A comparison of one-rate and two-rate inference frameworks for site-specific dN/dS estimation

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

Address:
[1]Department of Integrative Biology, Center for Computational Biology and Bioinformatics, and Institute of Cellular and Molecular Biology. The University of Texas at Austin, Austin, TX 78712, USA.
[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; mutation-selection models; evolutionary rate; sequence simulation; molecular evolution

Abstract

Two broad paradigms exist for inferring dN/dS, the ratio of nonsynonymous to synonymous substitution rates, from coding sequences: i) a one-rate approach, where dN/dS is represented with a single parameter, or ii) a two-rate approach, where dN and dS are estimated separately. These paradigms have been well-studied for positive-selection ($dN/dS > 1$) inference. By contrast, their relative merits for the specific purpose of dN/dS point estimation at individual sites remain largely untested. Here, we use sequence simulation to systematically compare how accurately each paradigm infers site-specific dN/dS ratios. In particular, we simulate alignments with mutation–selection models rather than with dN/dS-based models, thus addressing the reliability of dN/dS estimation when the simulation and inference model differ, i.e. when the inference model is mathematically misspecified. We find that one-rate frameworks universally infer more accurate dN/dS values. Surprisingly, we recover this result even when dS varies among sites. Therefore, even when extensive dS variation exists, modeling this variation substantially reduces accuracy. We attribute this finding to the increased statistical challenge of estimating dS relative to dN, which in turn is a natural result of the structure of the genetic code. A randomly chosen mutation is more likely going to result in a nonsynonymous than a synonymous change, and thus sequences are more informative for dN than for dS estimation. 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.

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. 2012c; 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 primarily been evaluated based on how well these methods 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. 2012c,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 (Echave et al. 2016). 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. In particular, we perform this analysis by simulating alignments using the mutation–selection (MutSel) modeling framework. Unlike dN/dS models, MutSel models are based on population genetics principles and describe the site-specific evolution-

ary 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, infer dN/dS on the simulated sequences using established approaches, and then compare inferred to expected dN/dS given the parameters of the MutSel model. Previously, we successfully used such an approach to identify biases in dN/dS inference approaches for whole-gene evolutionary rates (Spielman and Wilke 2015b). Here, we employ this approach to evaluate the performance of site-specific dN/dS inference approaches.

Because we perform dN/dS inferences on sequence data simulated under a MutSel model, the dN/dS model is mathematically misspecified. For example, while dN/dS models assume that all nonsynonymous substitutions occur at the same rate, MutSel models assume different rates for each type of nonsynonymous substitution, and similarly for synonymous substitutions in more complex scenarios. This misspecification directly implies that we can expect statistical biases in the performance of dN/dS inference frameworks when applied to data simulated with MutSel models. This study therefore provides insight into how dN/dS inference frameworks behave specifically when they are misspecified. Indeed, real genomes do not evolve according to a dN/dS model, and thus virtually all applications of dN/dS models will be misspecified to some degree.

Two primary questions motivate the present study: i) How accurate are various inference frameworks 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). Here, we do not benchmark positive-selection inference, but we instead ask 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. As a consequence, any dN/dS calculated from MutSel model parameters describes, by definition, the steady-state dN/dS. Since dN/dS is an inherently time-sensitive measurement (Rocha et al. 2006; Kryazhimskiy and Plotkin 2008; Wolf et al. 2009; 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

Materials and Methods

Derivation of MutSel simulation parameters

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). The rate matrix for this model is given by

$$Q_{ij}^{(k)} = \begin{cases} \mu_{x_{ij}y_{ij}} F_{ij}^{(k)} & \text{single nucleotide change} \\ 0 & \text{multiple nucleotide changes} \end{cases}, \tag{1}$$

where $\mu_{x_{ij}y_{ij}}$ is the site-invariant mutation rate where x_{ij} is the focal nucleotide before mutation and y_{ij} the focal nucleotide after mutation during the substitution from codon i to j. $F_{ij}^{(k)}$ is the fixation probability from codon i to j at site k and is defined as

$$F_{ij}^{(k)} = \frac{S_{ij}^{(k)}}{1 - e^{-S_{ij}^{(k)}}}, \tag{2}$$

where $S_{ij}^{(k)}$ is the scaled selection coefficient from codon i to j at site k (Halpern and Bruno 1998). Thus, this model is specified using a nucleotide-level mutation model (μ_{xy} parameters) and codon-level fitness values (F_{ij} parameters).

We simulated four sets of alignments, all of which assumed an HKY85 mutation model (Hasegawa et al. 1985) with the transition-tranversion bias parameter κ set to 4. Two alignment sets assumed equal nucleotide frequencies ($\pi_i = 0.25$ for $i \in \{A, C, G, T\}$), and the other two alignment sets assumed unequal nucleotide frequencies (arbitrarily set to $\pi_A = 0.32$, $\pi_T = 0.28$, $\pi_C = 0.18$, $\pi_G = 0.22$), to incorporate underlying nucleotide compositional bias. We refer to these parameterizations, respectively, as Π_{equal} and Π_{unequal}. For each mutational parameterization, we simulated an alignment set where all synonymous codons shared the same fitness value (no codon bias) and an alignment set where synonymous codons differed in fitnesses (codon bias). The Π_{equal} and Π_{unequal} simulations without codon bias used the same sets of fitness parameters, and similarly the Π_{equal} and Π_{unequal} simulations with codon bias used the same sets of fitness parameters.

It has previously been found that site-specific amino acid frequencies, in empirical alignments, tend to follow a Boltzmann distribution (Porto et al. 2004; Ramsey et al. 2011). Therefore, to derive realistic sets of codon fitnesses for each these simulations, we began by simulating 100 site-specific amino-acid frequency distributions from a Boltzmann distribution:

$$F(a) = \frac{\exp(-\lambda_a)}{\sum_b \exp(-\lambda_b)}, \tag{3}$$

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 ranking for all amino acids. These frequency calculations formed the basis of our derivation of fitness values used in all simulations.

Importantly, when a symmetric nucleotide mutation model is assumed (e.g. $\mu_{xy} = \mu_{yx}$), codon fitness values can be calculated directly as the logarithm of codon equilibrium frequency values (Sella and Hirsh 2005). Therefore, we directly computed codon fitness values from the derived

4

frequency values, under the assumption that synonymous codons shared the same fitness. These fitness parameters were employed for both Π_{equal} and $\Pi_{unequal}$ simulations without codon bias.

To derive fitness parameters for simulations 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 $\mathcal{U}(0.5, 0.8)$, to the preferred codon, and we assigned the remaining frequency $(1 - \gamma)F(a)$ evenly to all remaining synonymous codons. In this way, the overall amino-acid state frequency was unchanged, but its synonymous codons occurred with differing frequencies. Note that a single parameter γ was selected for each frequency distribution (i.e., each resulting alignment position), and not for each set of synonymous codons. Again, fitness distributions were directly computed from these resulting codon frequencies for use in both Π_{equal} and $\Pi_{unequal}$ simulations with codon bias.

Unlike the Π_{equal} simulations, the $\Pi_{unequal}$ simulations did not contain symmetric mutation rates. Therefore, we obtained stationary codon frequencies for the $\Pi_{unequal}$ simulations, for use in dN/dS calculations, numerically as the dominant eigenvector of each MutSel model's matrix, which was constructed from mutation rates and codon fitness values. In this way, all stationary codon frequency distributions incorporated, by definition, information regarding both codon-level fitness and nucleotide-level mutation.

We simulated heterogeneous 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. Throughout, we use N to refer to a given simulation's number of taxa and B to refer a given simulation's branch length. We simulated 50 alignment replicates for each combination of N and B. We additionally simulated alignments, using only the $\Pi_{unequal}$ parameterizations, along five different empirical phylogenies (Table 1), again with 50 replicates each. For these simulations, we directly used the original empirical branch lengths.

dN/dS inference

For each simulated codon frequency distribution, we computed dN/dS according to the method outlined in Spielman and Wilke (2015b). 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). 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/dnds_1rate_2rate`.

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, and we specified the default 20x20 grid for two-rate FUBAR inferences (Murrell et al. 2013). We left all other settings 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 model (SLAC2), and the latter calculations correspond to a one-rate model (SLAC1). We conducted all inferences using the true tree along

5

which we simulated each alignment.

As in Kosakovsky Pond and Frost (2005), we excluded all unreliable dN/dS inferences when correlating inferred and true dN/dS values. Specifically, we excluded FEL estimates where $dN/dS = 1$ and the P-value indicating whether the estimate differed significantly from 1 was also equal to 1. Such estimates represent uninformative sites where no mutation has occured (Meyer et al. 2015). In addition, we excluded SLAC2 estimates if the number of synonymous mutations counted was 0, and hence the resulting dN/dS was undefined. Finally, we excluded all FEL and FUBAR inferences for which the algorithm did not converge as uninformative.

Statistical Analysis

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 the `glht()` function in the R package multcomp (Hothorn et al. 2008). In particular, we built mixed-effects linear models in the lme4 package with the general code `lmer(X ~ method + (1|replicate) + (1|N:B))`, where X is either the Pearson correlation between inferred and true dN/dS or the root-mean-square deviation (RMSD) of the inferred from the true dN/dS. RMSD is calculated as $RMSD(\theta) = \sqrt{(\hat{\theta} - \theta)^2}$, where $\hat{\theta}$ is the true parameter value and θ is the estimated value. Note that, for all linear models, we excluded simulations under the $B = 0.0025$ branch length condition.

Data Availability

All code and results are freely available from the Github repository:
`https://github.com/sjspielman/dnds_1rate_2rate`.

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). Our simulation strategy is described in detail in *Methods and Materials*. Briefly, MutSel models are parameterized with a nucleotide-level mutation model and a distribution of codon fitness values. All simulations employed the HKY85 mutation model (Hasegawa et al. 1985) with the transition-transversion bias parameter $\kappa = 4.0$. We simulated data under four primary conditions: specifying either equal or unequal nucleotide frequencies in the HKY85 model, and specifying no codon bias or codon bias for the codon fitness values. Simulations without codon bias assumed that all synonymous codons had the same fitness, and simulations with codon bias assumed that synonymous codons differed in fitness values. We refer to simulations with equal nucleotide frequencies as Π_{equal} and to simulations with unequal nucleotide frequencies as $\Pi_{unequal}$.

Each simulated alignment contained 100 sites, and 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 combinations 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 N and B conditions directly comparable.

6

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. Throughout, we use *method* to refer to distinct algorithmic approaches (FEL, FUBAR, and SLAC), and we use *model* to refer to a one-rate or a two-rate parameterization. We use either *framework* or *approach* to more generally discuss one-rate vs. two rate methods.

We performed all dN/dS inference using the HyPhy batch language (Kosakovsky Pond et al. 2005). 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 quantified performance for all inference frameworks using several metrics, notably the Pearson correlation between true and inferred dN/dS and the root-mean-square deviation (RMSD) of inferred from true dN/dS. 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 this condition of stationary. When either the simulated divergence or number of sequences analyzed is low, then, it not necessarily possible to capture the true equilibrium distribution of codons. Therefore, to determine the relative performance of dN/dS inference methods, we considered the most accurate inferences as those with the highest dN/dS correlations, or conversely the lowest RMSD, within a given choice of N and B.

In Figure 1, we show resulting Pearson correlation coefficients, averaged across all 50 replicates, between inferred and true dN/dS for each inference framework, specifically for Π_{unequal} simulations. Results for Π_{equal} simulations showed virtually identical correlations ($P = 0.68$ ANOVA, Figure S1). In the absence of codon bias, dS was equal to 1 at all sites. As such, we expected that one-rate inference frameworks would outperform two-rate inference frameworks. We indeed found that one-rate inference frameworks showed the highest correlations when there was no synonymous selection (Figure 1A), in particular at low-to-intermediate divergence levels ($B \in \{0.01, 0.04, 0.16\}$). As the sequences became more diverged, and hence more informative, two-rate frameworks increas-

ingly performed as well as one-rate frameworks did. Even so, two-rate frameworks almost never outperformed one-rate frameworks.

In the presence of codon bias, both dN and dS varied at each site, and therefore we expected that two-rate frameworks would be more well-suited for these simulations. Surprisingly, however, one-rate frameworks still outperformed two-rate frameworks across N and B conditions, in spite of the pervasive site-wise dS variation across sites (Figure 1B and Figure S1). Moreover, correlation differences between one-rate and two-rate frameworks were more pronounced for simulations with codon bias than for simulations without codon bias. In other words, two-rate frameworks performed worse on data simulated with codon bias than they did on data simulated without codon bias.

To complement our correlation analysis, we calculated several additional metrics to quantify accuracy: i) root-mean-square deviation (RMSD) of the inferred dN/dS from the true dN/dS (Figure S2), ii) estimator bias for each inference framework (Figure S3), and iii) variance in residuals for a simple linear model regressing inferred on true dN/dS (Figure S4). These metrics displayed the same general trends as did correlation analysis: One-rate frameworks were generally more accurate and precise (lower RMSD, lower estimator bias, and lower residual variance) than were two-rate frameworks, and these overarching trends were more pronounced for simulations with codon bias (all $P < 2 \times 10^{-16}$, ANOVA). As divergence increased, each metric dropped substantially for both one- and two-rate frameworks, with error and/or bias for one-rate frameworks dissipating more quickly than for two-rate frameworks. These patterns were consistent between the Π_{unequal} and Π_{unequal} simulations for estimator bias and residual variance (both $P > 0.2$, ANOVA), although Π_{equal} displayed marginally smaller RMSD values compared to Π_{unequal} ($P = 0.04$ with an average difference of -4.45×10^{-3}, ANOVA). Thus dN/dS inference was robust to the presence of nucleotide compositional bias.

Rate parameterization affects dN/dS estimates more strongly than does inference method

We next quantified performance differences among inference frameworks more rigorously, using linear models. For each simulation set, we built mixed-effects linear models with either Pearson correlation or RMSD as the response, inference approach 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 inference approaches.

Linear model analysis confirmed prior observations that each one-rate method significantly outperformed its respective two-rate counterpart (Figure 2 for Π_{unequal} simulations, and Figure S5 for Π_{equal} simulations). In addition, correlations differences among one-rate methods were not statistically significant for any inferences performed on Π_{unequal} simulations (Figure 2A). For Π_{equal} simulations, SLAC1 yielded significantly higher correlations than did FUBAR1, although the effect magnitude size was minimal, with a mean difference of $r = 0.01$ (Figure S5A).

Results for RMSD linear models additionally showed, for both Π_{unequal} and Π_{equal} simulations, that one-rate methods yielded less error in point estimates than did two-rate methods (Figure 2B and S5B). Unlike results from linear models with correlation as the response, however, RMSD analysis showed some more substantial differences among one-rate methods. Overall, SLAC1 and FEL1 performed comparably, but FUBAR1 showed lower RMSD than both SLAC1 and FEL1, albeit with a very small effect magnitude. This result persisted across all simulation conditions. Together, these findings suggest the number of parameters used to model dN/dS mattered more than did the specific inference method chosen.

8

Data contain insufficient information for precise site-wise dS estimation

We next sought to determine why one-rate frameworks outperformed two-rate frameworks. To this end, we examined inference accuracy for individual dN and dS parameters, rather than their ratio, across inference frameworks. Given the broad similarly among inference methods and simulation sets, we considered only FEL inferences on Π_{unequal} simulations for these analyses.

To ascertain whether the added dS parameter might have affected dN estimation in two-rate models, we first confirmed that dN estimates between one-rate and two-rate models were comparable. We examined individual dN estimates between FEL1 and FEL2, and we found that, aside from low-information conditions (e.g. $B = 0.0025$ and/or $N = 128$), dN estimates were virtually identical between FEL1 and FEL2, for both simulations with and without codon bias (Figure S6). Therefore, differences in dN estimation did not distinguish one-rate vs. two-rate model performance.

We therefore compared how well FEL2 estimated dN relative to dS, specifically for simulations with codon bias, in which dS variation exists. We found that FEL2 consistently estimated dN more precisely than dS, as measured using both correlations and RMSD (Figure 3). Although accuracy for both dN and dS estimation increased as either B or N increased, dN estimates universally displayed higher correlations and lower RMSD than did dS estimates. As such, it appeared that dS was simply statistically more difficult to estimate than was dN.

We hypothesized that this result was a direct consequence of the relative amount of information in the alignments for nonsynonymous vs. synonymous changes. Indeed, due to the structure of the genetic code, 74% of all possible single-nucleotide changes are nonsynonymous, and the remaining 26% are synonymous. Therefore, under the simplest condition of neutral evolution, where both nonsynonymous and synyonymous changes are equally likely to go to fixation, a given evolutionary lineage should experience approximately three times more nonsynonyomus than synonymous substitutions.

We therefore asked whether our simulations displayed this kind of pattern of mutation accumulation. Using the simulated ancestral sequences within each simulated alignments, we directly counted the number of nonsynonymous and synonymous changes which had occurred across the phylogeny. Indeed, we observed that nonsynonymous changes occurred roughly twice as frequently over the course of simulation than did synonymous changes (Figure 4). This result was fully compatible with the notion that statistical estimation of dS was more challenging than dN because of sample size: Alignments contained nearly double the amount of information contributing to dN than to dS. As a consequence, dS estimation was less precise and noisier across simulation conditions, ultimately explaining why two-rate frameworks yielded less precise dN/dS compared to one-rate frameworks.

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

We additionally observed that dN/dS inference accuracy increased both as the number of sequences N and the branch lengths B (divergence) grew (Figures 1, S1–4), 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 whether N, B, or some combination of these conditions drove this trend. Therefore, we next assessed the relative importance of N and B on estimating the equilibrium dN/dS rate ratio.

We calculated the tree length for each N and B parameterization, specifically for Π_{unequal} simulations. Note that, in the context of our mutation–selection simulations, the tree length indicates

9

the expected number of substitutions per site across the entire tree, relative to the number of neutral substitutions (Spielman and Wilke 2015a; Tamuri et al. 2012). 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 the resulting correlations. For example, as shown in Figure 5, we compared dN/dS correlations and RMSD from FEL1 for three combinations of N and B conditions that all yielded virtually the same tree lengths (162–164). Simulations with lower N and higher B resulted in far more accurate dN/dS estimates, even though all simulations in Figure 5 experienced the same average number of substitutions. This increase was highly significant; for data simulated without codon bias, correlations increased an average of $\sim 28\%$, from $B = 0.04$ to $B = 0.64$, and similarly RMSD decreased an average of $\sim 50\%$ (both $P < 10^{-15}$, linear model). For data simulated with codon bias, correlations increased an average of $\sim 33\%$, from $B = 0.04$ to $B = 0.64$, and RMSD decreased an average of $\sim 52\%$ (both $P < 10^{-15}$, linear model). Therefore, the relative importance of divergence over number of taxa held for simulations with and without codon bias alike.

Simulations along empirical phylogenies recapitulate observed trends

While the balanced-tree simulations described above provided a useful framework for examining broad patterns in inference-framework behaviors, they did not necessarily reflect the properties of empirical datasets. We therefore assessed how applicable our results were to real data analysis by simulating an additional set of alignments along five empirical phylogenies (Table 1). Importantly, we considered the original empirical branch lengths for this analysis, so these phylogenies featured a range of number of taxa and divergence levels representative of empirical studies. We simulated alignments using the Π_{unequal} MutSel parameterizations, with both no codon bias and codon bias. We simulated 50 replicate alignments for each phylogeny and parameterization, and for simplicity we inferred dN/dS using only FUBAR1 and FUBAR2.

We identified the same general trends in these empirical simulations as we observed for simulations along balanced trees: FUBAR1 estimated dN/dS more precisely than did FUBAR2, and phylogenies with higher divergence levels (e.g. longer branch lengths) yielded more accurate estimates (Figure 6). Furthermore, FUBAR2 estimated dN more accurately than dS (Figure S8), again reflecting the relative difficulty in estimating dS compared to dN. Importantly, most empirical phylogenies showed mean correlations with true dN/dS of $0.4 < r < 0.6$, with the key exception of the biogenic amine receptor phylogeny ("amine"), whose extremely high number of taxa and divergence yield exceptionally high correlations with both FUBAR1 and FUBAR2. Therefore, we found that under more realistic conditions, estimated dN/dS will correlate with the equilibrium dN/dS ratio with only moderate strength.

In addition, results for simulations along the empirical phylogenies supported our findings regarding the relative importance of divergence vs. number of taxa, in particular through the juxtaposition of results for camelid and vertebrate rhodopsin ("vertrho") simulations. These phylogenies showed similar tree lengths, but the camelid tree length was driven by the number of sequences, and the vertebrate rhodopsin tree length was driven by its larger branch lengths (Table 1). Correlations and RMSD revealed a far higher inference accuracy for the vertebrate rhodopsin simulations than for the camelid simulations. On average, vertebrate rhodopsin correlations were 0.08 higher than were camelid correlations, and vertebrate rhodopsin RMSD was 0.1 lower than was camelid RMSD (both $P < 5 \times 10^{-11}$, linear model). These metrics were consistent between simulations with and without codon bias (both $P > 0.5$, linear model test for interaction effect of codon bias parameterization). Therefore, even for data simulated along empirical phylogenies, sequence diver-

gence proved more important than did the number of taxa for accurately estimating the equilibrium dN/dS rate ratio.

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, across a variety of inference algorithms. Strikingly, the presence of codon bias, i.e. synonymous rate variation, did not negate this result. In fact, the increased accuracy of one-rate compared to two-rate models was even more pronounced when codon bias was present (Figures 1 and 2). We attribute the improved performance of one-rate frameworks to the relative statistical difficulty in estimating dS compared to dN, due to the relative amount of information in the data informing these two quantities.

Our results demonstrate that two-rate frameworks did not properly accomplish their intended goal of modeling synonymous rate variation. 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. Instead, we recommend measuring site-specific evolutionary rates using frameworks which estimate only dN and implicitly fix $dS = 1$ or consider a global dS for the entire sequence. Alternatively, if one wishes to identity or quantify cases of strong synonymous constraint, we recommend approaches such as FRESCo (Sealfon et al. 2015), which employ hypothesis testing to compare models that fit dS globally with models that fit dS at individual sites and/or regions.

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 5). 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. More specifically, increasing the number of taxa in a given analysis may only be beneficial if the new sequences are substantially diverged from the existing sequences. 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.

For this study, we simulated fully heterogeneous sequences with each site evolving according to a unique MutSel model. While MutSel models have some obvious 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. For this reason, MutSel-based simulation approaches have been used to study the behavior of phylogenetic and evolutionary rate inferences (Holder et al. 2008; McCandlish et al. 2013; Spielman and Wilke 2015b; dos Reis 2015).

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 whose precise biological meaning is unclear.

11

However, our use of MutSel models for simulations raises several important caveats that directly impact how to interpret our results. First, our results are contingent on the fact that our simulation model did not match our inference model, and hence the inference model was mathematically misspecified. For example, comparable performance of SLAC, an approximate counting-based method, with FEL and FUBAR, both of which employ more rigorous statistical procedures, may be directly attributed to model misspecification. Indeed, previous studies using properly-specified models have suggested that SLAC may be a biased estimator when correctly specified, particularly at high divergences (Kosakovsky Pond and Frost 2005). Therefore, it is certainly possible that a two-rate dN/dS framework would outperform a one-rate dN/dS framework if the model were correctly specified. On the other hand, natural sequences certainly do not evolve according to a dN/dS model, and as such all dN/dS inferences on biological data are technically misspecified. Therefore, we contend that MutSel simulation strategy provides a more realistic analogy to empirical data analysis than would inferring dN/dS in a properly specified context.

Second, because MutSel models can only correspond to sites evolving under an evolutionary equilibrium [i.e. under either purifying selection or neutral evolution where $dN/dS \leq 1$ (Spielman and Wilke 2015b)], our results do not necessarily apply to contexts where sequences do not evolve under equilibrium, or for the specific application of positive-selection inference ($dN/dS > 1$). For example, dN/dS inference is perhaps most commonly used to study sequences evolving along a changing fitness landscape, as would be the case for viral and/or pathogen evolution (Delport et al. 2008; Murrell et al. 2012b; Demogines et al. 2013; Meyer et al. 2015; Meyer and Wilke 2015a) By contrast, the MutSel model used here assumes that the fitness landscape is static across the phylogeny. Therefore, we must emphasize that our results apply primarily to sequences evolving under equilibrium conditions, and not necessarily to sequences evolving under shifting selection pressures. Future work may be needed to fully understand how one-rate vs. two-rate models compare under different circumstances and specifically for positive-selection inference.

Finally, our simulations assumed that dS variation was driven by selection on synonymous codons. In other circumstances, synonymous rate variation might emerge when a given gene contains mutational hotspots, e.g. regions with a strongly elevated nucleotide mutation rate. In such circumstances, it is possible that a two-rate model would outperform a one-rate model if the variation in mutational processes were sufficiently large (Kosakovsky Pond and Frost 2005).

Our study additionally builds on the well-documented time-dependency of the dN/dS metric, a phenomenon studied primarily 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 and indicate 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, at which time the long-term evolutionary constraints will become apparent. It is therefore likely that most 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 (Echave et al. 2016). These metrics reflect the overarching biophysical constraints that influence protein evolutionary trajectories. Studies which have examined the correlations between site-specific dN/dS, or conversely amino-acid level evolutionary rates, and such structural quantities have recovered correlations with strengths widely ranging from 0.1–0.8 (Shih and Hwang 2012; Huang et al. 2014; Yeh et al. 2014b,a; Shahmoradi et al. 2014; Meyer and Wilke 2015b,c; Jackson et al. 2016). Recent work by Jackson et al. (2016) has attributed the source of this wide range of correlations directly to the extent of sequence divergence present in a given dataset, such

that more diverged datasets display higher correlations and less diverged datasets display lower correlations. Our findings that increased divergence levels contribute strongly to dN/dS inference accuracy are fully consistent with those of Jackson et al. (2016). Therefore, we suggest that future work examining the relationship between protein evolutionary rate and structure should focus on datasets with intermediate-to-high divergences, which are most likely to provide meaningful information about long-term evolutionary constraints.

Acknowledgments

Figures and Tables
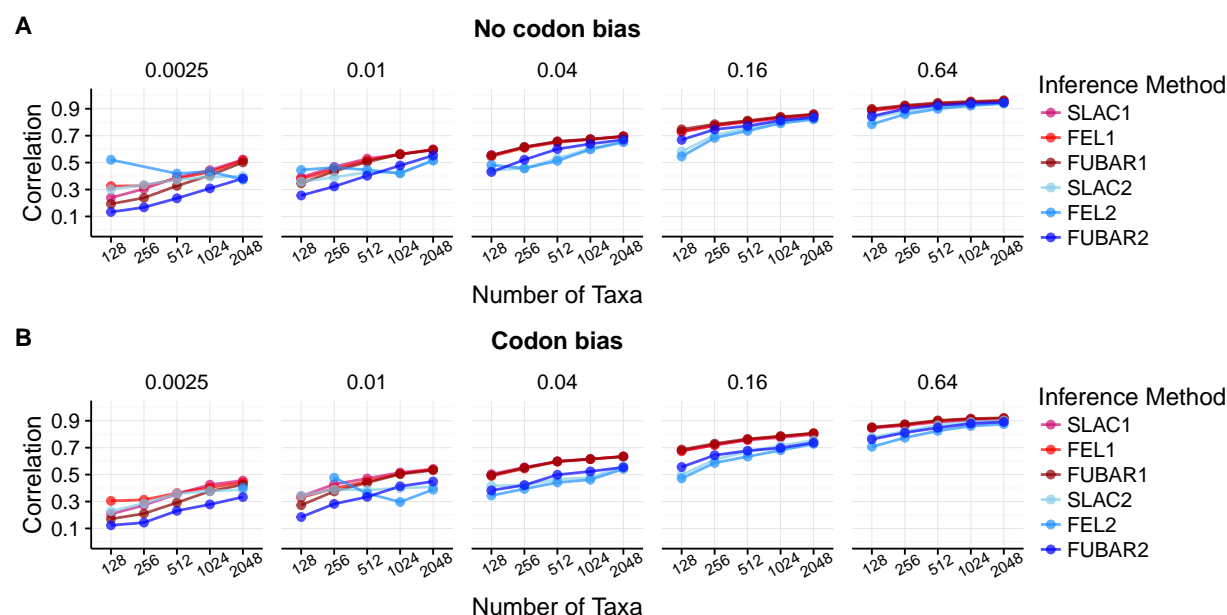
Figure 1: Pearson correlation coefficients between true and inferred dN/dS across inference approaches and N-B conditions, for Π_{unequal} simulations. A) Correlations for alignments simulated without codon bias. B) Correlations for alignments simulated with codon bias. The label above each sub-plot 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. A corresponding figure for Π_{equal} simulations is in Figure S1. Note that certain FEL2 points (at $B = 0.0025$ and $B = 0.01/N = 128$, for codon bias simulations) are not present because FEL2 generally failed to converge under these conditions.
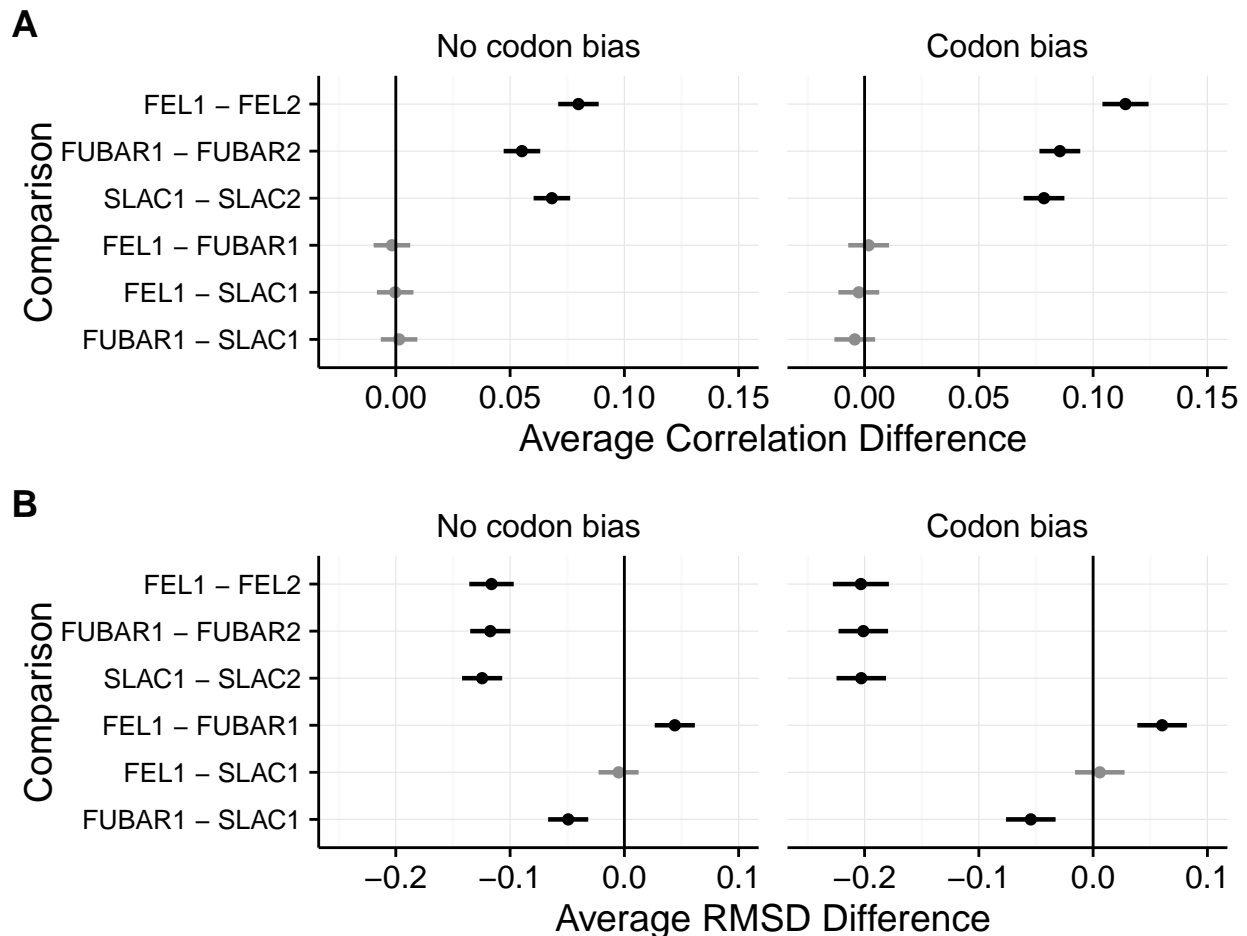
A

B

Figure 2: Pairwise comparisons of correlation strength, RMSD, and variance in residual across methods, as determined through multiple comparisons tests, for Π_{unequal} simulations. A) Results for multiple comparison tests of correlation strength. B) Results for multiple comparison tests of RMSD. Points indicate the estimated average difference between measurement for the respective inference approaches, 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 ($P > 0.01$). All P-values were corrected for multiple testing. Corresponding multiple comparison results for Π_{equal} simulations are in Figure S5.
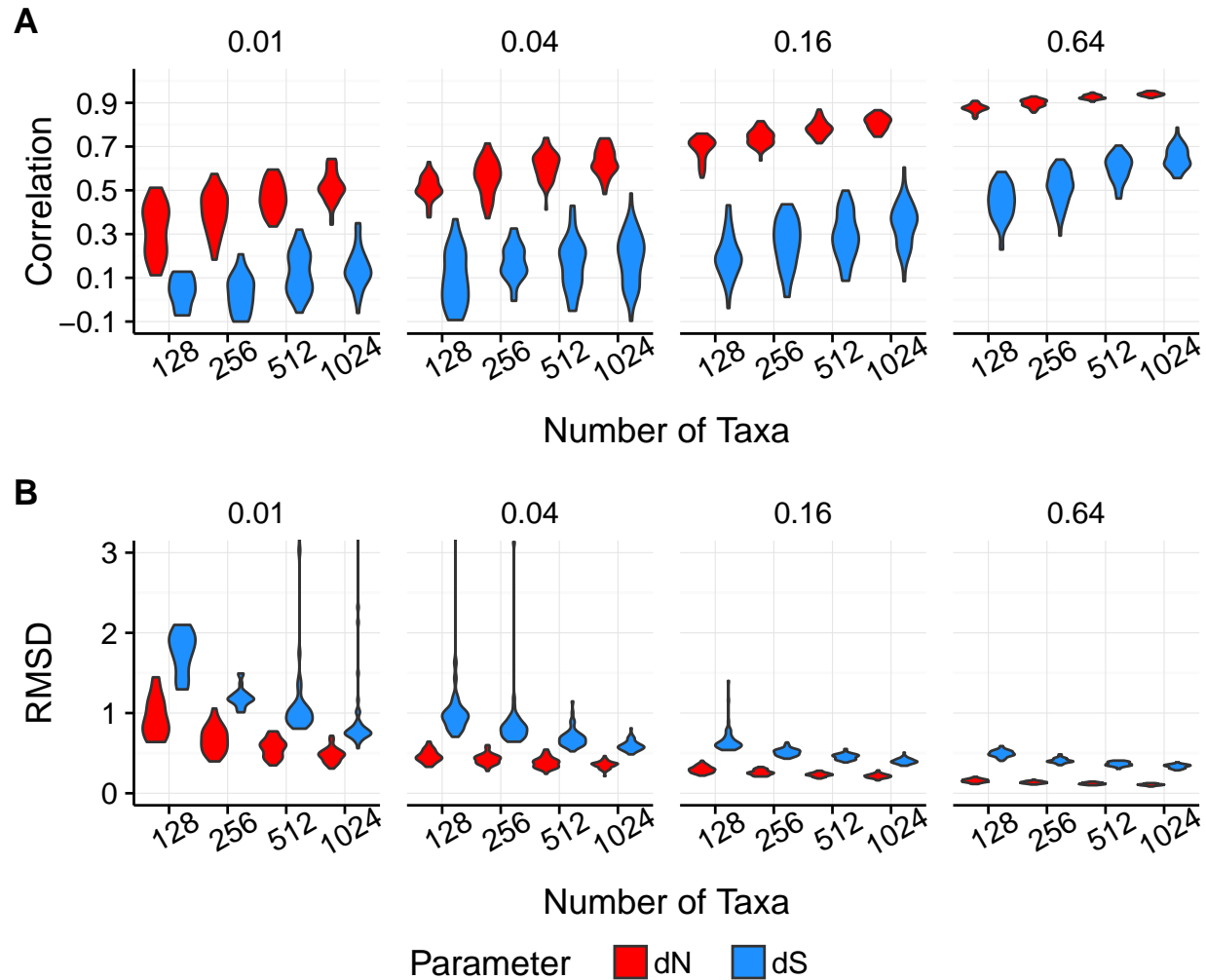
14

Figure 3: *dN* estimates are more precise than are *dS* estimates. Results in this figure are shown for a subset of conditions for Π_{unequal} simulations with codon bias, as inferred with FEL2. A) Violin plots of Pearson correlations between inferred and true *dN* and *dS* values. B) Violin plots of RMSD of inferred from true *dN* and *dS* values. Outlying points beyond the y-axis ranges have been removed for visualization.
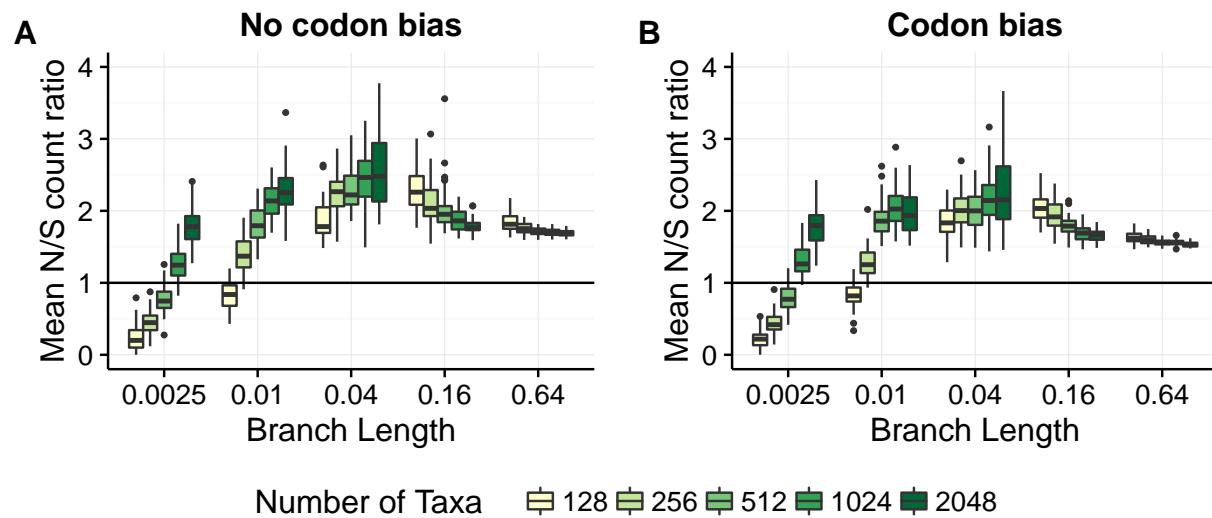
Figure 4: Ratio of the number of nonsynonymous to synonymous changes which occurred during simulations, as counted directly from simulated ancestral sequences. Each boxplot represents, across 50 replicates, the per-alignment ratio of the number of nonsynonymous changes to synonymous changes averaged across sites. Results shown here are from Π_{unequal} simulations.

Table 1: **Empirical phylogenies**

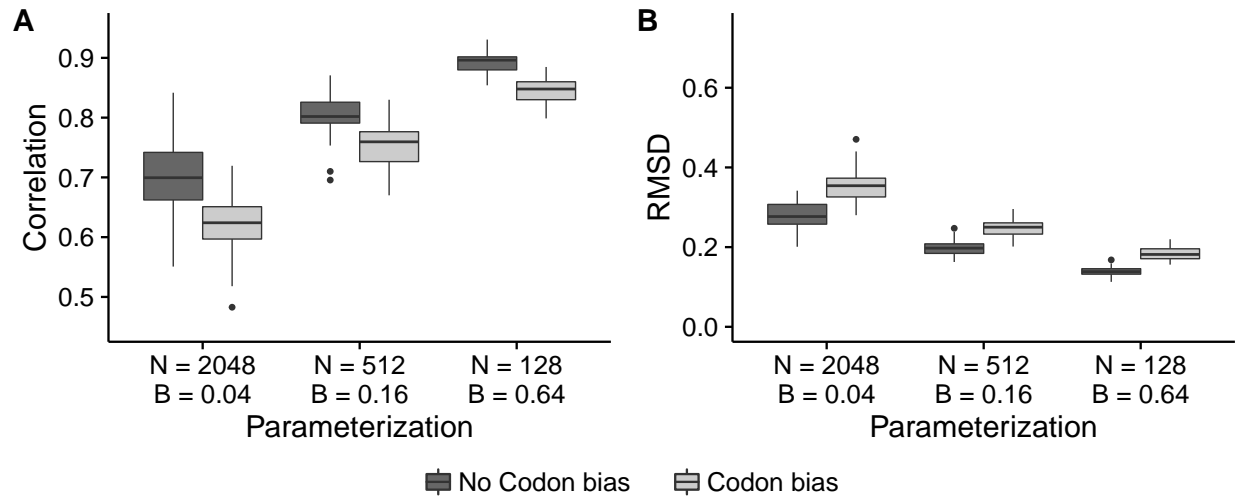| Dataset	N^1	Tree length	Mean branch length ($\pm$ sd.)	Reference
amine[2]	3039	440.861	0.073±0.149	Spielman et al. (2015)
camelid[3]	212	15.428	0.037±0.049	Kosakovsky Pond and Muse (2005); Murrell et al. (2012a, 2013)
vertrho[4]	38	12.815	0.183±0.159	Murrell et al. (2012a, 2013)
h3[5]	3854	5.951	0.0005±0.001	Meyer and Wilke (2015b)
hivrt[6]	383	5.642	0.007±0.01	Murrell et al. (2012a)

16

Figure 5: The amount of divergence is more important than the number of sequences is for obtaining the equilibrium dN/dS value. Boxplots represent either A) correlation or B) RMSD across the 50 respective simulation replicates. Results shown correspond to $\Pi_{unequal}$ simulations as inferred with FEL1. From left to right, tree lengths are equal to 163.76, 163.52, and 162.56. A corresponding figure for Π_{equal} simulations is in Figure S7.
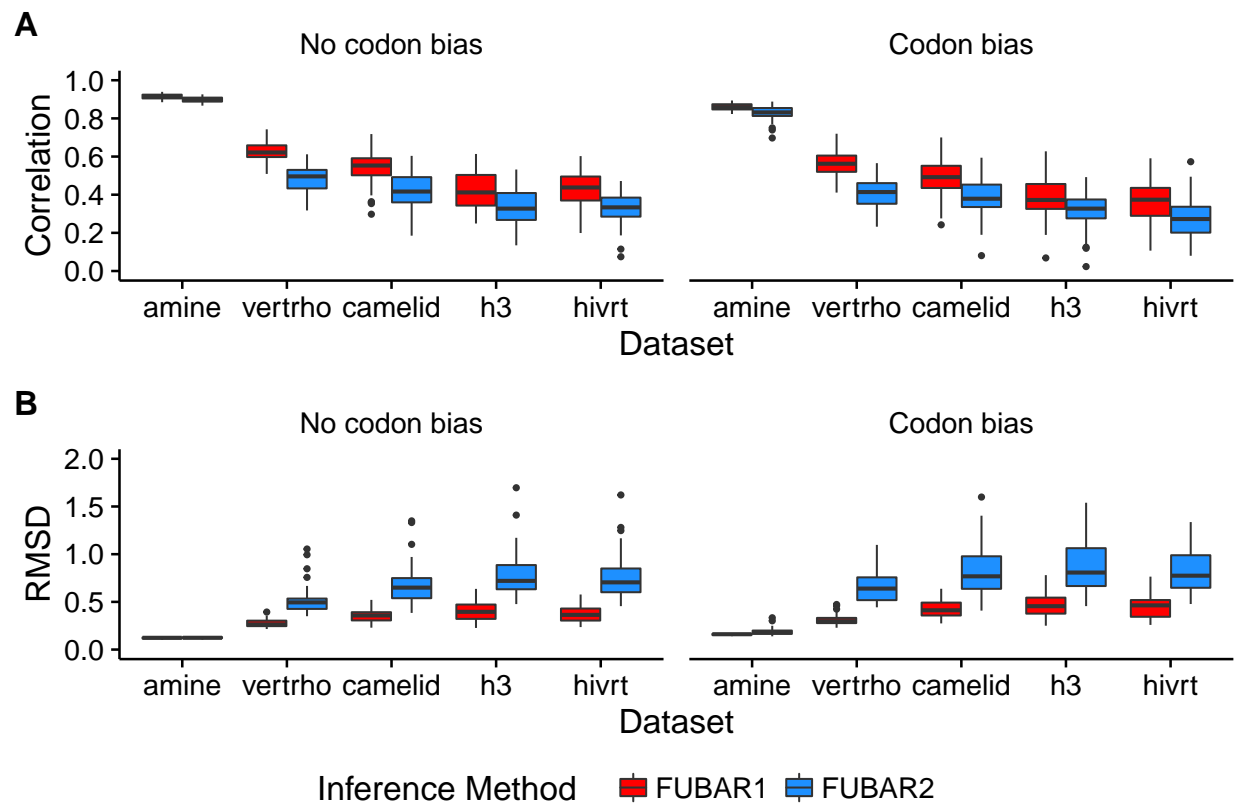

Figure 6: Inference results for simulations performed along empirical phylogenies (Table 1). A) Pearson correlations between inferred and true dN/dS values. B) RMSD of inferred from true dN/dS values.

17

References

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

Delport W, Scheffler K, Seoighe C. 2008. Frequent toggling between alternative amino acids is driven by selection in HIV-1. PLOS Pathog. 4:e1000242.

Demogines A, Abraham J, Choe H, Farzan M, Sawyer S L. 2013. Dual host-virus arms races shape an essential housekeeping protein. PLOS Biol. 11:e1001571.

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.

Echave J, Spielman S J, Wilke C O. 2016. Causes of evolutionary rate variation among protein sites. Nature Rev. Genet. 17:109–121.

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 J. 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.

Jackson E L, Shahmoradi A, Spielman S J, Jack B R, Wilke C O. 2016. Intermediate divergence levels maximize the strength of structuresequence correlations in enzymes and viral proteins. Prot. Sci. page *In Press*.

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.

McCandlish D M, Rajon E, Shah P, Ding Y, Plotkin J B. 2013. The role of epistasis in protein evolution. Nature 497:E1–E2.

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. 2015a. Geometric constraints dominate the antigenic evolution of influenza H3N2 hemagglutinin. PLOS Pathog. 11:e1004940.

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

Meyer A G, Wilke C O. 2015c. 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, de Oliveira T, Seebregts C, Kosakovsky Pond S L, Scheffler K, Southern African Treatment and Resistance Network-SATuRN Consortium. 2012b. Modeling HIV-1 drug resistance as episodic directional selection. PLOS Comput. Biol. 8: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. 2012c. 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.

Porto M, Roman H E, Vendruscolo M, Bastolla U. 2004. Prediction of site-specific amino acid distributions and limits of divergent evolutionary changes in protein sequences. Mol. Biol. Evol. 22:630–638.

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.

Sealfon R S, Lin M F, Jungreis I, Wolf M Y, Kellis M, Sabeti P C. 2015. FRESCo: finding regions of excess synonymous constraint in diverse viruses. Genome Biol. 16.

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

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, Kumar K, Wilke C O. 2015. Comprehensive, structurally-informed alignment and phylogeny of vertebrate biogenic amine receptors. PeerJ 3:e773.

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.