selection, at least on a per-site basis.

We suggest that error in dS estimates may explain the relatively poor performance of two-rate models, particularly on datasets which had codon bias. Indeed, the reason that the dN/dS ratio includes the dS denominator is to have a suitable normalization to dN that provides a baseline, neutral substitution rate (i.e. mutation rate). Statistically, the most robust way to obtain this baseline rate, if we assume that mutation rates are mostly constant across sites, is to compute an average dS across all sites (as in SLAC1, FEL2_1, and FUBAR2_1). Fixing dS to 1, as FUBAR1 and FEL1 do, yields dN/dS values that are essentially equivalent to those returned by this procedure (Figure 2). Site-specific rate inferences necessarily have high levels of noise, due to dataset size limitations, and thus estimating a separate dS at each site likely contributes substantial noise and ultimately reduces estimate reliability. One-rate methods avoid this statistical problem, and they do not appear to suffer dramatically when codon bias was present.

We additionally have found that high levels of sequence divergence are critically important for obtaining a reliable steady-state dN/dS value, moreso than the number of sequences analyzed (Figure 4). This finding has important implications for data set collection: It may be preferable to include fewer, more divergent sequences rather than as many sequences as one can obtain. Measuring dN/dS from thousands of sequences with low divergence may actually be less effective than analyzing fewer, more diverged sequences, even if the mean number of per-site substitutions would be the same. Increasing the number of taxa in a given analysis may only be beneficial if the new sequences are substantially diverged from the existing sequences. We emphasize that the number of taxa should still be sufficiently large ($\geq$100) to achieve reliable estimates, due to the inherent high level of noise in site-specific inferences.

These findings additionally build on the well-documented time-dependency of the dN/dS metric, a phenomenon studied largely in the context of polymorphic data (Rocha et al. 2006; Kryazhimskiy and Plotkin 2008; Wolf et al. 2009; Mugal et al. 2014; Meyer et al. 2015). Our results extend these findings, indicating that this time-dependency is more general and pertains also to circumstances where the data contain only fixed differences. This finding makes intuitive sense: As divergence

increases, sites will be more likely to visit the full range of selectively tolerated states, and therefore the long-term evolutionary constraints will become apparent. Importantly, even at exceptionally high divergence levels, inferred dN/dS estimates could never fully recapitulate the dN/dS that describes the steady-state distribution (Figure 1). For instance, the inferred dN/dS values for simulated alignment with the most divergence ($N = 2048$ and $B = 0.64$) had a correlation coefficient of 0.93 with the true dN/dS values, thereby explaining only 86% of the variation expected at evolutionary equilibrium. Such an empirical dataset would be difficult, if even possible, to obtain, and therefore we may not be able to recover the equilibrium dN/dS value from empirical data. Instead, it is most likely that all dN/dS measurements will be biased by time to some degree, even if all differences are fixed and not polymorphic.

Finally, our study has important implications for research that seeks to relate site-specific dN/dS ratios to protein structural properties, such as relative-solvent accessibility or weighted contact number (Spielman and Wilke 2013; Meyer and Wilke 2013; Meyer et al. 2013; Shahmoradi et al. 2014; Meyer and Wilke 2015a,b). These metrics reflect the overarching biophysical constraints that influence protein evolutionary trajectories. Studies which have examined the correlations between site-specific dN/dS and such structural quantities have recovered relatively low, although significant, correlations, generally ranging from 0.1–0.6 (Shahmoradi et al. 2014; Meyer and Wilke 2015a,b). Importantly, these studies analyzed viral sequence data, which contained relatively low levels of divergence. By contrast, other studies which considered more substantially diverged enzyme proteins (albeit using protein-sequence–derived evolutionary rates instead of dN/dS) recovered higher correlations, ranging mostly from 0.3–0.8 (Shih and Hwang 2012; Huang et al. 2014; Yeh et al. 2014b,a), between structural measures and evolutionary rate.

Our results suggest that this discrepancy is, in fact, not unexpected, and moreover that low structure–rate correlations recovered from highly similar viral sequence data are not worrisome. Just like structural quantities do, MutSel models describe the long-term evolutionary constraints acting at specific coding-sequence positions. Indeed, for data with minimal divergence, we did not recover particularly large correlations between inferred and true dN/dS (Figure 1). For example, the average correlation for simulations without codon bias for $N = 512$ and $B = 0.0025$, typical values for a virus sequence alignment, was $r = 0.414$ (inferred with SLAC1). This correlation falls well within the range of observed structure-rate correlations in empirical viral datasets. Likewise, correlations between true and inferred dN/dS were higher for more diverged data, just as observed in aforementioned studies of the structure-rate relationship in enzymes. Therefore, we suggest that future work examining the relationship between protein evolutionary rate and structure should focus on obtaining highly diverged datasets, which are more likely to provide meaningful information about long-term evolutionary constraints.

Methods

Alignment simulation

We simulated heterogeneous alignments, such that each site evolved according to a distinct distribution of codon state frequencies, according to the HB98 MutSel model (Halpern and Bruno 1998) using Pyvolve (Spielman and Wilke 2015a). We began by deriving site-specific MutSel model parameterizations. We simulated 100 site-specific amino-acid frequency distributions from a Boltzmann distribution:

$$F(a) = \frac{\exp(-\lambda_a)}{\sum_b \exp(-\lambda_b)}, \tag{1}$$

where $F(a)$ is the state frequency of amino-acid a, a and b index amino acids from 0–19, and the parameter λ increases with evolutionary rate (Ramsey et al. 2011). For each frequency distribution, we sampled a value for λ from a uniform distribution $\mathcal{U}(0, 3)$, and we selected a random fitness ranking for all amino acids.

Once amino acid frequencies were computed, we assigned frequencies to codons in two distinct ways: without selection for codon bias, and with selection for codon bias. To generate frequency distributions without codon bias, we assigned all synonymous codons the same frequency (summing to the corresponding amino-acid frequency). Alternatively, to generate frequency distributions with codon bias, we randomly selected a preferred codon for each amino acid. We assigned a state frequency of $\gamma F(a)$, where γ was drawn from a uniform distribution $\mathcal{U}(0.6, 0.9)$, to the preferred codon, and we assigned the remaining frequency $F(a) - \gamma F(a)$ evenly to all remaining synonymous codons. In this way, the overall amino-acid state frequency was unchanged, but the synonymous codons occurred with differing frequencies.

Using these site-specific MutSel model parameterizations, we simulated two sets of heterogenous alignments, one without and one with codon bias. All sites evolved according to the HKY85 (Hasegawa et al. 1985) mutation model, with $\kappa = 4$. We simulated heterogenous alignments across an array of balanced phylogenies, containing either 128, 256, 512, 1024, or 2048 sequences. For each number of taxa, we simulated sequences with varying degrees of divergence, with all branch lengths equal to either 0.0025, 0.01, 0.04, 0.16, or 0.64. We simulated 50 alignment replicates for each combination of these conditions.

dN/dS inference

For each simulated codon frequency distribution, we computed dN/dS according to the method outlined in Spielman and Wilke (2015b). We calculated dN/dS using this method in both a two-rate manner, in which dN and dS were calculated individually for each site and divided to obtain dN/dS, and in a one-rate manner, in which each site-specific dN is normalized by the mean dS across all sites.

For each simulated alignment, we inferred site-specific dN/dS values with the HyPhy software (Kosakovsky Pond et al. 2005) using several approaches: fixed-effects likelihood (FEL) (Kosakovsky Pond and Frost 2005), FUBAR (Murrell et al. 2013), and single ancestral counting (SLAC) (Kosakovsky Pond and Frost 2005). For all methods used, we specified the MG94xHKY85 (Muse and Gaut 1994; Kosakovsky Pond and Frost 2005) rate matrix with F1x4 codon frequencies, which has been shown to reduce bias in dN/dS estimation (Spielman and Wilke 2015b). We provide customized HyPhy batchfiles which enforce the F1x4 codon frequency specification in the github repository `https://github.com/sjspielman/sitewise_dnds_mutsel`.

For both FEL and FUBAR, we inferred dN/dS with both a one-rate model, in which dN/dS is represented by a single parameter and a two-rate model, in which dN and dS are modeled by separate parameters (Kosakovsky Pond and Frost 2005). For the one-rate FUBAR inferences, we specified 100 grid points to account for the reduced grid dimensionality caused by ignoring dS variation [as in Spielman et al. (2014)], and for the two-rate FUBAR inferences, we specified the default 20x20 grid (Murrell et al. 2013). All other settings were left as their default values. Similarly, for SLAC inference, we calculated dN/dS in two ways. As SLAC enumerates dN and dS on a site-specific basis, there exist two ways to calculate site-wise dN/dS: dS can be considered site-specific, or dS values can be globally averaged, and this mean can be used to normalize all site-specific dN estimates. The former calculations effectively correspond to a two-rate method (SLAC2), and the latter calculations correspond to a one-rate method (SLAC1). All inferences were conducted using the true tree along which each alignment was simulated.

As in Kosakovsky Pond and Frost (2005), we excluded all unreliable dN/dS inferences when correlating inferred and true dN/dS values. Specifically, estimates made by FEL were excluded if the estimated dN/dS equaled 1 and the P-value indicating whether the estimate differed significantly from 1 was equal to 1. Such measurements have been shown to indicate uninformative sites (Meyer et al. 2015). In addition, estimates made by SLAC2 were excluded if the number of synonymous mutations counted was 0, and hence the resulting dN/dS was undefined. Finally, all inferences with $dN/dS \geq 100$ were considered uninformative, as these high estimates likely reflect estimation error.

Data analysis and availability

Statistics were conducted in the R statistical programming language. Linear modeling was conducted using the R package lme4 (Bates et al. 2012). We inferred effect magnitudes and significance, which we corrected for multiple testing, using `glht()` function in the R package multcomp (Hothorn et al. 2008). In particular, each mixed-effects model described in the *Results* subsection *Modeling synonymous rate reduces inference accuracy* was built in the lme4 package with the general code `lmer(r ~ method + (1|replicate) + (1|N:B))`, where `r` is the Pearson correlation between inferred and true dN/dS. Relative error between inferred and true dN/dS values was calculated, for each condition, as $\text{abs}(dN/dS_{\text{inf}} - dN/dS_{\text{true}})/dN/dS_{\text{true}}$, where dN/dS_{inf} indicates the average inferred dN/dS and dN/dS_{true} indicates the true dN/dS.

All code and simulated data are freely available from `https://github.com/sjspielman/sitewise_dnds_mutsel`.

Acknowledgments

This work was supported in part by NIH grant F31 GM113622-01 to SJS, NIH grant R01 GM088344 to COW, ARO grant W911NF-12-1-0390 to COW, DTRA grant HDTRA1-12-C-0007 to COW, and NSF Cooperative Agreement No. DBI-0939454 (BEACON Center) to COW. Computational resources were provided by the University of Texas at Austin's Center for Computational Biology and Bioinformatics (CCBB) and the Stampede cluster at the Texas Advanced Computing Center (TACC). We thank Julian Echave for insightful discussion and helpful feedback.

References

Bates D, Maechler M, Bolker B. 2012. lme4: Linear mixed-effects models using S4 classes. R package version 0.999999-0.

dos Reis M. 2015. How to calculate the non-synonymous to synonymous rate ratio of protein-coding genes under the fisher-wright mutation-selection framework. Biol. Lett. 11:20141031.

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

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.

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

Hothorn T, Bretz F, Westfall P. 2008. Simultaneous inference in general parametric models. Biometrical Journal 50(3):346–363.

Huang T T, del Valle Marcos M L, Hwang J K, Echave J. 2014. A mechanistic stress model of protein evolution accounts for site-specific evolutionary rates and their relationship with packing density and flexibility. BMC Evol. Biol. 14:78.

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

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

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

Lemey P, Minin V N, Bielejec F, Kosakovsky Pond S L, Suchard M A. 2012. A counting renaissance: combining stochastic mapping and empirical Bayes to quickly detect amino acid sites under positive selection. Bioinformatics 28:3248–3256.

Liberles D A, Teufel A, Liu L, 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 haemagglutinin. Phil. Trans. R. Soc. B. 368:20120334.

Meyer A G, Spielman S J, Bedford T, Wilke C O. 2015. Time dependence of evolutionary metrics during the 2009 pandemic influenza virus outbreak. Virus Evolution 1:vev006–10.

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

Meyer A G, Wilke C O. 2015a. Geometric constraints dominate the antigenic evolution of influenza H3N2 hemagglutinin. PLOS Pathog. 11:e1004940.

Meyer A G, Wilke C O. 2015b. The utility of protein structure as a predictor of site-wise *dN/dS* varies widely among HIV-1 proteins. J. R. Soc. Interface. page *In Press*.

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, de Oliveira T, Seebregts C, Kosakovsky Pond S L, Scheffler K, Southern African Treatment and Resistance Network-SATuRN Consortium. 2012a. Modeling HIV-1 drug resistance as episodic directional selection. PLOS Comput. Biol. 8(5):e1002507.

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, Kosakovsky Pond S L. 2012b. Detecting individual sites subject to episodic diversifying selection. PLOS Genet 8(7):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.

Ramsey D C, Scherrer M P, Zhou T, Wilke C O. 2011. The relationship between relative solvent accessibility and evolutionary rate in protein evolution. Genetics 188:479–488.

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

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. U.S.A. 107:4629–4634.

Scheffler K, Murrell B, Kosakovsky Pond S L. 2014. On the validity of evolutionary models with site-specific parameters. PLOS ONE 9(4):e94534.

Shahmoradi A, Sydykova D K, Spielman S J, Jackson E L, Dawson E T, Meyer A G, Wilke C O. 2014. Predicting evolutionary site variability from structure in viral proteins: Buriedness, packing, flexibility, and design. J. Mol. Evol. 79:130–142.

Shih C H, Hwang J K. 2012. Evolutionary information hidden in a single protein structure. Proteins 80:1647–1657.

Spielman S J, Dawson E T, Wilke C O. 2014. Limited utility of residue-masking for positive-selection inference. Mol. Biol. Evol. 31:2496–2500.

Spielman S J, Wilke C O. 2013. Membrane environment imposes unique selection pressures on transmembrane domains of G protein-coupled receptors. J. Mol. Evol. 76:172–182.

Spielman S J, Wilke C O. 2015a. Pyvolve: A flexible python module for simulating sequences along phylogenies. PLOS ONE 10:e0139047.

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

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.

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

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.

Wolf J B W, Kunstner A, Nam K, Jakobsson M, Ellegren H. 2009. Nonlinear dynamics of nonsynonymous d_{N} and synonymous d_{S} substitution rates affects inference of selection. Genome Biol. Evol. 1:308–319.

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

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

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

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

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.

Yeh S W, Huang T T, Liu J W, Yu S H, Shih C H, Hwang J K, Echave J. 2014a. Local packing density is the main structural determinant of the rate of protein sequence evolution at site level. BioMed Res. Int. 2014:572409.

Yeh S W, Liu J W, Yu S H, Shih C H, Hwang J K, Echave J. 2014b. Site-specific structural constraints on protein sequence evolutionary divergence: Local packing density versus solvent exposure. Mol. Biol. Evol. 31:135–139.

Figures

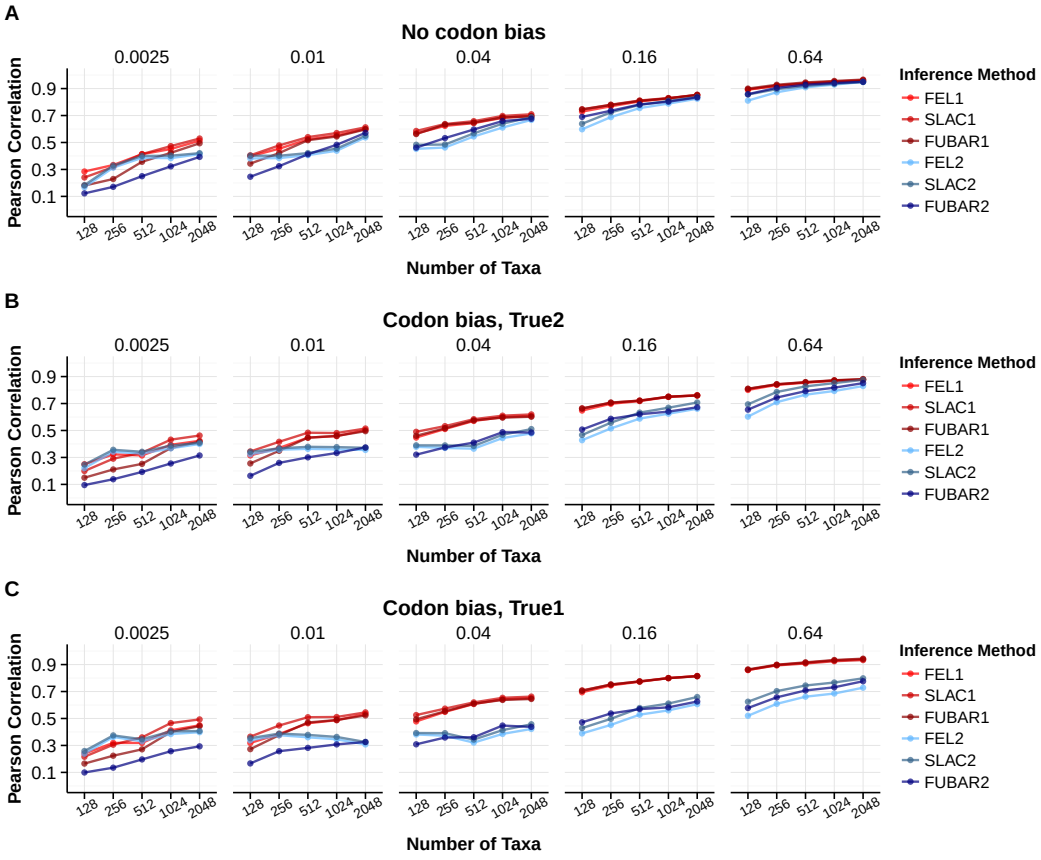

Figure 1: Pearson correlation coefficients between true and inferred dN/dS across inference approaches and N-B conditions. A) Correlations for alignments simulated without codon bias. B) Correlations with True2 for alignments simulated with codon bias. C) Correlations with True1 for alignments simulated with codon bias. The label above each sub-plot indicates the branch lengths B of the balanced phylogeny along which sequences were simulated, and the x-axes indicate the number of sequences N. Each point represents the correlation coefficient averaged across 50 replicates.

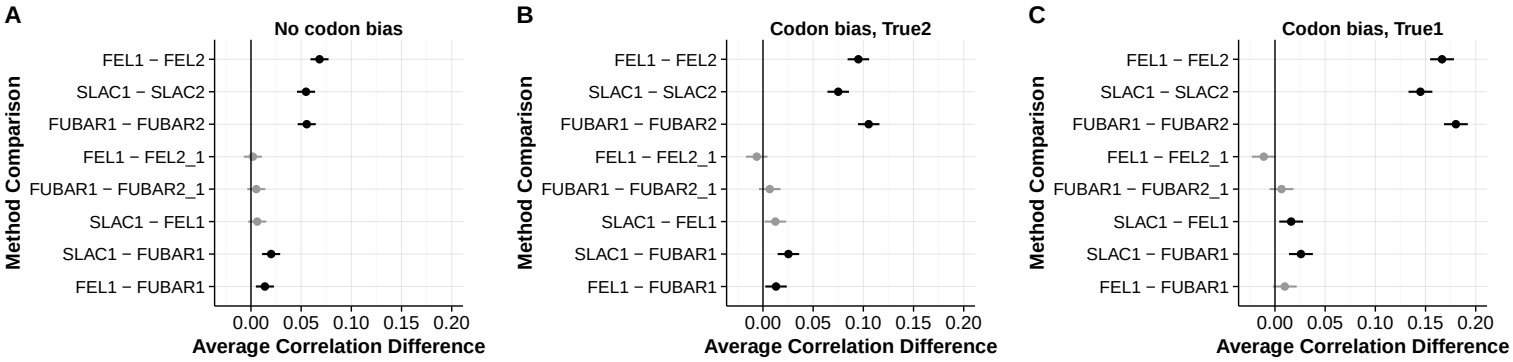

Figure 2: Pairwise comparisons of correlation strength across methods, as determined through multiple comparisons tests. A) Results for data simulated without codon bias. B) Results for data simulated with codon bias, as correlated with True2. C) Results for data simulated with codon bias, as correlated with True1. Points indicate the estimated average difference between correlations for the respective methods, and lines indicate 95% confidence intervals. Black lines indicate that the performance difference between methods differed significantly from 0 (all $P < 0.01$). Gray lines indicate that the difference was not statically significant ($P > 0.01$).

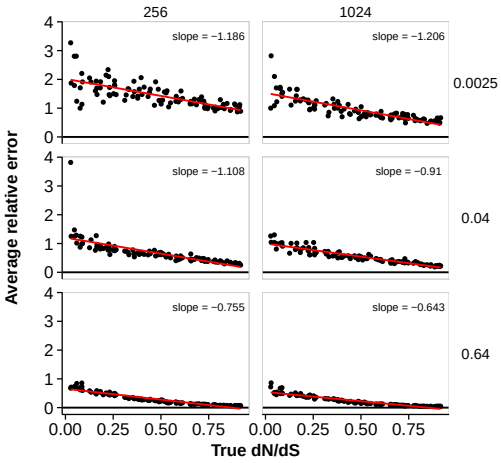

Figure 3: Average relative error of inferred dN/dS values by SLAC1 for a subset of N and B conditions. Each point represents the relative error averaged across 50 replicates. Labels above each column indicate the number of sequences N, and labels to the right of each row indicate the branch lengths B. Results are shown here for data simulated without codon bias. The horizontal line indicates an average relative error of 0, and the diagonal line is the regression line whose slope is indicated in each panel. All slopes shown are significant at $P < 10^{-15}$.

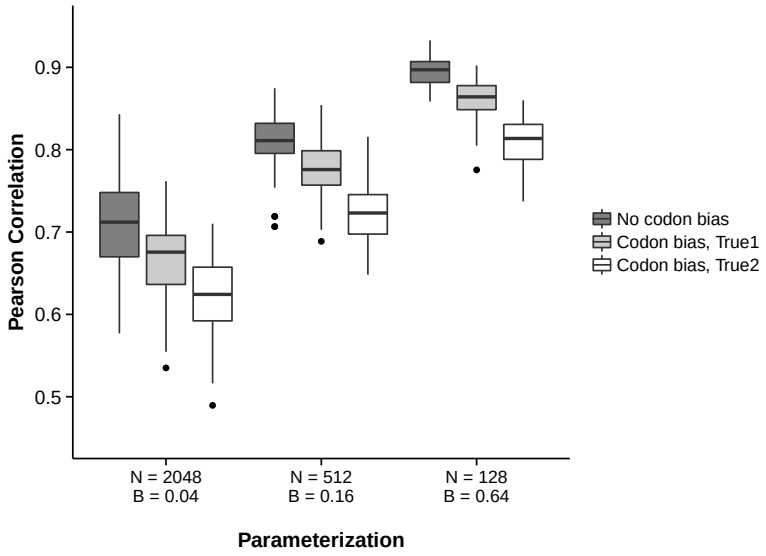

Figure 4: The amount of divergence is more important than the number of sequences is for obtaining the equilibrium dN/dS value. Each boxplot represents correlation coefficients, from SLAC1 inference, across the 50 respective replicates. From left to right, tree lengths are equal to 163.76, 163.52, and 162.56. Although mean number of per-site substitutions was therefore virtually equal among the conditions shown, higher divergence among sequences led to significantly higher accuracy than did a larger number of sequences.

Supplementary Figures

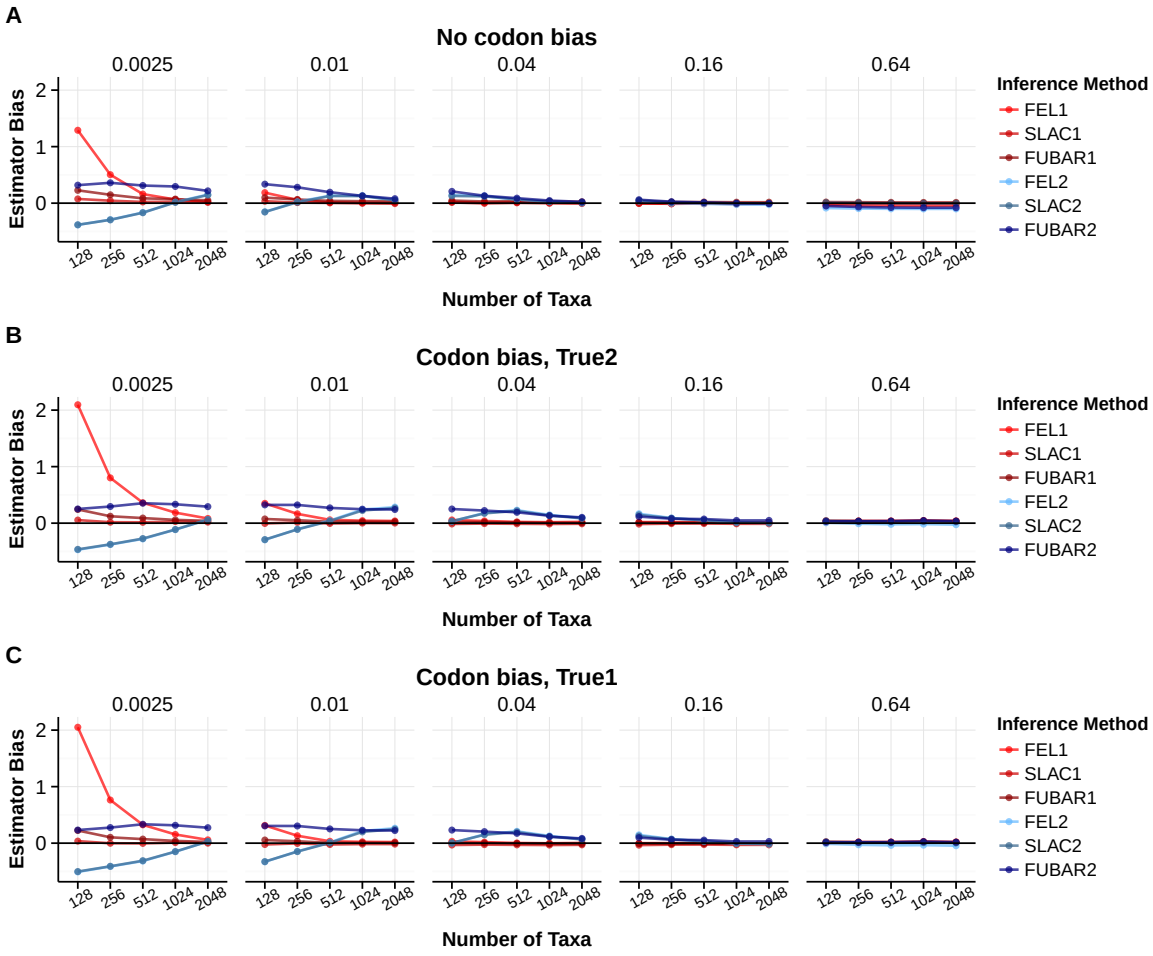

Figure S1. Estimator bias of inferred dN/dS relative to true dN/dS. A) Estimator bias for alignments simulated without codon bias. B) Estimator bias for alignments simulated with codon bias, using True2 as a reference. C) Estimator bias for alignments simulated with codon bias, using True1 as a reference. Each point represents the correlation coefficient averaged across 50 replicates. The label above each sub-plot indicates the branch lengths B of the balanced phylogeny along which sequences evolved, and the x-axes indicate the number of sequences N. Points not shown exist off the scale.

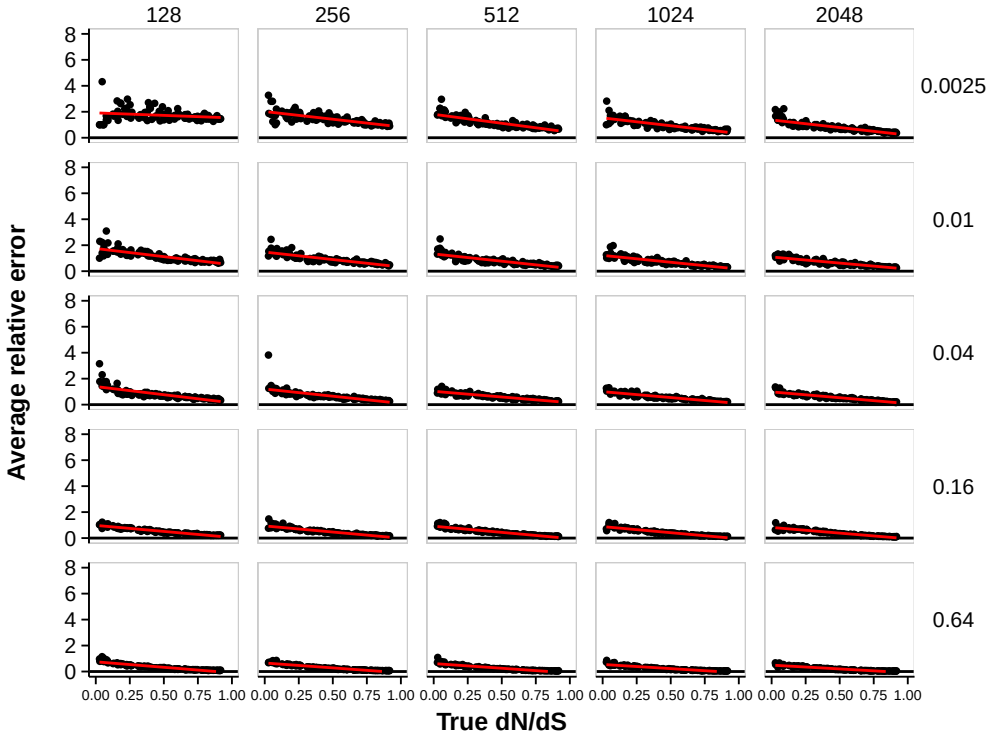

Figure S2. Average relative error of inferred dN/dS values by SLAC1 for simulations without codon bias. Each point represents the relative error average across 50 replicates. Labels above each column indicate the number of sequences N, and labels to the right of each row indicate the branch lengths B.

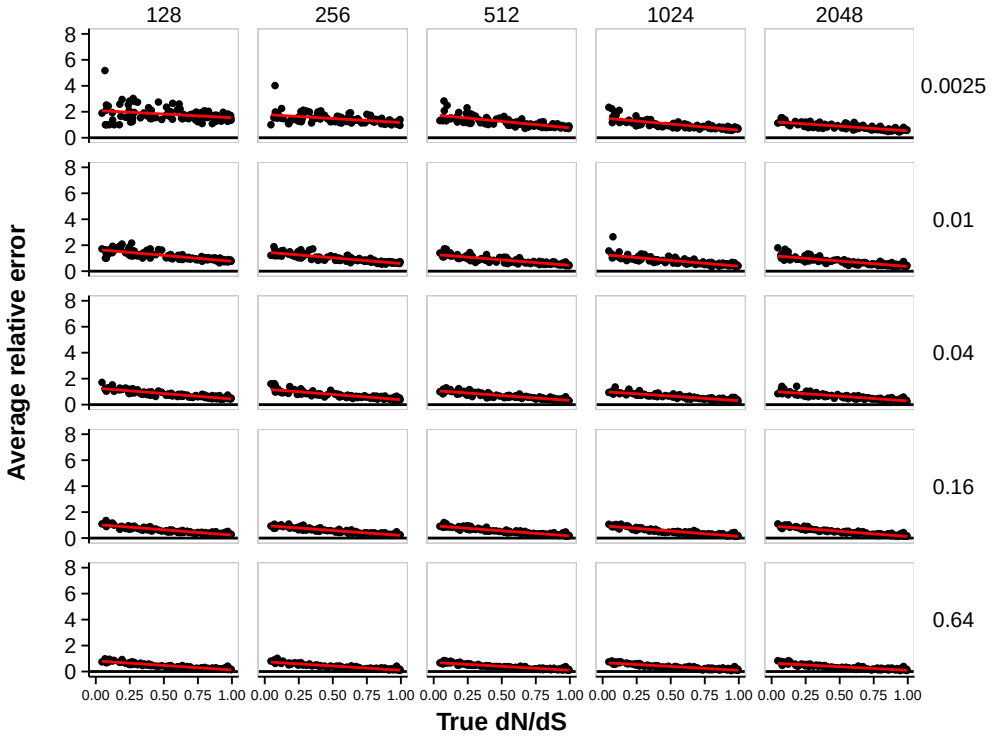

Figure S3. Average relative error of inferred dN/dS values with True2 by SLAC1 for simulations with codon bias. Each point represents the relative error average across 50 replicates. Labels above each column indicate the number of sequences N, and labels to the right of each row indicate the branch lengths B. The horizontal line indicates an average relative error of 0, and the diagonal line is the regression line.

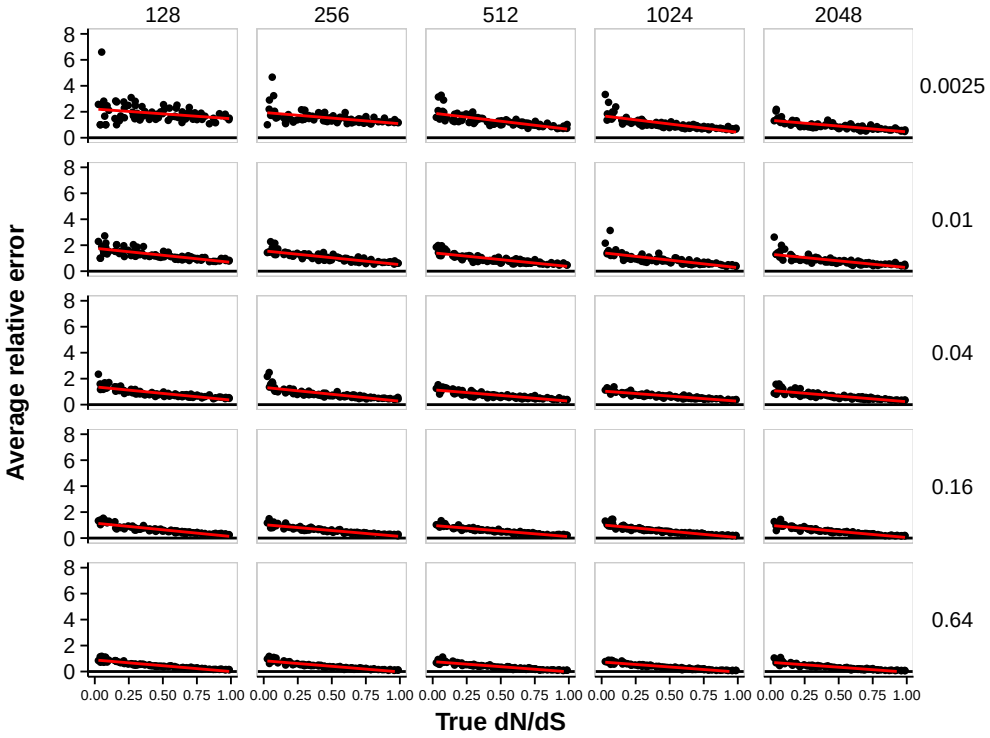

Figure S4. Average relative error of inferred dN/dS values with True1 by SLAC1 for simulations with codon bias. Each point represents the relative error average across 50 replicates. Labels above each column indicate the number of sequences N, and labels to the right of each row indicate the branch lengths B. The horizontal line indicates an average relative error of 0, and the diagonal line is the regression line.