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

2020-12-08

Kelly L. Sovacool[1], Sarah L. Westcott[2], M. Brodie Mumphrey[1], Gabrielle A. 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

† To whom correspondence should be addressed: pschloss@umich.edu

Abstract

Importance

Introduction

Results

- 16S rRNA gene sequence datasets from human gut, mouse gut, marine, and soil environments were processed with mothur and clustered *de novo* with mothur's OptiClust algorithm.
- All clustering was performed as a distance threshold of 0.03.
- The Matthews Correlation Coefficient was calculated to serve as a measure of OTU quality.

Reference clustering with public databases

- Public reference databases (greengenes, silva, and RDP) were clustered *de novo* using OptiClust, then query datasets were fit to the *de novo* OTUs using OptiFit.
  - In open-reference mode, OTU quality was similar between fitting the datasets to reference OTUs with OptiFit and clustering the datasets *de novo* with OptiClust.
  - However, in closed-reference mode, OTU quality was slightly worse when fitting to greengenes and silva, and much worse when fitting to RDP as compared to OptiClust. OptiFit was able to map more query sequences to reference OTUs created with the greengenes and silva databases than with RDP.
  - In terms of runtime, closed-reference OptiFit outperformed OptiClust, while OptiClust out-performed open-reference OptiFit.
  - These results held true for all four datasets and all three reference databases.

Reference clustering with split datasets

- Datasets were randomly split into a reference fraction and a query fraction. Reference sizes from 10% to 80% of the sequences were created, with the remaining sequences used for the query. Reference sequences were clustered *de novo* with OptiClust.

Query sequences were then fit to the *de novo* OTUs with OptiFit.

  - OTU quality from fitting split datasets was highly similar to that from *de novo* clustering the whole dataset.
  - Closed-reference OptiFit with split datasets was faster than OptiClust on whole datasets.
    - OptiClust performed faster than open-reference OptiFit only when the OptiFit reference fraction was 30% or less.
  - Different methods for selecting the sequences to be used as the reference were tested; a simple random sample, weighting sequences by relative abundance, and weighting by similarity to other sequences in the dataset.
    - OTU quality was similar with the simple and abundance-weighted sampling, but slightly worse with similarity-weighted sampling.
    - The fraction of query sequences that are able to be fit to the reference OTUs in closed-reference mode decreases as the reference fraction increases.

Comparison to vsearch

- To compare to existing software, vsearch was used to cluster OTUs *de novo* or with reference-based clustering to the greengenes database.
  - For all datasets and clustering methods (*de novo*, open reference, and closed reference), mothur's clustering algorithms produced higher quality OTUs than vsearch.
  - When closed-reference clustering against the greengenes database, vsearch was able to map more query sequences to the reference than mothur's OptiFit algorithm.
  - In terms of runtime, OptiFit genearlly performed faster than vsearch when reference clustering, while vsearch *de novo* clustering outperformed OptiClust.

Discussion

Materials and Methods

The OptiFit Algorithm

Analysis Workflow

Acknowledgements

KLS received support from the NIH Training Program in Bioinformatics (T32 GM070449).

The funders had no role in study design, data collection and interpretation, or the decision to submit the work for publication.

Author Contributions

References