October 8th, 2018
Department of Ecology and Evolutionary Biology
University of Tennessee
569 Dabney Hall
Knoxville, TN 37996-1610

Dear Dr. Kumar
Editor-in-Chief, *Molecular Biology and Evolution*:

We are pleased to present a revised version manuscript on "*Population Genetics Based Phylogenetics Under Stabilizing Selection for an Optimal Amino Acid Sequence: A Nested Modeling Approach*". We are grateful to the Associated Editor and the two reviewers for their constructive criticisms and positive assessment of our work. Below are our responses to their comments. The reviewer's comments are in bold, and our responses are in plain text underneath.

Sincerely,

Michael A. Gilchrist, on behalf of:
Jeremy M. Beaulieu
Brian C. O'Meara
Russell Zaretzki
Cedric Landerer
Juanjuan Chai

Editorial comments:
Associate Editor
Editors' comments to the author:
Dear authors,

I obtained two reviews from experts in the field. Both the reviewers and I were generally positive, as your approach is relatively original and does not follow the standard modeling scheme of previous related work. As with any new modeling approach, it is your task to convince that this novel approach is indeed beneficial over state-of-the-art alternative modeling choices. This seems to be the main criticism of both reviewers, especially reviewer #2.

See our response below. We now include a supplementary table that includes the fit of our data set to 195 different codon models implemented in IQtree. We note it is difficult to know how exactly the different models overlap with ours, with respect to implementation. Nevertheless, the likelihoods appear comparable between similar models, and SelAC still provides a rather sizeable improvement in fit of 187,000 to over 300,000 AICc units.

I also suggest that the new modeling approach should be placed as part of the Results rather than within the Methods section.

We agree that the way the results are presented is far too dependent on a careful reading of the methods section. In order to satisfy both the AE and Reviewer 1, we have added an introductory paragraph to open the Results that broadly summarizes the main the points of the model.

Reviewers' comments:
Reviewer: 1

Comments to the Author
The authors address the problem of detecting the strength and direction of negative selection and conducting phylogenetic analyses on protein-coding sequences. They make use of the well-studied mutation-selection framework. The authors introduce two primary innovations on this framework by fixing the number of categories to one for each amino acid and by devising a mechanistic model to predict individual fitnesses based on the energy costs required to maintain biologically required levels of protein function. This is based on the assumption that the fitness effect of a non-optimal amino acid is linearly related to its Grantham distance from the optimum, and that this effect is independent of the rest of the protein sequence. The authors' model is known as SelAC.

This paper makes an extremely important contribution to the further development of readily applicable mutation-selection models. The authors introduce a principled method for reducing the number of parameters involved in mutation-selection models, making them tractable for genome-scale analyses. More importantly, the authors specify a highly original model that predicts amino acid fitnesses based on measurable quantities such as gene expression and energy costs. Judging by the simulation results (Supplementary Figures 3, 5), the model is sound and able to successfully recover parameters on genomic data sets, as well as having some success in inferring gene expression rates from sequence data alone. This begins to fulfil a major unrealised promise of the mutation-selection framework in that it permits mechanistic modelling and inference of the factors that affect fitnesses at specific positions within a protein.

There is one relatively minor methodological issue with the paper, in that some weight is placed on the analyses of model fit under AIC, which are not as convincing as they might be. Despite the numerous innovations involved in the SelAC model, it is only compared to models that apply the same set of parameters to all sites. This means it is difficult to separate the contribution of the mechanistic aspects of SelAC from the effect of simply allowing variation in selection parameters among sites.

If the authors intend to argue for the use of SelAC on the basis of superior model fit and realism, then ideally SelAC should also be compared to a more parametric mutation-selection model that allowed site-specific amino acid fitnesses to vary arbitrarily (though this would certainly need to be done on a smaller data set for tractability). Alternatively, the comparison could be to a similar model that substituted a simple parametric distribution for some aspects of the mechanistic model, or to a version of SelAC with a greater number of category types, or a model with site-specific ω.

On the other hand, the authors could make a very strong argument for the use of SelAC based purely on their existing results showing the more realistic simulated behaviour of SelAC (Fig. 3b) and the validity of the model in successfully recovering branch lengths and model parameters (Fig. 3a, supplementary figs. 3 and 5). This approach would not require further analysis (although it could help to compare the performance of FMutSel etc. on the simulated data sets) and coheres well with their stated aims in the introduction.

As we state below, we have compared SelAC to 195 models in IQtree. The initial IQtree paper (Nguyen et al., 2015) has been cited over 600 times so far and implements recent models. We now include these fit as part of a Supplemental table. There are no perfectly overlapping models between SelAC and IQtree (i.e., our Goldman-Yang 94 implementation lacks rate heterogeneity,

unlike the models used in IQtree) and we certainly handle missing data differently (SelAC ignores missing taxa -- that is, that taxon is effectively deleted from that site for the tree, while most other software instead integrates the likelihood over all possible states for the missing taxon), however, the likelihood appear comparable between similar models. In all cases, our SelAC models provide a large improvement in fit over these models.

Aside from this comment, I have included some minor comments below by line.

Abstract
21-3: "Our results indicate there is great potential for more accurate inference of phylogenetic trees and branch lengths from already existing data…" – I don't think this argument is actually made in the results/discussion?

Good point. We have modified this sentence to focus on our tests of model adequacy.

Introduction
55-71: The main point is very clearly made. Perhaps note somewhere that the omega quantity makes more sense when applied as a gene-wide rather than a site-specific parameter as it averages over substitutions at different sites.

We have added this sentence at the end of this section, and we thank the reviewer for this suggestion.

111: "Beyond fitting the phylogenetic data better…" please clarify here what SelAC is being compared to.

We have clarified this sentence.

Results
122: Please insert a 1-2 sentence summary of your model to give context to the following paragraph.

We have added the following text to open this section:

"*The SelAC model requires the construction of gene and amino acid specific substitution matrices that uses a submodel nested within our substitution model. This requires only a handful of genome-wide parameters such as nucleotide specific mutation rates mu_ij, which are scaled by effective population size Ne, amino acid side chain physicochemical weighting parameters*

alpha_c, alpha_p, and alpha_v, and a gamma distribution shape parameter alpha_g describing the distribution of site sensitivities G. In addition to these genome-wide parameters, the model requires a gene-specific functionality expression parameter that describes the average rate at which the protein's functionality is produced by the organism or a gene's `average functionality production rate' psi. By linking transition rates q_ij to gene expression in the form of protein synthesis rate phi, our approach allows use of the same model for genes under varying degrees of stabilizing selection. Specifically, we assume the strength of stabilizing selection for an optimal sequence, aoptvec, is proportional to psi, which we can estimate for each gene."

124: "the strength of stabilizing selection for the optimal sequence, $\vec{aa}$*…" unclear on first reading whether the symbol refers to the strength of selection or the optimal sequence.

Perhaps the word "the" is the source of confusion? As a remedy, we replace "the" with "an" in hopes this clarifies that the term refers to the optimal sequence.

125-7: See the issue raised in the general discussion above. It is not clear that the improvement in model fit comes from the relation to translation rate/gene expression in particular.

See our detailed response above. But, we note that when broadly comparing against all available codon models, SelAC has an enormous improvement in fit with or without the addition of the site-specific parameter.

132-134: Perhaps note here the numbers of parameters involved in FMutSel vs SelAC.

We have added this detail.

135: Perhaps introduce this paragraph with a brief sentence describing the purpose of this analysis; the transition is a bit jarring.

We agree, and have added the following text to open this paragraph in the Results:

"*We note our use of AICc as opposed to AIC in the above model comparisons. At the outset of our study it was unclear what the appropriate sample size, n, when comparing models of sequence evolution. Building upon the work of Jhwueng et al. (2014), our simulations suggest that using the number of taxa times the number of sites as the sample size correction performs best as a small sample size correction for estimating Kullback-Liebler distance in phylogenetic models (Appendix 1).*"

148-9: "…using only codon sequences, our model can predict…" this phrasing seems to make a stronger claim than is warranted by the data. The inferred translation rate is correlated with gene expression data, but the high variability in the estimates shown in Figure 2 would seem to preclude actually being able to predict expression levels. Perhaps simply state that this correlation is remarkable given that it could be found with only codon sequences.

We agree, and have modified this sentence to incorporate the reviewer's suggestion.

150: Please specify here that the site-specific sensitivity is discrete Gamma-distributed.

We have incorporated this change.

155-6: The Grantham weights have not been mentioned yet – perhaps do so at the beginning of the section (see comment to line 122).

See our added text above.

172: Perhaps restate briefly how functionality is being measured here (the benefit function?).

We have added a brief statement defining functionality in this context.

Discussion
221-2: "…that ω is best explained as a proxy for gene expression…" or that both ω and gene expression are anticorrelated proxies for selective strength.

Good point. We have revised this sentence accordingly.

224: Perhaps briefly describe the difference between SelAC and these other models, and justify why SelAC is not compared to them in this study.

We appreciate this suggestion and have chosen to keep the text as it was in the previous submission. Most of the text in the discussion up until this particular line details the similarities and differences between our model and those of not just Lartillot et al. and Thorne et al., but also Halpern and Bruno. However, if the reviewer feels strongly that we add further text comparing and contrasting these model, we would consider doing so in the final revision.

238-50: Also worth mentioning here the role of misfolding toxicity in linking expression to evolutionary rate (e.g. Drummond & Wilke (2008) Cell 134(2))? This might suggest additional costs to upregulating expression of a less-functional protein.

We agree with the reviewer and have mentioned misfolding as a potential additional cost, citing Drummond and Wilke (2008).

256-7: Again, it is not clear whether the gene-specific values are more important than simply allowing variation in selection parameters among sites.

See response above. Whether or not we allow site-specific variation, the use of our SelAC has a rather sizeable improvement in fit compared to other models.

291-304: This attention paid to statistical methods in this section is deeply appreciated.

We thank the reviewer for the kinds words here. This is a topic that certainly needs deeper exploration, and is something we hope to continue looking into.

336-8: Difficult to understand this sentence – do you mean using a mixture over substitution models with different optimal amino acids?

We agree that this sentence lacks clarity and have therefore removed it entirely.

344-346: If this is referring to covarion-like models, perhaps cite Tuffley C. & Steel M. (1998), Mathematical Bioscience 147:63-91 and/or Whelan S. (2008), MBE 25(8):1683-1694.

Yes. With regards to a Hidden Markov model, what we had in mind is more in line with Tuffley and Steel (1998) and Penny et al. (2001), than the one suggested by Felsenstein and Churchhill (1996). This modelling approach would integrate easily into our existing framework -- in fact, we have already begun testing such a model. We have also added these citations, as well as the citation of Whelan (2008).

Materials & Methods
No specific comments.

Figures & Tables
Table 1. The caption should state the data set used for the comparison, alongside the number of sites and genes.

We have added this information to the caption.

Figure 3. It would help to demonstrate the efficacy of SelAC for phylogenetics if the branch lengths reconstructed under the other models were shown for comparison (with the existing branch lengths rescaled to expected substitutions/site).

Due to the differences in the scale of the branch lengths between the GTR and SelAC, for example, we have chosen to keep this figure unchanged. However, we have de-emphasized the importance of SelAC with regards to phylogenetic estimation, as this was not specifically evaluated in the present paper.

Supplementary figure 3, 5: These results are very important for validating the model; it would be good if at least some of these were moved to the main text.

Good point, and have moved some of the text about the simulation results from the Supplemental to the main text. We have also moved Figure S5 to the main text, as it is the most relevant text to the other analyses that appear later in this section. So, per the suggestions from the reviewer, we now open the Results section with a brief summary of our model, followed by brief remarks about the simulations.

Reviewer: 2

Comments to the Author
This manuscript describes a population-based substitution model, designed to both better model the evolutionary process in proteins as well as to provide direct access to quantities of biological and biophysical interest. The approach is novel and interesting, and the paper is generally clearly written. The authors demonstrate that their models fit phylogenetic data much better than so-called "standard models".

My first objection is that the field has moved substantially beyond the standard models. There is a plethora of better models, that are increasingly available in common software packages. (The CAT model of rate heterogeneity is, for example, an option in RAxML.) It is much easier to demonstrate an improvement over quite old models whose limitations are well known than it is to demonstrate an improvement over those who have been addressing some of the same issues as the authors. Does this new SelAC models do better than some of these other more recent, more mechanistic, less phenomenological models?

To address this, we have run SelAC vs 195 models in IQtree (though note "195" suggests more models than there actually are: 195 comes from combination of codon models with various models for heterogeneity: GY+F1X4 vs MGK+F1X4 vs GY+F3X4 all count as different models). The initial IQtree paper (Nguyen et al., 2015) has been cited over 600 times so far and implements recent models. We include this as supplemental table. There are no perfectly overlapping models between SelAC and IQtree (i.e., our Goldman-Yang 94 implementation lacks rate heterogeneity, unlike the models used in IQtree) and we likely handle missing data differently (SelAC ignores missing taxa -- that is, that taxon is effectively deleted from that site for the tree, while most other software instead integrates the likelihood over all possible states for the missing taxon), however, the likelihood appear comparable between similar models.

Under this, it appears that the selac models still perform better than competing codon models.

In addition, the model is both extremely specific and speculative. It is unclear how to relate the performance of the model and the correctness of the multiple hypotheses. If "X" is added to the theoretical model, and that improves the fidelity with observed data (as measured by AIC), does that mean that there is evidence that "X" is important? I know that this is a general problem, but models that involve so many assumptions are especially problematic in this regard, as it is certain that at least some of the assumptions are wrong, and in this case the relationship of adding "X" and the model fitting better is particularly ambiguous.

We have added a paragraph to the general caveats section of the Discussion to include the reviewer's important points about the specific assumptions of our model and how "X" improving fit does not mean "X" is a true parameter in reality.

For instance, the rationale for the 20 different substitution models, each representing a site with a propensity towards a different amino acid, seems forced. Some sites require a polar amino acid, others a small amino acid, others a specific amino acid. It seems more of a limitation to lump all of these different types of selective constraints together based on the fact that they all are most satisfied by a serine than an approach such as the CAT model which assumes less of a pre-specified structure. For this reason, I do not believe that "SelAC strongly supports the hypothesis that the strength of stabilizing selection against physicochemical deviations from \arrow{a}* varies between sites." I would rather believe that this reflects how \arrow{a}* varies between many different sites under different selective constraints that have been artificially grouped as a single type of site. In this case, relaxing the degree of selection at different sites compensates for the model misspecification caused by assuming that the nature of the selection is the same at all of the sites that favor a specific amino acid.

We agree and have softened the tone: more of "consistent with the hypothesis" language throughout. We have also somewhat expanded on the section "There are deficiencies in our implementation…" to explain further how a model focused on amino acid properties rather than identity could be more realistic.

Again, I am sensitive to the fact that the relationship between improving model fit and increasing model realism is always difficult, but I believe the authors need to be clearer about what their models can and cannot say about evolution, which I believe is significantly less than implied by their discussion.

Please see above.

I find a number of other assumptions troubling. First, there is the assumption that there is a set of relatively constant propensities for optimal amino acids at each site. This is demonstrated by the fact that the amino acid found in the wild type in one organism can represent a lethal mutation in another. I am not sure that the assumption that these propensities do not change is better than the omega-based models which explicitly represent changes in these propensities.

We have code in the R package associated with this paper to have a hidden Markov model, where the optimal amino acid itself can change, rather than be fixed per site. That's not part of this paper, and so of course is not relevant to its suitability to publish, but is an indication that we agree with the reviewer's concern in general, enough to create models to deal with it (we intend to test these as an extension to the basic model in the future).

The data we used in this paper suggest that the SelAC model fits better than models that utilize the omega parameter: it is not an absolute conclusion that SelAC's assumptions are better, but it is consistent with them. We imagine that for datasets where optimal amino acids change so much that the optimal one in one species becomes lethal in another, an omega model would be better (though there is a chance that SelAC would choose an "intermediate" amino acid as the optimal one instead).

(There is an interesting direction of recent research (see Shah et al. 2015 and Pollock et al. 2012) regarding how the substitutions both reflect and influence the selective environment at a site. In particular, if the rest of the protein adjusts to the new occupying amino acid through substitutions elsewhere ("entrenchment"), then it is not unexpected that there would be a change in selection that "…perfectly coincides with the fixation of a new amino acid."

We see the reviewer's point, but the entrenchment idea involves epistatic selection on a protein in general -- when one amino acid changes, selection on others changes as well. The "perfectly coincides" statement relates to the site specific omega models which does not include any epistatic terms. More specifically, in the omega models it is not that selection at other sites changes with a change at a focal site, but rather that under this model, for omega < 1, any time an amino acid is substituted the rate of going to any other at that site is lower. We have added statements to clarify this (including citing the relevant literature, which we thank the reviewer for pointing towards). We also point out the reality of the entrenchment effect in the caveats section of the discussion.

Similarly, omega > 1, in as much as it corresponds to a model of protein evolution, often corresponds to antagonistic co-evolution. Under these circumstances, it is again not unexpected that the shifts in the pessimal amino acid would correspond to changes at this site.)

We mention antagonistic co-evolution in this section to highlight this.

The concern is over functionality, with sub-optimal proteins requiring greater concentrations in order to satisfy the needs of the organism. More recent work by Shakhnovich's lab, for instance, indicates that resistance against aggregation might be more important – in this case, we would expect lower concentrations of sub-optimal proteins.

There could be multiple models besides functionality. Forming maladaptive aggregations is one. Others could be risk of harmful prion formation, or risks of catalyzing reactions with deleterious consequences, or of irreversibly binding costly substrates. Hopefully, the framework we develop here can be used to model selection under such models (replacing functionality with propensity to form maladaptive aggregations, for example) and then compared to ours and other models.

It is well known that gene expression levels are correlated with substitution rates. It would also seem that, in the model presented here, fidelity to \arrow{a}* would correlate with substitution rate. Does the fact that gene expression is related, in the context of this model, to fidelity to \arrow{a}* tell us more about the evolution of proteins beyond what we already knew from this previously observed correlation?

Under our model, substitution rate usually would go down as one approaches the optimal amino acid ("usually" because there can be local optima that also reduce substitution rate once there). When quite far from the optimum, many mutations are beneficial and so the probability of them

being fixed is higher; when at the optimum, all mutations are deleterious and so the fixation probability is lower. Expression also plays a role: the higher the gene expression, the greater the fitness differences, and thus the greater the substitution rate difference between proteins close to and far from the optimum. There is also a decrease in expression itself under our model as a protein approaches the optimum. Thus, rather than just an observed correlation, this model provides mechanisms for this, as well as novel predictions about how expression, substitution rates, and distance from the optimal sequence interact with each other.

We have opted not to include this in the manuscript (doing this well would require additional analyses that may make the manuscript unduly long and complex), but we could if the reviewers and editors find it of sufficient importance.

More minor comments:

I recognize that having the Methods section at the end of the article can be awkward, but sufficient details need to be present in the Results section for the results to be understood. This is not the case here.

Based on this comment, as well as those of Reviewer 1, we have added text throughout the Results section to improve clarity, and presentation of the work. We hope that the Results are much more self-contained than in the previous version, and rely less on a close reading of the details that appear in the Materials and Methods.

The discussion of codon models takes the form of these models more seriously than their inventors. Maybe it is more appropriate to say that it is difficult for the non-synonymous substitution rate to be greater than the synonymous substitution rate in the absence of some process such as antagonistic co-evolution. I would agree, however, that it is useful to point out that these codon models do not have fidelity to the evolutionary process.

This is a good point, and we added a sentence on page 3, line 85-89 to incorporate this point.

I might include a mention of site-wise mutation selection models such as Tamuri et al. 2009 that also represent a different substitution rate at each site using a model that incorporates both mutation and selection. I also mention a number of authors (e.g. McClellan, Mackiewicz, Yang, Zhang, Goldstein, Higgs) who have created substitution matrices that are explicit functions of amino acid physicochemical properties.

We appreciate this suggestion and have included citations to work by the above authors at least once in the manuscript.

The notion of sample size as "the number of taxa times the number of sites" is overly simplistic. If we have 137 near-identical samples linked by short branches, to count these as 137 independent samples is a gross over-estimation. More subtlety, adding taxa means changing the underlying tree topology, which means a change of model.

The reviewer has identified an interesting edge case where our suggested sample-size correction for AICc may not be ideal. While we generally agree that this particular situation would be a gross-overestimation, this is not the case for the yeast data set that was the focus our analysis. The purpose of this exercise was to clarify 1) the disconnect between how AICc is used in comparative biology (i.e., ancestral state reconstruction, phylogenetic least-squares, etc.) versus how it is used in molecular biology, and 2) what exactly the sample size corrected penalty should be for AICc when we are examining molecular sequences. We feel strongly that simply counting the number of sites would be a gross underestimation of the penalty, but we hasten to acknowledge that what we present in the paper is not the final word on the matter. We are currently conducting more in depth simulations to better understand what the constitutes the sample size for AICc under various scenarios when the underlying data is a matrix.