Dear Dr. Hyndman,

I appreciate the quick reviews of the second manuscript iteration. The changes described below will hopefully address the remaining comments.

Thank you for considering this revised version of the manuscript.

Erik

EDITOR'S LETTER

Dear Erik

Thank you for your submission to the R Journal, titled "Fast and Flexible Search for Homologous Biological Sequences with DECIPHER v3". I sent the revised version back to the same two reviewers, and their reports are attached. One has some further requests for minor revisions.

We would appreciate a revised version and point-by-point response to the reviewers comments within 6 weeks. Remember, that when responding to the reviewer's comments, your job is to persuade me, the editor, that either you've dealt with the issue, or that it's not relevant. To this end, please produce a single document that includes all the reviewers comments mingled with your responses.

When you have your revised article ready, please submit a zip containing all relevant files through the article submission at https://journal.r-project.org/submissions.html. You will need to include the article id listed in the Subject line into the form.

I look forward to seeing the next iteration.

Best wishes,
Rob Hyndman
(Editor, R Journal)

FIRST REVIEW

My comments have been adequately addressed.

A major change is removing the material on read alignment. This seems like an excellent decision to me. It clearly wasn't the main interest of the author, and there is already a lot of highly optimized software for this.

Correcting the main thing missing in the previous version, DECIPHER performance is now compared with the MMseqs2 and BLAST programs, and with Biostrings functions. I also appreciate the extra details on the algorithm.

I thank the reviewer for the constructive review, which I agree improved the manuscript.

SECOND REVIEW

The revised version looks much better, and the vast majority of my comments have been addressed well. This is largely due to the omission of the short- and long- read mapping sections from the previous version, which I think was a good choice.

I appreciate the reviewer's close look at the revised manuscript and constructive feedback.

One comment I feel was not adequately addressed is:

9. *Section 5 and Figure 4. It is not clear what point is being made by highlighting the number of matched annotation orthologs in Figure 4. These seem to be largely correlated with the total number of predicted orthologs.*

"The goal of Figure 4 was to show (1) improvements with lower k-mer length, (2) improvements with lower step size, and (3) improvements with amino acid over nucleotide search. In the absence of matched annotations, the improvements would be predicated on the assumption that more predictions are better – which is likely true but nonetheless an assumption. With the inclusion of matched annotations, it is clearer that the increased number of predictions are also due to more matching annotations. I edited the revised text to clarify this point."

I believe the authors' response and revised manuscript do not really address this point. I would like to see the author demonstrate that there are more “matching annotations” than expected by chance, perhaps as an odds ratio, and how changing k-mer length and step size affects this odds ratio.

My interpretation of this comment is that the reviewer would like me to compare the significance of drawing *M* matching annotations from *N* predicted pairs across conditions. For example, at a k-mer length of 7 amino acids and a step size of 5, there were *N*=9761 predicted pairs and *M*=2293 matching annotations. Among the predicted pairs, there were *K*=2297 annotation matches possible. Each of *K* trials had a 1/*N* chance of randomly landing on its pair partner. Therefore, the probability of M of K successes is approximately dbinom(*M*, *K*, 1/*N*).

I added p-values to Figure 3 in response to this comment. Given the rarity of encountering a matching annotation by chance (1/N per match), it is clear from the increase in significance that the additional matching annotations predicted are not due to chance alone. If the reviewer would prefer something else then more information is needed. Specifically, with respect to the suggestion of using log-odds it was unclear to me what the contingency table was meant to contain. An example would be appreciated if the addition of p-values was insufficient to assuage the reviewer's concern.

I also have a few suggestions for the newly added Figure 3.

1. Suggest adding titles or characters A) and B) to the left and right panels in the plot so it is clearer which part of the figure is being discussed in section 3 and described in legend.

Done.

2. State the step-size parameter for DECIPHER used in the right panel

This was already part of the figure legend. Note the two panels are also color matched.

3. In Fig3, The meanings of “AUC1”, “Average AUC1 up to the first false positive” (x axis of left panel), “probability of detecting true positives before false positives” and “Probability of detection” (y-axis of right panel) are confusing and should be better clarified. Is not “Average AUC1 up to the first false positive” a tautology based on the definition given? It is also not clear to me if “probability of detecting true positives before false positives” means the same thing as “Probability of detection” on the y axis of the right panel.

I clarified the text in response to this comment. The "AUC1" is a standard commonly used in benchmarking search programs, although I think this is a poor choice of name. The idea is simply to quantify the fraction of the possible true positives found before the first positive is found. Better search algorithms will have a better scoring of hits, so true positives will appear more often with higher scores than false positives.

4. (minor) The right panel of Fig 3 would be easier to read as a line graph than a bar graph.

I considered the option of converting the right panel **(B)** to a line graph, which is how I had originally plotted the data. My concern is that readers may confuse the left legend line types with the right panel lines, and there is no good way to distinguish the lines between **(A)** and **(B)** panels. Furthermore, it is difficult to see the high-end (~100%) because all the lines are on top of each other. Therefore, I chose to stick with a bar plot.