## Introduction, Rough

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.

Some have argued that MutSel models are more robust than $dN/dS$ and can better describe the evo process owing to their treatment of amino acid identities. More fine-grained modeling results than a $dN/dS$ analysis would yield. However, it is virtually unknown how these models really relate, so it remains unclear whether one model should be preferred over another. We don't know how parameter estimates even relate.

Although both $dN/dS$ and MutSel models describe the same fundamental process of protein evolution along a phylogeny, the relationship between these models is largely unknown. These two classes of models have largely been developed independently, and as a consequence we do not know whether parameters estimated from a $dN/dS$ model are similar, distinct, or even contradictory to those estimated from from a MutSel model of coding sequence evolution. 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 show that we can accurately estimate $dN/dS$ values using MutSel model parameter estimates, and these estimates 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 selection, in which $dN/dS > 1$. This result has important implications for circumstances under which MutSel model use is justified.

## Mathematical relationship between selection coefficients and omega

**At this time, this section is largely copied from the NIH proposal, except with unequal mu's. We can always replace this with the mu's that cancel out again, given that that's the proof.** 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

$$f_i = e^{s_i} \Big/ \sum_k e^{s_k}\,, \tag{1}$$

where the sum in the denominator runs over all 61 sense codons [12]. 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 [6,12]

$$\pi_{i \to j} = \frac{1 - (f_i/f_j)^{1/N_e}}{1 - f_i/f_j} \approx \frac{1}{N_e}\frac{\ln f_j - \ln f_i}{1 - f_i/f_j}\,, \tag{2}$$

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

$$K_\mathrm{S} = N_e \sum_i \sum_{j \in \mathcal{S}_i} f_i \pi_{i \to j} \mu_{ij}\,, \tag{3}$$

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 $L_\mathrm{S}$, which we can calculate as

$$L_{\mathrm{S}} = \sum_i \sum_{j \in \mathcal{S}_i} f_i \,. \tag{4}$$

Under the assumption that all synonymous codons have equal fitness (all synonymous mutations are neutral), we have $\pi_{i \to j} = 1/N_e$ [13], and thus we find for $dS$, the synonymous rate per synonymous site,

$$dS = \frac{K_{\mathrm{S}}}{L_{\mathrm{S}}} = \frac{\sum_i \sum_{j \in \mathcal{S}_i} f_i \mu_{ij}}{\sum_i \sum_{j \in \mathcal{S}_i} f_i}. \tag{5}$$

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

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

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. (3) and (4) but summing over $j \in \mathcal{N}_i$ instead of $j \in \mathcal{S}_i$.

Equations (1)–(6) establish a connection between the scaled selection coefficients $s_i$ (i.e., the primary parameters of a MutSel model) and the evolutionary rate ratio $dN/dS$.

# Methods

## Sequence simulation and omega inference

We simulated protein-coding sequences as a continuous-time Markov process [14] 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

$$Q_{ij} = \begin{cases} 0 & \text{multiple nucleotide changes} \\ \mu_{ij} f_{ij} & \text{single nucleotide transversion} \\ \kappa \mu_{ij} f_{ij} & \text{single nucleotide transition} \end{cases}\,, \tag{7}$$

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

$$f_{ij} = ln\left(\frac{\pi_j \mu_{ij}}{\pi_i \mu_{ji}}\right) \Big/ \left(1 - \frac{\pi_i \mu_{ji}}{\pi_j \mu_{ij}}\right), \tag{8}$$

where $\pi_i$ is the equilibrium frequency of codon $i$.

For each simulation, we derived steady-state amino acid frequencies according to the Boltzmann distribution,

$$F(a) = \frac{e^{s_a \beta}}{\sum_b e^{s_b \beta}} \tag{9}$$

, where $F(a)$ is the equilibrium frequency of amino acid $a$, and the denominator sums over all 20 amino acids [12, 15]. In this framework, $s_a$ represents the scaled selection coefficient for amino acid $a$, analogous to the primary parameter given by a MutSel model. Unless otherwise stated, $\beta$ **was set to 2.0 for all simulations (I MADE THIS UP, COME BACK TO IT ?????)**. We assigned these frequency values to amino acids as follows. For each set of frequencies, we determined the number of preferred amino acids as the number of frequency values greater than 0.05. 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).

All simulations were conducted along 2-taxon tress with branch lengths and $\mu$ values fixed at 0.005 and $10^{-6}$, respectively. Unless otherwise stated, we simulated alignments of one-million positions, and amino acid frequencies were derived using a $\beta = 2.0$. 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.

For each simulated alignment, we inferred $dN/dS$ values in two ways; first, we used selection coefficients to derive the $dN/dS$ value, and second we used the standard maximum likelihood M0 model, which uses the GY94 evolutionary model [2], as implemented in PAML [16]. The GY94 model includes two primary parameters, $\omega = dN/dS$ and $\kappa$. For each inference, we fixed $\kappa$ to the known simulated value, and we provided HyPhy with equal equilibrium codon frequencies, such that each codon had a frequency of 1/61. This codon frequency was necessary to achieve accurate $dN/dS$ maximum likelihood estimates, and is discussed more in depth in Results.

# Results

## $dN/dS$ 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 fixed all mutation rates to $10^{-6}$, and we selected steady-state amino acid 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 $a$. These amino acid frequencies are analogous to the scaled selection coefficient parameters which a mutation-selection model would produce. **Again, this was made up so return to this later. - Unless otherwise stated, we set** $\beta = 2.0$**.** 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).

### 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 (1)–(6), 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 (2), the sum of the probabilities going from codon $i$ to codon $j$ and from codon $j$ to codon $i$ is

$$xP(x,y) + y(P(y,x) = \frac{2xy[\ln x - \ln y]}{x - y} \tag{11}$$

, 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

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

. 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.

## Maximum Likelihood inferences strongly biased by equilibrium frequency specifications

An emerging result from this study was that specs really matter. In order to get good agreement between our derived $dN/dS$ values and ML estimates, we had to set equilibrium codon frequencies equal to 1/61 each. For instance, when you use other commonly used frequency specifications, there is tendency for ML methods to yield wildly inflated values for $dN/dS$ (Figure 3); as frequency specifications become more and more tailored to the data set, ML estimates become more and more distant from the true value.

This is because of where selection is represented. In mutsel models, selection is essentially manifested by the frequencies, whereas in dnds models, it is represented by omega. If you provide true frequencies, or ask for it to be calculated from data, then the frequencies hold the effects of selection, causing inflated estimates of dnds. However, if you set all codons to 1/61, this scenario effectively represents a neutral case with no effects of selection, which then allows the omega parameter to capture the effects.

However, this effect depends heavily on the specifications given for $\beta$, which controls the level of selection pressure. Higher levels of $\beta$ naturally lead to more skewed codon frequency distributions, whereas lower values of $\beta$ will gradually approach a flattened distribution. Thus, we see that this error caused by frequency specification decreases when the real codon distribution is flatter (Figure 4)

## Discussion

Several things emerge from this analysis - 1. If you fit a mutsel model and calculate dnds from it, it's just as good as if you had used a mech codon model. 2. we prove that dnds, when calc'd from mutsel, must be less than 1. although generally acknowledged that purifying selection is a feature, we demonstrate it precisely and reveal that, in cases of positive selection, mutsel models are likely not appropriate. 3. we also have a more realistic way to assess performance of dnds models. bias is introduced by codon frequencies used by the model. otherwise, freqs will likely capture selective pressure, rendering omega estimates bonkers. typically highly elevated. More work should be done in this field, for instance, in dnds analyses featuring high levels of codon bias may result in artifactually high dnds, although unclear the extent to which this will actually influence.

Discussion points: - Important insight is that dN/dS¿1 inherently cannot be described by mutation-selection models, and thus at those sites its results may be misleading. In particular, possibly a confounding factor in the Rodrigue implementation, as positively selected sites in the alignment could introduce bias. - Importance of examining intersections between models. Must understand how estimates from one relate to another. Helps to ensure robust results; model agreement is key, so we must formulate explicit relationships among them to systematically assess agreement. -

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

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

[14] Yang Z Computational Molecular Evolution. Oxford University Press.

[15] 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.

[16] Yang Z (2007) Mol Biol Evol 24: 1586–1591.

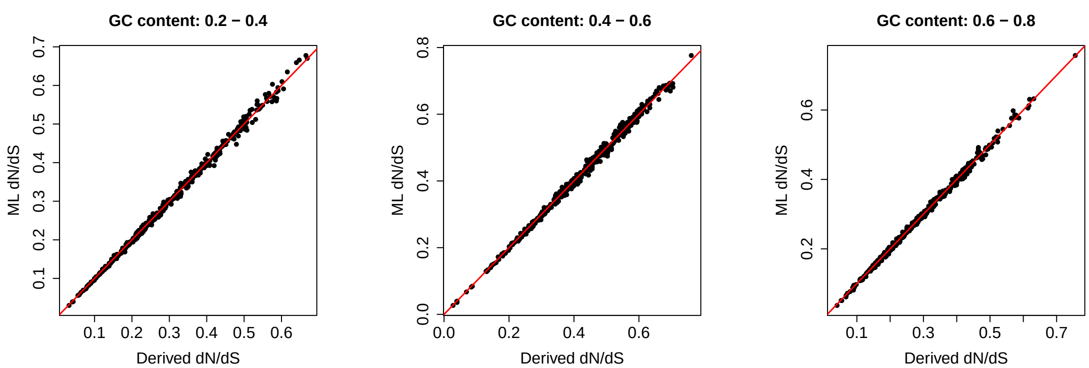


Figure 1: GC content additionally does not influence the relationship. 400 simulations per panel.

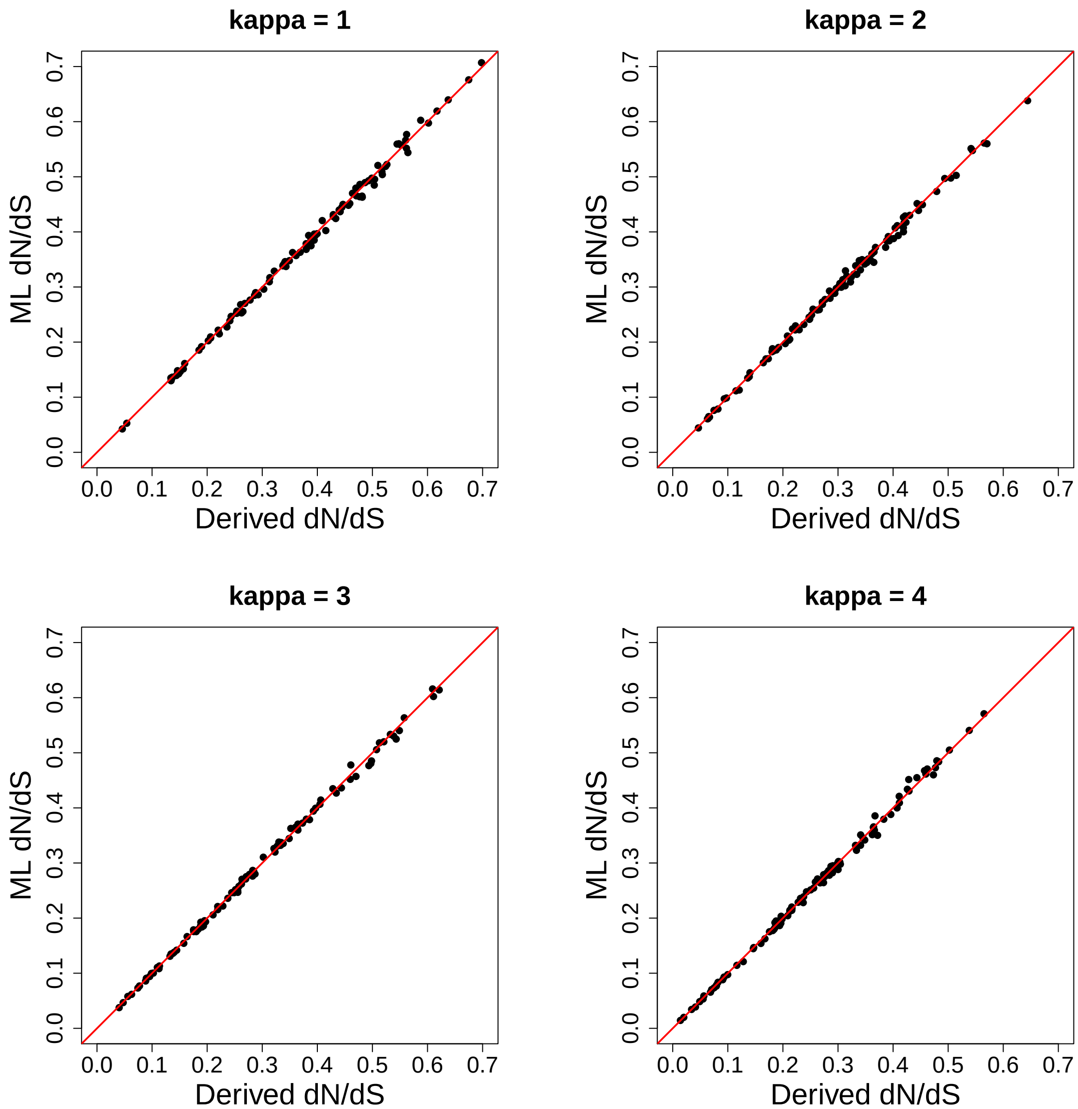


Figure 2: Value of kappa does not matter for agreement with derived and ML omegas. Relationship robust to differences in (symmetric) mutational spectrum. 100 simulations per panel.

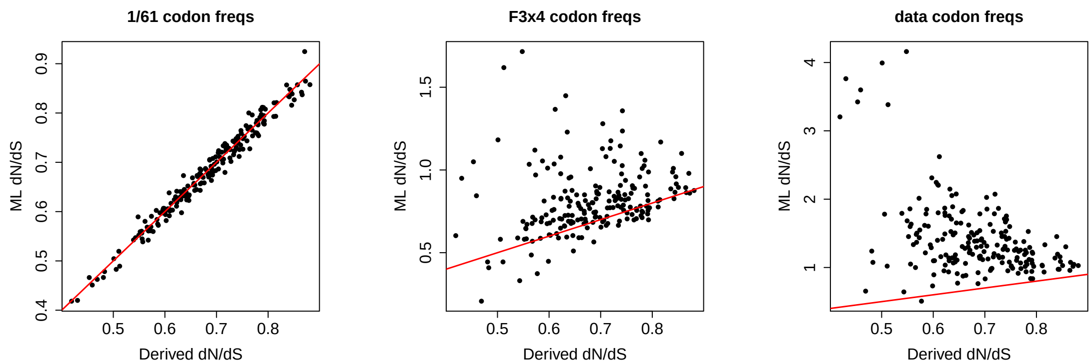


Figure 3: Equilibrium codon frequency specification to ML inference matters. Omega estimates agree when specify equal codon frequencies, and error increases as frequency specifications are more and more tailored to the data, ultimately resulting in wildly inflated values when the real frequencies are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

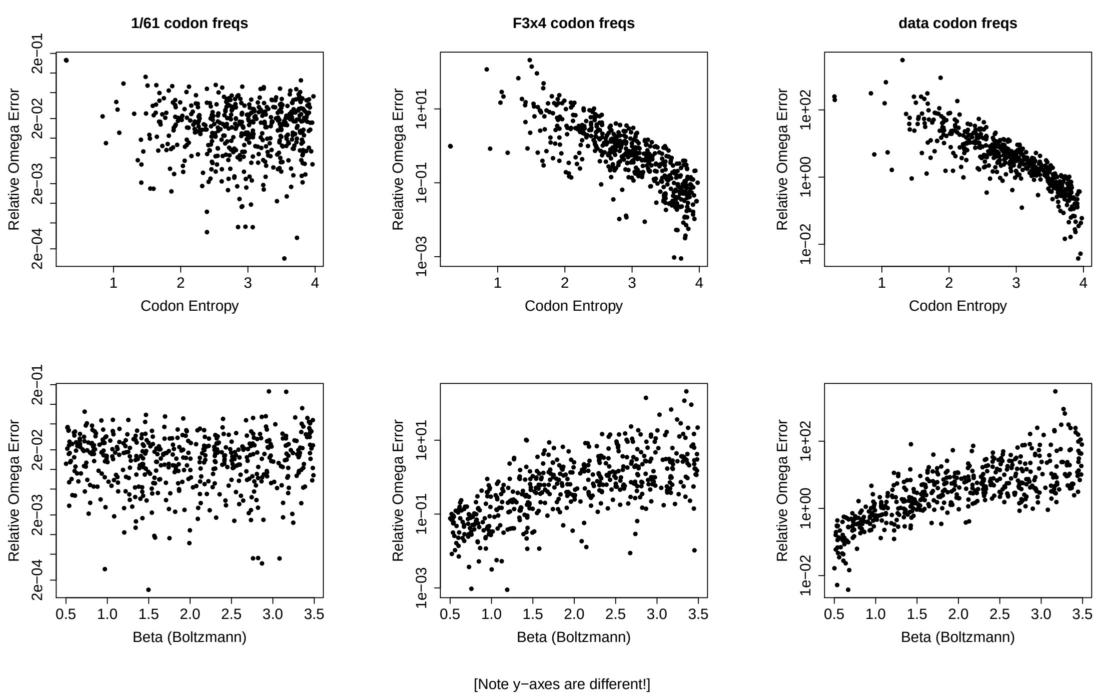


Figure 4: Issues with frequency specifications strongly related to the codon frequencies in the data set. Issue is more egregious when there are relatively few codons, based on entropy. As entropy increases (more permissive, and thus data set codon frequencies are flatter), the error decreases and ML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

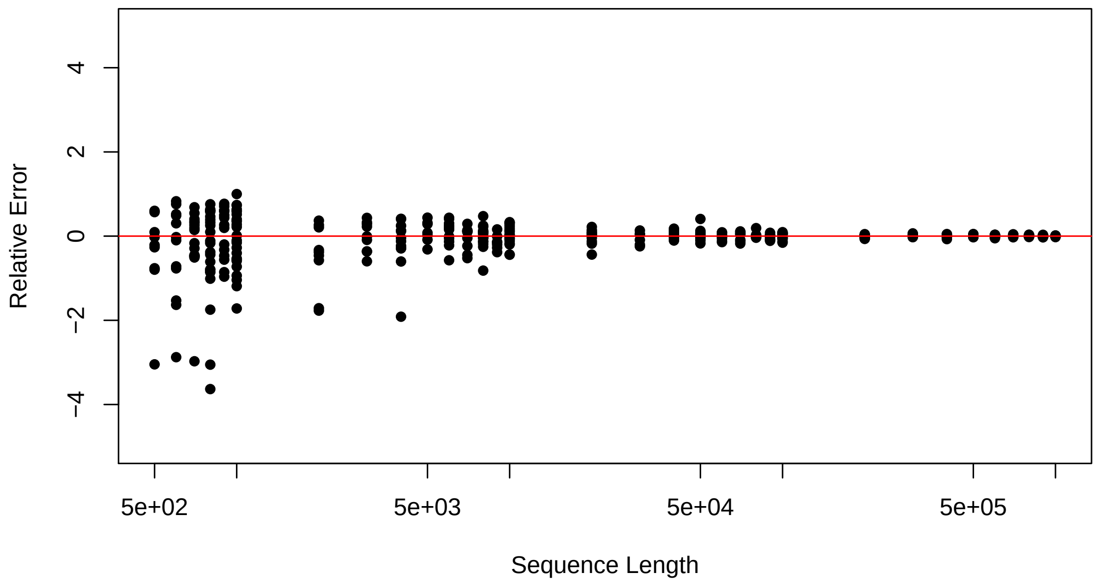


Figure 5: Convergence of derived $dN/dS$ and ML estimates of $dN/dS$. Each point represents results from a single simulation. The y-axis indicates relative error of the ML $dN/dS$ estimates, and the x-axis indicates sequence length on a log-scale. As the sequence length, or the data set size, increases,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

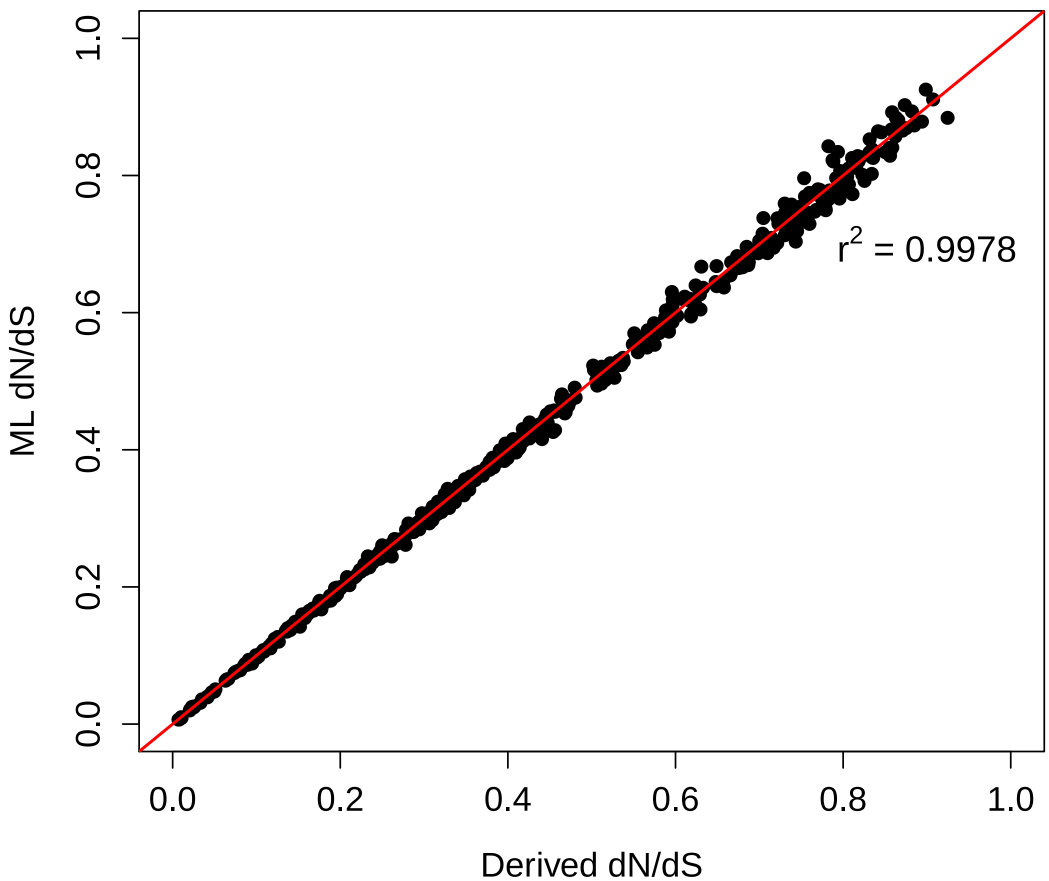


Figure 6: Relationship works exceedingly well. There are 500 points in this plot, each of which corresponds to a single simulation. Beta values for each simulation were randomly chosen between 0.5-3.5, so this plot does contain varying levels of selective constraint on amino acid distributions. Note 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.