

Review of manuscript “*Limited utility of residue masking for positive-selection inference*” by Spielman *et al.*

The manuscript explores how filtering poorly aligned sites from an alignment might impact inference of positively selected sites under codon models. There is some methodological novelty here, as the authors have implemented some new scoring algorithms, and a novel gap-penalization scheme. The main framework for the filtering is a reimplementaion of the methods originally provided in the Guidance software.

The manuscript is highly focused, and it is nice and short. As the primary result is that filtering offers little to no improvement on the true positive rate, there is not much point in reviewing the methods further. It's a straightforward finding, and it is derived from sequences simulated according to realistic levels of sequence divergence and indel rates.

Detection of positively selected sites remains a topic of wide interest; hence, there will be wide appeal to studies such as these. It's unfortunate that there was no real benefit to filtering, as it would have been a nice tool to have. Moreover, the results and conclusions are qualitatively similar to those of Jordan and Goldman (2012). Some additional issues could be covered to give the study some wider appeal; e.g., add more divergence, and higher indel rates, so that the new scoring and penalization methods could really shine, or broaden the scope and assess other inference tasks (e.g., branch length estimation) over a wider set of analytical goals (estimation of divergence dates). I think something like this might be needed for publication as a full article in MBE. However, this does not seem fair to authors who might not be interested in such issues; and, there is some real value to the study as it is currently focused. So, I recommend that the authors revise the manuscript so that it meets the criteria for the “Letter” category. Revision would largely involve reducing its length.

If the authors choose to move forward with a full article, I would like to see a little more discussion (and possibly some evaluation) of the difference they observed between FUBAR and PAML derived results (Figures 1-3). Are these entirely due to some of the approximations made by FUBAR in order to achieve large speedups, or is there also some relation to the way the models are formulated and the way the prior distribution is used. There seems to be some interesting issues there.

Recommendation: Accept a revised (shortened) version of the manuscript for publication as a Letter.