Reviewer: 1

Comments to the Author
The authors have made an effort to address the reviewers' comments and can now be accepted.

Reviewer: 2

Comments to the Author
This is a revised version, for which I was not reviewer on the first version. I have evaluated both the manuscript and the response to reviewers.

I would like to note that I am a specialist of molecular evolution but not of applied mathematics, and I did not evaluate the correctness of the mathematics, especially pp 14-17.

I agree with the previous reviewers that the impact of the proposed new model on ω is too small to be relevant in itself. On the other hand, the model does appear to make other interesting contributions than estimation of ω, but these are insufficiently highlighted by the authors. Thus my overall advice is to change the focus of the paper to what this model can contribute, and show how it can indeed help discover new insight into molecular evolution.

Specific examples of utility of the new model which should be highlighted:

p. 12 of the manuscript:
"the mutational bias at the nucleotide-level results in suboptimal amino-acid being overrepresented in the sequence, compared to what would be expected based on their fitness alone."

p. 1 of response to reviewers:
"this novel result is important, as it is reassuring for a certain number of previously published analyses, in particular correlating dN/dS with life-history traits, in a context where GC content also correlates with life-history traits (Figuet et al, 2016; Bolívar et al, 2016)."

And:
"by deriving a codon model where the underlying mutational process and the observed nucleotide frequencies are formally and correctly disentangled, we clarify the problem about 3x4 versus 1x4; which has been a long-standing problem. Many people are still using the 3x4 formalism, in spite of the fact that this has been shown to lead to inaccurate inference. Perhaps one reason is that people don't understand why a model such as 1x4, which does not correctly predict the nucleotide frequencies across positions, might do better than 3x4. In this respect, our work is important, because it gives the first clear explanation of how to correctly formalize this problem."

It is true that the response to reviewers can be leveraged to explain better what this model can contribute, and show how it can indeed help discover new insight into molecular evolution. We updated the discussion:

… current parametric codon models predict that the observed and underlying mutational biases should be equal. For that reason, they are inherently misspecified and are unable to tease apart opposing effects of mutation and selection correctly. As shown inour work, the misspecification of these models does not strongly impact the estimation of the net effect of selection on non-synonymous mutations (ω). This novel result is important, as it is reassuring for a certainnumber of previously published analyses, in particular correlating ω with life-history traits, in a context where GC content also correlates with life-history traits (Figuet et al., 2016; Bolívar et al., 2019). However, current parametric models don't estimate the mutational process accurately.

….

Inferring parameters on simulated alignments, we show that the model derived using this mean-field argument correctly estimates the underlying mutational bias and selective pressure. In this respect, our work gives the first clear explanation of how to correctly disentangle the underlying mutational bias and the observed nucleotide frequencies. Our model can predict the accurate nucleotide composition at first, second and third codon positions, while current parametric model fails to predict them. We argue that parametric codon models using three different mutational processes at the first, second and third coding positions (3x4 formalism) to accommodate for variation in observed nucleotide frequencies is not a theoretically sound modelling. Indeed this variation is an emerging property of the balance between mutation and selection as shown in our work. Moreover, the 3x4 formalism has been shown to lead to inaccurate inference of ω (Spielman and Wilke, 2015). Altogether, we concur in this direction that 3x4 formalism is inaccurate andnot mechanistically sound, and as a result should not be used to estimate ω.

Applying our model to empirical alignments, we also observe that there is a selection differential opposing the mutational bias. This observation also points to a fundamental property of natural genetic sequences, namely that they are not optimized but are the result of interactions between evolutionary forces (Sella and Hirsh, 2005). In the specific case highlighted in this work, the mutational bias at the nucleotide-level results in suboptimal amino-acids being overrepresented in the sequence, compared to what would be expected based on their fitness alone. For example, under a mutational bias toward AT, AT-rich amino acids might not necessarily be the fittest but are excessively generated by the mutational process, resulting in a stronger purifying selection against AT-rich amino acids. This was pointed out previously (Singer and Hickey, 2000), although never directly formalized in a phylogenetic codon model. One important consequence of this tradeoff between mutation and selection is that the

observed higher mean fixation probability toward GC is mimicking the effect of biased gene conversion toward GC (gBGC), although unlike gBGC, the phenomenon described here corresponds to a genuine selective effect.

Other comments:

Please clarify in the Results, Simulation experiments, that simulations were run with other topologies and branch lengths. I was concerned about this until I read the response to reviewers.

We added this precision to the result section:
More analyses are shown in supplementary materials with different sequence length (498, 996, 2490, 4980 and 9960 codon sites), different branch lengths (decreased by a factor 2 and increased by a factor 2, 4, 8) and a different topology (90 mammals). These analyses have shown that the number of sites does not influence the estimator's accuracy for mutational bias (λ), nor for selection pressure (ω).

Figure 2 is very difficult to read. I suggest making the y=x line less strong relative to the actual results lines.

Also it's odd that Figure 2 has numbered panels A-C, but that the top half of the figure isn't included. I suggest either labelling the top "A" and the bottom three panels B-D, or (preferably) separating this into two figures.

As suggested, we separated figure 2 in two figures, which is clearer and reduced the width of the y=x line relative to the actual results lines.

line 191: the *true* codon process doesn't have Q=0 for codons more than one mutation away, be more careful in the writing.

We edited the sentence to:
Specifically, at each site *i*, the underlying codon process is….

It would be interesting to discuss why beta-lactamase has a different bias with MG. Is it related to the strong bias on position 1 (Table 1)?

We added a succinct discussion to the manuscript:
Curiously, in β-lactamase, the MG model estimates a weaker underlying mutational bias than the observed bias (i.e. $AT/GC < \lambda < 1$). This effect could be due to the first, second and third positions having compositional biases in different directions, which is harder to disentangle (table 1).

line 97 "being able" -> "which is able".

line 280, should be "datasets".

line 304, should "amino-acids".

line 326, should be "disentangle".

Absolutely

p. 13 How were the models cited chosen relative to the very large literature on codon models?
"several other codon models (Rodrigue et al., 2008; Kosakovsky Pond et al., 2020) should be included in a broader comparison of the accuracy of the estimation of the underlying mutational bias and strength of selection on protein-coding DNA sequences"

We edited this sentence to:
In addition, several other parametrization of codon models as listed in Rodrigue et al. (2008) and Kosakovsky Pond et al. (2020) should be included in a broader comparison of the accuracy of the estimation of the underlying mutational bias and strength of selection on protein-coding DNA sequences.

In several places a codon or amino-acid is noted as "neutral". It is the change between amino-acids which is neutral. An amino-acid can be "equivalent" to another for fitness purposes.

We edited two instances of such approximations, which now reads:
By modelling fitness at the amino-acid level, we assume that all codons encoding for one particular amino acid are selectively equivalent.

Reviewer: 3

Comments to the Author
The paper is much more clear now and reviews were well-addressed. However, I realize one comment in my initial review was not sufficiently clear and therefore was not fully addressed.

Specifically I refer to my initial #2 review comment - "On page 10 where the authors claim support for their model based on LRT and AIC comparisons with MG94. However, this is potentially problematic, as explained in this paper - https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0011587."

In this revision, the authors have made some very good textual improvements to describe caveats of using AIC/LRT as indicators of a better-performing model. I agree these caveats improve the paper, but this wasn't the main point of my comment, and I apologize if I was not clear enough. My main concern isn't evaluation with AIC/LRT itself, but instead that the specific comparison performed was between the proposed model and the ****simplest MG94.*****

This quoted sentence from the linked paper Delport 2010 is the crux of my point: "We argue that, since the single rate model is so easy to improve upon, new codon models should not be validated entirely on the basis of improved model fit over this model."

All that said, it's not clear what other baseline model the authors might compare to besides the simple MG94 without dealing with an unreasonable amount of computation for a minor point like this. I therefore suggest a very quick minor addition to the paper to state the caveat that comparing the proposed model to the **simplest** model may overstate performance of the proposed model, but the observed performance of the proposed model provides more support for its use on top of the improved fit **compared to the simplest alternative.**

We edited the paragraph to reflect this addition:

The empirical fit to the data between the nested models, using AIC and Likelihood ratio test (Posada and Buckley, 2004), always favors favor the MF model compared to the MG model. Of note, owing to its very strong and unreasonable assumption that the dN/dS is the same across all amino-acid pairs, the MG model is in fact very easy to improve upon (Delport et al, 2010), and thus the higher fit of MF compared to MG is not in itself a very strong argument in favor of the use of MF. However, our simulations suggest that, in spite of the larger estimation error on the individual rates between all pairs of amino-acids on smaller alignments, the estimate of the mutation bias is always reasonably accurate, even on small alignments (supplementary materials).