OptiFit: a fast method for fitting amplicon sequences to existing OTUs

Kelly L. Sovacool[1], Sarah L. Westcott[2], M. Brodie Mumphrey[1], Gabrielle Dotson[1], Patrick D. Schloss[2]

1 Department of Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, MI 48109
2 Department of Microbiology and Immunology, University of Michigan, Ann Arbor, MI 48109

Assigning amplicon sequences to Operational Taxonomic Units (OTUs) is an important step in characterizing the composition of microbial communities across large datasets. OptiClust, a *de novo* OTU clustering method in the mothur program, has been shown to produce higher quality OTU assignments than other methods and at comparable or faster speeds (1, 2). However, in some cases one may wish to incorporate new samples into a previously clustered dataset without performing clustering again on all sequences, such as when deploying a machine learning model where OTUs are features (3). To provide an efficient & robust method to fit amplicon sequence data to existing OTUs, we developed the OptiFit algorithm as a new component of the mothur program. To benchmark the OptiFit algorithm against *de novo* clustering with the OptiClust algorithm, we used four published datasets isolated from soil, marine, mouse, and human samples. For each dataset, a subset of sequences was randomly selected and clustered into OTUs with OptiClust, then the remaining sequences were fit to the existing OTUs using the OptiFit algorithm. This was repeated with subsets of varying sizes ranging from 10 to 90% of sequences in order to evaluate the bounds of the dataset size required for OptiFit. Separately, all sequences were clustered with OptiClust to provide a baseline of OTU assignment quality and runtime performance. Each of these routines was repeated 10 times with different random seeds to produce results that are robust to random variation. OTU quality was evaluated using the Matthews Correlation Coefficient (MCC) with a sequence similarity threshold of 97% as described previously (4, 5). On average, fitting sequences into existing OTUs with OptiFit performed 10 times faster than *de novo* clustering with OptiClust, while the average MCC scores produced were nearly indistinguishable across each dataset. The OptiFit results across subset sizes ranging from 10 to 90% of sequences were also very similar, with slightly higher MCC scores for larger subset sizes. Thus, OptiFit is an efficient way to fit new sequences to existing OTUs yet without sacrificing the quality of OTU assignments.

References

1. **Westcott SL**, **Schloss PD**. 2017. OptiClust, an Improved Method for Assigning Amplicon-Based Sequence Data to Operational Taxonomic Units. mSphere **2**:e00073–17. doi:10.1128/mSphereDirect.00073-17.

2. **Schloss PD**, **Westcott SL**, **Ryabin T**, **Hall JR**, **Hartmann M**, **Hollister EB**, **Lesniewski RA**, **Oakley BB**, **Parks DH**, **Robinson CJ**, **Sahl JW**, **Stres B**, **Thallinger GG**, **Horn DJV**, **Weber CF**. 2009. Introducing mothur: Open-Source, Platform-Independent, Community-Supported Software for Describing and Comparing Microbial Communities. Appl Environ Microbiol **75**:7537–7541. doi:10.1128/AEM.01541-09.

3. **Topçuoğlu BD**, **Lesniak NA**, **Ruffin M**, **Wiens J**, **Schloss PD**. 2019. Effective application of machine learning to microbiome-based classification problems. bioRxiv 816090. doi:10.1101/816090.

4. **Schloss PD**. 2016. Application of a Database-Independent Approach To Assess the Quality of Operational Taxonomic Unit Picking Methods. mSystems **1**:e00027–16. doi:10.1128/mSystems.00027-16.

5. **Westcott SL**, **Schloss PD**. 2015. De novo clustering methods outperform reference-based methods for assigning 16S rRNA gene sequences to operational taxonomic units. PeerJ **3**:e1487. doi:10.7717/peerj.1487.