Machine learning classification by fitting amplicon sequences to existing OTUs

Running title: self-reference-based OTU clustering for ML classification

Courtney R. Armour[1], Kelly L. Sovacool[2], William L. Close[1,*], Begüm D. Topçuoğlu[1,#], Jenna Wiens[3], Patrick D. Schloss [1,†]

[1] Department of Microbiology and Immunology, University of Michigan, Ann Arbor, Michigan, USA

[2] Department of Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, Michigan, USA

[3] Department of Electrical Engineering and Computer Science, University of Michigan, Ann Arbor, Michigan, USA

[*] Current Affiliation: Bio-Rad Laboratories, Hercules, California, USA

[#] Current Affiliation: Bristol Myers Squibb, Summit, New Jersey, USA

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

Observation format

Abstract

Machine learning classification using the gut microbiome relies on assigning 16S rRNA gene sequences into operational taxonomic units (OTUs) to quantify microbial composition. OTU abundances are then used to train a classification model that can be applied to classify new samples. The standard approaches to clustering sequences include reference-based and *de novo* clustering. Reference-based clustering requires a well-curated reference database that may not exist for all systems. *De novo* clustering tends to produce higher quality OTU assignments than reference-based, but clusters depend on the sequences in the dataset and therefore OTU assignments will change when new samples are sequenced. This lack of stability complicates machine learning classification since new sequences must be reclustered with the old data and the model retrained with the new OTU assignments. The OptiFit algorithm addresses these issues by fitting new sequences into existing OTUs. While OptiFit produces high quality OTU clusters, it is unclear whether this method for fitting new sequence data into existing OTUs will impact the performance of classification models trained with the older data. We used OptiFit to cluster sequences into existing OTUs and evaluated model performance in classifying a dataset containing samples from patients with and without colonic screen relevant neoplasia (SRN). We compared the performance of this model to standard methods including *de novo* and database-reference-based clustering. We found that using OptiFit performed as well or better in classifying SRNs. OptiFit can streamline the process of classifying new samples by avoiding the need to retrain models using reclustered sequences.

Importance

There is great potential for using microbiome data to aid in diagnosis. A challenge with OTU-based classification models is that 16S rRNA gene sequences are often assigned to OTUs based on similarity to other sequences in the dataset. If data are generated from new patients, the old and new sequences must all be reassigned to OTUs and the classification model retrained. Yet there is a desire to have a single, validated model that can be widely deployed. To overcome this obstacle, we applied the OptiFit clustering algorithm to fit new sequence data to existing OTUs allowing for reuse of the model. A random forest model implemented using OptiFit performed as well as the traditional reassign and retrain approach. This result shows that it is possible to train and apply machine learning models based on OTU relative abundance data that do not require retraining or the use of a reference database.

There is increasing evidence for an association between the composition of the gut microbiome and a variety of diseases, such as crohn's disease and colorectal cancer (1, 2). There is great potential to diagnose disease with gut microbiome sequence data and machine learning. Taxonomic composition of microbial communities can be assessed using amplicon sequencing of the 16S rRNA gene, which is the input to classification models. Analysis of 16S rRNA gene sequence data generally relies on assigning sequences into operational taxonomic units (OTUs). The process of OTU clustering can either be reference-based or *de novo*. The quality of OTUs generated with reference-based clustering is generally poor compared to those generated with *de novo* clustering (3). While *de novo* clustering produces high-quality OTU clusters where sequences are accurately grouped based on similarity thresholds, the resulting OTU clusters depend on the sequences within the dataset and the addition of new data has the potential to redefine OTU cluster composition. The unstable nature of *de novo* OTU clustering complicates deployment of machine learning models since integration of additional data requires reclustering all the data and retraining the model. The ability to integrate new data into a validated model without reclustering and retraining could allow for the application of a single model that can continually classify new data. Recently, Sovacool *et al.* introduced OptiFit, a method for fitting new sequence data into existing OTUs (4). While OptiFit can effectively fit new sequence data to existing OTU clusters, it is unknown if the use of OptiFit will have an impact on classification performance. Here, we tested the ability of OptiFit to cluster new sequence data into existing OTU clusters for the purpose of classifying disease based on gut microbiome composition.

We compared the ability of several approaches for assigning 16S rRNA gene sequences to OTUs including, *de novo* and reference-based clustering. For reference-based clustering, we used closed-reference clustering to a public database (database-reference-based) and to OTUs generated from a subset of the samples (self-reference-based). To test how the model

performance compared between these approaches, we used a publicly available dataset of 16S rRNA gene sequences from stool samples of healthy subjects (n = 226) as well as subjects with screen-relevant neoplasia (SRN) consisting of advanced adenoma and carcinoma (n = 229) (5). For the *de novo* workflows, all the 16S rRNA sequence data was clustered into OTUs. The OTU clustering was conducted using two common algorithms: 1) the OptiClust algorithm in mothur (6) and 2) the VSEARCH algorithm used in QIIME2 (7, 8). For both algorithms, the resulting abundance data was then split into training and testing sets, where the training set was used to tune hyperparameters and ultimately train and select the model. The model was applied to the testing set and performance evaluated (Figure 1A). We also conducted reference-based OTU clustering using OptiFit to fit the sequence data into OTUs based on the greengenes reference database. To compare with another commonly used method, we also used the VSEARCH algorithm to fit the sequence data to the greengenes reference (Figure 1B). In the OptiFit self-reference workflow, the data was split into a training and a testing set. The training set was clustered into OTUs and used to train a classification model. The OptiFit algorithm was used to fit sequence data of samples not part of the original dataset into the existing OTUs, and used the same model to classify the samples (Figure 1C). For each of the workflows the process was repeated for 100 random splits of the data to account for variation caused by the choice of the random number generator seed.

We first examined the quality of the resulting OTU clusters from each method using the Matthews correlation coefficient (MCC). MCC is a metric used to measure OTU cluster quality based on the similarity of all pairs of sequences and whether they are appropriately clustered or not (3). MCC scores range between negative one and one, and measure how well clustering assignment correlates with the distance between sequences. To ensure that OptiFit appropriately integrated new sequence data into the existing OTUs, we expected the MCC scores produced by the OptiFit workflow to be similar to that of *de novo* clustering using the

OptiClust algorithm. In the OptiFit workflow the test data was fit to the clustered training data for each of the 100 data splits resulting in an MCC score for each split of the data. In the remaining workflows, the data was only clustered once and then split into the training and testing sets resulting in a single MCC score for each method. Indeed, the MCC scores were similar between the OptiClust *de novo* (MCC = 0.884) and OptiFit self-reference workflows (average MCC = 0.879, standard deviation = 0.002). Consistent with prior findings, the reference-based methods produced lower MCC scores (OptiFit Greengenes MCC = 0.786; VSEARCH Greengenes MCC = 0.531) than the *de novo* methods (OptiClust *de novo* MCC = 0.884; VSEARCH *de novo* MCC = 0.641) (4). Another metric we examined for the OptiFit workflow was the fraction of sequences from the test set that mapped to the reference OTUs. Since sequences that did not map to reference OTUs were eliminated, if a high percentage of reads did not map to an OTU we expected this loss of data to negatively impact classification performance. We found that loss of data was not an issue since on average 99.8% (standard deviation = 0.68%) of sequences in the subsampled test set mapped to the reference OTUs. This number is higher than the average fraction of reads mapped in the OptiFit Greengenes workflow (96.8% +/- 3.5). These results indicate that the OptiFit self-reference method performed as well as the OptiClust *de novo* method and better than using an external database.

We next assessed model performance using OTU relative abundances from the training data from the workflows to train a model to predict SRNs and used the model on the held out data. Using the predicted and actual diagnosis classification, we calculated the area under the receiver operating characteristic curve (AUROC) for each data split. During cross-validation (CV) training, the performance of the OptiFit self-reference and OptiClust *de novo* models were not significantly different (p-value = 0.066; Figure 2A), while performance for both VSEARCH methods was significantly lower than the OptiClust *de novo*, OptiFit self, and OptiFit Greengenes methods (p-values < 0.05). The trained model was then applied to the test data

classifying samples as either control or SRN. The VSEARCH Greengenes method performed slightly worse than the OptiClust *de novo* method (p-value = 0.030). However the performance on the test data for the OptiClust *de novo*, OptiFit Greengenes, OptiFit self-reference, and VSEARCH *de novo* approaches were not significantly different (p-values > 0.05; Figures 2B and 2C). These results indicate that new data could be fit to existing OTU clusters using OptiFit without impacting model performance.

We tested the ability of OptiFit to integrate new data into existing OTUs for the purpose of machine learning classification using OTU relative abundance. A potential problem with using OptiFit is that any sequences from the new samples that do not map to the existing OTU clusters will be discarded, resulting in a possible loss of information. However, we demonstrated that OptiFit can be used to fit new sequence data into existing OTU clusters and it could perform as well in predicting SRN compared to *de novo* clustering all the sequence data together. In this instance, the performance of OptiFit was equivalent to using a database-reference-based method despite the lower quality of the OTU clusters in the database-reference-based approach. This likely indicates that the sequences that are important to the model are well characterized by the reference database. However, a less well studied system may not be as well characterized by a reference-database which would make the ability to utilize one's own data a reference an exciting possiblility. The ability to integrate data from new samples into existing OTUs enables the implementation of a single machine learning model. This is important for model implementation because not all of the data needs to be available or known at the time of model generation. A robust machine learning model can be implemented as part of a non-invasive and low-cost diagnostic for SRN and other diseases.

Materials and Methods

Dataset. Raw 16S rRNA gene sequence data from the V4 region were previously generated from human stool samples. Sequences were downloaded from the NCBI Sequence Read

Archive (accession no. SRP062005) (5, 9). This dataset contains stool samples from 490 subjects. For this analysis, samples from subjects identified in the metadata as normal, high risk normal, or adenoma were categorized as “normal”, while samples from subjects identified as advanced adenoma or carcinoma were categorized as “screen relevant neoplasia” (SRN). The resulting dataset consisted of 261 normal samples and 229 SRN samples.

Data processing. The full dataset was preprocessed with mothur (v1.47) (10) to join forward and reverse reads, merge duplicate reads, align to the SILVA reference database (v132) (11), precluster, remove chimeras with UCHIME (9), assign taxonomy, and remove non-bacterial reads following the Schloss Lab MiSeq standard operating procedure described on the mothur website (https://mothur.org/wiki/miseq_sop/). 100 splits of the 490 samples were generated where 80% of the samples (392 samples) were randomly assigned to the training set and the remaining 20% (98 samples) were assigned to the test set. Using 100 splits of the data accounts for the variation that may be observed depending on the samples that are in the training or test sets. Each sample was in the training set an average of 80 times (standard deviation = 4.1) and the test set an average of 20 times (standard deviation = 4.1).

Reference-based workflows.

1. OptiFit Self: The preprocess data was split into the training and testing sets. The training set was clustered into OTUs using OptiClust, then the test set was fit to the OTUs of the training set using the OptiFit algorithm (4). The OptiFit algorithm was run with method open so that any sequences that did not map to the existing OTU clusters would form new OTUs. The data was then subsampled to 10,000 reads and any novel OTUs from the test set were removed. This process was repeated for each of the 100 splits resulting in 100 training and testing datasets.

2. OptiFit Greengenes: Reference sequences from the Greengenes database v13_8_99 (12) were downloaded and processed with mothur by trimming to the V4 region and clustered *de novo* with OptiClust (6). The preprocessed data was fit to the clustered reference data using OptiFit with the method open to allow any sequences that did not map to the existing reference clusters would form new OTUs. The data was then subsampled to 10,000 reads and any novel OTUs from the test set were removed. The dataset was then split into two sets where 80% of the samples were assigned to the training set and 20% to the testing set. This process was repeated for each of the 100 splits resulting in 100 training and testing datasets.

3. VSEARCH Greengenes: Preprocessed data was clustered using VSEARCH v2.15.2 (7) directly to unprocessed Greengenes 97% OTU reference alignment consistent with how VSEARCH is typically used by the QIIME2 software for reference-based clustering (8). The data was then subsampled to 10,000 reads and any novel OTUs from the test set were removed. The dataset was then split into two sets where 80% of the samples were assigned to the training set and 20% to the testing set. This process was repeated for each of the 100 splits resulting in 100 training and testing datasets.

De novo workflows.

1. OptiClust *de novo*: All the preprocessed data was clustered together with OptiClust (6) to generate OTUs. The data was subsampled to 10,000 reads per sample and the resulting abundance tables were split into the training and testing sets. The process was repeated for each of the 100 splits resulting in 100 training and testing datasets.

2. VSEARCH *de novo*: All the preprocessed data was clustered using VSEARCH v2.15.2 (7) with 97% identity and then subsampled to 10,000 reads per sample. The process

was repeated for each of the 100 splits resulting in 100 training and testing datasets for both workflows.

Machine Learning. A random forest model was trained with the R package mikrompl (v 1.2.0) (13) to predict the diagnosis (SRN or normal) for the samples in the test set for each data split. The training set was preprocessed to normalize OTU counts (scale and center), collapse correlated OTUs, and remove OTUs with zero variance. The preprocessing from the training set was then applied to the test set. Any OTUs in the test set that were not in the training set were removed. P-values comparing model performance were calculated as previously described (14). The averaged ROC curves were plotted by taking the average and standard deviation of the sensitivity at each specificity value.

Code Availability.

The analysis workflow was implemented in Snakemake (15). Scripts for analysis were written in R (16) and GNU bash (17). The software used includes mothur v1.47.0 (10), VSEARCH v2.15.2 (7), RStudio (18), the Tidyverse metapackage (19), R Markdown (20), the SRA toolkit (21), and conda (22). The complete workflow and supporting files required to reproduce this study are available at: https://github.com/SchlossLab/Armour_OptiFitGLNE_mBio_2023

Acknowledgments

This work was supported through a grant from the NIH (R01CA215574).

References

1. **Morgan XC**, **Tickle TL**, **Sokol H**, **Gevers D**, **Devaney KL**, **Ward DV**, **Reyes JA**, **Shah SA**, **LeLeiko N**, **Snapper SB**, **Bousvaros A**, **Korzenik J**, **Sands BE**, **Xavier RJ**, **Huttenhower C**. 2012. Dysfunction of the intestinal microbiome in inflammatory bowel disease and treatment. Genome Biol **13**:R79.

2. **Sobhani I**, **Tap J**, **Roudot-Thoraval F**, **Roperch JP**, **Letulle S**, **Langella P**, **Corthier G**, **Tran Van Nhieu J**, **Furet JP**. 2011. Microbial dysbiosis in colorectal cancer (CRC) patients. PLoS One **6**:e16393.

3. **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.

4. **Sovacool KL**, **Westcott SL**, **Mumphrey MB**, **Dotson GA**, **Schloss PD**. 2022. OptiFit: An improved method for fitting amplicon sequences to existing OTUs. mSphere **7**:e00916–21. doi:10.1128/msphere.00916-21.

5. **Baxter NT**, **Ruffin MT**, **Rogers MAM**, **Schloss PD**. 2016. Microbiota-based model improves the sensitivity of fecal immunochemical test for detecting colonic lesions. Genome Medicine **8**:37. doi:10.1186/s13073-016-0290-3.

6. **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.

7. **Rognes T**, **Flouri T**, **Nichols B**, **Quince C**, **Mahé F**. 2016. VSEARCH: a versatile open source tool for metagenomics. PeerJ **4**:e2584. doi:10.7717/peerj.2584.

8. **Bolyen E**, **Rideout JR**, **Dillon MR**, **Bokulich NA**, **Abnet CC**, **Al-Ghalith GA**, **Alexander H**, **Alm EJ**, **Arumugam M**, **Asnicar F**, **Bai Y**, **Bisanz JE**, **Bittinger K**, **Brejnrod A**, **Brislawn CJ**, **Brown CT**, **Callahan BJ**, **Caraballo-Rodríguez AM**, **Chase J**, **Cope EK**, **Da Silva R**, **Diener C**, **Dorrestein PC**, **Douglas GM**, **Durall DM**, **Duvallet C**, **Edwardson CF**, **Ernst M**, **Estaki M**, **Fouquier J**, **Gauglitz JM**, **Gibbons SM**, **Gibson DL**, **Gonzalez A**, **Gorlick K**, **Guo J**, **Hillmann B**, **Holmes S**, **Holste H**, **Huttenhower C**, **Huttley GA**, **Janssen S**, **Jarmusch AK**, **Jiang L**, **Kaehler BD**, **Kang KB**, **Keefe CR**, **Keim P**, **Kelley ST**, **Knights D**, **Koester I**, **Kosciolek T**, **Kreps J**, **Langille MGI**, **Lee J**, **Ley R**, **Liu Y-X**, **Loftfield E**, **Lozupone C**, **Maher M**, **Marotz C**, **Martin BD**, **McDonald D**, **McIver LJ**, **Melnik AV**, **Metcalf JL**, **Morgan SC**, **Morton JT**, **Naimey AT**, **Navas-Molina JA**, **Nothias LF**, **Orchanian SB**, **Pearson T**, **Peoples SL**, **Petras D**, **Preuss ML**, **Pruesse E**, **Rasmussen LB**, **Rivers A**, **Robeson MS**, **Rosenthal P**, **Segata N**, **Shaffer M**, **Shiffer A**, **Sinha R**, **Song SJ**, **Spear JR**, **Swafford AD**, **Thompson LR**, **Torres PJ**, **Trinh P**, **Tripathi A**, **Turnbaugh PJ**, **Ul-Hasan S**, **Hooft JJJ van der**, **Vargas F**, **Vázquez-Baeza Y**, **Vogtmann E**, **Hippel M von**, **Walters W**, **Wan Y**, **Wang M**, **Warren J**, **Weber KC**, **Williamson CHD**, **Willis AD**, **Xu ZZ**, **Zaneveld JR**, **Zhang Y**, **Zhu Q**, **Knight R**, **Caporaso JG**. 2019. Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. Nature Biotechnology **37**:852–857. doi:10.1038/s41587-019-0209-9.

9. **Edgar RC**, **Haas BJ**, **Clemente JC**, **Quince C**, **Knight R**. 2011. UCHIME improves sensitivity and speed of chimera detection. Bioinformatics **27**:2194–2200. doi:10.1093/bioinformatics/btr381.

10. **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**, **Van Horn DJ**, **Weber CF**. 2009. Introducing mothur: Open-source, platform-independent, community-supported software for describing and comparing microbial communities. Applied and Environmental Microbiology **75**:7537–7541. doi:10.1128/AEM.01541-09.

11. **Quast C**, **Pruesse E**, **Yilmaz P**, **Gerken J**, **Schweer T**, **Yarza P**, **Peplies J**, **Glöckner FO**. 2013. The SILVA ribosomal RNA gene database project: Improved data processing and web-based tools. Nucleic Acids Research **41**:D590–D596. doi:10.1093/nar/gks1219.

12. **DeSantis TZ**, **Hugenholtz P**, **Larsen N**, **Rojas M**, **Brodie EL**, **Keller K**, **Huber T**, **Dalevi D**, **Hu P**, **Andersen GL**. 2006. Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. Applied and Environmental Microbiology **72**:5069–5072. doi:10.1128/AEM.03006-05.

13. **Topçuoğlu BD**, **Lapp Z**, **Sovacool KL**, **Snitkin E**, **Wiens J**, **Schloss PD**. 2021. mikropml: User-Friendly R Package for Supervised Machine Learning Pipelines. Journal of Open Source Software **6**:3073. doi:10.21105/joss.03073.

14. **Topçuoğlu BD**, **Lesniak NA**, **Ruffin MT**, **Wiens J**, **Schloss PD**. 2020. A framework for effective application of machine learning to microbiome-based classification problems. mBio **11**:e00434–20. doi:10.1128/mBio.00434-20.

15. **Koster J**, **Rahmann S**. 2012. Snakemake–a scalable bioinformatics workflow engine. Bioinformatics **28**:2520–2522. doi:10.1093/bioinformatics/bts480.

16. **R Core Team**. 2020. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

17. **GNU Project**. Bash reference manual.

18. **RStudio Team**. 2019. RStudio: Integrated development environment for r. RStudio, Inc., Boston, MA.

19. **Wickham H**, **Averick M**, **Bryan J**, **Chang W**, **McGowan LD**, **François R**, **Grolemund G**, **Hayes A**, **Henry L**, **Hester J**, **Kuhn M**, **Pedersen TL**, **Miller E**, **Bache SM**, **Müller K**, **Ooms J**, **Robinson D**, **Seidel DP**, **Spinu V**, **Takahashi K**, **Vaughan D**, **Wilke C**, **Woo K**, **Yutani H**.

2019. Welcome to the Tidyverse. Journal of Open Source Software **4**:1686.

doi:10.21105/joss.01686.

20. **Xie Y**, **Allaire JJ**, **Grolemund G**. 2018. R Markdown: The Definitive Guide. Taylor &

Francis, CRC Press.

21. SRA-Tools - NCBI. http://ncbi.github.io/sra-tools/.

22. 2016. Anaconda Software Distribution. Anaconda Documentation. Anaconda Inc.

Figure Legends

Figure1: Overview of clustering workflows. The *de novo* and database-reference-based workflows were conducted using two approaches: OptiClust with mothur and VSEARCH as is used in the QIIME pipeline.

Figure 2: Model performance of OptiFit self-reference workflow is as good or better than other methods. A) Area under the receiver operating characteristic (AUROC) curve during cross-validation (train) for the various workflows. **B)** AUROC on the test data for the various workflows. The mean and standard deviation of the AUROC is represented by the black dot and whiskers in panels A and B. The mean AUROC is printed below the points. **C)** Averaged receiver operating characteristic (ROC) curves. Lines represent the average true positive rate for the range of false positive rates.