Introduction

Over the years, a variety of models have been proposed to describe the effects of natural selection on protein-coding sequences. Traditionally, the focus has been on estimating the evolutionary rate ratio, $\omega = dN/dS$, the rate of nonsynonymous to synonymous substitutions, which indicates how quickly a protein's constituent amino acids change. Following early counting methods for estimating dN/dS (e.g. refs [1] and [2]), mechanistic codon substitution models (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 so-called ω models have seen great success in the field of molecular evolution and are widely used to examine the strength of selection pressure in protein-coding sequences.

More recently, a second class of models, known as mutation-selection-balance (MutSel) models, has emerged as a popular alternative to ω models. MutSel models explicitly model the dynamic balance between mutation and selection, lending a potentially more precise and realistic description of the evolutionary process than do ω models, which merely model the final outcome (e.g. substitution) of the underlying mutation-selection interplay [6–9]. Unlike ω models, MutSel models yield estimates of amino acid selection coefficients, which indicate the extent to which natural selection favors, or disfavors, particular amino acids at protein positions. These selection coefficients, which can in turn be scaled relative to a focal amino acid, the primary parameters of interest that MutSel models produce. 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 [10, 11], allowing for the first time the potential for widespread use.

Although both ω 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, ω and scaled amino acid 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 ω values from selection coefficients alone. Using a simulation approach, we verify that ω values estimated using selection coefficients alone correspond precisely to those inferred using the standard maximum likelihood (ML)ω inference approach. 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 $dN/dS > 1$. This result has important implications for circumstances under which MutSel model use is justified.

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 [12] according to the MutSel model proposed by [6]. This model's instantaneous rate matrix $Q = q_{ij}$, which describes the probability of substitution from codon i to codon j, is given by

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

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

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

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 $(N) \sim (0, \sigma)$, where $\sigma\ U(0.5, 1.5)$. We assigned these coefficients to codons such that all synonymous codons had the same scaled selection coefficient. Following the theory developed by [13], we determined steady-state codon frequencies π_i according to

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

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 $\mathcal{U} \sim (1, 5.5)$. A single evolutionary model was applied to all positions in the simulated sequences, meaning that we did not incorporate any site-wise variation into the evolutionary process. 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.

Note that we additionally verified that the system was evolving under a state-state process by comparing the true codon frequencies with empirical codon frequencies calculated from the simulated alignment. The two sets of frequencies were virtually identical.

For each simulated alignment, we inferred ω in two main ways; first, we calculated ω using the mathematical framework described in (**??**)–(8), and second, we inferred ω used the standard maximum likelihood M0 model [14], which uses the GY94 rate matrix [4], as implemented in HyPhy [15]. The GY94 matrix includes the primary parameters ω, κ, and equilibrium codon frequencies. As it is well-known that the manner of specification for the latter two parameters strongly influences ω estimates [12, 16], we inferred ω under a variety of model parameterizations, including three κ parameterizations (κ fixed to 1, κ fixed to its true value, and κ as a free parameter), and four codon frequency specifications (equal codon frequencies, F3x4 codon frequencies [5], CF3x4 codon frequencies [17] and empirical, or F61, codon frequencies [4]). These different specifications yielded twelve maximum likelihood ω inferences per 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 amino acids [13]. 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 [6, 13]

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

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

For example, we can write the synonymous rate $K_{\rm S}$ as

$$K_{\rm S} = N_e \sum_i \sum_{j\in\mathcal{S}_i} \pi_i u_{ij} \mu_{ij}\,, \tag{6}$$

where $\mathcal{S}_i$ is the set of codons that are synonymous to codon i and differ from it by one nucleotide substitution. To normalize $K_{\rm S}$, we divide it by the number of synonymous sites, which we calculate according to the mutational opportunity definition of a site [4, 12] as

$$L_{\rm S} = \sum_i \sum_{j\in\mathcal{S}_i} \pi_i \mu_{ij},\,. \tag{7}$$

Under the assumption that all synonymous codons have equal fitness, and hence all synonymous mutations are selectively neutral, we have $u_{ij} = 1/N_e$ [18], and thus we find that dS, the synonymous rate per synonymous site, is equal to 1.

Similarly, we can derive an expression for dN, the non-synonymous rate per non-synonymous site, and we find

$$dN = \frac{K_{\rm N}}{L_{\rm N}} = \frac{N_e \sum_i \sum_{j\in\mathcal{N}_i} \pi_i u_{ij} \mu_{ij}}{\sum_i \sum_{j\in\mathcal{N}_i} \pi_i \mu_{ij}}, \tag{8}$$

where $\mathcal{N}_i$ is the set of codons that are nonsynonymous to codon i and differ from it by one nucleotide substitution. The quantities $K_{\rm N}$ and $L_{\rm N}$ are defined as in Eqs. (6) and (7) but summing over $j \in \mathcal{N}_i$ instead of $j \in \mathcal{S}_i$. Equations (**??**)–(8) establish a connection between the scaled amino acid selection coefficients s_a (i.e., the primary parameters of a MutSel model) and the evolutionary rate ratio dN/dS.

Scaled selection coefficients fully encapsulate ω

To validate our derived relationship between scaled selection coefficients and ω, we simulated protein-coding sequences along a 10-taxon phylogeny according to the Halpern-Bruno mutation-selection model [6], assuming no codon bias and symmetric nucleotide mutation rates. We calculated an ω for each simulated alignment using the derivations given in equations (**??**)–(8) as well as standard (ML) methods, according to the GY94 [4] rate matrix, as implemented in the HyPhy batch language [15].

As shown in Figure 1A, ω 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 the evolutionary rate ratio, ω, and that the results from MutSel and ω models are largely in agreement.

Moreover, as seen in Figure 1A, estimates for ω never exceed 1, but rather all reflect a regime of purifying selection. In fact, in SuppMat, we prove that, when calculated using scaled selection coefficients, ω is necessarily always less than or equal to 1. This important insight reveals that, while MutSel models fully agree with ω models, this 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 ($\omega > 1$).

Influence of ML model parameterizations

ML 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, it is well-known that different ML parameterizations can influence its ω estimation [12,16,19]. Therefore, we additionally inferred ML ω values with a total of 12 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 [17] 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 ω values as calculated using equations (**??**)–(8).

Results in Table 1 yield several important insights into the behavior of ML ω models. First, it is

clear ML methods only estimate ω accurately when equilibrium codon frequencies are set as equal (i.e., each codon has a frequency of 1/61). Under this frequency specification, ML also estimates κ fairly accurately (r = 0.913) while maintaining an extremely high ω correlation of 0.996. Second, the F3x4 and CF3x4 frequency estimators perform nearly identically to one another (p=0.54), and both display weak, negative correlations with the ω derived from selection coefficients. Finally, when empirical codon frequencies are used, ω estimates have a moderately strong, negative correlation. Thus, as the model's codon frequency parameters were more and more tailored to the given data set, error between derived and ML ω values increased.

However, it is interesting to note that, while ML yields the most accurate κ estimates when codon frequencies are set to 1/61 each, all frequency specifications yield κ estimates which correlate relatively strongly with the true 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.

The reason that these latter three frequency specifications yield ω values that negatively correlate with our derived ω values is that they all strongly overestimate omega, in particular the empirical frequency specification, as illustrated in Figure 2 (plots showing all regressions are in Figure S1). As we have proven that, when calculated from scaled selection coefficients, ω is necessarily less than or equal to 1, we interpret the use of F3x4 or empirical codon frequencies effectively as ML model mis-specifications, leading to grossly elevated dN/dS values.

Importantly, however, our simulated data sets contained relatively constrained amino acid, and thus codon, frequencies, as a single MutSel parameterization was applied to all positions in the simulated alignment. Therefore, we examined the extent to which codon constraints influenced this tendency for frequency specifications to yield spurious results. For each frequency specification, we calculated the Shannon 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, for sense codons, the maximum $H(i) = 4.11$ value is reached when all codons have a frequency of 1/61. We then examined how well the ML error scaled with codon entropy for the input frequencies, results in Figure 3. The x-axis here is abs(freqentropy - equalentropy)/equalentropy, and y-axis is log of the mlw error, relative to the derived omega value. The relationship is very clear - the farther your freq specification gets from 1/61 the more error you will have.

Thus, the question emerges of why specifying equal codon frequencies leads to precisely accurate dN/dS estimates, whereas the other, more commonly used specifications produce incorrect and

biased estimates. The basis for including codon frequencies in mechanistic codon substitution models is to capture biases in the underlying nucleotide base frequencies [12, 16]. The motivation for using frequency estimators such as F3x4 is that the resulting equilibrium codon frequencies will reflect those which would have been produced solely by mutational processes, and not from selection. However, without experimental evidence, there is no clear way to know *a priori* the source of nucleotide biases. Indeed, our simulations yielded alignments with a wide array of GC-content values, ranging from 0.22 - 0.79, in spite of an underlying symmetric mutation matrix. Thus, all compositional biases in our simulations emerged strictly from natural selection acting favoring particular amino acids. Importantly, as we adopted symmetric mutation rates in our simulations, these varying degrees of GC-content were not caused by any bias towards unequal base frequencies, but rather by codon fitness differences alone. Thus, the correct frequencies to specify in ML inference should reflect those frequencies which would be present *in the absence* of selection, which, the case of symmetric mutation rates, corresponds to equal codon frequencies.

So, it is critical to parameterize models properly in order to be confident that ω is the sole parameter representing selection. If, as is the case when non-equal freqs were specified, other parameters in the model take on an interpretation of selection, then ω is not going to mean what we naively assume it must. Highlights potential issues with the ω class of models. While the framework for dN/dS is one rooted in popgen, these markov processes are not. MutSel models, on the other hand, use a markov process explicitly based on the underlying population genetics principles governing the evolutionary process. When either F3x4, CF3x4, or empirical frequencies were used, information about the strength of natural selection was incorrectly placed in these parameters, which produced artifactually elevated measures of dN/dS, because this parameter was no longer the only one measuring selection.

Finally, we must ask whether this issue could be problematic in real sequence analyses. Our alignments were simulated through a homogenous evolutionary process, whereas in real sequence data, different amino acid (codon) frequencies would be naturally expected across sites. The median entropy in our simulated data set was 3.261, but ranged from 1.3 - 3.91. In real analyses, codon frequencies are treated as a global parameter, with the notion that averaging across sites will give a selection-proof estimate. We looked at e.coli genome which is know to have a lot of codon bias and thus may have low codon entropy. e coli ranges from 3.2-3.9, and half of our data is in that range. Thus, it is possible that incorrect frequency specs are producing false positives or getting the extent of selection basically wrong. Unless you are sure that codon frequency biases are attributed

to mutation instead of selection, we don't recommend this parameter.

These results reflect a broad misinterpretation of how the equilibrium codon frequencies parameters should be specified in mechanistic codon models. Typically, these values are taken directly from the given data and fixed in the model to reduce the number of parameters inferred. However, the codon frequencies observed in the data set represent those which exist *after* natural selection has acted on the sequence. Instead, the model should use the codon frequencies which would exist *in the absence of natural selection*. Selection, alternatively, acts to tune this frequencies to increase protein fitness. If one specifies equilibrium frequencies which exist after natural selection has acted on the protein sequence (i.e., frequencies measured directly from the data), then the influence of selection pressure is incorporated into these values. The desirable outcome, however, is that the dN/dS parameter measures the selective strength. If other parameters in this model contain selection information, estimates for dN/dS will not accurately capture selective effects.

Discussion

Here, we have derived a formal mathematical relationship between the parameter estimates of MutSel and mechanistic codon models. Through a simulation approach, we validated this relationship and demonstrated that these models yield nearly identical results. However, we additionally found that MutSel models necessarily cannot describe scenarios in which positive or diversifying selection occurs, or when $dN/dS > 1$. Although it is generally acknowledged that MutSel models carry the assumption of purifying selection, we formalize and prove this result. These findings have important implications for how and when these models should be applied. In cases of purifying selection or neutral evolution, these two competing models are virtually no different from one another, and thus use of either model is fully justified. Alternatively, if positive selection is occurring, MutSel models cannot be used as they mathematically cannot capture the actual evolutionary process.

This study highlights the importance of examining the similarities and differences among different evolutionary model classes. 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.

Now, for the frequency specification: does this really matter? Typically, when one measures codon frequencies from the data set, codon frequencies are treated as global parameters rather than site-specific. Thus, equilibrium codon frequency parameters effectively represent an average across

the entire data set, naturally lead to a flatter distribution of codon frequencies. Our simulated data sets, on the other hand, were evolved according to a single parameter site, leading to more constrained global codon frequencies. Therefore, the extent to which bias introduced by incorrect frequency specifications was driven by how far from equal we're talking.

Caveat about how our math relies on symmetric mutation rates and/or equal mutation rates, depending on which one ends up happening.

Benchmarking strategy is something important that has emerged from this study. Typically, inference framework performance is assessed through simulating datasets 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.

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] 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.

[8] 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.

[9] 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.

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

[11] 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.

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

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

[14] 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.

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

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

[17] 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.

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

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

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 [17], 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.

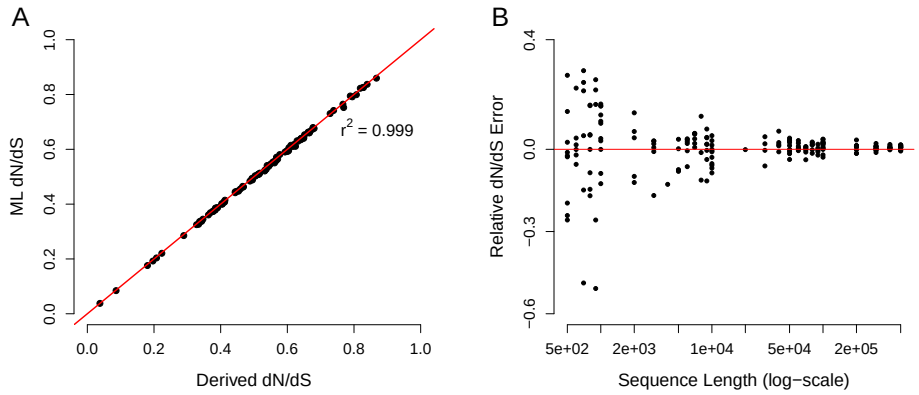

Figure 1: Relationship works exceedingly well. Left panel shows 100 points, each of which corresponds to single simulation. Note that here the ml inference is shown for equal codon frequency specs and kappa fixed to true value (a similar plot for free kappa is shown in suppfigs, but results are qualitatively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

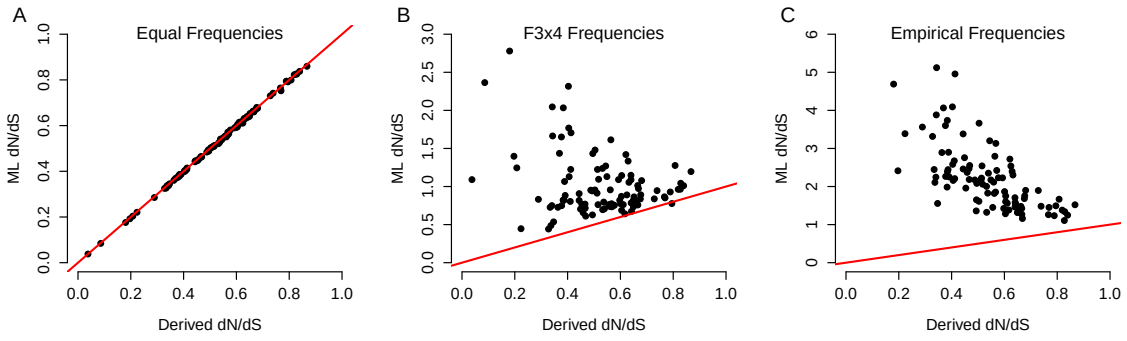

Figure 2: Issues with frequency specifications abound. In each plot, red line indicates 1:1 agreement, so note the y-axis differences. Relationship between omega values only really exists when equal codon frequencies are specified. When f3x4 or true freqs used, there is the potential to end up with dramatically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

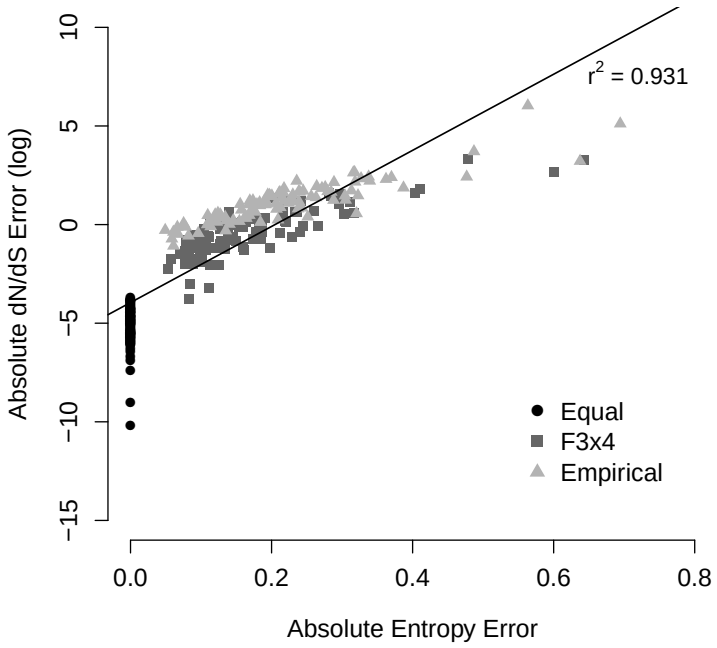

Figure 3: The farther the freqspec is from 1/61, the worse the ML inference is. Here the xaxis is the entropy of the codon frequencies provided to ml.

Supplementary Figures

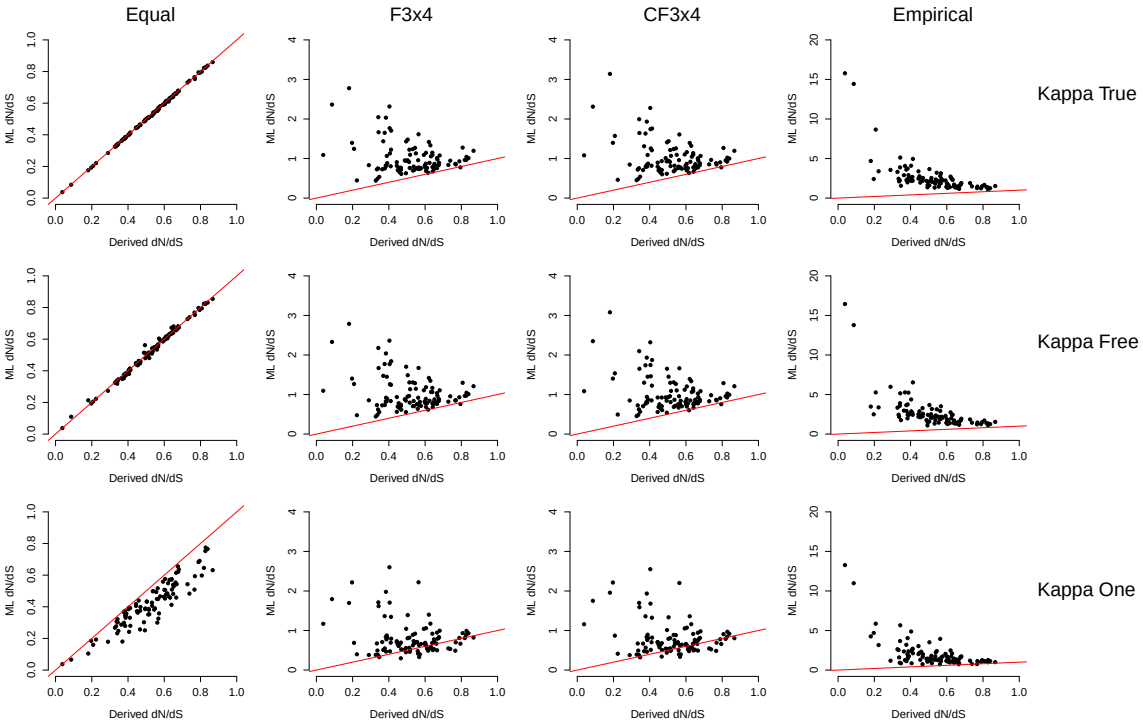

Fig. S1 Regression for all ML parameterizations.