## Introduction

Over the years, a variety of models have been proposed to describe the effects of natural selection on protein-coding sequences, in a phylogenetic context. Traditionally, the focus has been on mechanistic codon-substitution models (see ref. [1] for a comprehensive review). Since their introduction in the 1990s, these models have seen great success in inferring protein evolutionary rates, or the nonsynonymous/synonymous rate ratio ($dN/dS$). This metric indicates how quickly a protein's constituent amino acids change [2–4], allowing for the identification of positively-selected regions in protein sequences [4, 5].

More recently, a second class of models, known as mutation-selection-balance (MutSel) models, has emerged as a popular alternative to $dN/dS$ models. Unlike mechanistic codon models, MutSel models explicitly model the dynamic balance between mutation and selection, rather than merely the final outcome (e.g. substitution) of this process [6–9]. Moreover, these 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 $dN/dS$ 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 $dN/dS$ and MutSel models by examining the extent to which their focal parameters, $dN/dS$ and scaled amino acid selection coefficients, yield overlapping information about the evolutionary process. To this end, we derive a mathematical relationship between these two parameter classes, and we demonstrate that MutSel models fully embody the $dN/dS$ values. Using a simulation approach, we verify that $dN/dS$ values estimated

using selection coefficients alone correspond precisely to those inferred using a traditional $dN/dS$ maximum likelihood inference approach. Importantly, we additionally show that this relationship holds only under regimes of purifying selection or neutral evolution ($dN/dS \leq 1$). Therefore, MutSel models are inherently unable to describe 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.

# 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

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

where $\mu_{ij}$ 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 $\pi_i$ is the equilibrium frequency of codon $i$.

For each simulation, we simulated selection coefficients $s_a$ for each amino acid $a$ by drawing from a normal distribution $\mathcal{N} \sim (0, 1)$. Using these selection coefficients, we derived steady-state amino acid frequencies $F(a)$ according to

$$F(a) = \frac{e^{s_a \lambda}}{\sum_b e^{s_b \lambda}} \tag{3}$$

, where the denominator sums over all 20 amino acids [13]. Previous work has shown that this formulation accurately represents site-specific, steady-state amino acid frequencies. The parameter $\lambda$ scales linearly with RSA (how to describe this??). For each simulation, we selected a coefficient $\lambda$ from $\mathcal{U} \sim (0.5, 2.0)$, which represent reasonable such values as found by [13].

Once a set of 20 steady-state frequencies were derived, we assigned them to amino acids as follows. For each set of frequencies, we determined the number of preferred amino acids, defined as

the number of frequency values greater than 0.05, or the frequency one would expect under neutral evolution. We then selected a set of preferred amino acids such that the mean pair-wise Grantham scores among these amino acids was $\leq 100$. These amino acids were then randomly assigned to have the frequencies above 0.05, and the remaining amino acids were assigned randomly to all frequencies below 0.05. The resulting amino acid frequencies were then converted to codon frequencies such that all synonymous codons shared the same frequency (i.e., there was no codon bias).

We initiated each simulation by selecting a root sequence based on the steady-state codon frequencies. We then evolved this root sequence along a two-taxon phylogeny, effectively representing a single evolutionary trajectory of protein evolution. For all simulations, we fixed branch lengths and all $\mu_{ij}$ values to 0.005 and $10^{-5}$, respectively, and we selected a value for each simulation's $\kappa$ from $\mathcal{U} \sim (1, 5)$. Unless otherwise stated, we simulated alignments of one-million codon positions. Moreover, 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 scenario 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.

We inferred $dN/dS$ values in two ways; first, we derived a $dN/dS$ value using the mathematical relationships described in (4)–(9), and second, we used the standard maximum likelihood M0 model, which uses the GY94 evolutionary model [2], as implemented in HyPhy [14]. The GY94 model includes the primary parameters $\omega = dN/dS$, $\kappa$, and equilibrium codon frequencies. **Not true, but still need explanation for why we aren't doing it:For each inference, we fixed $\kappa$ to the known simulated value**. Unless otherwise stated, we specified that HyPhy use equal equilibrium codon frequencies, such that each codon had a frequency of 1/61. This frequency specification was necessary to achieve accurate $dN/dS$ maximum likelihood estimates, and is discussed more in depth in Results. 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; 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

$$f_i = e^{s_i} \Big/ \sum_k e^{s_k} \,, \tag{4}$$

where the sum in the denominator runs over all 61 sense codons [15]. Here, $s_i$ is the *scaled selection coefficient* for codon $i$; larger $s_i$ correspond to higher frequencies of codon $i$. The fixation probability for a mutation from $i$ to $j$ is [6, 15]

$$\pi_{i\to j} = \frac{1-(f_i/f_j)^{1/N_e}}{1-f_i/f_j} \approx \frac{1}{N_e}\frac{\ln f_j - \ln f_i}{1-f_i/f_j} \,, \tag{5}$$

where $N_e$ is the effective population size. We can calculate an evolutionary rate by summing over all fixation probabilities weighted by the frequency of the originating codon. For example, we can write the synonymous rate $K_\mathrm{S}$ as

$$K_\mathrm{S} = N_e \sum_i \sum_{j\in\mathcal{S}_i} f_i \pi_{i\to j} \mu_{ij} \,, \tag{6}$$

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 $K_\mathrm{S}$, we divide it by the number of synonymous sites, which we calculate according to the mutational opportunity definition of a site [2, 12] as

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

Under the assumption that all synonymous codons have equal fitness (all synonymous mutations are neutral), we have $\pi_{i\to j} = 1/N_e$ [16], and thus we find for $dS$, the synonymous rate per synonymous site,

$$dS = \frac{K_\mathrm{S}}{L_\mathrm{S}} = \frac{\sum_i \sum_{j\in\mathcal{S}_i} f_i \mu_{ij}}{\sum_i \sum_{j\in\mathcal{S}_i} f_i \mu_{ij}} = 1. \tag{8}$$

Similarly, for $dN$, the non-synonymous rate per non-synonymous site, we find

$$dN = \frac{K_\mathrm{N}}{L_\mathrm{N}} = \frac{N_e \sum_i \sum_{j\in\mathcal{N}_i} f_i \pi_{i\to j} \mu_{ij}}{\sum_i \sum_{j\in\mathcal{N}_i} f_i \mu_{ij}} \,, \tag{9}$$

where $\mathcal{N}_i$ is the set of codons that are not synonymous to codon $i$ and differ from it by one nucleotide substitution. The quantities $K_\mathrm{N}$ and $L_\mathrm{N}$ are defined as in Eqs. (6) and (7) but summing over $j \in \mathcal{N}_i$ instead of $j \in \mathcal{S}_i$.

Equations (4)–(9) establish a connection between the scaled selection coefficients $s_i$ (i.e., the primary parameters of a MutSel model) and the evolutionary rate ratio $dN/dS$.

## $dN/dS$ values fully encapsulated by scaled selection coefficients

To validate our derived relationship between $dN/dS$ values and scaled selection coefficients, we simulated protein-coding sequences along a lineage according to the Halpern-Bruno mutation-selection model [6]. For each simulation set, we selected steady-state amino acid frequencies, which we then converted to codon frequencies, according to

$$F(a) = \frac{e^{s_a \beta}}{\sum_b e^{s_b \beta}} \tag{10}$$

, where $F(a)$ corresponds to the frequency of amino acid $a$ and $s_a$ corresponds to the scaled selection coefficient for this amino acid. These $s_a$ values are analogous to the selection coefficient parameters which a MutSel model would infer from a given data set. For each simulation, we drew selection coefficients $s_a$ for each amino acid from $\mathcal{N} \sim (0, 1)$, and we selected the coefficient $\beta$ from $\mathcal{U} \sim (0.5, 3.0)$. Note that higher values of $\beta$ indicate stronger constraint on the amino acid distributions; as $\beta$ approaches infinity, effectively only a single amino acid will be allowed.

Following simulation, we calculated a $dN/dS$ value using both standard ML methods, according to the GY94 [2] model, and our relationship. As shown in Figure 6, $dN/dS$ 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 5). We confirmed, using simulations, that this relationship holds under different model parameterizations, including different specifications for $\kappa$ (Figure 2) and GC content (Figure 1). These results clearly show that MutSel model parameters fully encapsulate information regarding the evolutionary rate ratio, $dN/dS$, and that the results from MutSel and $dN/dS$ models are largely in agreement.

## Maximum Likelihood $dN/dS$ estimates strongly biased by equilibrium frequency parameterization

In verifying the derived relationship between selection coefficients and $dN/dS$, we encountered some biases in ML inference methods. In particular, the specific codon frequency specification that the ML inference used had a substantial effect on its accuracy. Only when specifying equal codon frequencies (an equilibrium frequency of 1/61 for each codon, regardless of that codon's actual frequency in the given data set) were we able to achieve agreement between our derived and ML $dN/dS$ estimates (Figure 3) . On the other hand, when more commonly used codon frequency specifications, included the popular F3x4 estimator and simply using frequencies as measured from

the data, ML yielded strongly inflated $dN/dS$ estimates. Indeed, as the model's codon frequency parameters were more and more tailored to the given data set, error between derived and ML $dN/dS$ values increased.

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 simulated alignment $i$, we calculated the Shannon entropy,

$$H(i) = -\sum_j P_j \ln P_j \tag{11}$$

, where $P_j$ is the frequency of codon $j$ 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 the relationship between error in $dN/dS$ estimates caused by different codon frequency specifications and alignment codon entropy, as shown in Figure 4. Clearly, as entropy increases, thus approaching a flatter distribution of codon frequencies, error in ML estimates does indeed decrease. Even so, specifying equal codon frequencies will nearly always minimize the error.

We contend that 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.

## Mutation-selection-balance models are only valid for purifying selection

**I really have no idea how to write up proofs, so I've virtually just latex'd the mathematica document. But at least the equations are there! ... next day: I'd better stop. This section might have to be yours, unfortunately.**

To show that mutation-selection models only corresponds to $dN/dS \leq 1$, we make use of that fact that each calculation of $dN/dS$, as described in equations (4)–(9), entails summing the forward and backward fixation probabilities between codons, which are in turn divided by the codon frequency sums. We additionally assume that all mutation rates $\mu_{ij}$ are equal, and as all values for $N$ will ultimately cancel, we exclude them from the following proof. We will deal with the case of

two nonsynonymous codons, $i$ and $j$, and we write their frequencies as $x$ and $y$, respectively.

As follows from (5), the sum of the probabilities going from codon $i$ to codon $j$ and from codon $j$ to codon $i$ is

$$xP(x,y) + y(P(y,x) = \frac{2xy[\ln x - \ln y]}{x-y} \tag{12}$$

, and we will demonstrate that this value is necessarily $\leq x + y$ for $x, y \geq 0$ and $x \leq y$.

To this end, we define the function

$$F(x,y) = x + y - \frac{2xy[\ln x - \ln y]}{x-y} \tag{13}$$

. Thus, we show that $F(x,y) \geq 0$. For the condition $x = y$, this is straightforward to show, as $\lim_{x \to y} F(x,y) = 0$. We now show that the first derivative of $F(x,y)$ is negative throughout $x \in (0,y)$, thus proving that $F(x,y)$ has to be monotonically decreasing, and hence $\geq 0$, in this interval.

## References


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

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

[3] 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.

[4] 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.

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

[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 Computational Molecular Evolution. Oxford University Press.

[13] Ramsey DC, Scherrer MP, Zhou T, Wilke CO (2011) The relationship between relative solvent accessibility and evolutionary rate in protein evolution. Genetics 188: 479–488.

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

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

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

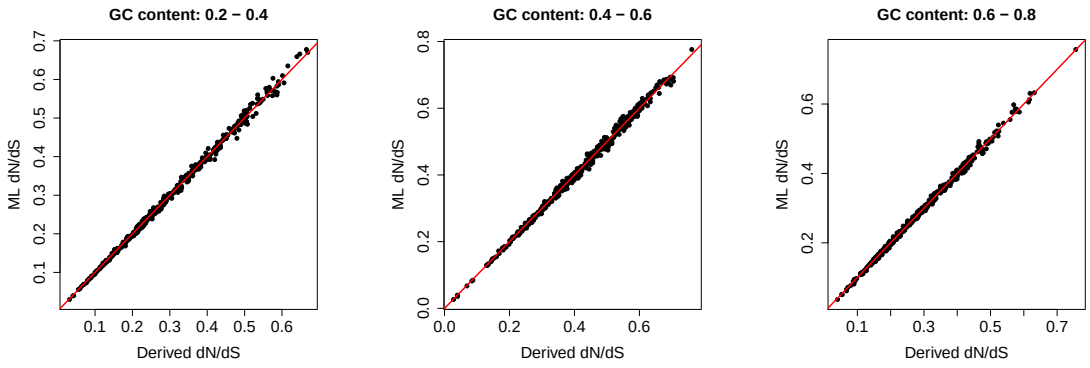


Figure 1: GC content additionally does not influence the relationship. 400 simulations per panel.

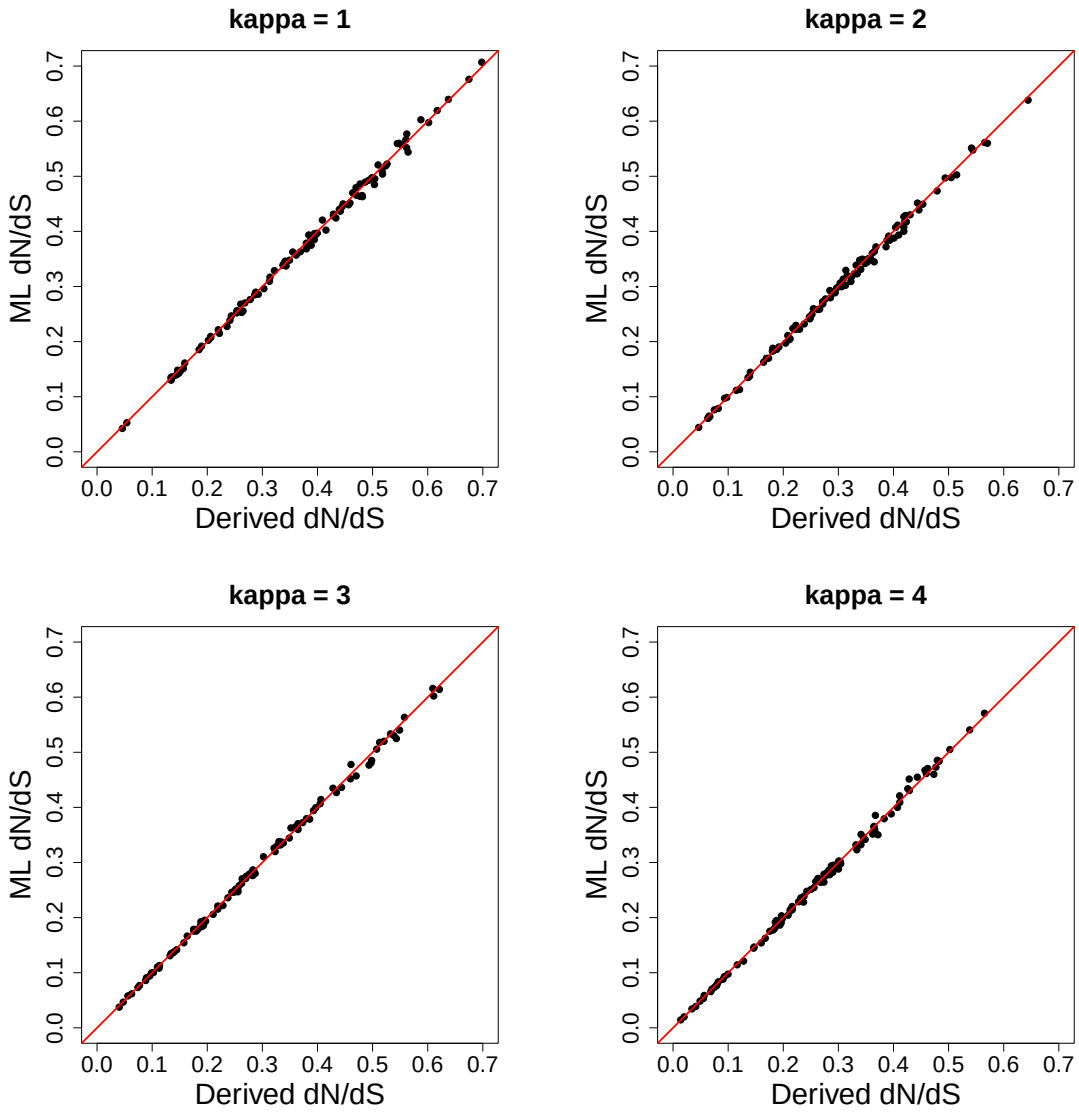


Figure 2: Value of kappa does not matter for agreement with derived and ML omegas. Relationship robust to differences in (symmetric) mutational spectrum. 100 simulations per panel.

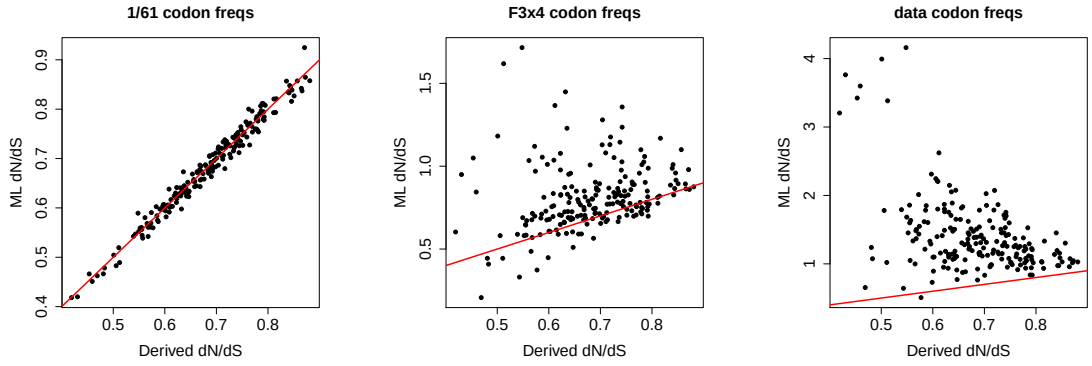


Figure 3: Equilibrium codon frequency specification to ML inference matters. Omega estimates agree when specify equal codon frequencies, and error increases as frequency specifications are more and more tailored to the data, ultimately resulting in wildly inflated values when the real frequencies are used.

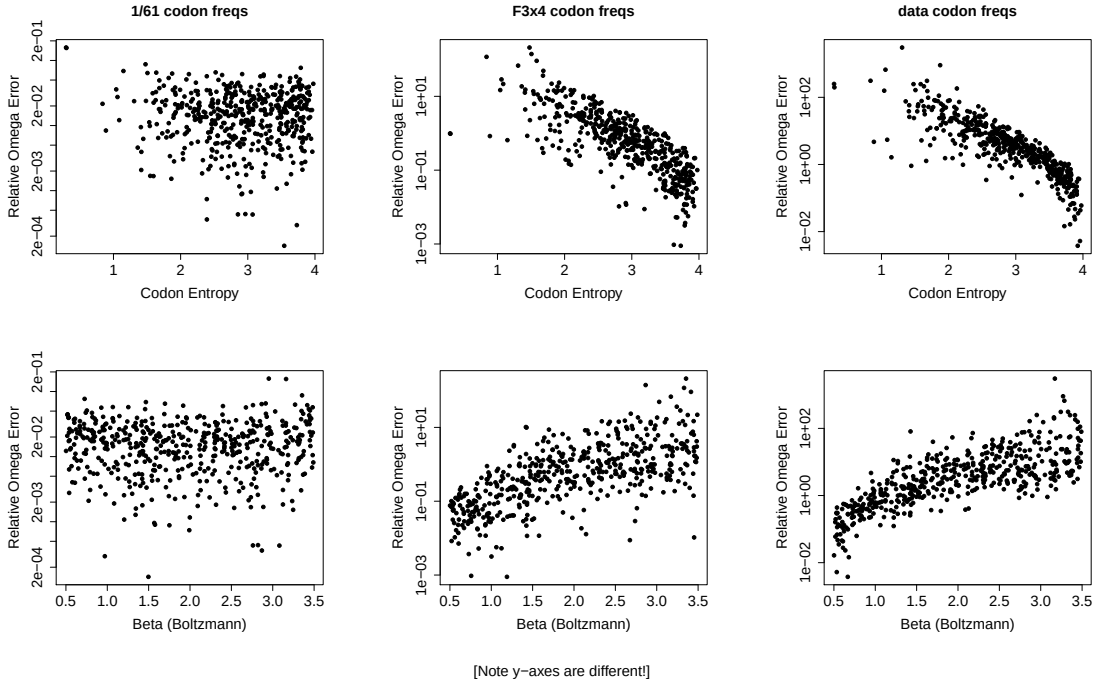


Figure 4: Issues with frequency specifications strongly related to the codon frequencies in the data set. Issue is more egregious when there are relatively few codons, based on entropy. As entropy increases (more permissive, and thus data set codon frequencies are flatter), the error decreases and ML more approximates the true omega value.

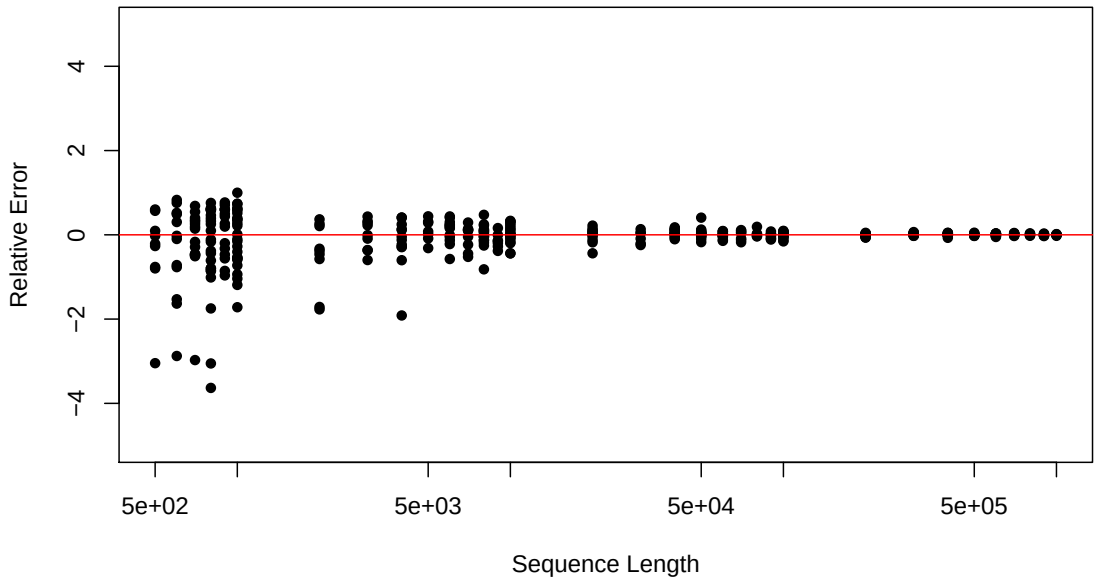


Figure 5: Convergence of derived $dN/dS$ and ML estimates of $dN/dS$. Each point represents results from a single simulation. The y-axis indicates relative error of the ML $dN/dS$ estimates, and the x-axis indicates sequence length on a log-scale. As the sequence length, or the data set size, increases, the two $dN/dS$ estimates converge to the same value. Note that this simulated data used a beta of 2.5.

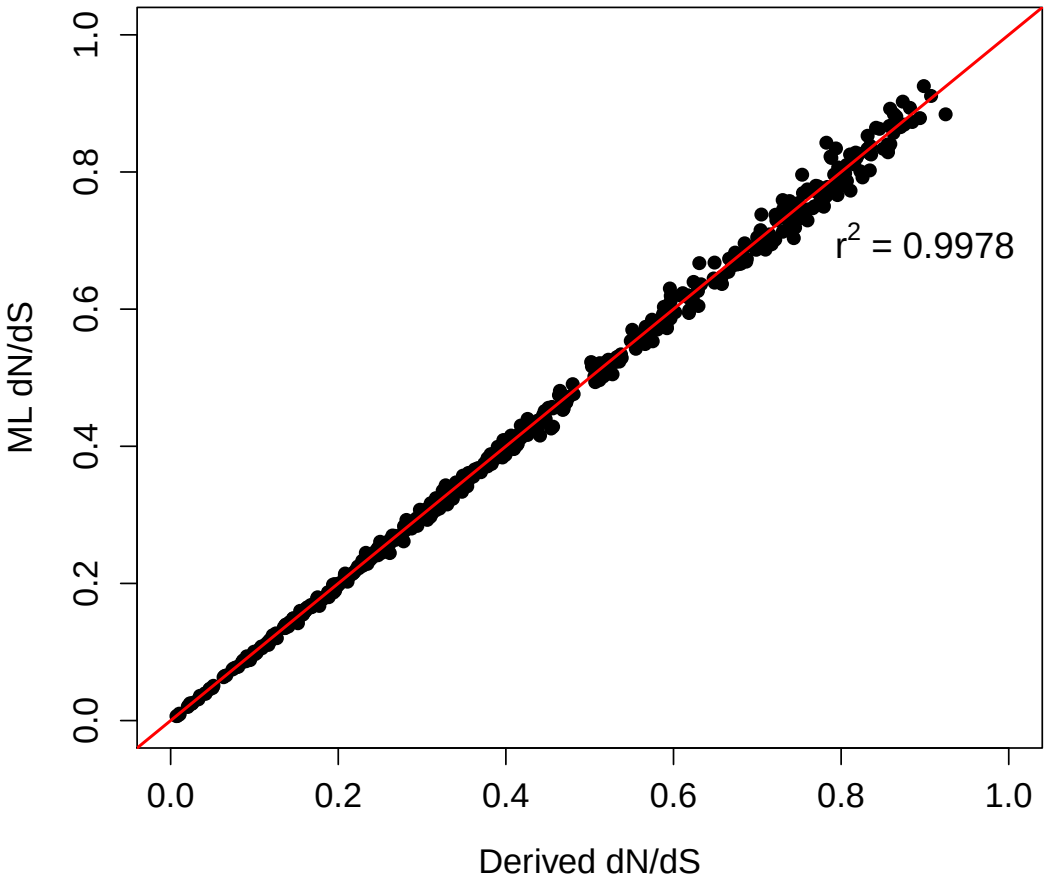


Figure 6: Relationship works exceedingly well. There are 500 points in this plot, each of which corresponds to a single simulation. Beta values for each simulation were randomly chosen between 0.5-3.5, so this plot does contain varying levels of selective constraint on amino acid distributions. Note that beta is not significant in a regression (either additive or interaction model) so the extent of constraint doesn't appear to have any influence. Moreover, kappa=1.0 here.