Correcting the Bias of Empirical Frequency Parameter Estimators in Codon Models

Sergei Kosakovsky Pond[1]*, Wayne Delport[2], Spencer V. Muse[3], Konrad Scheffler[4]

1 Department of Medicine, University of California San Diego, San Diego, California, United States of America, **2** Department of Pathology, University of California San Diego, San Diego, California, United States of America, **3** Department of Statistics, North Carolina State University, Raleigh, North Carolina, United States of America, **4** Computer Science Division, Department of Mathematical Sciences, Stellenbosch University, Stellenbosch, South Africa

Abstract

Markov models of codon substitution are powerful inferential tools for studying biological processes such as natural selection and preferences in amino acid substitution. The equilibrium character distributions of these models are almost always estimated using nucleotide frequencies observed in a sequence alignment, primarily as a matter of historical convention. In this note, we demonstrate that a popular class of such estimators are biased, and that this bias has an adverse effect on goodness of fit and estimates of substitution rates. We propose a "corrected" empirical estimator that begins with observed nucleotide counts, but accounts for the nucleotide composition of stop codons. We show via simulation that the corrected estimates outperform the *de facto* standard $F3\times4$ estimates not just by providing better estimates of the frequencies themselves, but also by leading to improved estimation of other parameters in the evolutionary models. On a curated collection of 856 sequence alignments, our estimators show a significant improvement in goodness of fit compared to the $F3\times4$ approach. Maximum likelihood estimation of the frequency parameters appears to be warranted in many cases, albeit at a greater computational cost. Our results demonstrate that there is little justification, either statistical or computational, for continued use of the $F3\times4$-style estimators.

Citation: Kosakovsky Pond S, Delport W, Muse SV, Scheffler K (2010) Correcting the Bias of Empirical Frequency Parameter Estimators in Codon Models. PLoS ONE 5(7): e11230. doi:10.1371/journal.pone.0011230

Editor: Thomas Mailund, Aarhus University, Denmark

Received April 6, 2010; **Accepted** June 1, 2010; **Published** July 30, 2010

Funding: This research was supported by the Joint Division of Mathematical Sciences/National Institute of General Medical Sciences Mathematical Biology Initiative through Grant NSF-0714991, the National Institutes of Health, AI47745 and by a University of California, San Diego Center for AIDS Research/NIAID Developmental Award to S.L.K.P. (AI36214). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: spond@ucsd.edu

Introduction

Virtually all codon models in wide use today (see [1,2] for recent reviews) are members of the class of finite-state, continuous time reversible Markov chains, each defined by an instantaneous rate matrix Q. Transition matrices for finite amounts of time are found via the matrix exponential of Q, so the probability that a position initially occupied by codon I is occupied by codon J after t units of time is $P_{IJ}(t)=\left(e^{Qt}\right)_{IJ}$ (throughout the manuscript we will use upper-case letters to index codons and lower-case letters to index nucleotides). If M is a model in this class, the individual entries of its rate matrix can be written in the canonical form $Q_{IJ}=\theta_{IJ}\pi_J^M$. The θ_{IJ} can be thought of as "rate parameters" that govern the relative rates of substitutions between different codons, while parameters π_J^M induce the equilibrium frequencies of the codons. The choice of π_J^M is the primary distinction between the two popular families of codon models: MG (introduced in [3]) and GY (introduced in [4]). How to best estimate the π_J^M— or more precisely, how to estimate model parameters that actually determine the π_J^M— from sequence alignments is the focus of this note. In order to frame this discussion we need to define what we mean by empirical *frequencies*, model *parameters* and *equilibrium* frequencies (Figure 1). Given an observed alignment, the position-specific empirical nucleotide frequencies, e_a^p where a is a nucleotide (A,C,G,T) and p the codon position (1,2,3), can be estimated directly by counts from the data, and the empirical codon frequencies, e_J, can be estimated by counts as well (the latter gives rise to the F61 codon frequency estimator [4]). Either of these estimates can be used to set model parameters, however typical alignments have insufficient information for the direct estimation of empirical codon frequencies with a sufficient degree of confidence. Rather, the empirical nucleotide frequencies are used to set the nucleotide frequency parameters, ϕ_a^p, and by multiplication of their constituents, the codon frequency parameters, π_J^M. For example, in the original MG94 model of codon evolution [3], the equilibrium frequency of codon $J=xyz$ is given by $\left(\phi_x\phi_y\phi_z\right)/\left(1-\Pi_{stop}\right)$, where $\Pi_{stop}=\phi_T\phi_A\phi_G+\phi_T\phi_A\phi_A+\phi_T\phi_G\phi_A$. A common extension of this model, referred to as MG94 F3×4, allows the three codon positions to have their own nucleotide frequency parameters and leads to equilibrium codon expressed as:

$$\pi_{xyz}=\left(\phi_x^1\phi_y^2\phi_z^3\right)/\left(1-\Pi_{stop}\right). \qquad (1)$$

In this expression the superscripts indicate the position, and the equation for Π_{stop} is modified in the obvious way. If we set all of the model nucleotide frequency parameters to be equal, i.e. $\phi_a^p=0.25$, the result is equal equilibrium frequencies for all codons, i.e. $\pi_J=1/61$ for all J. This vector of codon equilibrium frequencies allows us to easily tabulate, via marginalization, the equilibrium

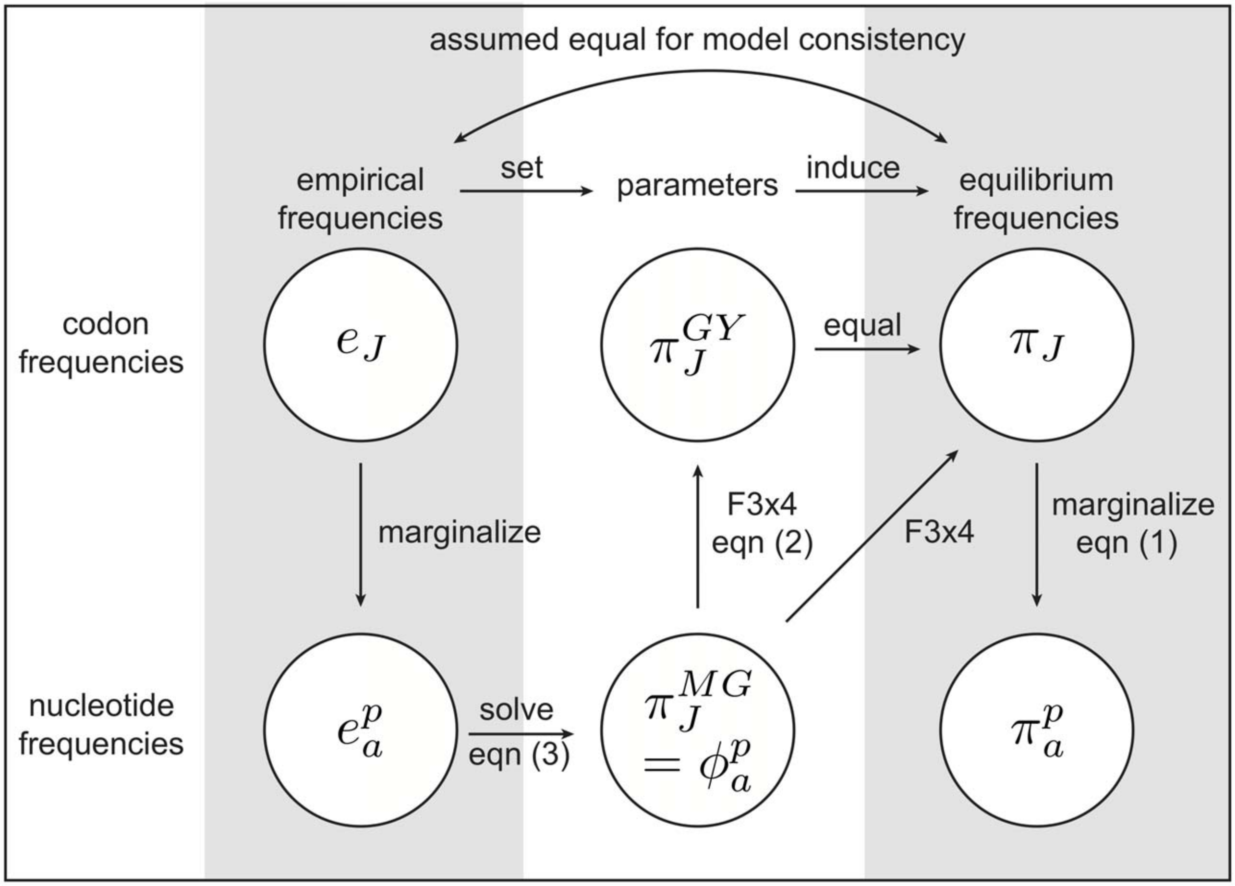

Figure 1. Relationships between empirical frequencies, frequency parameters and equilibrium frequencies in codon models.
doi:10.1371/journal.pone.0011230.g001

frequencies of each nucleotide at each position:

$$\frac{1}{61}\begin{pmatrix} A:16 & 14 & 14 \\ C:16 & 16 & 16 \\ G:16 & 15 & 15 \\ T:13 & 16 & 16 \end{pmatrix} = \begin{pmatrix} A:0.262 & 0.230 & 0.230 \\ C:0.262 & 0.262 & 0.262 \\ G:0.262 & 0.246 & 0.246 \\ T:0.213 & 0.262 & 0.262 \end{pmatrix}. \tag{2}$$

Note that there are only 13 occurrences of T in the first position, 14 of A in the second position, etc because the model explicitly disallows (TAG,TAA,TGA) as is standard for all other codon models. The finding from this exercise is that when one sets all the $\phi_a^p = 0.25$, each of the codon equilibrium frequencies, π_J takes the anticipated value of 1/61. However, remarkably, the equilibrium nucleotide frequencies generated by this model are *not* the anticipated 0.25. For instance, the equilibrium frequency of A at the first position is $1/61 \times 16 = 0.262$. Traditionally, the empirical nucleotide frequencies are used to set nucleotide frequency parameters, and it is therefore assumed that the induced equilibrium nucleotide frequencies are equal to those observed in the alignment. However, given that the nucleotide composition of stop codons is not accounted for, this practice is flawed, because $\phi_a^p \neq \pi_a^p$. The conflation of frequency parameters (ϕ_a^p) and equilibrium nucleotide (π_a^p) frequencies results in incorrect estimates of equilibrium nucleotide (and codon) frequencies as demonstrated in (2) above. This phenomenon is not restricted to the MG family of models. It is simple to demonstrate the exact same behavior for the GY family of models, again because of the incorrect designation of nucleotide frequency parameters in the rate matrix as equal to empirical nucleotide frequencies. We show that the traditional identification of frequency parameters and observed nucleotide frequencies leads to a cascade of problems. Model frequency parameters are estimated with bias, which leads to biased estimation of the equilibrium codon frequencies, which leads to compensatory biased estimation of the substitution rate parameters. We propose a correction, and a maximum likelihood frequency parameterization and show that both these approaches are not similarly biased, and therefore advocate their use in codon models.

Materials and Methods

To ensure clarity of presentation, we first carefully introduce the necessary notation (summarized in Figure 1). For a given substitution model, let π_J be the frequency of sense codon J ($J = 1,2,3,\ldots,61$) in its equilibrium distribution, and π_a^p, $a = 1,2,3,4$ be the equilibrium frequency of nucleotide a in codon position $p = 1,2,3$. When necessary, we will indicate specific models via a superscript (ie, MG or GY). The position specific nucleotide equilibrium frequencies, π_a^p, are uniquely determined by the codon equilibrium frequencies, π_J, through marginalization, e.g. π_T^1 is simply the sum of frequencies of the 13 sense codons that have a T in their first position, e.g. as in equation (2).

These equilibrium frequencies, of both nucleotides and codons, have traditionally been assumed equal to empirical frequencies observed in a sequence alignment, e_J or e_a^p, and used to set model parameters. If the specified model is correct, e_J converges to π_J and e_a^p to π_a^p as the sequence length N increases. (However, note that this result requires that the evolutionary process itself be at equilibrium; many important biological mechanisms— notably directional positive selection— are likely to disrupt equilibrium; see [5–7]).

Because the simple example in equation (2) demonstrated that the empirical and equilibrium nucleotide frequencies are not synonymous, we strive to obtain an expression that relates the equilibrium nucleotide frequencies to the model nucleotide frequencies, ϕ_a^p, and through extension –to the observed empirical frequencies. Even though the MG and GY models treat equilibrium codon frequencies differently, it is a fortunate coincidence that in either case the π_J have identical forms when written in terms of ϕ_a^p. Given twelve MG nucleotide frequency parameters, only 9 of which are independent because $\sum_a \phi_a^p = 1$ for each position p, the equilibrium frequency of codon $J = xyz$ induced by their values is as in equation (1).

By using e_a^p to directly estimate ϕ_a^p in equation (1), one obtains the popular $F3\times4$ estimator of codon equilibrium frequencies – by far the most common estimator used in literature for both MG and GY classes of models. The statistical and computational appeal of $F3\times4$ lies in its use of only 9 nucleotide parameters to describe 61 codon frequencies. However, the key shortcut— direct estimation of nucleotide frequency *parameters* with empirical nucleotide *frequencies* from the data— is flawed. The empirical nucleotide frequencies *are* unbiased estimates of the true equilibrium frequencies; unfortunately, the model parameters they are being used to estimate are something different. Thus, a fundamental problem with current practices is that use of the $F3\times4$ estimators with either MG or GY models leads to biased estimates of the ϕ_a^p, and in turn the π_J. As we will show below, the problems do not end there, and lead to biased estimation of other model parameters.

We first present two approaches for correcting these estimation errors. The obvious, but more computationally demanding method is to estimate the ϕ_a^p by maximum likelihood along with other model parameters. We dub this approach $MLF3\times4$. Theory suggests that estimates from this methodology will have all the desirable properties of maximum likelihood estimation. Maximum likelihood estimation of these values has been available in some software packages, e.g. in HyPhy [8], for a number of years, but to our knowledge it has rarely been used.

The second strategy, described here for the first time, relies on finding an expression for the induced equilibrium frequency of nucleotide a at codon position p (π_a^p) as a function of ϕ_a^p. Since the

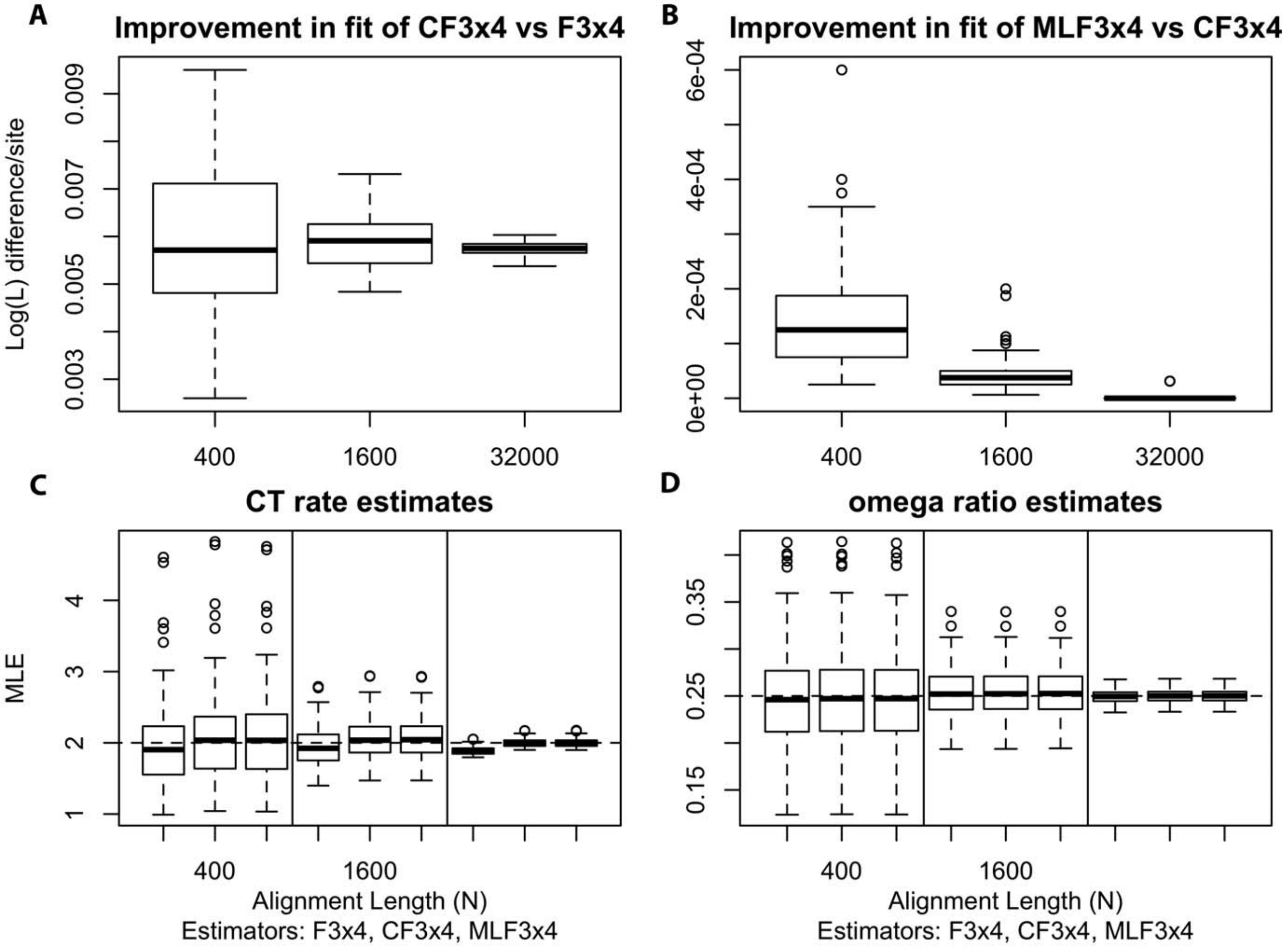

Figure 2. Comparison of frequency parameterizations fitted to simulated alignments. The top row (A,B) shows the comparison of $\log L$ scores on simulated data obtained with different corrected frequency estimates; C) Bias in the estimate of the substitution rate $\theta_{CT}=2.0$ in near-asymptotic regime ($L=32000$) is apparent under $F3\times4$, but does not exist for the other two estimators; D) variance of the $CF3\times4$ estimate for θ_{CT} is reduced with increasing sample size.
doi:10.1371/journal.pone.0011230.g002

ϕ_a^p define codon equilibrium frequencies (equation 1), we can readily obtain such equations by marginalization:

$$\begin{aligned}\pi_a^1 &= \phi_a^1\left(1-\sum\nolimits_{ayz\in X}\phi_y^2\phi_z^3\right)/(1-\pi_X)\\ \pi_a^2 &= \phi_a^2\left(1-\sum\nolimits_{xaz\in X}\phi_x^1\phi_z^3\right)/(1-\pi_X)\\ \pi_a^3 &= \phi_a^3\left(1-\sum\nolimits_{xya\in X}\phi_x^1\phi_y^2\right)/(1-\pi_X).\end{aligned} \quad (3)$$

Here, $1-\pi_X$ is simply scaling for the absence of stop codons: $\pi_X=\sum_{xyz\in X}\pi_x^1\pi_y^2\pi_z^3$, and $X=\{TAA,TAG,TGA\}$ defines the set of stop codons. The *corrected* $F3\times4$, or $CF3\times4$ estimator equates π_a^p with observed nucleotide frequencies e_a^p, and then solves the nonlinear system (3) for ϕ_a^p to obtain estimates of the latter. Because $\sum_{a=1}^4\phi_a^p=\sum_{a=1}^4\pi_a^p=1$, the above system of 12 nonlinear equations relate 9 independent observed statistics (e_a^p, *e.g.* for $a\in\{A,G,T\}$) with 9 independent model parameters ϕ_a^p. We were unable to obtain a closed form solution to the system, but it can be easily solved numerically at a negligible computational cost.

We conducted simulations to further investigate the effects of biases in the equilibrium frequencies on parameters typically estimated using phylogenetic models. We generated two-sequence codon alignments with uniform codon frequency composition ($\phi_a^p=0.25$). We used $\theta_{AC}=0.5$, $\theta_{AG}=1$, $\theta_{AT}=0.8$, $\theta_{CG}=0.3$, $\theta_{CT}=2.0$, $\theta_{GT}=0.1$ as substitution bias parameters in the MG94xREV model [9], and set the nonsynonymous/synonymous substitution rate ratio ω to 0.25. The two sequences were 10% divergent on average, and the length of the alignment, N, was one of 400, 1,600 or 32,000 codons. 100 replicates were generated for each value of N. We compared the fits of $F3\times4$, $CF3\times4$ and $MLF3\times4$ on simulated data sets, and furthermore compared simulated to inferred parameter estimates with each of the three frequency parameterizations. In addition to the simulated data, we fitted all three frequency parameterizations to a sample of 856 alignments from the carefully curated Pandit database [10]. All alignments were chosen to contain between 10 and 20 sequences and at least 200 reliably aligned codon sites. Given that each estimator has the same number of independent parameters (9), an improvement in log-likelihood under one of the models is considered as evidence in favor of the better fitting model, *e.g.* under the BIC [11] criterion. All new estimators for the MG94 class of models are implemented in HyPhy.

Results and Discussion

We simulated data with a uniform codon frequency composition and fitted all three frequency parameterizations for alignments of various sequence lengths. The suboptimal nature of the $F3\times4$ estimator is immediately apparent from Figure 2a, where the improvement in $\log L$ scores of the model equipped with the corrected estimator $CF3\times4$ is shown. For all replicates, the $CF3\times4$ estimator yielded better $\log L$, with median improvements of 2.29, 9.46, and 184 (for 400,1,600, and 32,000 codons respectively), or approximately 0.006 likelihood points per codon site. Note that as the sample size increased, the estimators from (3) effectively matched the performance of the maximum likelihood estimator (Figure 2b). Even more importantly, the use of the $F3\times4$ frequency estimator led to biased inference of other model parameters. Maximum likelihood estimates of some substitution rates were biased under the $F3\times4$, and the bias was progressively more pronounced with increasing sample size (Figure 2c). Indeed, for $N=32{,}000$, a simple likelihood ratio test rejected the (true) null of $\theta_{CT}=2.0$ at $p<0.05$ for all 100 replicates. Biased MLEs of the substitution rate parameter θ_{CT} is a result of the under/overestimates of π_T^p and π_C^p using $F3\times4$. Similar results were seen for the other θ_{IJ}. To our relief, the maximum likelihood estimate (MLE) for the ω ratio was not noticeably affected even for the largest sample size (mean 0.2494, median 0.2495, IQR $0.2445-0.2539$ under $F3\times4$; mean 0.2500, median 0.2501, IQR 0.2452,0.2545 under $CF3\times4$, Figure 2d).

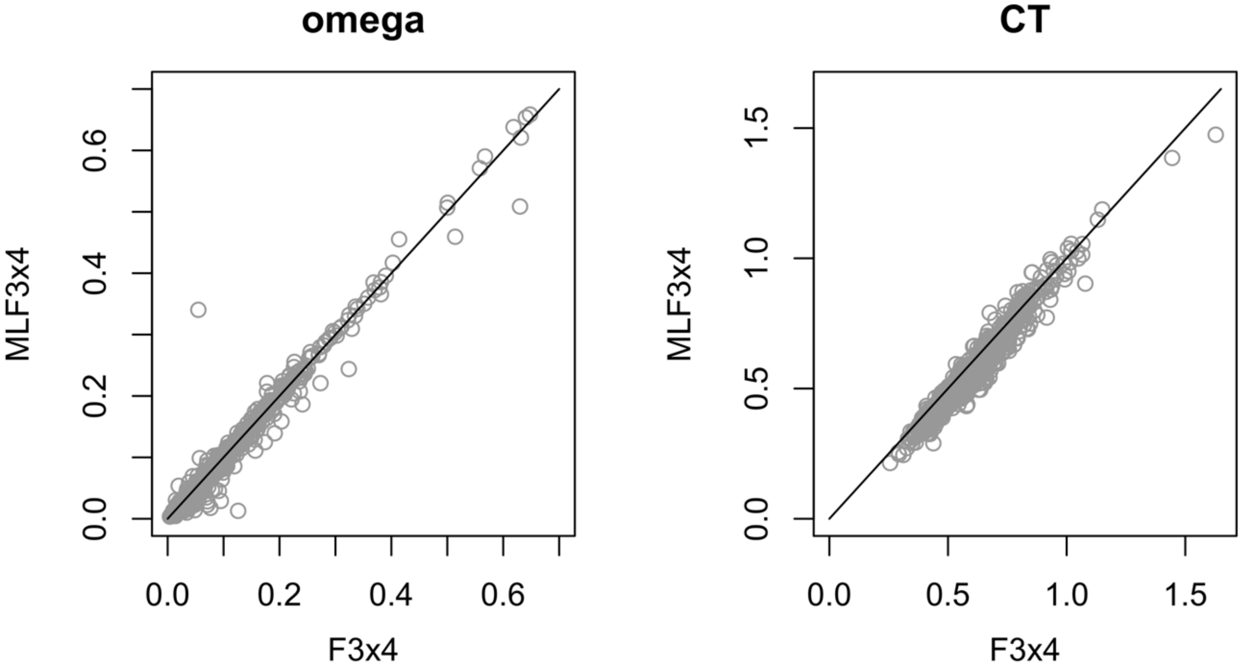

Figure 3. The effect of the frequency estimator on the inference of ω and θ_{CT} (relative to the θ_{AG} rate) substitution rate from 856 alignments sampled from the Pandit database [10]. The estimate of θ_{CT} under $F3\times4$ is biased downwards relative to $MLF3\times4$.
doi:10.1371/journal.pone.0011230.g003

For the Pandit alignments $\log L$ values were, of course, higher for the models estimated using $MLF3\times4$ than for those using $F3\times4$. However, the magnitudes of the differences were impressive (median 17.59, IQR $10.29-27.55$, max 453.2). The $CF3\times4$ estimator improved the $\log L$ score of the $F3\times4$ estimator for over 80%(692/856) of the alignments by a median of 7.4 points; in the remaining cases the median decrease in $\log L$ score was 2.9 points. As with the simulated data, the MLEs of ω were largely unaffected by the choice of frequency estimators (but there were some datasets where the difference was large), while some substitution rate estimates appeared biased (Figure 3). For example, the estimates of θ_{CT} were strongly linearly correlated between $MLF3\times4$ and $F3\times4$ methods ($r^2=0.952$), but the regression line was estimated as $F3\times4=0.073+0.930MLF3\times4$, which recapitulates the downward bias observed on simulated data (if the estimates were unbiased, we would expect an intercept of zero and slope of one).

We have demonstrated through simulations that the almost universally used $F3\times4$ estimator of equilibrium frequencies in codon substitution models is biased, and we have pointed out how a misinterpretation of standard codon model parameters is responsible for these biases. Although this bias appears to have little effect on estimation of "composite" parameters such as the nonsynonymous/synonymous rate ratio (ω) and branch lengths (results not shown), the bias has considerable damaging effects on the estimation of substitution rate parameters in the instantaneous rate matrix. This problem will become acutely relevant as researchers pursue finer-scale studies of the evolutionary process, such as developing substitution models with protein residue-dependent codon substitution rates [12,13]. Since the computational burden of the $F3\times4$ estimator is virtually identical to that of our proposed $CF3\times4$ estimator, which in turn is only marginally faster than $MLF3\times4$, we recommend the use of either of the alternatives offered in this manuscript over the $F3\times4$ estimator. Our current recommendation is to obtain $CF3\times4$ estimates and use them to initialize the optimization procedure for $MLF3\times4$ to speed up convergence.

Author Contributions

Conceived and designed the experiments: SLKP WD SVM KS. Performed the experiments: SLKP. Analyzed the data: SLKP. Contributed reagents/materials/analysis tools: SLKP. Wrote the paper: SLKP WD SVM KS.

References

1. Anisimova M, Kosiol C (2009) Investigating protein-coding sequence evolution with probabilistic codon substitution models. Mol Biol Evol 26: 255–271.
2. Delport W, Scheffler K, Seoighe C (2009) Models of coding sequence evolution. Brief Bioinform 10: 97–109.
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. Goldman N, Yang Z (1994) A codon-based model of nucleotide substitution for protein-coding DNA sequences. Mol Biol Evol 11: 725–736.
5. Seoighe C, Ketwaroo F, Pillay V, Scheffler K, Wood N, et al. (2007) A model of directional selection applied to the evolution of drug resistance in HIV-1. Mol Biol Evol 24: 1025–1031.
6. Kosakovsky Pond SL, Poon AFY, Leigh Brown AJ, Frost SDW (2008) A maximum likelihood method for detecting directional evolution in protein sequences and its application to influenza a virus. Mol Biol Evol 25: 1809–1824.
7. Lacerda M, Scheffler K, Seoighe C (2010) Epitope discovery with phylogenetic hidden Markov models. Mol Biol Evol.
8. Kosakovsky Pond SL, Frost SDW, Muse SV (2005) HyPhy: hypothesis testing using phylogenies. Bioinformatics 21: 676–9.
9. Kosakovsky Pond SL, Muse SV (2005) Site-to-site variation of synonymous substitution rates. Mol Biol Evol 22: 2375–2385.
10. Whelan S, de Bakker PIW, Quevillon E, Rodriguez N, Goldman N (2006) Pandit: an evolution-centric database of protein and associated nucleotide domains with inferred trees. Nucleic Acids Res 34: D327–31.
11. Schwarz G (1978) Estimating the dimension of a model. Ann Stat 6: 461–464.
12. Kosiol C, Holmes I, Goldman N (2007) An empirical codon model for protein sequence evolution. Mol Biol Evol 24: 1464–1479.
13. Conant GC, Stadler PF (2009) Solvent exposure imparts similar selective pressures across a range of yeast proteins. Mol Biol Evol 26: 1155–1161.