

Reviewer 1

Major comments

Optimizing a criterion is a good idea. It would be useful to clarify if each of the other algorithms aims to optimize some other criterion and if so which one. There is some text about this in the discussion, but this information might work better earlier in the manuscript. It would also be helpful to see how each algorithm performs on these other metrics (e.g., minimizing FPs). You can argue that MCC is a better criterion to use, but it is also good to show if the methods that aim to optimize another criterion actually do so and how MCC-based OptiClust does with regards to these metrics. In addition, some further discussion about the relative costs of FPs and FNs in typical metagenomics applications and downstream studies would provide some context for your argument to use MCC.

For the simulations, it would be nice to show the true values of some of the output measures (e.g., number of clusters) versus the estimated values.

A systematic evaluation of the influence of additional variables on performance (i.e., accuracy metrics, RAM, and speed) would provide better insight into expected performance on other data sets. The analyzed data sets do cover a broad range of environments with different types of communities. Even with those data, it would be nice to see performance metrics plotted versus phylogenetic diversity, percentage of sequences in the reference database (for pre-assignment analyses), and sequencing variables (e.g., read length, quality scores). In the simulations, you could additionally introduce realistic sequencing errors evaluate how performance changes.

This may be beyond the scope, but it would be very interesting to see the performance of OptiClust with other criteria as the target of optimization. In other words, how much of the performance is due to targeting MCC as the criterion versus the algorithm itself as an alternative to hierarchical clustering. You could investigate this by comparing the OptiClust algorithm minimizing FPs, for example, as compared to hierarchical clustering minimizing

FPs. If this analysis is not performed, I recommend discussing that the algorithm for building clusters and the criterion that is being optimized are evaluated together in your analyses and not disentangled.

Providing the analysis pipeline is great. Users who want to apply the method may like to have more description of the implementation and a stable link to the code (even if this is the same as mothur). To enable applications outside OTU clustering, you should also have a code base that is not embedded in mothur.

Minor comments

Line 69: Typo: “to be recapitulate”

Line 132: Not clear why the number of OTUs is mentioned here

Line 138: With partial convergence, does the output include the last value for change in MCC?

Lines 140-150: Mention if agglomerative hierarchical methods are also better than divisive (cite literature or own data).

Lines 151-153: Processing the sequences/OTUs deterministically is a good option for alleviating dependence on data order.

Line 161: Typo: “values considerably smaller”

Line 170: Is disk or CPU also useful to measure?

Lines 182-185: Clarify that this is assuming the pairwise similarities are similar to the analyzed data set. Can you say something about what happens as diversity increases or decreases?

Line 187: Remind the reader that this is “open reference”

Line 189: Typo: “reduce” should be past tense

Line 197: Are all of these with taxonomic assignment first? Clarify.

Line 204: Explain why (i.e., what types of errors the extra clusters represent).

Line 232: Typo: “effectively ask”

Line 239: Typo: “are encourage”

Reviewer 2

Major comments

This is a well described, self-enclosed, well implemented piece of work that does not require major revision. However, there are minor issues which should be addressed.

Minor comments

On page 6 and 7 the authors compare the speed of OptiClust to other methods; however, I do not think Figure 1 really conveys this, mostly due to the log-scaled y-axis for the time panel. Whilst I recognise why the log-scale has been used, the authors may consider a different way to visually present the data that better conveys the message “The OptiClust algorithm was considerably faster than the hierarchical algorithms and somewhat slower than the heuristic-based algorithms”

The comparison of RAM used (at the top of page 7) may also be better demonstrated visually.

It wasn’t clear to me which software implementation was used for “Nearest Neighbour” and “Furthest Neighbour”.

The authors describe the potential implications of optimising a metric other than MCC - I

wonder if they have actually done this?