My impression of the revised work is unchanged from that of the original version:  it contains excellent ideas, but is statistically misconstrued.  One particular statement left me puzzled:  "[...] the likelihood for a given site is the sum of the probabilities of each observed state at each tip."  This makes it sound like the likelihood is a sum over a number of terms corresponding to the number of taxa, which is obviously incorrect.  To me, a better explanation would be something like:  "The likelihood for a given site is calculated on the basis of internal node states as well as the leaf states, where the product of state transitions probabilities across all branches is computed.  Since internal node states are unknown, this product of transition probabilities is taken for all possible internal state configurations, and their sum is computed."  So, the sum is over all possible internal node configurations, not over all taxa.  The important point to note here is that the likelihood for a datum (site) is a function of the number of taxa and the tree topology.  If we were to add taxa, we would then be working with a different likelihood function (i.e., a different instance of Felsenstein's dynamic programming algorithm).  
  
Perplexingly, the authors now propose that the number observations is not the number of elements on which one computes the likelihood function (sites), but rather the number of sites times the number of taxa.  This is erroneous.  If anything, counting the number of observations as the number of sites is probability an over-estimate, given that sites are not, in fact, independent from one another (Posada & Buckley, Syst. Biol., 2004).  Sequences taken from different taxa are obviously non-independent, which is why we work in a phylogenetic framework in the first place.  It is certainly not always true that "more taxa allow the fitting of more complex models", such as working with a 50 taxa alignment of EF-1alpha sampled across primates, versus an alignment of the same gene broadly sampled across 20 eukaryotes; in spite of having fewer taxa, the latter alignment is clearly more information rich.  
  
In any case, the fundamental issue with the models proposed by the authors is the "infinitely-many-parameters" problem, where one is left without any asymptotic conditions.  It seems that the authors have completely ignored my remarks on this point, in both the revision and in the response, which makes me wonder if my explanations were too cryptic.  I therefore expand on this below.  
  
Let us suppose, for the moment, that we are adopting a maximum likelihood framework and invoking a GTR nucleotide substitution process with gamma-distributed rates across sites (i.e. GTR+Gamma), and working in the context of a fixed tree with 64 tips (taxa).  Such a model involves 6 exchangeability parameters (5 degrees of freedom), 4 nucleotide frequency parameters (3 degrees of freedom), 1 parameter controlling the variance of the gamma distribution (its mean is constrained to 1), and 125 branch length parameters, for a total of 134 estimated parameters.  We can clearly envisage asymptotic conditions for such a model, which consist of presenting it with more and more sites; no matter how many sites are added, we are still estimating 134 parameters, and we can thus, at least in principle, study the model under progressively larger sample size conditions.  
  
However, if we were to change the context to working with a 128 taxa, we now have 253 branch length parameters, for a total of 262 estimated parameters.  With 256 taxa, we have 518 estimated parameters.  With 512 taxa, we have 1030 estimated parameters.  And with 1024, we have 2054 estimated parameters.  Contrasting inferences of parameters could be interesting  However, such an assessment is not one of the asymptotic behavior of an overall statistical model; these contexts imply different statistical models, with different numbers of estimated parameters.  When we speak of a model's asymptotic behavior, we are referring to how a \*given\* model behaves with increasing amounts of data, not how different models behave.  As I've mentioned already, it is only in adding sites that we maintain the same overall statistical model.  
  
This point also holds for approaches that invoke variables at each site optimized to their ML values.  For the latter, however, adding sites also changes the form of the model, by increasing the number of (site-specific) parameters to be estimated yet again.  In this sense, we always end up with a moving target; the model changes form no matter how we increase the size of the data set, and it is for this reason that I state that one is left without any asymptotic conditions to envisage; in the site-specific ML framework, we simply have no way of presenting a \*given\* model with more data, since whichever way we try, we end up with a \*different\* model, having a different number of estimated parameters.  
  
Thus, to me, the main problem with this work is the idea of invoking an asymptotic estimator (i.e., AICc) in a context without asymptotic conditions.  
  
The AE suggests that perhaps cross-validation could be considered.  It's a good idea to consider this approach to model comparison, but only because I believe it further highlights the problem with the model construction proposed by the authors.  The cross-validation method consists of setting aside a bit of data (sites) as a 'test set', and fitting a model of interest to the data that remains, often called the 'training set'.  One then uses the parameter values obtained from the training set to compute the likelihood score on the test set.  With well-constructed statistical models, this is straightforward to do, and provides an unbiased measure of fit, which implicitly penalizes for the model's dimensionality.  But with models that have optimized site-specific parameters, it is problematic.  What would be the optimal amino acid at each site of the test data?  With their model, the authors can not compute the likelihood for previously unseen data, as one can normally to with a statistical model.  
  
I wonder why the authors don't take the usual solution to modeling across-site heterogeneity, such as overall rates, as in the gamma-distributed-rates model of Yang.  With Yang's model, one is not trying to fit by ML the rate at each datum, but rather, one is fitting the distribution of rates, treating the rate at each site as a random variable.  The likelihood is calculated for each site as a weighted average of the likelihood using the different possible rates (in practice, the gamma distribution is discretized, typically into 4 rates, and the discretization is done so as to have equal weights).  Why can't that authors do the same, estimating a distribution of optimal amino acid across sites?  In short, their likelihood function would be  
  
p(D\_p | theta, T) = \sum\_{a=1}^{20}  w\_a p(D\_p | a, theta, T),  
  
where 'theta' is the set of parameters other than the optimal amino acid, and were 'w\_a' is a weight (w is a 20-dimensional, 19 df weight vector), corresponding the proportion of sites of the alignment having amino acid 'a' as optimal.  With such a formulation, one does not estimate a\_p^\* for each site at all, but rather simply estimates w.  This would rescue the model from the infinitely-many-parameters trap.  
  
Additional Questions:  
Directions for Reviewers: The authors will appreciate detailed comments on the manuscript. Please write comments for the authors in a separate file, numbering all items that should be addressed before the manuscript is acceptable for publication, and attach your file at the bottom of this form (if you’ve inserted comments on an electronic copy of the manuscript, please attach that file as well). Reviewers are reminded that Systematic Biology is interested in publishing well-written papers of high scientific quality and of general interest. Thus, in your review, please address both the appropriateness of the paper for the journal as well as its scientific strengths and weaknesses. Please note that our instructions for authors are available on our website, [systbiol.org](http://systbiol.org/" \t "_blank). Use the buttons above to access the manuscript files. The HTML and PDF buttons link to the entire manuscript. Individual submitted files (such as data files) are available under the “Supplementary Files” button. We encourage reviewers to make comments directly on an electronic copy of the ms. If you do not have software that would allow you to make comments on the pdf version, please check under “Supplementary Files” to see if a Word version is available.:  
  
Do you wish to remain anonymous?: Yes  
  
How significant is this work?: Moderately  
  
Is the author aware of the background and source material to the problems set forth?: Yes  
  
Are the conclusions justified by the evidence presented and the assumptions involved?: No  
  
Are the illustrations and tables clear and understandable?: Yes  
  
In number are they: Sufficient  
  
The following people would be appropriate to review this MS further:  
  
  
Reviewer: 2  
  
Recommendation: Reject; resubmission encouraged  
  
Comments:  
Even after reading carefully the responses, my main concern remains about the  
realism (and then the applicability) of the model. I find the  
responses (as well as the model comparisons) still not convincing.  
This is why I recommend to reject this manuscript (in spite of many good ideas, as I commented in the first review).  
  
  
Comments about the responses:  
============================  
  
First, I would have appreciated the corrections to be highlighted in  
the manuscript.  
  
  
I agree that any model is more of less false, but it is reasonnable to  
expect that the rationale behind its construction is founded. If the  
model is purely phenomenological, this rationale is beyond the real  
process, but when it is purely mechanistic, as Selac is, the logic of  
the construction is to mimic the realism of the evolution process.  
  
In the case of Selac, the logic is based on the idea of dosage  
compensation in protein expression to correct non-optimality of  
protein structure. I am still not convinced by the realism of this  
hypothesis. The reference provided by the authors reports mostly cases  
when default in a gene provides up-regulation of related genes (such  
as paralogues), but not of the gene itself. Following this observation  
of global genetic stability, it is even more straightformard to  
imagine that non-optimality of a protein induces a lower expression of  
the gene and higher expression of related (and more fit) genes.  
  
  
The authors argue that a with a more generic model, it is more  
difficult to interpret parameters. But when the model is founded on  
much hypothetic mechanisms, the parameters provide informations that  
may lead to mis-interpretation of real mechanisms. By the way, the  
genericity was more on using a distance per se (that is, without  
linkage to the cost/benefit dynamics), and not about the  
additive-effect: this genericity provides a new way to explore the  
phenotype landscape, in which selection is seen as a factor of  
exploration.  
  
  
  
  
The comparison with SelAC\_M is interesting about the importance of  
optimizing a\*. But "majority rule" does not consider phylogenetic  
inertia, and an infered ancestral amino acid may be different from the  
most frequent. Moreover, an in-between would be to restrain the  
candidates to optimal a\* to the ones that are observed, which would be  
logical given the strong stability of a\*.  
  
  
  
  
About Fig3b, the answer of the authors is not more convincing about  
the biological realism of the model. SelAc has the protein evolve  
towards its observed functionality, and not the other models, because  
SelAc is made to consider this functionality. To say it differently,  
with another definition of functionality, you would observe the same  
behaviour, since the criterium is ad hoc to the model.  
  
  
  
  
  
Additional Questions:  
Directions for Reviewers: The authors will appreciate detailed comments on the manuscript. Please write comments for the authors in a separate file, numbering all items that should be addressed before the manuscript is acceptable for publication, and attach your file at the bottom of this form (if you’ve inserted comments on an electronic copy of the manuscript, please attach that file as well). Reviewers are reminded that Systematic Biology is interested in publishing well-written papers of high scientific quality and of general interest. Thus, in your review, please address both the appropriateness of the paper for the journal as well as its scientific strengths and weaknesses. Please note that our instructions for authors are available on our website, [systbiol.org](http://systbiol.org/" \t "_blank). Use the buttons above to access the manuscript files. The HTML and PDF buttons link to the entire manuscript. Individual submitted files (such as data files) are available under the “Supplementary Files” button. We encourage reviewers to make comments directly on an electronic copy of the ms. If you do not have software that would allow you to make comments on the pdf version, please check under “Supplementary Files” to see if a Word version is available.:  
  
Do you wish to remain anonymous?: Yes  
  
How significant is this work?: Moderately  
  
Is the author aware of the background and source material to the problems set forth?: Yes  
  
Are the conclusions justified by the evidence presented and the assumptions involved?: No  
  
Are the illustrations and tables clear and understandable?: Yes  
  
In number are they: Sufficient  
  
The following people would be appropriate to review this MS further:  
  
  
Reviewer: 3  
  
Recommendation: Reject; resubmission encouraged  
  
Comments:  
This ms. describes a model of codon substitution that assumes that every amino acid position in the protein sequence has an optimal amino acid.  This assumption may be reasonable but is probably too simplistic as well.  For example, one would expect the preference or selective pressure to vary considerably along the protein sequence.  Overall I am not very excited by the paper.  It seems like the right kind of papers for MBE or Genetics, but not Syst Biol.  
  
I have two major comments.  
  
A.  What are the utilities of the new model: more reliable phylogenetic tree reconstruction, better estimation of branch lengths, or more accurate estimation of parameters for a better understanding of the evolutionary process?  The authors claim that the model fits real datasets better than previous models, and has the potential for more accurate inference of phylogenetic trees and branch lengths.  But no evidence for this potential is supplied.  My intuition is that phylogenetic tree reconstruction is not that sensitive to the model assumed.  At least the authors need to show that the new and old models often produce different trees, and that the trees from the new model are often better.  Similarly for parameter estimation can we demonstrate the superiority of the new model (without simulating under the new model).  
  
B. I found many of the statements to be simplistic and over-loaded.  Example include terms like “average functionality production rate”, the benefit-cost discussion of p.8-9, the use of parameter φ' to represent protein synthesis rate or gene expression level (when the correlation between the two was only 0.3-0.4), etc.  I do not think our understanding of the evolutionary process of specific proteins is good enough to make such claims.  The use of the Grantham’s matrix to modify codon substitution rates was done many times before, for example in the first codon model of Goldman & Yang 1994, but previous studies at least acknowledge the caveats.  
  
My minor comments are more or less by page.  
  
I am surprised that the model is implemented using R, instead of C/C++.  Those codon models involve a lot of computation, and it is hard for one to take a model implemented in R seriously.  
  
p.3 The authors spent a whole page discussing the interpretation or ‘misinterpretation’ of w in current codon models.  Similar points were made by Bustamante (2005), Nielsen and Yang (2003), etc.  Perhaps cite the early papers and shorten the discussion.  
  
Bustamante CD. 2005. Population genetics of molecular evolution. Pp. 63-99 in Nielsen R, ed. Statistical Methods in Molecular Evolution. Springer, New York.  
Nielsen R, Yang Z. 2003. Estimating the distribution of selection coefficients from phylogenetic data with applications to mitochondrial and viral DNA. Mol Biol Evol 20:1231-1239.  
  
p.5 Use one of “Fisher-Wright” and “Wright-Fisher” process.  According to AWF Edwards, this should be called the Fisher model.  
  
p.7 “The nucleotide mutation matrix is also scaled by a diagonal matrix π whose entries, πi,i = πi, correspond to the equilibrium frequencies of each base.”  I don’t understand this sentence.  If “scaled” mean “multiplied”, the statement will be incorrect.  Note that with the UNREST model, you should not multiply the rate by πj.  
  
p.13 fixation probability.  Check Fisher (1930).  Perhaps there is no need to cite so many later papers.  
  
p. 17 Yang & Nielsen 2008 describes FMutSel as their main model, while FMutSel0 was a special case that was rejected in real data analyses.  Can you use FMutSel instead of FMutSel0?  
  
p.17 The comparison of codon and nucleotide models may not be as straightforward as it seems.  One could argue that for a fairer comparison, the nucleotide model should also make use of the knowledge that the DNA sequences are coding and thus do not contain stop codons.  Disallowing stop codons will immediately improve the log likelihood of the nucleotide model, even though no computer program implements this.  
  
p. 18. “Beaulieu et al. (2013) assessed whether parsimony scores and the size of monomorphic clades of empirical data were within the distributions of simulated under a new model and the best standard model”  
Hard to understand.  Please rephrase.  
  
p. 18 “A marginal reconstruction of the likeliest sequence across all remaining nodes”  
Should “remaining nodes” be “remaining interior nodes”?  
  
p.19 “we argue that for phylogenetic data, a good estimate of data set size is number of taxa multiplied by number of sites”  
This is really an extreme view.  
  
p.20 “when considering how likelihood is calculated, the likelihood for a given site is the sum of the probabilities of each observed state at each tip,...”  
This does not sound correct, but I found the sentence confusing.  
  
p.25 “The idea that quantitative information on gene expression is embedded within intra-genomic patterns of synonymous codon usage is well accepted;”  
I don’t think this is the case.  It is generally accepted that codon usage may be under selection and is thus related with gene expression in bacteria or fruit flies, but I don’t think there is strong evidence or general agreement that that is also the case in mammals.  
  
p.27 “Although this is a rather extreme example, it seems prudent for researchers to use a simulation based approach similar to the one we take here to determine the appropriate means for calculating the effective number of data points in their data.”  
The suggestion sounds strange, and I have not followed the authors’ argument well.  Sure every fool can do simulation but some knowledge of statistical theory should also be useful.  
  
  
Additional Questions:  
Directions for Reviewers: The authors will appreciate detailed comments on the manuscript. Please write comments for the authors in a separate file, numbering all items that should be addressed before the manuscript is acceptable for publication, and attach your file at the bottom of this form (if you’ve inserted comments on an electronic copy of the manuscript, please attach that file as well). Reviewers are reminded that Systematic Biology is interested in publishing well-written papers of high scientific quality and of general interest. Thus, in your review, please address both the appropriateness of the paper for the journal as well as its scientific strengths and weaknesses. Please note that our instructions for authors are available on our website, [systbiol.org](http://systbiol.org/" \t "_blank). Use the buttons above to access the manuscript files. The HTML and PDF buttons link to the entire manuscript. Individual submitted files (such as data files) are available under the “Supplementary Files” button. We encourage reviewers to make comments directly on an electronic copy of the ms. If you do not have software that would allow you to make comments on the pdf version, please check under “Supplementary Files” to see if a Word version is available.:  
  
Do you wish to remain anonymous?: Yes  
  
How significant is this work?: Slightly  
  
Is the author aware of the background and source material to the problems set forth?: Yes  
  
Are the conclusions justified by the evidence presented and the assumptions involved?: No  
  
Are the illustrations and tables clear and understandable?: Yes  
  
In number are they: Sufficient  
  
The following people would be appropriate to review this MS further: