The relationship between dN/dS and scaled selection coefficients

Stephanie J. Spielman[1] and Claus O. Wilke[1]

Address:
[1]Department of Integrative Biology, Center for Computational Biology and Bioinformatics, and Institute of Cellular and Molecular Biology. The University of Texas at Austin, Austin, TX 78712, USA.

*Corresponding author
Email: ??????????

Keywords: mutation-selection-balance models, mechanistic codon models, dN/dS, scaled selection coefficients, natural selection, protein evolution, models of sequence evolution

Abstract

Two models which investigate strength of selection in protein-coding sequences are mechanistic codon models and mutsel models. We have measures of dN/dS and amino acid/codon "propensities", which correspond to equilibrium frequencies. Are they the same? Are they different? We don't know! But now we do. And they're the same. This approach is really nice because it allows us to uncover properties of these metrics previously unidentified, etc. We found that codon-level models and metrics do not play nicely with amino-acid level models, and we found some interesting behaviors of the M0 model. Our study represents a very useful strategy - benchmarking and investigating model behavior be examining the intersection/relationship between distinct approaches.

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, diversifying selection ($dN/dS > 1$) [1–4]. Following early counting methods for estimating dN/dS (e.g. refs [5] and [6]), mechanistic codon models, which assume an explicit Markov-process model of sequence evolution (see ref. [7] for a comprehensive review), have taken a leading role as the inference method of choice since their introduction in the 1990s [1, 8, 9]. These models yield maximum likelihood estimates (MLEs) for the parameter ω, which represents the quantity dN/dS, and have seen great success in the field of molecular evolution.

A second class of models, known as mutation-selection-balance (MutSel) models, have emerged recently as a popular alternative to mechanistic codon models. The MutSel framework, couched firmly in population genetics theory, models the dynamic interplay between mutation and selection in a protein-coding sequence. MutSel models yield estimates of site-wise scaled selection coefficients, which indicate the extent to which natural selection favors, or disfavors, particular codons or amino acids at a given protein position. Although MutSel models were first introduced over 15 years ago [10], they have seen virtually no use due to their high computational expense. Recently, however, several computationally tractable model implementations have emerged [11, 12], allowing for the first time the potential for widespread use.

Although both mechanistic codon 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. Whether dN/dS values have any correspondence with scaled selection coefficients remains an open question. Therefore, while certain rhetorical arguments may be made in favor of using one method over another, there is currently no formalized, concrete rationale to guide researchers in their methodological choices.

Here, we formalize the relationship between mechanistic codon 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 between these models' primary parameters, allowing us to precisely infer dN/dS values from scaled selection coefficients. Using a simulation approach, we verify that these derived dN/dS values correspond precisely to ω MLEs inferred using standard mechanistic codon models.

Methods

We simulated protein-coding sequences as a continuous-time Markov process [13] according to the MutSel model proposed by [10]. This model's instantaneous rate matrix Q is given by

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

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

$$f_{ij} = \frac{2Ns_{ij}}{1 - e^{2Ns_{ij}}}, \quad (2)$$

where the value $2N_e s_{ij}$ represents the scaled selection coefficient for a mutation from codon i to codon j [10,14]. As shown by [10], the fixation probability

$$f_{ij} \propto ln\left(\frac{\pi_j \mu_{ij}}{\pi_i \mu_{ji}}\right) \Big/ \left(1 - \frac{\pi_i \mu_{ji}}{\pi_j \mu_{ij}}\right). \quad (3)$$

In this approximation, π_i is the steady-state, or equilibrium, frequency of codon i. Importantly, these equilibrium frequency values are those which result from the joint effects of both mutation and selection.

All alignments presented here were simulated along a 4-taxon phylogeny, beginning with a root sequence selected using steady-state codon frequencies. Unless otherwise stated, all simulated alignments contained 500,000 codon positions. A single evolutionary model was applied to all positions in the simulated sequences. While this lack of site-wise heterogeneity is unrealistic for real sequence evolution, it allows us to verify our derived relationship between equilibrium codon frequencies and dN/dS with a sufficiently sized data set.

To demonstrate the relationship between dN/dS and scaled selection coefficients, we simulated 100 sequences in which all synonymous codons have equal fitness (no codon bias), and 100 alignments in which synonymous codons featured different equilibrium frequencies (codon bias). For both sets of simulations, we assumed symmetric nucleotide mutation rates of $\mu_{xy} = 10^{-6}$ and $\kappa \sim \mathcal{U}(1, 6)$. We generated relative amino acid scaled selection coefficients S_a for each simulation, by fixing one coefficient to 0 and drawing the remaining 19 values from a normal distribution $\mathcal{N}(0, \sigma^2)$, where $\sigma^2 \sim \mathcal{U}(0, 4)$. Here, σ^2 effectively represents the strength of natural selection; larger values of σ^2 will correspond to greater fitness differences among amino acids, and thus more selective pressure. Moreover, these S_i values correspond to the relative amino acid fitness parameters as inferred by currently available MutSel inference methods [11, 12]. For simulations without codon bias, we directly assigned S_a values to codons such that all synonymous codons had the same scaled selection coefficient, and thus the same fitness. For simulations with codon bias, we randomly selected a preferred codon for each amino acid. We then assigned the preferred codon a selection coefficient of $S_a + \lambda$ and all non-preferred codons a selection coefficient of $S_a - \lambda$. For each codon bias simulation, we drew λ from $\mathcal{U}(0, 2)$.

Finally, we computed equilibrium frequencies for all codons according to a Boltzmann distribution,

$$\pi_i = \frac{e^{S_i}}{\sum_k e^{S_k}}, \quad (4)$$

where the denominator runs over all sense codons. Equation (4) directly relates codon equilibrium frequencies and ... using theory developed by Sella and Hirsh [15] Moreover, according to theory

developed by Sella and Hirsh [15], these equilibrium frequencies are directly related to scaled selection coefficients according to

We calculated a global dN/dS for each alignment using the mathematical framework outlined in (4)–(9) as well using standard maximum likelihood methods. Specifically, inferred dN/dS using the M0 mechanistic codon model [2], as implemented in the HyPhy batch language [16]. The M0 models uses the GY94 instantaneous rate matrix [1, 8], which includes the primary parameters ω, κ, and equilibrium codon frequencies. For simulations inferences, we inferred ω both by fixing κ to its true value, and maintaining κ as a free parameter of the model. We used the Fequal equilibrium codon frequency model parameterization, which assigns equal frequencies of 1/61 to all sense codons [13]. Codon frequency parameters, unlike the steady-state codon frequencies of the underlying evolutionary model, are meant to capture mutational biases, and these parameters should correspond to the equilibrium codon frequencies which would be expected in the absence of selection [17], and a symmetric mutation process would produce equal frequencies of 1/61.

Additionally, we simulated alignments which made use of experimentally-determined amino acid fitness and mutation rate data. We used site-wise influenza nucleoprotein (NP) amino acid preferences from Bloom 2014 [18] and nucleotide mutation rates for either NP [18], yeast [19], or polio virus [20]. Note that all of these experimental mutation rate matrices were asymmetric. We combined each the 498 amino acid preference distributions with each set of nucleotide mutation rates to determine a total of $493 \times 3 = 1494$ unique experimental evolutionary Markov models, using the approach in Bloom [18], wherein the Metropolis acceptance criterion [21] was used to calculate amino acid fixation rates. We calculated each model's equilibrium, or steady-state, codon frequencies such that detailed balance $\pi_i \mu_i j = \pi_j \mu_j i$ and $\sum \pi_i = 1$, where the sum runs across all 61 sense codons, was satisfied. Finally, for each set of equilibrium codon frequencies, we simulated alignments according to equation (1).

We inferred ω for the simulations which employed experimental data with 5 different M0 model parameterizations. All inferences considered *kappa* a free parameter of the model, but 5 different equilibrium codon frequency parameterizations were used. First, we inferred ω using the Fequal [13] parameterization, which assigns equal codon frequencies of 1/61 each. Second, we inferred ω by specifying codon frequencies which would arise strictly from mutational processes in the absence of natural selection. We computed these codon frequency values using the same approach as we did in calculating the true steady-state codon frequencies, except instead of using the experimental amino acid preference data, we all amino acids the same preference value of 0.05, thus eliminating any amino-acid level fitness differences. We term this frequency parameterization "Fnull." Finally, we used the common frequency estimators F3x4 [9], CF3x4 [22], and F61 [8]. As typical analyses consider model frequency parameters as protein-wide, not site-specific, parameters, we computed these parameter values by pooling, for each set of mutation rates, all 498 steady-state codon frequencies to derive average codon frequencies. This approach yielded three distinct sets of averaged codon frequencies, from which we directly calculated the parameters for F3x4, CF3x4, and F61.

Results

Mathematical relationship between dN/dS and scaled selection coefficients

We describe here how to calculate dN/dS from scaled selection coefficients. 1. can report all scaled selection coefficients as relative, as is done by MutSel implementations. 2. In the presence of a symmetric mutation scheme, we can derive freqs precisely using Sella and Hirsh theory. 3. Other-

wise, we can easily determine frequencies from the scaled amino acid or codon fitness parameters in a MutSel model using equations in methods

The fixation probability for a mutation from codon i to codon j is [10, 15]

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

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 [8, 13] as

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

and thus we find that

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

Similarly, for dS, the synonymous evolutionary rate K_{S} per synonymous site L_{S}, we find

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

where $\mathcal{S}_i$ is the set of codons that are synonymous to codon i and differ from it by one nucleotide substitution. The quantities K_{S} and L_{S} are defined as in Eqs. (6) and (7) but summing over $j \in \mathcal{S}_i$ instead of $j \in \mathcal{N}_i$.

Equations (5)–(9) establish a connection between the equilibrium codon frequencies and the evolutionary rate ratio dN/dS. Moreover, we note that, if we assume that all synonymous codons have equal fitness (e.g. synonymous mutations are neutral), the synonymous fixation rate $f_{ij} = 1/N_e$ [23]. Under this circumstance, the value for dS reduces to 1.

dN/dS can be accurately predicted from scaled selection coefficients

To validate the mathematical relationship between steady-state codon frequencies and dN/dS described in equations (5)–(9), we simulated protein-coding sequences along a 4-taxon phylogeny according to a mutation-selection model framework [10, 15]. We simulated 100 alignments in which synonymous codons had equal fitness values, and 100 alignments with codon bias, e.g. where the fitness values, and hence equilibrium frequencies, differed among synonymous codons (see Methods for details). All simulations assumed an underlying symmetric nucleotide rate matrix, with the transition-tranversion bias ratio $\kappa \sim \mathcal{U}(1, 6)$. For each alignment, we calculated dN/dS using equations (5)–(9) as well as using the M0 mechanistic codon model [1], as implemented in the HyPhy batch language [16].

The relationship between dN/dS measurements is shown in Figure 1A (for simulations with no codon bias) and Figure 1B (for simulations with codon bias). It is clear that dN/dS values derived using codon frequencies agree nearly perfectly with those inferred using standard maximum likelihood methods, and frequency differences among synonymous codons do not influence this robust relationship. Additionally, in Figure 1C, we demonstrate convergence of dN/dS estimates as the size of the data set, represented by simulated alignment length, increases. Taken together, these results demonstrate that MutSel model parameters fully encapsulate information regarding dN/dS, and that the results from MutSel and mechanistic codon models are in complete agreement.

Moreover, the strength of selection pressure scales fairly well with dN/dS. Figure 2 displays the relationship between dN/dS and the standard deviation, σ^2, of the distribution of amino acid selection coefficients. Higher values of σ^2 indicate larger fitness differences among amino acids, ultimately leading to stronger selection pressure acting on nonsynonymous substitutions. Figure 2 demonstrates that when fitness differences among amino acids are very high, dN/dS takes on lower values, properly reflecting stronger purifying selection. As expected, this trend is more robust for alignments without codon bias (Figure 2A, $r^2 = 0.83$) than for alignments with codon bias (Figure 2B, $r^2 = 0.45$). This difference emerges from the fact that fitness differences among synonymous codons will obscure the underlying amino acid fitness differences.

Importantly, Figure 2A shows that, in the limiting case when σ^2 approaches 0, and thus amino acids have virtually the same fitness values, dN/dS converges to a value of 1. This result properly reflects the case of neutral evolution. In fact, in **SI proof**, we prove that, when synonymous codons have equal fitness values, dN/dS is necessarily always less than or equal to 1. This restriction does not, however, hold in the face of codon bias, which can readily yield dN/dS values greater than 1 (Figures 1B and 2B), even though the protein sequence is evolving under equilibrium conditions. We discuss the implications of these findings in depth in *Discussion*.

Insights into behavior of dN/dS metric

Use of realistic data is also ok, and other exciting subsection headers

Results reported in the previous subsections were obtained from fully-simulated equilibrium codon frequencies, along with a symmetric mutation matrix. The latter assumption of a symmetric mutation process (e.g. where $\mu_{xy} = \mu_{yx}$) may not be entirely realistic. Indeed, mutational bias, typically $C/G \rightarrow T/A$, is known to contribute to biased nucleotide compositions [19, 20, 24, 25]. Therefore, we performed additional simulations which made use of realistic amino acid fitness and nucleotide mutation parameters. In particular, we used influenza nucleoprotein (NP) site-specific amino acid preference values, given by Bloom [18]. These data consisted of experimentally-determined fitness values for each individual amino acid across all sites in NP, yielding 498 distinct amino acid propensity distributions. We combined these experimental fitness parameters with three sets of experimentally determined mutation rates, either for NP [18], yeast [19], or polio virus [20]. Importantly, all of these mutation matrices are asymmetric, but feature differing degrees of asymmetry, with NP mutation rates being the most symmetric and polio mutation rates the most asymmetric. For each of the 498 amino acid fitness distributions, we calculated steady-state codon frequencies π_i under detailed balance, such that the relationships $\pi_i \mu_{ij} = \pi_j \mu_{ji}$ and $\sum \pi_i = 1$, where the sum runs across all 61 sense codons, were satisfied, as described in [18, 26].

For each resulting set steady-state codon frequencies, we again computed dN/dS using equations (5)–(9) and simulated alignments using equation (1). We inferred ω MLEs using the M0 mechanistic codon model for each alignment according to five different codon frequency model parameterizations. These parameterizations included Fequal [13] and the common frequency es-

timators F3x4 [9], CF3x4 [22], and F61 [8]. Additionally, we inferred ω using a fifth frequency parameterization which consisted of the codon frequencies which would arise strictly from mutational processes in the absence of natural selection. As this specification is the intended purpose for this parameter, we term it Fnull.

Resulting ω inferences correlate extremely well with dN/dS values, but, as expected, the strengths of these correlations differ among codon frequency model parameterizations. Figure **??** shows the resulting relationships between dN/dS values and ω MLEs for each set of mutation rates (NP, yeast and polio), across M0 model codon frequency parameterizations. Figure **??**A displays the bias, or systematic deviation from a 1:1 relationship, between dN/dS and ω, and Figure **??**B displays r^2 values between dN/dS and ω. Several trends emerge from these results. Note that a bias of 0 indicates that there is a perfect correlation between dN/dS and ω MLEs. 1. Fnull and Fequal give ω MLEs with the overall best equivalence to dN/dS values. Moreover, Fnull has the least noise but Fequal shows substantial noise, particularly as the mu's become increasingly asymmetric. Thus as expected, Fnull performed overall the best. 2. the estimators do a fairly good job of approximating Fnull. F3x4 and CF3x4 perform very similarly, but F61 performs marginally worse, although particularly for NP. This is an illustrative case - ILLUSTRATIVE: Interestingly, F61 for NP mutation rates performs relatively worse than others. We attribute this to the fact that these mutation rates were only slightly asymmetric, with the average factor between muxy/muyx=1.04 or something tiny. Thus, these F61 frequencies inadvertetly captured selective effects, and thus omega didn't end up representing dN/dS. Instead, as some info about selection was encapsulated in the freq params, there was less selection to place into omega and it came back artificially elevated, clearly see in the A panel where the bias shows that omega was an overestimate and higher dnds=weaker selection pressure. 3. Performance generally goes down as mutation rates become increasingly asymmetric. This suggests that either our dN/dS calculations are wrong for asymmetric, or that ω is wrong. More likely the second, as we explicitly include all mu's in our dN/dS calculation. It seems that the codon frequency parameters which are supposed to capture this asymmetric process

It's also really nice to see that bias for the NP is the highest. This is because np had the least amount of asymmetry in its mutation rates, and thus the frequencies more closely represent the actual fitnesses. Thus, all selection info is inside the F61, effectively a total model misspecification. While null performs the best, the ω -dN/dS show a marked decrease as asymmetry increases. This suggests that the codon frequency parameters do not entirely capture what they are meant to. Remember, the only way for ω to equal dN/dS is if the rest of the parameters are correct. Otherwise, it represents some quantity ω. The asymmetry-induced trends are also apparent in the relationship between codon entropy and dN/dS. Recall, when mutation rates are symmetric, equilibrium codon frequencies correspond precisely to codon fitness values [15], and thus selective strength. The relationship between entropy and dN/dS gets noisier as mutation rates become more asymmetric. This result perfectly reflects the dual influences of selection and mutation on equilibrium frequencies. The more that uneven mutation rates influence frequencies, the more noise there is.

Discussion

The oldest and most-widely used method to infer selection pressure in protein-coding genes calculates the evolutionary rate ratio of non-synonymous (dN) to synonymous (dS) substitution rates. In turn, dN/dS is commonly used to identify proteins or protein sites that experience negative selection ($dN/dS < 1$), evolve neutrally ($dN/dS \approx 1$), or that experience positive, diversifying selection ($dN/dS > 1$) [1–3]. By contrast, MutSel models equilibrium amino acid and/or codon steady state

frequencies and selection coefficients [**?**,10,12,27,28], for codons [14], or for both. Thus, while mechanistic codon models describe the how quickly a protein's constituent amino acids change, MutSel models calculate the strength of natural selection operating on the specific amino-acid changes.

Until now, however, it has been an open question how these two modeling frameworks relate to one another. Some have argued that MutSel models, given their firm grounding in population genetics theory and attention to site-specific amino acid fitness differences, offer a more fine-grained approach to studying protein evolution than do mechanistic codon models [10, 27]. Recent phylogenetic studies have also demonstrated that evolutionary models which explicitly consider amino acid fitness values offer dramatic improvements over other models, including mechanistic codon models, suggesting that MutSel models may more aptly represent the process of coding-sequence evolution [18, 26].

Here, we have derived a formal mathematical relationship between the quantities dN/dS and scaled codon selection coefficients, the primary parameters of mechanistic codon and MutSel models, respectively. Through a simulation approach, we find that these two models are in full agreement, and that the value for dN/dS can be precisely calculated from selection coefficients alone. Furthermore, this mathematical equivalency between selection coefficients and dN/dS values is robust to fitness differences among synonymous codons. However, it is important to note that our implementation of codon bias explicitly assumed that frequency differences among synonymous codons resulted from fitness differences alone. In other words, the sole source of codon bias in our simulations was selection, not mutation. This implementation might not be entirely biologically realistic, as both mutational and selective forces likely contribute to codon bias in real genomes [29–32]. However, the key finding that we present is that fitness differences among synonymous codons do not affect the robust mathematical equivalency between scaled selection coefficients and dN/dS.

Our results rest on the key assumptions that the protein sequence is evolving under steady-state, or equilibrium, conditions, and the nucleotide mutation rates are symmetric (e.g. $\mu_{xy} = \mu_{yx}$). The first assumption recapitulates the population genetics theory behind MutSel models, which assume that selection coefficients remain constant over the phylogeny, and therefore the protein is evolving along a static fitness landscape [10,27,28]. While the latter assumption of symmetric mutation rates may not be biologically realistic **cite papers which show asym mu**, it allowed us to investigate the mathematical relationship between dN/dS and selection coefficients under broad conditions. In particular, we have proven that, when synonymous codons have equal fitness and mutation rates are symmetric, dN/dS will always be less than 1. However, when synonymous codons were allowed to have different selection coefficients, dN/dS can easily be greater than 1, and indeed frequently was (Figures 1B and 2B). In fact, when synonymous codons have different fitnesses, it is possible to have arbitrarily high dN/dS values; in the most extreme case of codon bias, in which only a single codon per amino acid is selectively tolerated, the number of synonymous sites $L_S = 0$, and thus the value for dN/dS approaches infinity. We additionally expect that an asymmetric mutation rate matrix could yield $dN/dS > 1$.

In other words, even if the protein is evolving along to a static fitness landscape, it is possible to that the sequence will feature dN/dS values greater than 1. This finding seems paradoxical to classic interpretations of $dN/dS > 1$. Typically, these values are viewed as hallmarks of positive or diversifying selection, which is assumed to occur when the protein sequence experiences strong selective pressure to change its constituent amino acids. Positive selection, therefore, necessarily implies that the protein does not evolve at equilibrium, but rather a shift in selective constraint caused new amino acids to be favored. However, one must also recognize that the contention that $dN/dS > 1$ represents positive selection assumes that synonymous substitutions are selectively neutral, which is likely not the case when codon bias exists. Thus, what is classically termed positive selection can result simply from strong synonymous fitness differences. Therefore, it is

entirely possible that estimates of positive selection as inferred via dN/dS estimates in species with high levels of codon bias, such as many bacterial, *Drosophila*, or certain mammalian species [30–33], may not be true cases of positive selection, but rather simply signals of strong codon bias.

In proteins evolving under equilibrium, a fundamental assumption of MutSel models, the terms purifying and positive selection may not generally apply. Instead, we suggest that these terms be strictly reserved for the fitness effects of individual amino acid changes, rather than applying them as overall terms for the protein's evolutionary trajectory. Indeed, while it may easily be said that certain amino acid changes are acted on by pos or pur selection, the protein, or residue/position, itself is either evolving under equilibrium or non-equilibrium conditions. Equilibrium evolution can involve pur and pos, but the expectation is that, at the end of the day, it will average out to purifying selection since the fitness landscape is static. Everybody seems to be very imprecise about this. MutSel papers all state that an assumption is purifying selection, and then they go on to spend half the paper identifying instance of purifying vs. positive selection. This is a little bit ridiculous, semantically, so I'm writing something about how we need increased precision in our wording. SellaHirsh theory enters this discussion very well, in that they contend that equilibrium evolution necessarily involves just as many adaptive as deleterious substitution events. Thus, equilibrium evolution cannot be confused with purifying selection, and positive selection does not always represent the introduction of a novel selection pressure.

Maybe future work could get at distinguishing between positive selection when it arises the context of steady-state evolution vs when it is induced by a novel fitness landscape.

As a consequence of our study relating different modeling frameworks, we emphasize the importance of clarity. Here, dN/dS is necessarily less than 1, under the assumption that syn codons have the same fitness and mu's are symmetric (note that sym mu necessarily will produce eq freqs when same fitness). Even so studies have gone out to find instances of positive selection. How do we reconcile this, given the classic dN/dS definitions? We suggest that there needs to be a distinction between equilibrium and non-equilibrium processes. Both can experience pos, pur, neutral, but noneq comes along with changes in fitnesses. That one stokes shift paper shows that it might be possible, and certainly adaptive evolution would lead to novel fitnesses for novel amino acids (hughes, single amino acid changes. find more references like hoekstra or rubisco).

Incidentally, our study recovered that mechanistic codon models can produce strongly biased inferences when parameters are incorrectly specified. In particular, ω MLE values only corresponded to the true dN/dS value when the equilibrium codon frequency parameters were specified as equal (e.g. each codon had an equilibrium frequency of 1/61). Alternatively, the common approaches of using F61 (empirical) frequencies [8]) or frequency estimators such as F3x4 [9] and CF3x4 [22] always yielded incorrect and highly elevated ω MLEs. We explain this phenomenon by recognizing that the rationale for including codon frequency parameters in mechanistic codon models is to account for unequal nucleotide frequencies specifically caused by mutational and not selective forces [13,17]. The proper values for these parameters, then, should be the codon frequencies which would exist *in the absence of natural selection*. This approach is the only way to ensure that ω is the sole model parameter which contains information about natural selection. Otherwise, the ω parameter will no longer represent the true dN/dS evolutionary rate ratio.

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

proper equilibrium frequency parameterization for our alignments was indeed equal codon frequencies, which would be expected in the absence of natural selection and when mutation rates are symmetric.

These results emphasize that it is crucial to parameterize mechanistic codon models properly. Indeed, their primary parameter ω will only truly represent dN/dS when all other parameters are properly specified. If codon frequencies are not properly specified, which we suspect is the case in most analyses, then the ω MLEs are virtually meaningless and do not represent selective pressure. Therefore, we contend that there is hardly ever a justification to specify empirical codon frequencies, also known as the F61 frequency estimator [8], as natural selection has clearly produced the observed frequencies. Unfortunately, the F61 frequencies are the default parameterization in the widely-used PAML software's codeml implementation [34], so we strongly recommend that users take great care when using this package. In addition, the only robust way to ensure that codon frequencies are properly specified is through experimentally calculating mutation rates. Luckily, this data already exists for a variety of taxa, including **citations for papers uncovering mutation rates**. We recommend that, if experimental data is absent, users err on the side of caution and specify equal codon frequencies to reduce the possibility of false positives.

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

References

[1] 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.

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

[3] Kosakovsky Pond S, Frost SD (2005) Not so different after all: A comparison of methods for detecting amino acid sites under selection. Mol Biol Evol 22: 1208–1222.

[4] Huelsenbeck JP, Jain S, Frost SWD, Kosakovsky Pond SL (2006) A Dirichlet process model for detecting positive selection in protein-coding DNA sequences. Proc Natl Acad Sci USA 103: 6263–6268.

[5] 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.

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

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

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

[9] 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.

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

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

[12] 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.

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

[14] 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.

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

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

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

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

[19] Zhu YO, Siegal ML, Hall DW, Petrov DA (2014) Precise estimates of mutation rate and spectrum in yeast. Proc Natl Acad Sci USA : FORTHCOMING.

[20] Acevedo A, Brodsky L, Andino R (2014) Mutational and fitness landscapes of an RNA virus revealed through population sequencing. Nature 505: 686 – 690.

[21] Metropolis N, Rosenbluth A, Rosenbluth M, Teller A, Teller E (1953) Equation of state calculations by fast computing machines .

[22] 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.

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

[24] Hernandez RD, Williamson SH, Zhu L, D BC (2007) Context-dependent mutation rates may cause spurious signatures of a fixation bias favoring higher GC-content in humans. Mol Biol Evol 24: 2196 – 2202.

[25] Hershberg R, Petrov DA (2010) Evidence that mutation is universally biased towards AT in bacteria. PLoS Genet 6: e1001115.

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

[27] 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.

[28] 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.

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

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

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

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

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

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

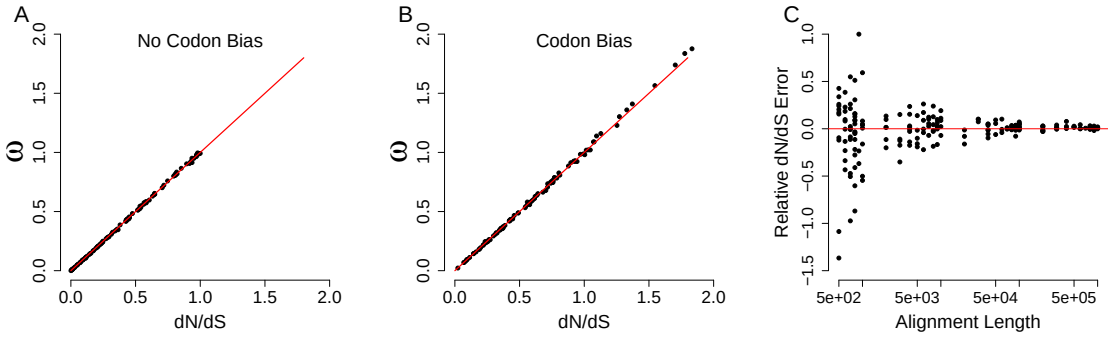

Figure 1: Relationship works exceedingly well. Left panel shows 100 points, each of which corresponds to single simulation. Note that here the ml inference is shown for equal codon frequency specs and kappa fixed to true value (a similar plot for free kappa is shown in suppfigs, but results are qualitatively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

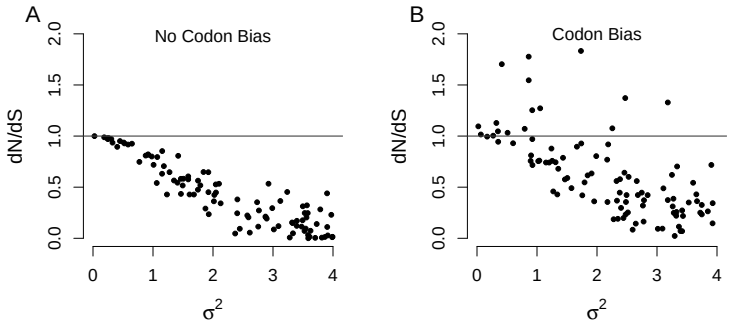

Figure 2: strength of selection scales well with dnds but the strength of the relationship diminishes with codon bias as synonymous now have frequency differences, so dnds is less of a reliable indicator of selection strength.

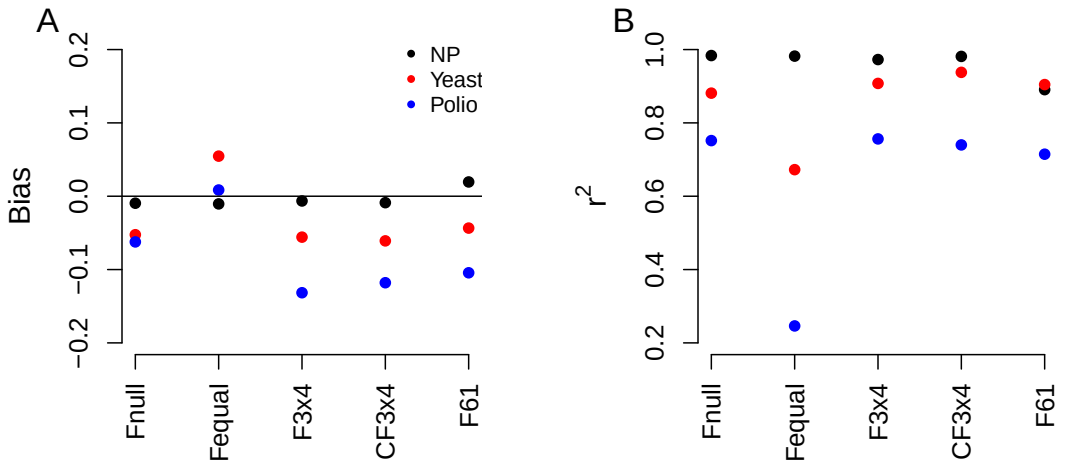

Figure 3: Global outperform, obviously. Clear that the null is probably the best bet. The question is, how well to commonly used estimators approximate this null specification? Decently, but clearly F3x4 and CF3x4 do better than F61. Also, Fequal does reasonably well, but it gets noisy as the asymmetry in mutation rates grows.

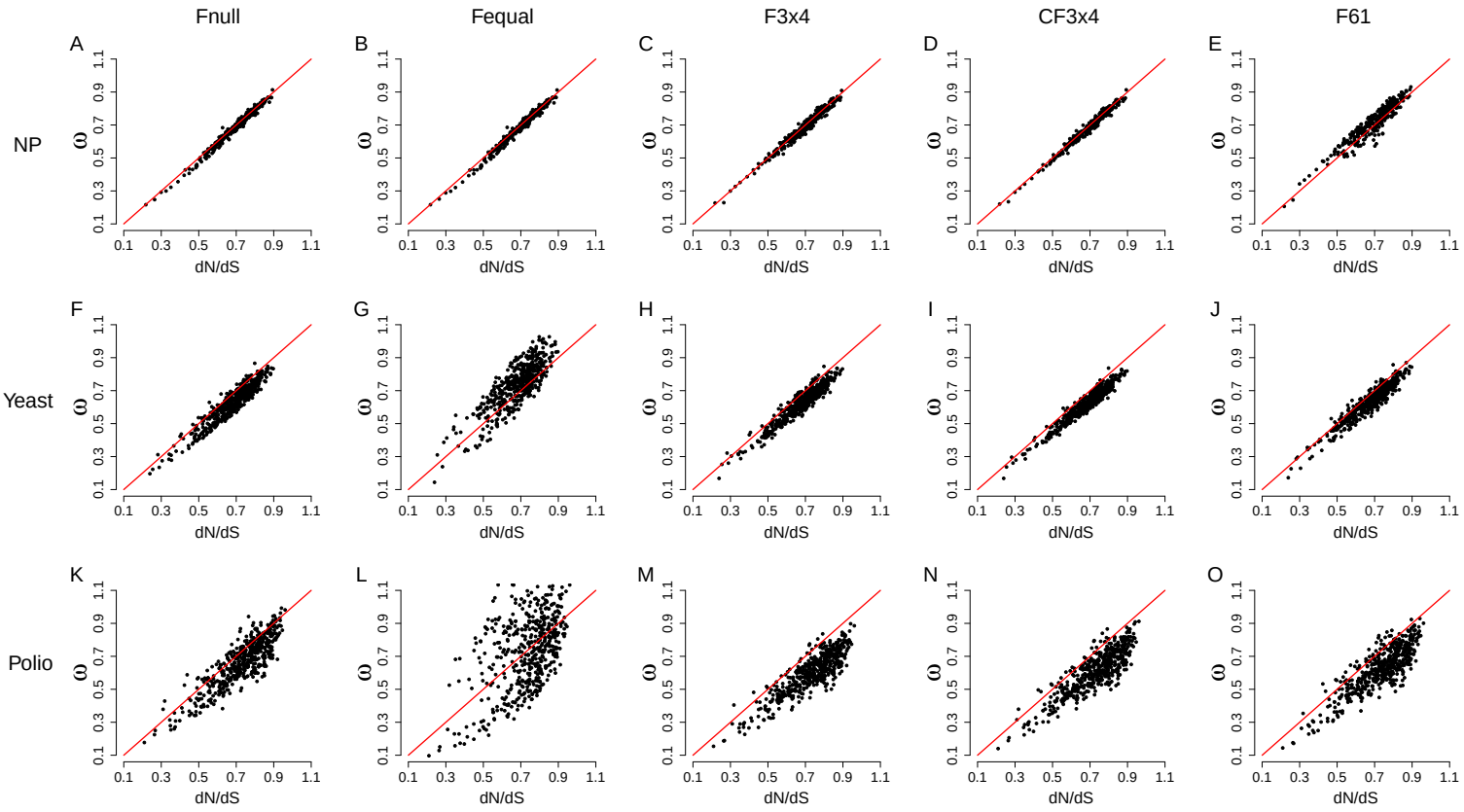

Fig. S1 Omega regression for all ML parameterizations for the np, yeast, and polio mutation rates.