In this paper, the authors investigate the utility of masking (or

removing) poorly supported columns of multiple sequence alignments (MSA)

prior to performing phylogenetic dN/dS analyses for selection. This

assessment is carried out using simulated data, selected to reflect a

reasonable subset of evolutionary parameters encountered in practice. The

authors also implemented several novel extensions to previous algorithms

for MSA filtering, and published their software online (github).

The main conclusions of the paper, and my assessment of how well they

are supported by the paper are these:

1). At realistic levels of divergence and indel variation there seems to

be little effect of residue filtering. I largely agree with this claim

and would only like to see some minor changes, and a few more

simulations (see below) to firm it up even further (major comments 1 and

3)

2). FUBAR and PAML M8 are the two downstream computational tools used to

analyze the data. The authors perform the comparison of the methods and

conclude that PAML is a more accurate approach on larger data, with very

low rates of reported false positives. I do not think that this

inference is particularly well supported by the analyses in the paper.

First, the data were simulated under the assumptions much closer to

those made M8 than FUBAR -- see major comments 4 and 2 -- hence FUBAR

was at the disadvantage to begin with. Second, the simulation setup is

suboptimal for studying the statistical performance of the method (see

below).

The paper is written clearly, motivates the study well, and largely

does a good job addressing point (1) above.

Overall, I think the paper is appropriate for MBE, since many of the

papers on the topic have been published in MBE (most notably the Jordan

and Goldman) paper, the topic of interest to many researchers using

dN/dS analyses, and there are conflicting recommendations in existing

literature. My recommendation is to ask the authors for major revisions.

As much as I hate to request additional simulations, I think this is

essential here.

Major suggestions/comments.

1. I simply don't have a good intuition for why alignment filtering

would improve the POWER to detect positive selection. Sure, by removing

variable misaligned sites, I expect filtering to reduce FPR. But

filtering does not IMPROVE the alignment, it simply REDUCES it. For a

FIXED FRP, this could boost power, but for a fixed detection threshold

(Figure 2), removing sites should largely reduce TPR. Does filtering the

alignment change the inferred distribution of dN/dS to make a dramatic

difference (I can buy that if a sufficiently large proportion of sites

is filtered)? For example, what happens to the MLE of the omega > 1

class in M8 when comparing filtered and unfiltered data? My intuition is

that this estimate will be generally LOWER for filtered alignments,

allowing the detection of sites simulated with omega closer to 1 more

reliably. The same can be extended to FUBAR, by examining how much

weight is assigned to each dN/dS value.

Another explanation for the apparent boost in power is that the mapping

between an aligned site and the corresponding simulated site is

influenced by the filtering. This can be examined in the following

simple way: for a fixed replicate, consider whether or not the simulated

(true) dN/dS assignments change for ANY sites in the filtered alignment.

For example, do we have a situation like this:

site X is present in both filtered and unfiltered alignments but it maps

to DIFFERENT underlying INdelible sites and hence different dN/dS rates?

2. I don't think that a discussion of FPR rates of FUBAR or M8 is

particularly meaningful unless the alignment contains some sites that

are exactly neutral (i.e. with dN/dS = 1); having no null data in the

simulation is going to make the test appear more conservative that it

is. For example simulating under the log-normal (mean 0.37) could

generate very skewed distributions (for completeness, the authors should

provide either the skewness (sigma parameter) of the log-normal, or the

actual 50-bin distribution used to generate the data). Thus, if FPR is a

quantity of interest, it is essential to include a non-trivial

proportion of sites simulated at or near neutrality, otherwise a false

positive would only occur if a site simulated under reasonably strong

constraint (if the log-normal is peaked), is misclassified as POSITIVELY

selected. It should be easy to simply add the category of dN/dS = 1 for

INdelible simulations.

3. I also don't think that is is enough to pick a single biological

selection profile (IAV HA) and draw generalized conclusions from it. IAV

HA is a classic case of a low average dN/dS gene with patches of

selection. In addition, the biological process of selection in IAV HA is

most likely that of EPISODIC selection, when discrete antigenic shifts

change antibody neutralization capacity dramatically. This leads to

counterintuitive results when analyzing smaller alignments yields

apparent greater power to detect selection when looking at dN/dS

averaged over time.

It would be instructive to consider a different selective profile,

something like an antiviral factor (APOBEC3G or TRIM5alpha, see the

corresponding papers from Harmit Malik's group), or the well studies

sperm lysin (or a self-incompatibility locus). The distribution of dN/dS

would quite different (mean closer to one), and a more of a challenge to

classify.

4. INdelible generates data WITHOUT synonymous rate variation (FUBAR

assumes dS varies from site to site as well); using the GY94 frequency

parameterization (FUBAR uses MG94 CF3x4); and assuming the HKY85

nucleotide model (FUBAR uses GTR). M8 is essentially the same model as

what is being used to generate the data, except the dN/dS distribution

is a 10-bin discretized beta + a point mass at > 1.

Running FUBAR on data without synonymous rate variation is essentially

constraining it to use 20 a-propri fixed points to model the variation

in dN (dS is fixed), while M8 is free to fit a whole parametric

distribution to it. FUBAR is more or less wasting 95% of it's dS, dN

grid in this scenario.

Since INdelible does not support dS variation, it is probably easiest to

"balance" the comparison, by running a version of FUBAR which does

assumes a constant dS but puts a more dense grid on dN (at the same

computational cost). It is actually fairly easy to do by making small

modifications to the FUBAR batch file (I'll be happy to tell the authors

how to do it in < 15 minutes) to assume dS = 1.

That said, I would argue (and I know that the Wilke lab published

papers drawing similar conclusions) that ALL biological data contain

SOME level of dS variation.

Minor suggestions/comments.

1. I don't have a good sense of how many alignment sites were being

filtered out on average; are we talking about only a few per simulation?

Perhaps this information can be included in Table 1, e.g. for each

filtering strategy and sequence count, add a column with median and IQR

or percent of sites filtered per replicate.

2. In Table 1, it would be useful to see the TPR under the \_correct\_

alignment (i.e. using the truly homologous positions as generated by

INdelible). I don't know if that's trivial to do, since I have not used

INdelible and don't know if it outputs gapped aligned or unaligned

sequences

3. In Figure 1 panels B and C can be combined since the curves don't

overlap and can be labeled without ambiguity.

4. Looking at Figure 3 I was somewhat intrigued by the "flipped"

behavior of TPR beween methods: FUBAR has a convex curve when using

filtered alignments and a sigmoid - like curve when using unfiltered

alignments, and for PAML the opposite behavior holds. Why may that be?

Also, the labeling seems off -- both black and grey line seem to be

plotting "Unfiltered" alignments.

5. Larger trees can lead to "difficult alignments" even if mean root

path length is reasonably low, e.g. simply because it is harder to infer

a good guide tree from pairwise sequence distances or other "crude"

metrics for many sequences. I wonder if this is a possible area where

filtering may be beneficial? For example, in Table 1 FUBAR was starting

to show improvement for 60 and 150 sequences when using filtering.

6. Unless I am mistaken, including the simulation count as a random

effect in the mixed effects model relies on the sample size of two to

estimate the random effect regression coefficients; does this lead to

model overfitting, because there effectively is a separate model

parameter per simulation?