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 inference method of choice since their introduction in the 1990s [4, 5]. 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, has 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 [6], they have seen virtually no use due to their high computational expense. However, recently, several computationally tractable model implementations have emerged [7,8], 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 aim to 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 and omega inference

We simulated protein-coding sequences as a continuous-time Markov process [9] according to the MutSel model proposed by [6]. 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.

For each simulation, we generated 20 scaled selection coefficients, S_a, by fixing one coefficient to 0 and drawing the remaining 19 values from a normal distribution $(N) \sim (0, x)$, where $x \sim U(1, 2)$. We assigned these coefficients to codons such that all synonymous codons had the same scaled selection coefficient. Following the theory developed by [10], we determined steady-state codon frequencies π_i according to

$$\pi_a = \frac{e^{S_i}}{\sum_k e^{S_k}}, \quad (3)$$

where S_i is the scaled selection coefficient for codon i and the denominator sums over all 61 sense codons.

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. Unless otherwise stated, we simulated alignments of 500,000 codon positions with a global, symmetric mutation rate $\mu_{xy} = 10^{-6}$, and a value for for κ was drawn from $\mathcal{U} \sim (1, 6)$. 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 ω with a sufficiently sized data set.

For each simulated alignment, we inferred ω in two main ways; first, we calculated ω using the mathematical framework described in (4)–(8), and second, we inferred ω used the standard maximum likelihood M0 model [11], which uses the GY94 rate matrix [4], as implemented in HyPhy [12]. The GY94 matrix includes the primary parameters ω, κ, and equilibrium codon frequencies. As different κ and equilibrium codon frequency parameterizations can change ω estimates [9, 13], 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, simulated value), and four codon frequency specifications (equal codon frequencies, F3x4 codon frequencies [5], CF3x4 codon frequencies [14] and empirical codon frequencies [4]). These different specifications yielded twelve maximum likelihood ω inferences 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. We assume the following: (i) the mutational process is symmetric, such that $\mu_{xy} = \mu_{yx}$ for all nucleotide pairs xy; (ii) all synonymous codons for a given amino acid have the same fitness (and therefore the same scaled selection coefficient); there is no synonymous rate variation or codon bias.

In the framework of a MutSel model, we can write the steady-state frequency of codon i as

$$\pi_i = \frac{e^{S_i}}{\sum_k e^{S_k}}, \tag{4}$$

where the sum in the denominator runs over all 61 sense codons [10]. 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 [10]

$$f_{ij} = \frac{1-(\pi_i/\pi_j)^{1/N_e}}{1-\pi_i/\pi_j} \approx \frac{1}{N_e}\frac{\ln \pi_j - \ln \pi_i}{1-\pi_i/\pi_j}\,, \tag{5}$$

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

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

where $\mathcal{N}_i$ is the set of codons that are nonsynonymous to codon i and differ from it by one nucleotide. To normalize K_N, we divide it by the number of nonsynonymous sites, which we calculate according to the mutational opportunity definition of a site [4, 9] as

$$L_\mathrm{N} = \sum_i \sum_{j\in\mathcal{N}_i} \pi_i \mu_{ij}, \tag{7}$$

and thus we find that

$$dN = \frac{K_\mathrm{N}}{L_\mathrm{N}} = \frac{N_e \sum_i \sum_{j\in\mathcal{N}_i} \pi_i f_{ij} \mu_{ij}}{\sum_i \sum_{j\in\mathcal{N}_i} \pi_i \mu_{ij}}. \tag{8}$$

We can similarly calculate the synonymous evolutionary rate K_S as

$$K_\mathrm{S} = N_e \sum_i \sum_{j\in\mathcal{S}_i} \pi_i f_{ij} \mu_{ij}, \tag{9}$$

where $\mathcal{S}_i$ is the set of codons that are synonymous to codon i and differ from it by one nucleotide. Under the assumption that all synonymous codons have equal fitness, and hence all synonymous mutations are selectively neutral, we have $f_{ij} = 1/N_e$ [15], yielding

$$K_\mathrm{S} = \sum_i \sum_{j\in\mathcal{S}_i} \pi_i \mu_{ij}. \tag{10}$$

We can subsequently normalize K_S by the number of synonymous sites

$$L_\mathrm{S} = \sum_i \sum_{j\in\mathcal{S}_i} \pi_i \mu_{ij}, \tag{11}$$

thus revealing that the synonymous rate per synonymous site dS is equal to 1. Equations (4)–(11) establish a connection between the scaled selection coefficients and the evolutionary rate ratio dN/dS.

Scaled selection coefficients fully encapsulate ω

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 [6, 10]. For these simulations, we assume no codon bias (i.e. all synonymous codons share the same selection coefficient) and symmetric nucleotide mutation rates. We calculated an dN/dS for each simulated alignment using the derivations given in equations (4)–(11) as well as using a standard maximum likelihood (ML) approach, according to the GY94 [4] rate matrix, as implemented in the HyPhy batch language [12].

As shown in Figure 1A, dN/dS values derived using selection coefficients agree nearly perfectly with those inferred using standard maximum likelihood methods. We additionally demonstrate convergence of these values with increasing amounts of data, represented by simulated alignment length (Figure 1B). Taken together, these results clearly show that MutSel model parameters fully encapsulate information regarding dN/dS, and that the results from MutSel and ω models are in complete agreement.

Moreover, as seen in Figure 1A, 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. This important insight reveals that, while MutSel and mechanistic codon models fully agree, their relationship only holds under conditions of purifying selection or neutral evolution. MutSel models, therefore, are inherently unable to describe protein evolution under positive, diversifying selection ($dN/dS > 1$).

Influence of ML model parameterizations

MLE ω values 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 [9,13,16], we inferred ω according to a total of 12 distinct model parameterizations; we inferred ω when κ was either a free parameter of the model, fixed to its true value, or fixed to 1, and we examined different equilibrium codon frequency specifications, including equal codon frequencies, frequencies calculated using either the F3x4 [5] or the CF3x4 [14] 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 (4)–(11).

Results in Table 1 yield several important insights into the behavior of mechanistic codon models. First, it is clear that ML methods 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 (regression plots for all ML parameterizations are shown in Figures S1 and 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.

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,

$$H(i) = -\sum_i \pi_i \ln \pi_i \tag{12}$$

, 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.

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 [**?**, 6, 8, 17, 18], for codons [19], 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. Here, we have derived a formal mathematical relationship between the parameter estimates of ω and MutSel models, and through a simulation approach, we find that these two models are in full agreement. Importantly, 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.

Moreover, we found that the widely-used maximum likelihood approach to estimate ω was only accurate under certain model parameterizations. In particular, ω MLE values were only correct when equilibrium codon frequency parameters were specified as equal (e.g. each codon had an equilibrium frequency of 1/61), regardless of the true codon frequency distribution, while using common frequency estimators such as F3x4 [5] and CF3x4 [14] or empirical codon frequencies (also known as the F61 estimator) always yielded strongly inflated ω estimates.

We explain this phenomenon by recognizing that the rationale for including codon frequency parameters in mechanistic codon substitution models is to capture biases in the underlying nucleotide base frequencies [9, 13]. It is further assumed that any nucleotide composition bias results strictly from mutational processes, and not from natural selection. 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, these compositional biases resulted entirely from natural selection favoring particular codons, not by any bias towards unequal base frequencies. Therefore, the proper equilibrium frequency specification for our alignments was indeed equal codon frequencies, which would result from a symmetric mutation scheme in the absence of natural selection. Unfortunately, however, even the commonly-used F3x4 and CF3x4 frequency estimators were unable to properly specify codon frequencies, resulting in elevated ω values. Thus, it is critical to parameterize mechanistic codon models properly in order to ensure that the ω parameter is the sole parameter measuring the strength of natural selection. Otherwise, other model parameters may contain information about the selection process, and the resulting ω MLE will not accurately represent the true dN/dS value.

This study highlights the importance of examining the While some may prefer one model over another, it is important to have a full understanding of why one model might be applied over another. Our results demonstrate that, for circumstances of purifying selection or neutral, the models are robust and in agreement. However, be careful, because sometimes they don't agree, and this is equally important.

Methods which characterize evolutionary dynamics of protein-coding sequences are among the most widely-used tools in comparative sequence analysis, with applications ranging from identifying key functional protein residues to predicting the evolutionary trajectories and virulence of disease. Traditional codon substitution models used to study protein evolution assess evolutionary rate, or how quickly a protein's constituent amino acids change over phylogenetic time-scales.

Several studies have pointed out that ω models suffer from

However, these models overlook a key aspect of protein evolutionary dynamics: natural selection favors distinct, site-specific distributions of amino acids across positions in proteins. To address this gap, a class of computational models, known as mutation-selection models, have appeared in the literature. These models explicitly account for the dynamic interplay between the mutational and selective forces shaping protein evolution, and in doing so prioritize site-specific amino acid propensities. Although mutation-selection models were first proposed over 15 years ago, it is only within the past year that our computational power has reached a point where such models are tractable. Therefore, there exists critical need for the development of tools which can assess the validity of and test hypotheses regarding these models.

Many have pointed out that ω models suffer from certain

However, while widely-applied, codon-substitution models have several serious drawbacks.

Results presented here rely on two main assumptions: there is no codon bias (e.g. synonymous codons have equal fitness), and underlying mutation rates are symmetrical. In principle, the mathematical equivalency between scaled selection coefficients and dN/dS values should hold regardless of either of these phenomena. The main caveat we note is that our proof that this derived $dN/dS < 1$ may not hold under cases of asymmetric mutations rates, e.g. where $\mu_{AT} \neq \mu_{TA}$.

Therefore, it is possible that, if these conditions were violated, dN/dS as calculated from selection coefficients might indeed be greater than 1. While these are certainly simplifying assumptions, we contend that ...

This study moreover reveals 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.

Here, by deriving an explicit relationship between dN/dS and selection coefficients, we incidentally uncovered some biases in mechanistic codon model inference.

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

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

[8] 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.

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

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

[11] 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.

[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] 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.

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

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

[17] 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.

[18] 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.

[19] 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.

Table 1: Effect of ML parameterizations on inference.

Codon frequencies	κ parameterization	ω correlation	ω error	κ correlation	κ error
Equal	True	0.999	0.008		
Equal	1	0.916	0.195		
Equal	Free	0.996	0.023	0.913	0.106
F3x4	True	-0.276	1.696		
F3x4	1	-0.233	1.317		
F3x4	Free	-0.278	1.727	0.929	0.141
CF3x4	True	-0.301	1.718		
CF3x4	1	-0.259	1.317		
CF3x4	Free	-0.301	1.747	0.932	0.136
Empirical	True	-0.648	10.09		
Empirical	1	-0.629	7.992		
Empirical	Free	-0.656	10.29	0.804	0.227

Codon frequency specifications were either set as equal (1/61 per codon), calculated from the F3x4 estimator [5], calculated from the CF3x4 estimator [14], 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.

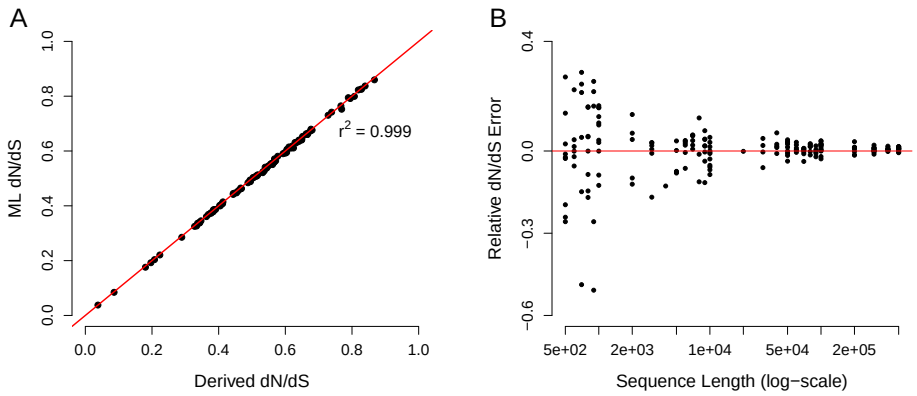

Figure 1: Relationship works exceedingly well. Left panel shows 100 points, each of which corresponds to single simulation. Note that here the ml inference is shown for equal codon frequency specs and kappa fixed to true value (a similar plot for free kappa is shown in suppfigs, but results are qualitatively 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.

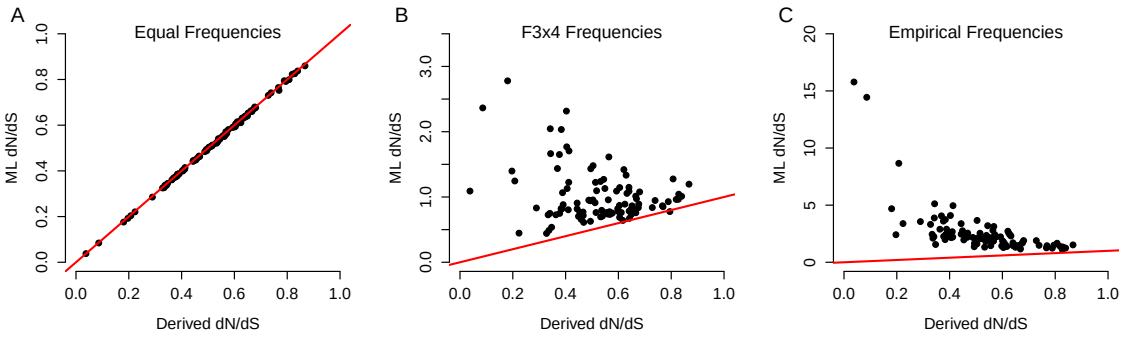

Figure 2: Issues with frequency specifications abound. In each plot, red line indicates 1:1 agreement, so note the y-axis differences. Relationship between omega values only really exists when equal codon frequencies are specified. When f3x4 or true freqs used, there is the potential to end up with dramatically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

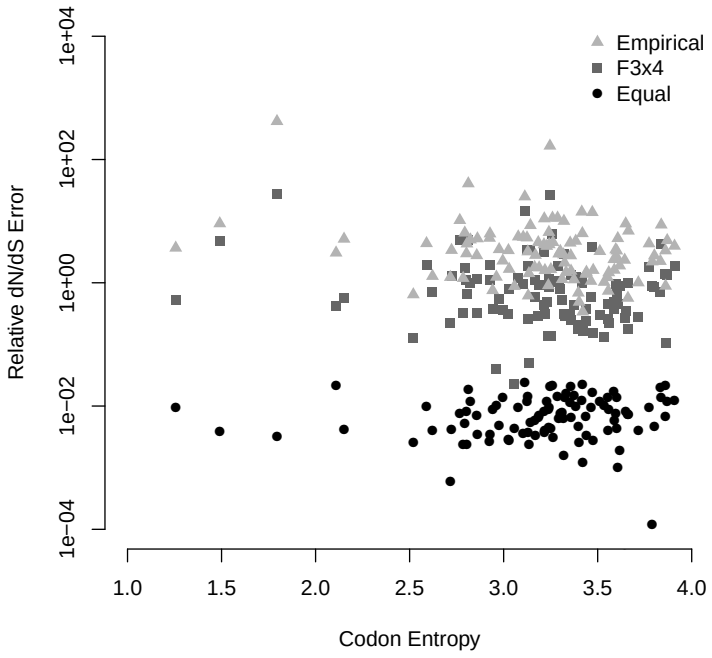

Figure 3: Entropy has no independent effect here. Xaxis is data entropy, Yaxis is logscale absolute value of relative error.

Supplementary Figures

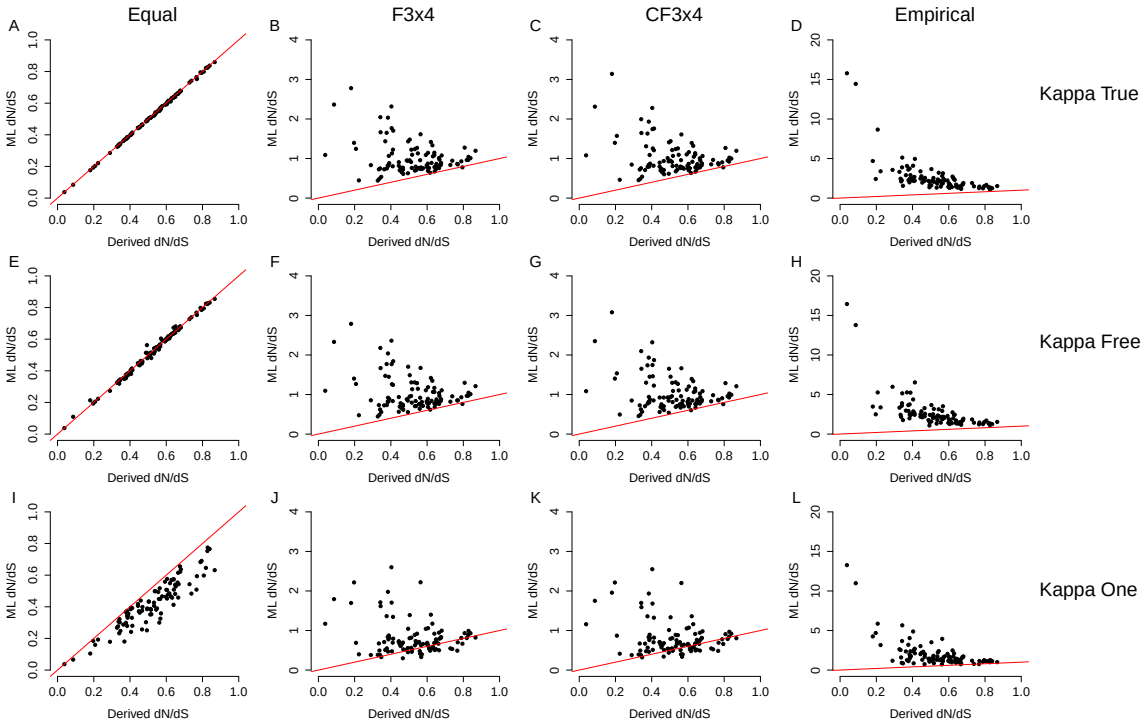

Fig. S1 Omega regression for all ML parameterizations.

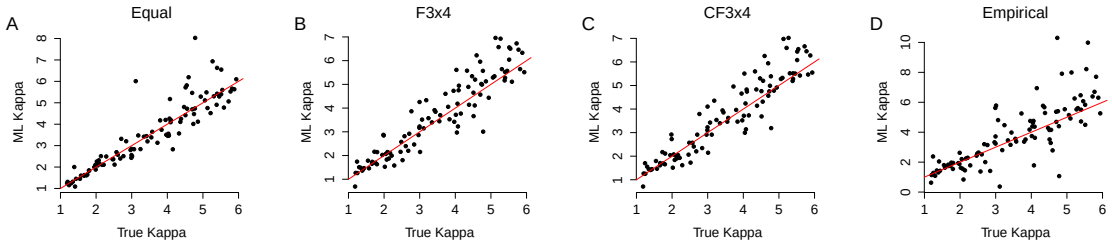

Fig. S2 Kappa regression for all ML freqspec parameterizations where kappa is a free parameter.