October 17th, 2017

Department of Ecology and Evolutionary Biology

University of Tennessee

569 Dabney Hall

Knoxville, TN 37996-1610

Dear Dr. Near

Editor-in-Chief, *Systematic Biology*:

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 you, 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

**Response to the Associate Editor:**

**Dear Authors,**

**Thank you for submitting to Systematic Biology your manuscript « Population Genetics Based**

**Phylogenetics Under Stabilizing Selection for an Optimal Amino Acid Sequence: A Nested Modeling Approach». It has been reviewed by two distinguished referees, who recommend rejection, and acceptance subject to major changes. Your manuscript presents appealing ideas, but there are some critical problems indicated by the first referee. I recommend rejection, but encourage resubmission if you are able to address all the points raised by the referees.**

**Notably, AIC is not appropriate to compare a parameter-rich model like yours, as it’s an asymptotic formula assuming that n (number of statistical units) is much larger than k (number of parameters). In your model, these two numbers are similar, which poses the infinitely-many-parameters problem underlined by the first referee. Note, moreover, that AICc is not appropriate when k>n, as the penalization term becomes negative. Cross-validated likelihood could be an option, if applicable to your model (?). Or Bayes factor, if computable (?). Model comparison in such setting is definitely a difficult task…**

**Hope this helps, sincerely,**

**Olivier Gascuel**

We appreciate these comments, and we agree that our use of the standard AIC as a guide for model selection may have been inappropriate, given how parameter-rich our model is. In fact, we agree with the AE as well as Reviewer 1, that the sample-size corrected AIC, AICc, is probably more ideal. Unfortunately, as the reviewer points out, given that the number of parameters in the model exceeds the sample size, AICc cannot be calculated using the number of sites as the number of observations. However, to address this issue, we point out that with respect to the sample size, *n*, the field as a whole has been rather inconsistent.

Papers appearing in our field often report two variants of AICc. With comparative methods (e.g., Butler and King 2004, O'Meara et al. 2006, Beaulieu et al. 2013) that examine the evolution of, say, a trait, the sample size in the AICc calculation is taken as the number of taxa. In other words, more taxa allow the fitting of more complex models, given more data. When modelling sequence evolution, however, which is basically the same as a discrete character model used in comparative methods, the sample size is just the number of sites (as in ModelTest). Obviously, both cannot be correct, and we argue that for phylogenetic data, a good estimate of data set size is the *number of taxa* multiplied by *number of sites*. We recognize, of course, that our point of view in this matter may come across as self-serving. So, in our revision, we also demonstrate through a simulation approach that the number of taxa multiplied by the number of sites is the best measure of sample size through a simulation approach. We now include an Appendix 1, which describes our simulation approach and results. With this sample size correction, AICc can certainly be calculated and the support for our parameter-rich model remains strongly supported. Regardless, model adequacy measures also suggest our model is better, and it allows inference of meaningful population genetics parameters.

**Responses to Reviewer 1:**

**Recommendation: Reject**

**Comments:**

**In this work, the authors devise codon substitution models that incorporate considerations of a cost-benefit function of a gene (based on the inferred optimal amino acid at each position) and the gene products' rate of synthesis. The formulations are grounded in populations-genetics theory, allowing for straightforward interpretation of parameters, and tests of the role of different factors in governing the substitution process.**

**I think that there are some great ideas in this work. However, there are also some critical problems, which prevent me from recommending acceptance.**

**The first major issue is something known as the "infinitely many parameters" problem. Essentially, one generally wishes to be able to study the asymptotic behaviour of any particular statistical model; that is to say, we wish to study how a given model reacts (i.e., what happens to our parameter estimates) as we present it with more and more data (more columns in the multiple sequence alignment). But with models such as those proposed in this work, whenever we increase the size of the data set, we also change the form of the model (adding more parameters, i.e., optimal amino acids at each position). In sum, there are no asymptotic conditions for their models.**

**The second major problem is that we cannot compare models across different state spaces with information criteria approaches. This is because "data" means something different when dealing with nucleotide versus codons, and therefore comparisons of likelihood (or AIC) scores across data-types are not on equal footing. Please see Seo & Kishino, Syst. Biol., 2009, for more details on this point.**

**Another major problem is the model comparisons based on the AIC. Here, I would first point out that the authors have miscalculated the AIC for their proposed models (but not for the previously proposed models). The equation for the AIC is:**

**AIC = -2lnL + 2K,**

**where 'lnL' is the natural log of the likelihood, and 'K' is the number of parameters. The above equation was used for existing models, but for their new models, the authors computed:**

**AIC (sic)= -2lnL - 2K.**

**Even with the correctly computed AIC, their models still \*appear\* to perform much better than traditional codon models, in spite of the fact that the true AIC values are actually about 200,000 units higher than reported in the manuscript.**

We thank the reviewer for pointing this out. Indeed, we miscalculated AIC in the original submission as a consequence of a coding error. This has now been corrected in the revision as well as the R package that accompanies the manuscript.

**This brings be to my last point. As an asymptotic estimator, the AIC operates under the assumption that the number of observations presented to the model greatly exceeds the number of parameters estimated from these observations. But for their models, the number of estimated parameters is actually greater than the number of observations (I counted 49,881 data columns in all for their 100 genes); note that this is related to the 'infinitely many parameters' problem. I'm not certain, but it might be appropriate to use what is sometimes referred to as AICc (with an extra lower case 'c'). This is a modified (second-order correction) version of the AIC, which is meant for cases where the amount of data is low, relative to the number of estimated parameters:**

**AICc = -2lnL + 2Kn/(n-K-1),**

**where 'n' is the number of observations. The AIC and AICc are almost the same for the GY94 model. For the FMutSel0 model, the AICc is a few hundred units higher than the AIC. So, in these latter two cases, using the AIC or AICc wouldn't make much difference. However, for their proposed models, the AICc is around 40 million units lower than the AIC. Thus, according to the AICc, their models' fit are astronomically poorer (not better) than popular models.**

We appreciate these comments. As we point out in our response to the AE, we agree with the reviewer that the use of AICc is preferred over the standard AIC (which only penalizes the number of parameters, regardless of sample size) due the parameter richness of our model. However, in our view, there is an inconsistency in the way we treat sample size in phylogenetics that we address in our revision. Specifically with comparative methods, such as those that examine the evolution of a trait arrayed onto a phylogeny, the sample size in the AICc calculation is taken as the number of taxa. In other words, more taxa allow the fitting of more complex models, given more data. When modelling sequence evolution, however, the sample size is just the number of sites, as the reviewer points out. We argue that for phylogenetic data, a good estimate of data set size should actually be the *number of taxa* multiplied by *number of sites*. When doing so AICc can be calculated with our model and the support for our parameter-rich model remains strongly supported. We also now include an extensive simulation approach, and examine a large set of different sample sizes to determine which best approximate the remaining portion of Kullback-Liebler (KL) divergence between two models after accounting for the deviance (NB: our replicates vary the number of characters and number of sites). These results have been added to main text, with a detailed description of our approach now included as part of a new Appendix 1.

**Responses to Reviewer 2:**

**Recommendation: Accept with major revisions**

**Comments:**

**Beaulieu and colleagues propose a manuscript entitled "Population**

**Genetics Based Phylogenetics Under Stabilizing Selection for an**

**Optimal Amino Acid Sequence: A Nested Modeling Approach".**

**In this work, they present a model of evolution of codon sequences,**

**that follows the logic of previous mutation-selection models (as**

**initiated by Halpern & Bruno 1998). In this modeling, the mutation**

**layer is the most general nucleotide substitution model (a point which**

**is not discussed), but actually the main point of the manuscript is**

**the selection layer.**

We assume the reviewer is referring to our use of the unrestricted nucleotide (UNREST; sensu Yang 1994) as a means of populating our mutation matrix. We have added a sentence expanding on our choice of the UNREST model (see lines 137-140), but essentially, it is more consistent with our model by being more flexible with respect to codon biases.

**Their concern is to model the most mechanistically as possible the**

**link between protein fitness and translation process. The rationale is**

**that usual models consider synonymous and non-synonymous**

**substitutions, with no interest in the functional role of the encoded**

**amino-acids, nor in the similarities between the substituted**

**amino-acids. And what they call the "nature of selection" is not taken**

**into account in those models, which is quite relevant.**

**So, their idea is to relate this nature to a distance to a ideal**

**protein, expected to perform an ideal "functionality". If the real**

**protein is not the ideal one, more proteins are needed to be produced**

**to accomplish this level of functionality, with entails a higher cost**

**due to protein production, and then a lower fitness.**

**This work has the merit to link concepts of fitness and phenotype to**

**sequences in the context of an evolutive process, which is quite**

**ambitious. Here, the phenotype is assumed constant, and evolution is**

**entirely constrained to preserve this constant phenotype.**

**This hypothesis restrains considerably the applicability of the model,**

**but as the authors say it could be extended in the future.**

**Actually I think the core modeling idea that founds the model**

**interesting (although I see it differently, as explained further), but**

**explanation about the direction of the model is missing.**

We appreciate the comments of the reviewer. In the discussion we are upfront about the simplifying assumptions made by our model, all of which were due to a commitment to computation feasibility. We note that we are currently working on several extensions, most notably a version of our model that uses a Hidden Markov modelling approach (HMM), in order to relax the assumption of a single global optimal amino acid. [Search, for example, for “HMM” in <https://github.com/bomeara/selac/blob/master/R/selac.R>, code for the next version of SelAC to see our implementation of this]. We would argue this is later work for another venue (as it still needs much testing) and the current model still useful enough to publish. All models are wrong, but many are publishable.

**---------------------------------------**

**Model:**

**- All the work relies on the strong hypothesis that the only**

**consequence of a protein change is to make the protein less**

**efficient to perform a function -- which I agree -- but that with**

**more proteins the desired function will then be fulfilled.**

**I am not at all convinced about the realism of this hypothesis.**

**Functionality is not a one-dimension feature, and a change in a**

**protein sequence can reduce its functionality in a way that even**

**with many instances of the protein it is not possible to get the**

**optimal functionality.**

**Perhaps you could provide some references of this kind of dosage**

**compensation?**

This is a good suggestion. We have added a reference to this. Studies of compensation are still coarse-scale (silencing of an entire gene), and we assume a degree of finer evolutionary control, but this does suggest there are such mechanisms.

**Actually, given the resulting formula page 12, I think it would be**

**much simpler to say that the reduction in fitness is due to the**

**distance to the optimal (ie the d(a,a\*)), and that this reduction is**

**correlated with the expression level of the gene, psi.**

**Perhaps for you it would be disappointing because it removes all the**

**cost/benefit modeling part of the manuscript, but it seems much more**

**realistic.**

The comment as written implies that replacing a specific formulation with a more general, but mathematically equivalent, formulation increases the realism of the model. This revised statement makes sense since the cost/benefit function we use ultimately results in a generic, additive effect. Nevertheless, this is a valid and valuable point which we address at the end of the Introduction and in the Discussion (see lines 529-539) where we point out that making models too generic results in their misinterpretation (e.g. GY94’s omega) and makes it difficult to identify implicit assumptions made in such models (e.g. Yang and Nielsen 2008).

**Moreover, in the formula, the "n" factor can by simplified since it**

**is usually much larger than 1.**

The 1/n term we believe the reviewer is referring to in the functionality equation is necessary to ensure that the functionality of an optimal peptide is 1. Nevertheless, we acknowledge that there are many potential simplifications of our model. However, we leave such explorations for later studies.

**- Concerning the definition of the distance, it is quite interesting**

**to build it from parameterized linear combination of given**

**physico-chemical properties. As exemplified in the results, it is**

**then a good way to compare the relative influence of those features**

**on the process (and eventually on the selection constraints).**

We are glad that reviewer agrees.

**- From this point of view (which is not exclusive from yours), the**

**main evolutive modeling innovation is about the exchangeability**

**matrix between amino-acids. The idea of defining this matrix as a**

**fitness landscape around a specific target seems to me very**

**interesting and innovative, when other mutation-selection models**

**consider equiprobable exchangeability rates, or at most general ones**

**(such as WAG or LG models).**

**On the contrary, amino-acid site-specificity of those models is**

**considered in the equilibrium of the process (such as in CAT**

**models). Those models then model evolution towards a target with no**

**explicit consideration on the variance around this target, when**

**SELAC models both.**

**- Actually, there is no discussion about the equilibrium of the**

**Process. From the formula in page 12, we can compute that for**

**the amino-acid part, the equilibrium frequency of amino-acid a\_i**

**is proportional to (W\_i)^{2N\_e-b}. From this formula, it is easy**

**to compute the influence of distance d and psi on the equilibrium,**

**and this would shed some interesting qualitative light on the role**

**of constraining selection around optimal amino-acids. Furthermore,**

**since other site-specific models are defined through equilibrium**

**frequencies, it gives some hints of comparison between SelAC and**

**those models. Also, there is no mention of the root distribution of the**

**process. I guess the process is assumed stationary.**

We agree with (see Sella and Hirsh 2015), and appreciate, the reviewer’s insight about the equilibrium frequency. we agree that further analysis would lead to further insight. However, this is beyond the scope of the work we present here.

Regarding the assumption of stationarity, because we estimate the ancestral state and then model the transition from this state, we are explicitly not assuming a stationary process. The process does eventually approach the stationary distribution as the branch lengths approach infinity. We have added sentences on lines 236-239 to highlight these points.

**- Is the whole model normalized to ensure one substitution per unit of**

**time at equilibrium? Since mutation matrix is normalized, I assume**

**it is not. In this case, there is no point in comparing the**

**evolution through SELAC with normalized other models on a same tree**

**(as in Figure 3), and you should normalize it.**

For SelAC, we normalize with one mutation per unit time; with other models, we assume one substitution per unit time. Thus, for the same unit time, SelAC will have fewer changes (under most conditions). However, for Figure 3 it is all normalized to the proportion of missing branch length under the two models. Moreover, if we did rescale in some other way, it would have the effect of stretching or compressing the yellow line relative to the others, but with the same overall trend. In other words, under SelAC fitness increases slightly to the observed value, while fitness under the other models radically drops.

**- I guess the most time-consuming computational problem is the need to**

**infer all the optimal amino-acids? Since the idea behind the model**

**is that amino-acids are well conserved, a way to accelerate this**

**would be instead to take on each site the most conserved amino-acid**

**as the optimal one, or at least on the most conserved sites.**

**Otherwise, with ancestral amino-acid reconstruction using a usual**

**distance matrix (or even the distances inferred by your model to**

**make it more coherent), you could restrain considerably the number**

**of candidate optimal amino-acids on each site.**

This is true. However, in our implementation, a user can speed up the calculation by making various assumptions about the optimal amino acids that does not require full optimization. In our revision, we now include the fit or models where we assumed the most conserved amino acid is optimal (which we refer to as the “majority-rule”). This notably decreased the necessary computation time, but it was generally a worse overall fit when compared to a model where the optimal amino acids were optimized. We added a sentence pointing to the improvement in the computational time by assuming the most conserved amino acid is optimal (lines 345-349).

Our long term plan (for which we have NSF funds), once this is in press (we don’t want to act before peer review), is to have a hackathon to incorporate the SelAC model into faster, more popular applications like RAxML and RevBayes.

**Results: -------**

**- p17 l325 : It is argued that linking the strength of stabilizing**

**selection to gene expression improves the model fit. It should be**

**tested on real data through LRT by fixing the parameter of gene**

**Expression.**

**- The comparison of SelAC against other models is unfair, since those**

**other models (GTR+Gamma) and FMutSel0 do not include any**

**site-specificity. What your result says is that in strongly**

**conserved proteins, amino-acids are site specific.**

**The comparison would be more much more interesting with**

**site-specific models, such those of Rodrigue & Lartillot, or Bruno &**

**Halpern.**

For this dataset, or even ones with many more taxa, those would be infeasible (i.e., there are far too many parameters).

**- I do not see why Figure 3b highlights the improvement in biological**

**realism in amino-acid sequence evolution. It means that SelAC**

**prevents any evolution too far away from a given protein (as**

**expected), but it is not more realistic.**

The important thing is the direction of change, not the slow rate -- SelAC has the protein evolve towards its observed functionality (generally) while the others quickly cause a loss of functionality. That is, towards an empirical summary statistic for the observed sequence or far away from it.

**- On the yeast sequence, the inferred tree (Fig 3) is quite long. If**

**the model is not normalized (actually it is not said), it is**

**difficult to interpret these lengths. Otherwise, they are**

**prohibitively long in comparison with usual branch lengths,**

**especially with these species. This could be interpreted as the need**

**for the model to build very long branches to explain the observed**

**divergence between the sequences (that are in these species not so**

**divergent). In this case I see this as a drawback of the model,**

**since it does not infer realistic branch lengths.**

This concern may arise from ambiguity about branch lengths in the initial manuscript. Under our SelAC model the mutation matrix is rescaled, which is then multiplied by our fixation rate matrix. In the end, the branch lengths represent not substitutions per site, but rather represent “proposed mutation” per site. We have also modified the scale bar Figure 3 so that it lists the units to avoid further confusion with edge lengths estimated under standard nucleotide substitution models and those provided by our SelAC model.

**And finally, among the minor remarks, there a many mistakes in the**

**text and formula. A final proof-reading before submission is missing.**

We thank the reviewer for highlighting all of the typos. In the revised version, we have made all the necessary corrections listed below, as well as checked for any new ones that may have been missed or could have been added based through the revision process.

**keywords: "super cool", really? (I saw it!)**

We apologize for the inclusion of this particular keyword. This was an inside joke among authors that did not get removed prior to submission. It has been removed in this new version.

**p.3 l 34 : Their examples are not appropriate, since they cite only**

**nucleotide models, usually used for non-coding regions.**

We could have made our point a bit clearer. The point we are making is that most practitioners rely on simple models, most often nucleotide models such as GTR, to model sequence evolution and/or estimate a phylogeny. However, when the underlying sequence data is coding, as it often is the amino acid that the sequence actually codes for is completely ignored. A good example comes from plants, where a GTR-based model is routinely applied to coding regions of the chloroplast, such *rbc*L and *mat*K. In any case, we have added clarity to this point.

**p.3 l.38 : mutation bias ARE far ...**

We have made this change.

**p.3 l.40 : of the model AS is**

We have made this change.

**p.6 l.103-104 : the arrow over a\* is episodic through the manuscript.**

We thank the reviewer for highlighting this inconsistency. Throughout the manuscript, we now use a\* when discussing a single site and \vec{a\*} when discussing the peptide.

**p.7 l.129 & 142 : there is no need to set G->T rate to 1 if**

**afterwards the mutation rate is normalized.**

**p.7 l.133 & 142 : pi is different in both lines (ie nucleotide versus**

**codon frequencies).**

We have corrected this.

**p.7 l.148 : to meEt its**

We have made this change.

**p.7 l.150 : It seems that rather eta=C/B and not the inverse.**

**Otherwise I do not understand the formulas in p. 11 l.**

**180.**

The reviewer is correct, and we appreciate them pointing out this mistake.

**p.8 formula 1 : a parenthesis is missing**

We have fixed this equation.

**p.8 l167-168 : each siteS sensItivity**

We correct the spelling

**P.9 formula 3 : n\_g but sometimes it is only "n" further. It is**

**episodic too.**

IBID

**p.9 l169 : any indirect costs**

We have corrected this.

**P.10 l.180 : there is no need to introduce a new variable phi.**

We appreciate the suggestion. The reviewer does not suggest an alternative to our use of \*phi\*, and thus we have decided to not incorporate any changes.

**p.10 l185-188 : I do not understand the example. Perhaps more detailed**

**formulas would help.**

We apologize for our poor explanation. We have tried to make our examples simpler and our language clearer. (lines 198-199).

**p.11 l191 : expliciTLy**

We have fixed this spelling mistake.

**p.11 formula 6 : G\_p has disappeared (and in formula page 12).**

**Moreover, there is "single site p" above, and then a**

**sum over p in the formula.**

We agree the lack of a G\_p was a mistake, however, the notation for the fitness W\_i and the fitness ratio W\_i/W\_j are at the scale of the entire amino acid sequence (hence the arrow over the c indicating a vector of codons),

**p.12 l197 : "few" parameters contradicts with the final 50004**

**parameters in Table 1. Btw, why 50004?**

We see the reviewers point, however, the word “few” in this context is in reference to our matrices containing up to 26,880 unique rates, generated from only a handful of parameters. In general, however, our model is rather parameter rich when the number of estimated optimal amino acids are taken into account. This is why there are 50,004 parameters in total -- i.e., 49,881 optimal amino acids, one composite parameter for phi for each of the 100 genes, two for Grantham weights, and 11 for the nucleotide substitution.

**p.14 l253 : HÖhna**

We have fix this reference.

**p.16 l294 : SelAC is in lower case.**

We have fixed this issue.

**p.17 l328 : SelAC+Gamma I guess what it is, but not defined explicitly.**

Good point. We have added a parenthetical statement that defines what we mean by SelAC+Gamma.

**p.18 l336 : Figure 1 and not Figure S1**

This change has been incorporated.

**p.18 l341 : 1.40.**

We are unclear as to issue the reviewer is referring to, and therefore, we have not altered the text.

**p.18 l352 : varIation**

This has been fixed.

**p.20 l399 : I do not understand distribution of effects of what?**

We can appreciate the confusion caused by the vague use of the term ‘effects’ and have replaced it with “distribution of the effects of amino acid substitutions” in order to be clearer.

**p.21 l426 : quantiTative**

**p.22 l451 : domiNance**

We have corrected the spelling.

**p.22 l455 : I do not understand why CUB could not be included in the**

**model (as in Yand&Nielsen, 2008 or Pouyet&al, 2016). It**

**there is epistatic interaction between sites, there are**

**similar epistatic interactions between amino-acids, that**

**you do not consider in SelAC.**

Because most researchers believe that changes in codon usage lead to changes in the assembly cost of the protein, combining these cost effects with the effects of amino acid substitution results in non-additive interactions between sites or epistasis. Because the work the reviewer cites are non-mechanistic, they fail to appreciate this fact and, thus, implicitly make additional unstated assumptions. While we are confident we can ignore these epistatic effects under certain conditions, we are not comfortable with doing so at this point. We’d also like to point out that both Yang and Nielsen (2008) and Pouyet et al (2016) suffer from the same weaknesses as the GY94 with respect to non-synonymous substitutions. We now mention this point in at the end of the introduction (e.g., 72-80) and discuss it in more detail in the Discussion (e.g., lines 529-539).

**p.23 l458 : can likely BE ignored**

We have added this word.

**supp mat:**

**l628 : Perhaps the strong downward bias in alpha\_G is due to the**

**difficulty for the model to account for changing amino-acids?**

Good point. If it is ok with the reviewer, we have added this statement with reference to the reviewer.

**l635 : the the gamma**

Fixed.

**Fig S3 : "(d)" is missing at the end of the legend. In (c), what is**

**Gene index?**

We have added the (d) to the legend, as well as an explicit statement as to what we mean by “gene index” on the x-axis.