From: **Sebastian Schmidt** sebastian.schmidt@imls.uzh.ch
Subject: Re: Decision on your PeerJ submission: "De novo clustering methods out-perform reference-based methods for assigning 16S rRNA gene sequences to operational taxonomic units" (#2015:10:7412:0:0:REVIEW)
Date: November 12, 2015 at 6:42 PM
To: A. Murat Eren meren@mbl.edu
Cc: Patrick Schloss pschloss@umich.edu, PeerJ peer.review@peerj.com

SS

Dear Patrick,

thank you for your message, and for coming back to me directly about this point. I agree: this is what open review can be great for, and of course I also agree that this exchange be public if that is technically possible / in line with PeerJ policy.

To the point. I was indeed undecided on whether or not to include the point on the metrics in the review, for two reasons. First, as I said, I feel a bit uneasy about promoting work I was involved in so bluntly. Second, I completely agree with what you said: He et al also used the MCC for these kinds of tests, so your replication of their experiments is of course valid.

However, there is a little more to the story. When we prepared our said study (2015, Env Microbiol), we also considered using the MCC for the tests we conducted. Of course, I had read your very nice 2011 paper where you introduced it for the problem of testing OTU quality. Eventually, we decided against the MCC for the types of tests we conducted, and went with AMI, NMI & ARI instead, for the reasons I gave in the review, and after consulting an expert in statistics. Some of the tests we ran were similar in spirit to the downsampling experiments that you do in your present study, albeit of course much less systematic (with only two data points of artificial subsets). As it happens, the corresponding author of the He study, Hong-Wei Zhou, and I both gave a talk in the same session at a conference in 2014, and discussed for some time afterwards. Part of that exchange was also about whether or not their types of tests and our types of tests made sense for assessing OTU “stability”, “reproducibility” or “robustness” – whatever you may call it. So you could say that the issue of MCC vs AMI/NMI/ARI has already been on my mind for some time.

I do believe that the MCC is an appropriate measure of OTU quality, in the sense in which you introduced it in 2011 and also use it in the present study. But as I said, I’m having a slightly hard time understanding what the MCC values actually mean in the down-sampling experiments; it certainly has a very different interpretation in this framework than in the more “classical” setup based on sequence distances. The metrics I suggested are also far from perfect, but they were designed for exactly this kind of problem: to compare different clusterings and assess their “agreement” or “overlap” in an as unbiased as possible way. We settled on the AMI as preferred measure eventually, because it seems to suffer the least from OTU count and size distribution effects. Nevertheless, unlike the MCC, these metrics cannot give any idea of “quality”, only of “stability” or “reproducibility”. I am completely prepared to believe that the MCC can be interpreted as a “stability” measure in the way that you compute it; I simply have a hard time figuring it out and interpreting it, honestly.

Also, it may very well be that the MI-based measures would give very similar trends as you observe with the MCC anyway, making the whole exercise slightly redundant. However, my intuition, based also on the very limited tests on down-sampling that I have done using AMI, would be that the heuristic methods tend to suffer more, as they are generally noisier and less deterministic. In our two-datapoint test, AL was by far the most “stable”, whereas UCLUST was very noisy indeed. In any case, I also realised after sending the review that the way you (and He et al) calculate the MCC in downsampling inherently favours the closed-ref and open-ref methods. USEARCH does a good job at reproducibly matching a sequence to the same “centroid” when you run it multiple times with identical (or randomized) input. But database matching is a very different problem from clustering. And given the strong effect of reference database sequence order that you show later in the paper, I guess it would be a “fairer” comparison to also randomize the sequence order of the database, not only of the queries. This would obviously apply, regardless of whether you stick with the MCC or whether you re-run some tests with other metrics.

OK, after a lot of text and explanations, here’s what I think. I agree with you that the He et al concept of “stability” is problematic, and that OTU stability is not the most important feature by itself. You can be very consistent in making very bad clusters. If critically reproducing the He study and pointing out problematic points was your main goal, you have indeed done a very good job (as you said, both reviewers agree on that). Nevertheless, I contend that “stability”, “robustness” or “reproducibility”, the way that I would define them, are certainly desirable features of OTU clusterings, and OTU “quality” seems to often go with “stability” (for example, AL produced the most stable OTUs in all our tests and wins by MCC in your 2011 tests). It’s great if you make good OTUs, but even better if you are consistently doing so. My suggestion to repeat some of the down-sampling tests using alternative metrics was motivated by all the reasons I gave above, plus the feeling that the study could (further) gain by providing an alternative assessment of “stability” which (at least to me) is more sensible than the one proposed by He et al.

I would leave the decision on whether or not to include such alternative tests to you, and of course primarily to Murat. Given all the reasons above, I see this as a recommendation, but I certainly do not insist that it must be done, and I would of course not recommend a rejection should you choose to not add these tests. As I said, the manuscript stands and has merit as is; it is mostly due to the MCC-vs-AMI/NMI/ARI question being on my mind for a long time beforehands that I pointed all this out.

With very best regards (also to Murat!),

Sebastian

PS: Just to avoid confusion. Yes, I did sign with “Sebastian”, not “Thomas”. This is a very long story, and mostly my parents are to blame…
PPS: Should you decide to give AMI / NMI / ARI a shot, NX Vinh, the original author of the AMI paper, uploaded a short R script to compute these to his homepage (https://sites.google.com/site/vinhnguyenx/softwares)

Sebastian Schmidt, PhD
Postdoctoral Researcher
Institute of Molecular Life Sciences
University of Zurich
Winterthurerstrasse 190
Y55-L56
CH-8057 Zurich

Phone: +41 44 63 53 148

Phone: +41 44 63 551 48
sebastian.schmidt@imls.uzh.ch

On 12Nov, 2015, at 19:57 , A. Murat Eren <meren@mbl.edu> wrote:

Hi,

I just wanted to chime in very quickly in case PeerJ staff, and Thomas would like to know my stance:

I am absolutely fine with this exchange, and furthermore very happy to witness a practical example of how openness can improve communication and make the scientific evaluation / peer-review process a collaborative endeavour. If everyone agrees, I think the exchange between the two parties should be a part of the review record as Pat also suggested.

Pat, thank you very much for including PeerJ staff and me in this conversation.

Best wishes,

--

A. Murat Eren (meren)
http://meren.org :: *gpg*

On Thu, Nov 12, 2015 at 12:33 PM, Patrick Schloss <pschloss@umich.edu> wrote:

> Hi Thomas (and Meren),
>
> I hope it isn't inappropriate to email you regarding your review. Feel free to tell me to go fly a kite. I hope that since you signed your review that this signaled an openness to be contacted. I hope this is fine with Meren! I would be open to you adding this to the peer-review record if the reviews are co-published with the manuscript.
>
> I'm in the process of working through your comments and that of the other reviewer. I could judge from your review that you were a bit on the fence about whether to recommend we use the ARI/NMI/AMI in place of the MCC for the measurement of stability in the first part of the manuscript where we try to replicate the He paper's work. My initial inclination is to stick with the MCC approach, which was the method used by the original He study. If we were to do one of the other methods instead, then I feel like that would confuse the issue and would be a different type of replication. I feel that indicating that they perhaps picked the wrong metric would add unnecessarily to the figures and text when the point of the first section of the Results section is that their idea of "stability" is subservient to the quality of OTU assignments. Furthermore, the consistency of that quality over multiple random seeds is probably a better metric of "stability" than to compare the subsampled OTU assignments to the full dataset assignments. That being said, I feel like we've done a good job of addressing the quality/stability issue and the reviewers both seemed to agree. If you both feel like this would be an important issue, then we'll do it, but it just seems like piling on when we've already made the point that their concept of "stability" is problematic. Your thoughts?
>
> Thanks again for providing an open review and I have no problem with people pushing their own work when it fits!
>
> Sincerely,
> Pat Schloss

> On Nov 9, 2015, at 7:44 PM, PeerJ <peer.review@peerj.com> wrote:
>
> **PeerJ**
>
> Thank you for your submission to PeerJ. I am writing to inform you that in my opinion as the Academic Editor for your article, your manuscript "De novo clustering methods out-perform reference-based methods for assigning 16S rRNA gene sequences to operational taxonomic units" (#2015:10:7412:0:0:REVIEW) requires a number of major revisions before we could accept it for publication.
>
> The comments supplied by the reviewers on this revision are pasted below. My comments are as follows:
>
> ## Editor's comments
>
> Dear authors,
>
> Thank you very much for your effort to make your analyses easy to reproduce and follow!