Introduction

Over the years, various methods have been used to calculate the strength of natural selection acting on protein-coding sequences. Traditionally, the focus has been on estimating the evolutionary rate ratio, dN/dS, the rate of nonsynonymous to synonymous substitution rates. This metric indicates how quickly a protein's constituent amino acids change, and is widely used to identify cases of positive selection ($dN/dS > 1$). Following early counting methods for estimating dN/dS (e.g. refs [1] and [2]), mechanistic codon substitution models, which assume an explicit Markov-process model of sequence evolution (see ref. [3] for a comprehensive review), have taken a leading role as the inferenqce method of choice since their introduction in the 1990s [4–6]. These models yield maximum likelihood estimates (MLEs) for the parameter ω, which represents the quantity dN/dS, and have seen great success in the field of molecular evolution.

A second class of models, known as mutation-selection-balance (MutSel) models, have emerged recently as a popular alternative to mechanistic codon models. The MutSel framework, couched firmly in population genetics theory, models the dynamic interplay between mutation and selection in a protein-coding sequence. MutSel models yield estimates of site-wise scaled selection coefficients, which indicate the extent to which natural selection favors, or disfavors, particular codons or amino acids at a given protein position. Although MutSel models were first introduced over 15 years ago [7], they have seen virtually no use due to their high computational expense. However, recently, several computationally tractable model implementations have emerged [8,9], allowing for the first time the potential for widespread use.

Although both dN/dS models and MutSel models describe the same fundamental process of protein-coding sequence evolution along a phylogeny, it is largely unknown how these two classes of models relate to one another. In particular, as these inference methods have been developed independently, it remains an open question whether or not parameter estimates from one model are comparable to those of the other model. Therefore, while certain rhetorical arguments may be made in favor of using one method over another, there is currently no formalized, concrete rationale to guide researchers in their methodological choices.

Here, we formalize the relationship between ω and MutSel models by examining the extent to which their focal parameters, dN/dS and scaled selection coefficients, yield overlapping information about the evolutionary process. To this end, we derive a mathematical relationship these models' primary parameters from which one can infer dN/dS values from selection coefficients alone. Using a simulation approach, we verify that dN/dS values estimated using selection coefficients alone correspond precisely to ω MLEs inferred using standard mechanistic codon models.

(???)Further, we prove that, under conditions of symmetric mutation rates, this relationship holds only under regimes of purifying selection or neutral evolution ($dN/dS \leq 1$). This proof reveals that MutSel models are inherently unable to describe accurately protein evolution under a regime of positive diversifying selection, or when $dN/dS > 1$.

Moreover, our analyses incidentally have revealed certain biases inherent in the ML ω inference approach. (...)

Methods

Sequence simulation

We simulated protein-coding sequences as a continuous-time Markov process [10] according to the MutSel model proposed by [7]. This model's instantaneous rate matrix Q is given by

$$Q_{ij} = \begin{cases} f_{ij}\mu_{ij}\kappa & \text{single nucleotide transition} \\ f_{ij}\mu_{ij} & \text{single nucleotide transversion} \\ 0 & \text{multiple nucleotide changes} \end{cases}, \quad (1)$$

where μ_{ij} is the symmetric nucleotide mutation rate and f_{ij}, the fixation probability from codon i to j, is defined as

$$f_{ij} = ln\left(\frac{\pi_j\mu_{ij}}{\pi_i\mu_{ji}}\right) \Big/ \left(1 - \frac{\pi_i\mu_{ji}}{\pi_j\mu_{ij}}\right), \quad (2)$$

where π_i is the equilibrium frequency of codon i.

We simulated protein-coding sequences along a 10-taxon phylogeny, with all branch lengths equal to 0.01, beginning with a root sequence selected using steady-state codon frequencies. All simulations featured a symmetric mutation rate $\mu_{xy} = 10^{-6}$, and a value for for κ was drawn from $\mathcal{U} \sim (1, 6)$, resulting in a fully symmetric mutation matrix. Unless otherwise stated, all simulated alignments contained 500,000 codon positions. A single evolutionary model was applied to all positions in the simulated sequences. While this lack of site-wise heterogeneity is unrealistic for real sequence evolution, it allows us to verify our derived relationship between selection coefficients and dN/dS with a sufficiently sized data set.

Assigning scaled selection coefficients

We simulated 100 alignments under the assumption that all synonymous codons have equal fitness (e.g. no codon bias), and a second set of 100 alignments which incorporated codon bias. For both frameworks, we generated amino acid scaled selection coefficients, S_a, for each simulation, by fixing one coefficient to 0 and drawing the remaining 19 values from a normal distribution $(N) \sim (0, x)$, where $x \sim U(1, 2)$.

For simulations without codon bias, we assigned these amino acid selection coefficients to codons such that all synonymous codons had the same scaled selection coefficient. For simulations with codon bias, we randomly selected a preferred codon for each amino acid. We then assigned the preferred codon a selection coefficient of $S_a + x$, where $x = ln(2)$, and we assigned non-preferred codons selection coefficients of $S_a - x/(k-1)$, where k is the total number of synonymous codons for the given amino acid. In this way, the average selection coefficient for this amino acid remained unchanged, whereas selective strength was partitioned such that a single codon was most selectively favored.

dN/dS Inference

We calculated a global dN/dS for each alignment using the mathematical framework outlined in (3)–(8) as well using standard maximum likelihood methods. Specifically, inferred dN/dS using the M0 mechanistic codon model [11], as implemented in the HyPhy batch language [12]. We used the GY94 instantaneous rate matrix [4,6], which includes the primary parameters ω, κ, and equilibrium codon frequencies. As different κ and equilibrium codon frequency parameterizations can change ω estimates [10, 13, 14], we inferred ω under a variety of model parameterizations, including three κ parameterizations (κ fixed to 1, κ as a free parameter, and κ fixed to its true value), and four codon frequency specifications (equal codon frequencies, F3x4 codon frequencies [5], CF3x4 codon frequencies [15] and empirical codon frequencies), ultimately resulting in 12 ω MLEs per simulated alignment. Note that we additionally verified that the system was evolving under a state-state process by verifying that the true, simulated codon frequencies were the same as the empirical

codon frequencies calculated from the simulated alignment. All code used is freely available at **github**.

Results

Mathematical relationship between selection coefficients and omega

We describe here how to calculate dN/dS from the parameters of a MutSel model. Under the assumption that the mutational process is symmetric, e.g. $\mu_{xy} = \mu_{yx}$ for all nucleotide pairs xy, we can write the steady-state frequency of codon i as

$$\pi_i = \frac{e^{S_i}}{\sum_k e^{S_k}}, \tag{3}$$

where the sum in the denominator runs over all 61 sense codons [16]. Here, S_i is the scaled selection coefficient for codon i; larger S_i values correspond to higher frequencies of codon i.

The fixation probability for a mutation from codon i to codon j is [7, 16]

$$f_{ij} = \frac{1 - (\pi_i/\pi_j)^{1/N_e}}{1 - \pi_i/\pi_j} \approx \frac{1}{N_e} \frac{\ln \pi_j - \ln \pi_i}{1 - \pi_i/\pi_j}\,, \tag{4}$$

where N_e is the effective population size. Using this framework, we can calculate an evolutionary rate by summing over all substitution probabilities weighted by the frequency of the originating codon. Further, we can establish specific expressions for nonsynonymous and synonymous evolutionary rates, and then divide them in order to obtain a value for the evolutionary rate ratio dN/dS.

To begin, we can write the nonsynonymous rate K_N as

$$K_\mathrm{N} = N_e \sum_i \sum_{j \in \mathcal{N}_i} \pi_i f_{ij} \mu_{ij}\,, \tag{5}$$

where $\mathcal{N}_i$ is the set of codons that are nonsynonymous to codon i and differ from it by one nucleotide. To normalize K_N, we divide it by the number of nonsynonymous sites, which we calculate according to the mutational opportunity definition of a site [4, 10] as

$$L_\mathrm{N} = \sum_i \sum_{j \in \mathcal{N}_i} \pi_i \mu_{ij}\,, \tag{6}$$

and thus we find that

$$dN = \frac{K_\mathrm{N}}{L_\mathrm{N}} = \frac{N_e \sum_i \sum_{j \in \mathcal{N}_i} \pi_i f_{ij} \mu_{ij}}{\sum_i \sum_{j \in \mathcal{N}_i} \pi_i \mu_{ij}}\,. \tag{7}$$

Similarly, for dS, the synonymous evolutionary rate K_S per synonymous site L_S, we find

$$dS = \frac{K_\mathrm{S}}{L_\mathrm{S}} = \frac{N_e \sum_i \sum_{j \in \mathcal{S}_i} \pi_i f_{ij} \mu_{ij}}{\sum_i \sum_{j \in \mathcal{S}_i} \pi_i \mu_{ij}}\,, \tag{8}$$

where $\mathcal{S}_i$ is the set of codons that are synonymous to codon i and differ from it by one nucleotide substitution. The quantities K_S and L_S are defined as in Eqs. (5) and (6) but summing over $j \in \mathcal{S}_i$ instead of $j \in \mathcal{N}_i$.

Equations (3)–(8) establish a connection between the scaled selection coefficients and the evolutionary rate ratio dN/dS. Moreover, we note that, if we assume that all synonymous codons have equal fitness (e.g. synonymous mutations are neutral), the synonymous fixation rate $f_i j = 1/N_e$ [17]. Under this circumstance, the value for dS would reduce to 1.

dN/dS can be accurately predicted from scaled selection coefficients

To validate our derived relationship between scaled selection coefficients and dN/dS, we simulated protein-coding sequences along a 10-taxon phylogeny according to a mutation-selection model framework [7, 16]. We simulated 100 alignments which gave equal fitness values to synonymous codons, and 100 alignments which incorporated codon bias (see Method for details). All simulations assumed a symmetric nucleotide rate matrix. For each alignment, we calculated dN/dS using equations (3)–(8) as well as using the M0 mechanistic codon model [4,6,11], as implemented in the HyPhy batch language [12].

The relationship between dN/dS calculations is shown in Figure 1A (for simulations with no codon bias) and Figure 1B (for simulations with codon bias), and it is clear that dN/dS values derived using selection coefficients agree nearly perfectly with those inferred using standard maximum likelihood methods. Fitness differences among synonymous codons do not influence this robust relationship. Additionally, in Figure 1C, we demonstrate convergence of dN/dS estimates as the size of the data set, represented by simulated alignment length,increases. Taken together, these results demonstrate that MutSel model parameters fully encapsulate information regarding dN/dS, and that the results from MutSel and mechanistic codon models are in complete agreement. As the results the datasets simulated with and without codon bias are largely the same, we focus the rest of our results on simulations without codon bias.

Moreover, as seen in Figure 1A-B, dN/dS values are always less than 1, reflecting a universal regime of purifying selection. In fact, **in SuppMat,** we prove that, when calculated using scaled selection coefficients, dN/dS is necessarily always less than or equal to 1, although this proof holds only when all synonymous codons have the same fitness value.

Influence of ML model parameterizations

The maximum likelihood dN/dS estimates (the model's ω parameter) reported in the previous subsection were obtained by fixing κ parameter in the GY94 rate matrix to its true simulated value, and specifying equal codon frequencies. However, as different model parameterizations can influence the resulting ω MLE [10, 13, 14], we inferred ω according to a total of 12 distinct model parameterizations, including three parameterizations for κ (fixed to its true value, fixed to 1, or a free parameter of the model) as well as four equilibrium codon frequency specifications (equal codon frequencies, frequencies calculated using either the F3x4 [5] or the CF3x4 [15] estimators, or empirical codon frequencies as taken from the simulated alignment). Table 1 shows how the ω values estimated according to these different ML parameterizations relate to the dN/dS values as calculated using equations (3)–(8), specifically for the simulations without codon bias (Table S1 gives the equivalent results for simulations with codon bias).

Results in Table 1 yield several important insights into the behavior of mechanistic codon models. First, it is clear that these models only estimate dN/dS accurately when equilibrium codon frequencies are set as equal (i.e., each codon has a frequency of 1/61). Second, the F3x4 and CF3x4 frequency estimators perform nearly identically to one another (p=0.54), yielding ω estimates with weak, negative correlations with the dN/dS derived from selection coefficients. Finally, when empirical codon frequencies are used, ω estimates have a moderately strong, negative correlation with derived dN/dS values. In other words, as the ML codon frequency parameters were more and more tailored to the given data set, ω MLEs decreased in accuracy. In fact, the negative correlations observed for the latter three frequency specifications actually reflect strongly overestimated ω MLEs; indeed, when empirical frequencies were specified (along with κ fixed to its true value), ω estimates were universally above 1. Figure 2 displays regressions between derived

and ML inferred ω values across the equal, F3x4, and empirical codon frequency specifications with κ fixed to its true value (dN/dS regression plots for all ML parameterizations are shown for simulations without codon bias in Figure S1 and for simulations with codon bias in Figure S2).

However, it is interesting to note that, while ML yields the most accurate κ estimates when equal codon frequencies are specified, all frequency specifications yield κ estimates which correlate fairly well with the true, simulated κ values. Thus, the ML model's ability to accurately estimate κ appears much more robust to codon frequency specifications is than its ability to estimate ω. Regression plots displaying the relationship between true κ values and κ MLEs are shown in Figures S3 and S4, for simulations without and with codon bias, respectively.

Importantly, however, our simulated alignments contained relatively constrained codon frequency distributions, given that each position in a given alignment followed the same evolutionary model. Therefore, we examined whether these frequency constraints influenced the error in ω MLEs. For each alignment, we calculated the codon entropy,

$$H(i) = -\sum_i \pi_i \ln \pi_i \tag{9}$$

, where π_i is the frequency of codon i and the sum runs over all sense codons. Note that the maximum $H(i) = 4.11$ value is reached when all codons have a frequency of 1/61. In Figure 3, we show the relationship between each alignment's codon entropy and the error between selection coefficient-derived dN/dS values and ω MLE values, when inferred across equal, F3x4, and empirical frequency specifications. Here, codon entropy does not independently affect the ω MLE error (p=0.78), but rather the error stems primarily from the specific codon frequency parameterizations in the ML model. Again, these results carry over precisely to simulations with codon bias (Figure **??**)

Discussion

The oldest and most-widely used method to infer selection pressure in protein-coding genes calculates calculates the ratio of non-synonymous (dN) to synonymous (dS) substitution rates dN/dS to identify sites that experience negative selection ($dN/dS < 1$), sites that evolve neutrally ($dN/dS \approx 1$), and sites that experience positive diversifying selection ($dN/dS > 1$). By contrast, MutSel models estimate scaled selection coefficients, either for individual amino acids, [**?**, 7, 9, 18, 19], for codons [20], or for both. Thus, while mechanistic codon models describe the how quickly a protein's constituent amino acids change, MutSel models calculate the strength of natural selection operating on the specific amino-acid changes.

Until now, however, it has been an open question how these two modeling frameworks relate to one another. Some have argued that MutSel models, given their firm grounding in population genetics theory and specific attention to site-specific amino acid fitness differences, offer a more fine-grained approach to studying protein evolution than do mechanistic codon models [7, 18]. Recent phylogenetic studies have also demonstrated that evolutionary models which explicitly consider amino acid fitness values offer dramatic improvements over other models, including mechanistic codon models, suggesting that MutSel models may more accurately represent the process of coding-sequence evolution [21, 22].

Here, we have derived a formal mathematical relationship between the quantities dN/dS and scaled codon selection coefficients, the primary parameters of mechanistic codon and MutSel models, respectively. Through a simulation approach, we find that these two models are in full agreement, and that the value for dN/dS can be precisely calculated from selection coefficients alone.

Our results rest on the key assumptions that the protein sequence is evolving under steady-state, or equilibrium, conditions, and the nucleotide mutation rates are symmetric (e.g. $\mu_{xy} = \mu_{yx}$). The first assumption recapitulates the population genetics theory behind MutSel models, which assume that selection coefficients remain constant over the phylogeny, and therefore the protein is evolving along a static fitness landscape [7, 18, 19]. We make the latter assumption of symmetric mutation rates because it allows us to derive precise quantities for equilibrium codon frequencies from selection coefficients, according to theory relating statistical physics to evolutionary biology under steady-state conditions [16, 23]. These frequency calculations are non-trivial in cases of asymmetric mutation rates (e.g. mutational which generate nucleotide compositional bias), and therefore merit further study. Regardless, the mathematical relationship between selection coefficients and dN/dS in equations (4) - (8) should hold under any mutational framework.

The equivalency between selection coefficients and dN/dS values we have demonstrated is robust to differences in synonymous codon fitnesses. However, it is important to note that our implementation of codon bias explicitly assumed that frequency differences among synonymous codons resulted from fitness differences alone. In other words, the sole source of codon bias in our simulations was selection, not mutation. This implementation might not be entirely biologically realistic, as both mutational and selective forces likely contribute to codon bias in real genomes [24–27]. However, the key finding that we present is that fitness differences among synonymous codons do not affect the robust mathematical equivalency between scaled selection coefficients and dN/dS.

Incidentally, our study recovered that mechanistic codon models can produce strongly biased inferences when parameters are incorrectly specified. In particular, ω MLE values only corresponded to the true dN/dS value when the equilibrium codon frequency parameters were specified as equal (e.g. each codon had an equilibrium frequency of 1/61). Alternatively, the more common approaches of using empirical codon frequencies (also known as the F61 estimator [4]) or frequency estimators such as F3x4 [5] and CF3x4 [15] always yielded strongly dramatically inflated ω MLEs. We additionally demonstrated that the narrow distribution of codon frequencies in our alignments did not cause these errors (Figures 3 and S5), but rather the error stemmed specifically from the ML frequency parameterizations themselves.

We explain this phenomenon by recognizing that the rationale for including codon frequency parameters in mechanistic codon models is to account for unequal nucleotide frequencies specifically caused by mutational and not selective forces [10,13]. The proper values for these parameters, then, should be the codon frequencies which would exist *in the absence of natural selection*. This approach is the only way to ensure that ω is the sole model parameter which contains information about natural selection. Otherwise, the ω parameter will no longer represent the true dN/dS evolutionary rate ratio.

Moreover, frequency estimators such as F3x4 and CF3x4, which use positional nucleotide frequencies to calculate codon frequencies, must make the implicit assumption that observed unequal base frequencies result from biased mutation rates. This assumption, however, may not be fully justified. Indeed, our simulated alignments featured a wide array of nucleotide compositions, with GC-contents ranging from 0.22-0.79. Given that we simulated sequences according to a symmetric mutation matrix, all compositional biases in our data sets resulted entirely from natural selection favoring particular codons, not by any bias towards unequal base frequencies. Therefore, the proper equilibrium frequency parameterization for our alignments was indeed equal codon frequencies, which would be expected in the absence of natural selection and when mutation rates are symmetric.

These results emphasize that it is crucial to parameterize mechanistic codon models properly. Indeed, their primary parameter ω will only truly represent dN/dS when all other parameters are properly specified. If codon frequencies are not properly specified, which we suspect is the

case in most analyses, then the ω MLEs are virtually meaningless and do not represent selective pressure. Therefore, we contend that there is hardly ever a justification to specify empirical codon frequencies, also known as the F61 frequency estimator [4], as natural selection has clearly produced the observed frequencies. Unfortunately, the F61 frequencies are the default parameterization in the widely-used PAML software's codeml implementation [28], so we strongly recommend that users take great care when using this package. In addition, the only robust way to ensure that codon frequencies are properly specified is through experimentally calculating mutation rates. Luckily, this data already exists for a variety of taxa, including **mutation citations**. We recommend that, if experimental data is absent, users err on the side of caution and specify equal codon frequencies to reduce the possibility of false positives.

Finally, we contend that this methods presented in this paper reveal a promising future avenue for methodological benchmarking. Typically, researchers assess the performance of a given inference framework through simulations which adhere to the underlying model's assumptions. However, this strategy can only confirm that inference methods are behaving as expected; it cannot confirm that the underlying model accurately represents the evolutionary process. Instead, we suggest an alternate approach to benchmark inference methods, and indeed evolutionary models: assessing the extent to which distinct models agree may serve as a novel, robust strategy to determine the accuracy of different modeling frameworks.

Still working on placement

This relationship holds only if the protein evolves accordingly to a strictly steady-state process, otherwise known as purifying selection. Alternatively, under non-equilibrium conditions, (e.g. positive selection, when $\omega > 1$), MutSel models are inherently unable to describe protein evolution. These findings have important implications for when the use of each model is justified; if positive selection has occurred along the protein's evolutionary trajectory, MutSel models will likely yield spurious results. This proof, however, only holds when mutation rates are symmetric and when synonymous codons have equal fitnesses.

Moreover, results obtained when fitness differences among synonymous codons were considered raise an additional issue with dN/dS; even though all sequences were simulated according to a strictly steady-state process, these sequences frequently featured dN/dS values greater than 1. Indeed, our derivation of dN/dS from selection coefficients shows that, when fitness differences exist among synonymous codons, dN/dS need not be less than 1. In contrast, the most extreme case of codon bias, in which only a single codon per amino acid is selectively tolerated, the number of synonymous sites $L_S = 0$, and thus the value for dN/dS approaches infinity. This finding seems paradoxical to classic interpretations of $dN/dS > 1$. Typically, these values are viewed as hallmarks of positive or diversifying selection, which is assumed to occur when the protein sequence experiences strong selective pressure to change its constituent amino acids. Positive selection, therefore, necessarily implies that the protein does not evolve at equilibrium, but rather a shift in selective constraint caused new amino acids to be favored.

However, one must also recognize that the contention that $dN/dS > 1$ represents positive selection assumes that synonymous substitutions are selectively neutral, which is clearly not the case when codon bias exists. Therefore, it is entirely possible that estimates of positive selection as inferred via dN/dS estimates in species with high levels of codon bias, such as many bacterial, *Drosophila*, or certain mammalian species [25–27, 29], may not be true cases of positive selection, but rather simply signals of strong codon bias.

However, the calculations here necessitate that the protein is evolving according to a strictly equilibrium process, which seems paradoxical to having positive selection. In other words, even if

the protein is evolving according to a static fitness landscape, sufficient levels of codon bias may yield values for dN/dS which are well above 1. Thus, it seems as though what is classically termed positive selection can result simply from strong synonymous fitness differences.

References

[1] Li WH, Wu CI, Luo CC (1985) A new method for estimating synonymous and nonsynonymous rates of nucleotide substitution consider the relative likelihood of nucleotide and codon changes. Mol Biol Evol 2: 150–174.

[2] Nei M, Gojobori T (1986) Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. Mol Biol Evol 3: 418–426.

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

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

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

[6] 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.

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

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

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

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

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

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

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

[14] Zhang Z, Yu J (2006) Evaluation of six methods for estimating synonymous and nonsynonymous substitution rates. Geno Prot Bioinfo 4: 173–181.

[15] Kosakovsky Pond SL, Delport W, Muse SV, Scheffler K (2010) Correcting the bias of empirical frequency parameter estimators in codon models. PLoS One 5: e11230.

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

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

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

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

[20] 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.

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

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

[23] de Vladar HP, Barton NH (2011) The contribution of statistical physics to evolutionary biology. Trends Ecol Evol 26: 424 – 432.

[24] Blumer M (1991) The selection-mutation-drift theory of synonymous codon usage 129: 897–907.

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

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

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

[28] Yang Z (2007) PAML 4: Phylogenetic analysis by maximum likelihood. Molecular Biology and Evolution 24: 1586–1591.

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

Table 1: Effect of ML parameterizations on inference.

Codon frequencies	κ parameterization	dN/dS correlation	dN/dS error	κ correlation	κ error
Equal	True	1	0.008		
Equal	Free	0.996	0.023	0.913	0.106
Equal	1	0.916	0.195		
F3x4	True	-0.276	1.696		
F3x4	Free	-0.278	1.727	0.929	0.141
F3x4	1	-0.233	1.317		
CF3x4	True	-0.301	1.718		
CF3x4	Free	-0.301	1.747	0.932	0.136
CF3x4	1	-0.259	1.317		
Empirical	True	-0.648	10.085		
Empirical	Free	-0.656	10.288	0.804	0.227
Empirical	1	-0.629	7.992		

Codon frequency specifications were either set as equal (1/61 per codon), calculated from the F3x4 estimator [5], calculated from the CF3x4 estimator [15], or set equal to the simulated alignment's empirical frequencies. κ was specified as either a fixed value, its true simulated value or 1, or as a free parameter of the model. Correlations given are between the ML ω estimate and our derived ω values. Error refers to the mean absolute error between these two ω estimates. Similar values for κ are shown for those inferences where κ was a free parameter of the model. Note that all is significant.

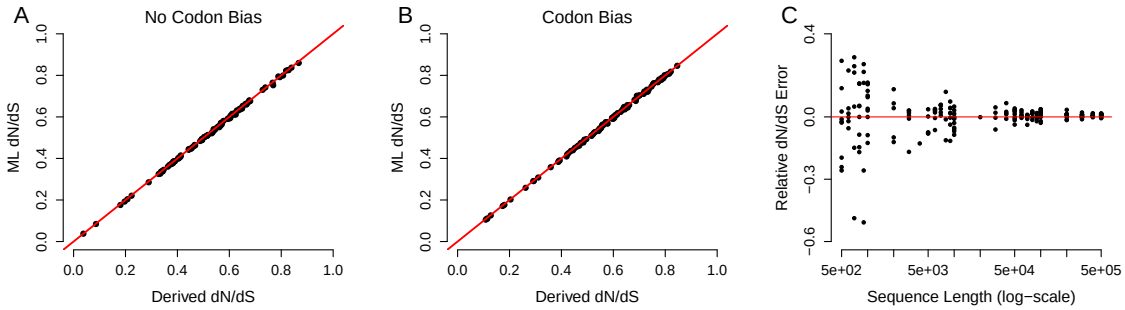

Figure 1: Relationship works exceedingly well. Left panel shows 100 points, each of which corresponds to single simulation. Note that here the ml inference is shown for equal codon frequency specs and kappa fixed to true value (a similar plot for free kappa is shown in suppfigs, but results are qualitatively identical.) Right panels shows convergence of omega values as data set size (represented as simulated alignment length) increases. The y-axis indicates relative error of the ML dN/dS estimates, and the x-axis indicates sequence length on a log-scale. As the sequence length, or the data set size, increases, the two dN/dS estimates converge to the same value.

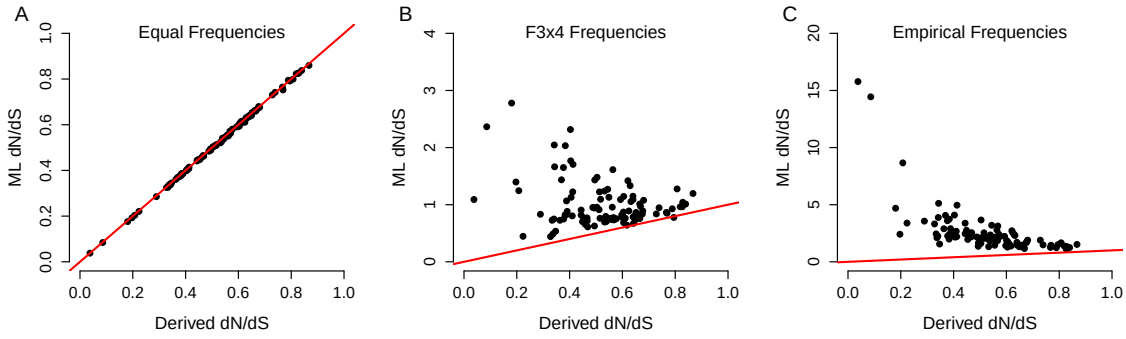

Figure 2: Issues with frequency specifications abound. In each plot, red line indicates 1:1 agreement, so note the y-axis differences. Relationship between omega values only really exists when equal codon frequencies are specified. When f3x4 or true freqs used, there is the potential to end up with dramatically inflated values. cf3x4 not shown because its results are statistically the same as f3x4.

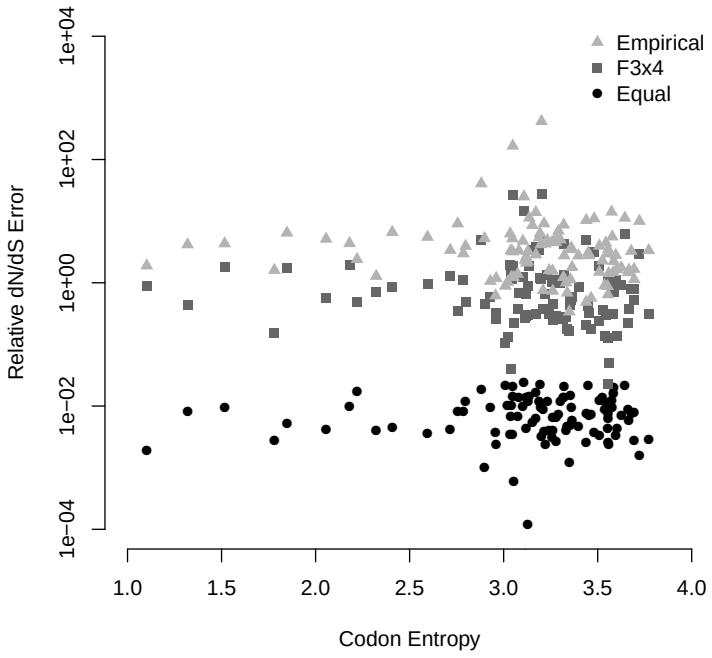

Figure 3: Entropy has no independent effect here. Xaxis is data entropy, Yaxis is logscale absolute value of relative error.

Supplementary Figures and Tables

Codon frequencies	κ parameterization	dN/dS correlation	dN/dS error	κ correlation	κ error
Equal	True	1	0.008		
Equal	Free	0.998	0.019	0.93	0.098
Equal	1	0.928	0.202		
F3x4	True	-0.053	1.026		
F3x4	Free	-0.036	1.03	0.878	0.154
F3x4	1	0.067	0.724		
CF3x4	True	-0.055	1.02		
CF3x4	Free	-0.034	1.016	0.87	0.157
CF3x4	1	0.056	0.723		
Empirical	True	-0.566	4.735		
Empirical	Free	-0.59	4.728	0.826	0.206
Empirical	1	-0.496	3.126		

Results from runs with codon bias. Codon frequency specifications were either set as equal (1/61 per codon), calculated from the F3x4 estimator [5], calculated from the CF3x4 estimator [15], or set equal to the simulated alignment's empirical frequencies. κ was specified as either a fixed value, its true simulated value or 1, or as a free parameter of the model. Correlations given are between the ML ω estimate and our derived ω values. Error refers to the mean absolute error between these two ω estimates. Similar values for κ are shown for those inferences where κ was a free parameter of the model. Note that all is significant.

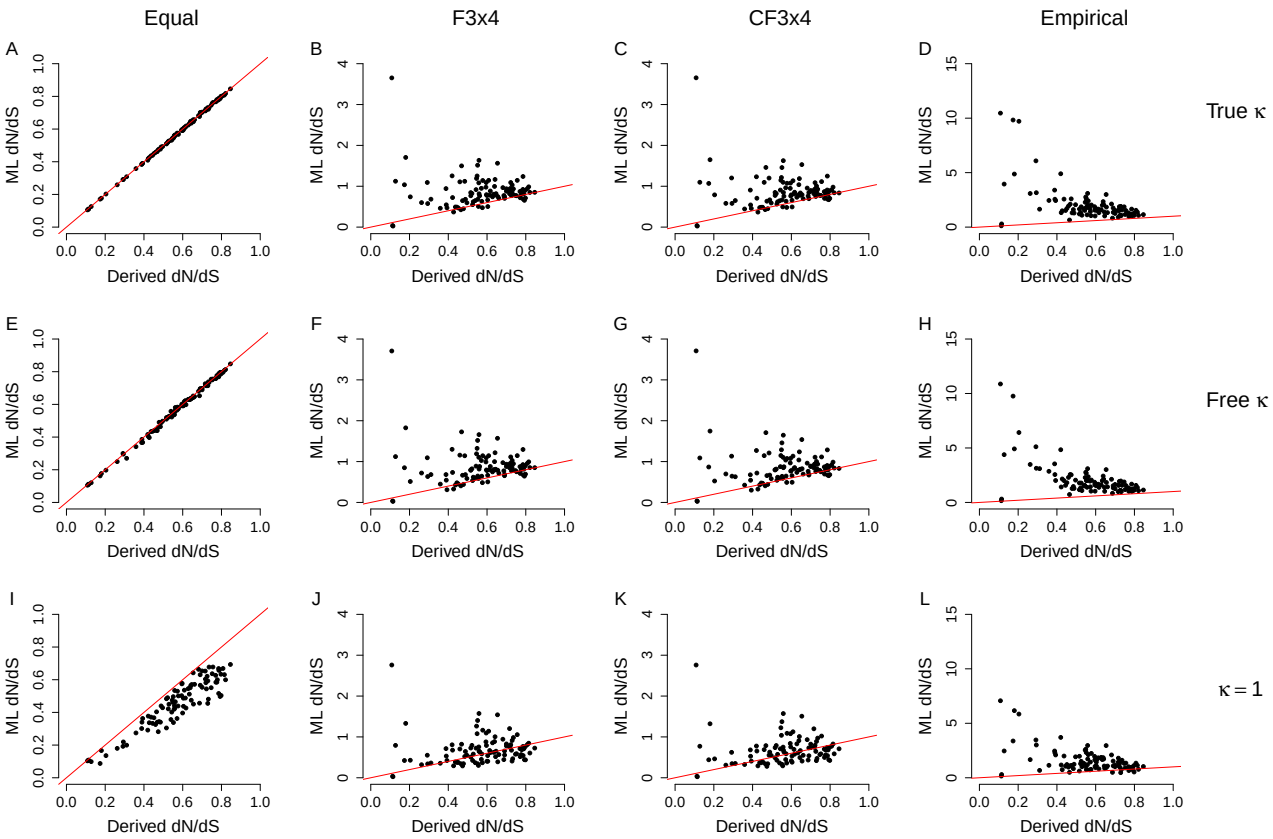

Fig. S1 Omega regression for all ML parameterizations, without codon bias.

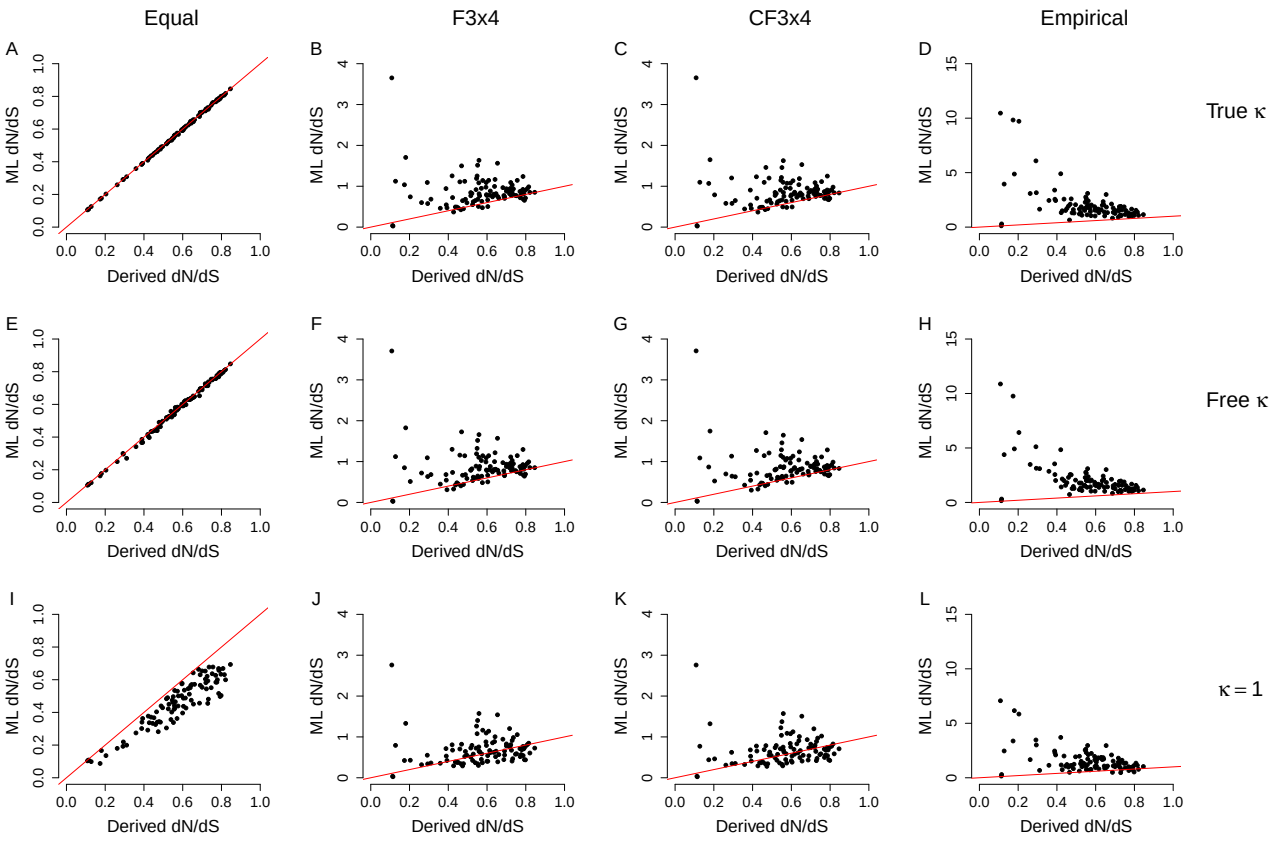

Fig. S2 Omega regression for all ML parameterizations, with codon bias.

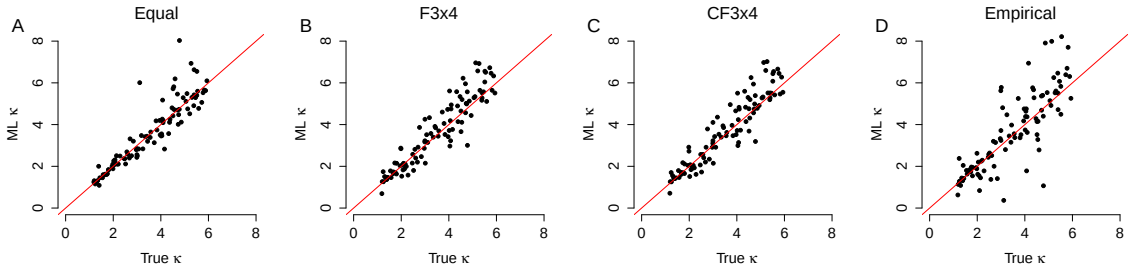

Fig. S3 Kappa regression for all ML freqspec parameterizations where kappa is a free parameter, without codon bias.

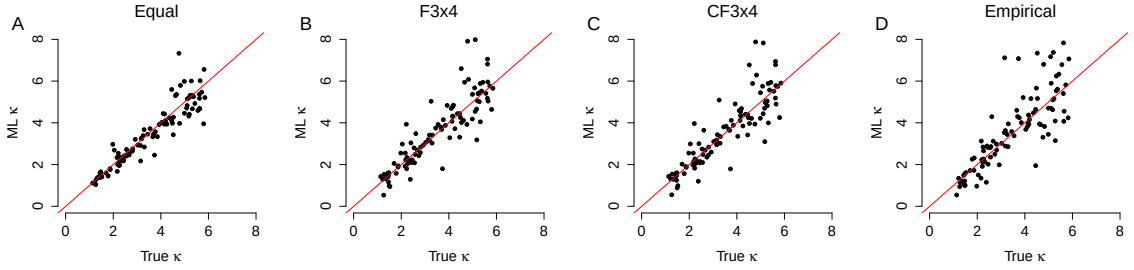

Fig. S4 Kappa regression for all ML freqspec parameterizations where kappa is a free parameter, without codon bias.

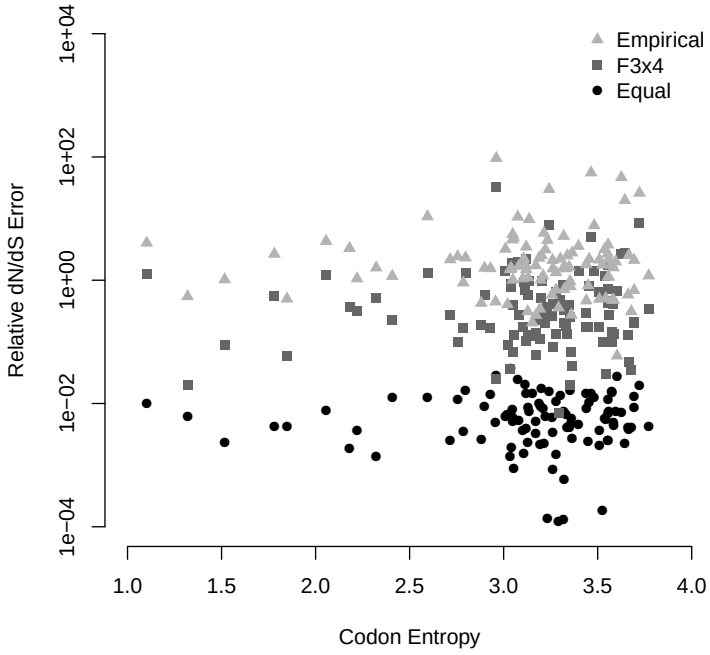

Fig. S5 Entropy and error for the codon bias data set.