W4671 Final Project
Using DNABERT to Enhance OTU Clustering

George Austin, gia2105@columbia.edu

December 2020

Grouping all distinct reads from a microbial community into a manageable number of variables is a significant step in any mircobiome analysis. However, there is still no perfect approach. In this work, we analyzed the how using a DNABERT (1) deep learning model can improve operational taxonomic unit (OTU) clustering methods. While we found that the current version of the model did not outperform existing methods, our clustering approach gave similar results, suggesting it could potentially outperform existing methods if fine-tuned on related training tasks.

1 Introduction

Recent research has identified strong links between human gut microbiome communities and many serious conditions, including inflammatory diseases, depression and cancer (2), illustrating the importance of further work in this field. However, there are still limitations in the current data processing techniques. A key step in most microbiome research is clustering reads into operational taxonomic units based on sequence similarity. This delineation is necessary since there are generally too many distinct reads to consider them all separately. However, most common clustering methods have been shown to suffer from serious limitations (3). These problems often stem from the way read similarities are measured, which is often comparing the proportion of matching kmers across reads. This approach, while logical and computationally efficient, doesn't help determine which differences in kmers are more important, and as a result ties are often decided at random. This randomness makes most established methods' results difficult to reproduce and susceptible to small changes in datasets/hyperparameters.

One potential solution is to transfer information from deep learning models into an OTU clustering approach, potentially allowing for better interpretations of which reads should be clustered together through pre-trained vectorized representations. To do this, we used the BERT model structure (4), applying it to reads from microbiome datasets. We used the 6-mer DNABERT model (1), which was pretrained by predicting masked kmers from genomic sequences. By doing this, the model learned to create meaningful vectorized representations of reads, which we can leverage as a novel way to measure read similarity for hierarchical OTU clustering. Due to computational costs of the DNABERT model, this clustering method can only be applied to at most $\mathcal{O}(10000)$ distinct reads, so it must be pre-processed by an alternative method if there are too many reads.

We applied this clustering framework to three colorectal cancer datasets with lower-dimensional OTUs, analyzing how futher clustering via DNABERT impacted the signal in downstream analysis, through both random forest and lasso logistic regression modelling. We compared the DNABERT clusters to both clusters resulting from reads' BLAST phylogenetic assignments and sumaclust (5), a common clustering method which is shown to perform comparably to state of the art methods (6). We found that while the DNABERT clustering did not significantly outperform sumaclust, it did significantly better than phylogenetic clusters, implying there is meaningful information in the read's vectorized representations. Given how DNABERT's model was pretrained on a very different prediction task, it is possible we could see significant improvements from this clustering method when using a fine-tuned model. All code used in this analysis can be found in the project's github repository: `https://github.com/gaustin15/W4671_project`

2 Methods

2.1 The DNABERT Clustering Algorithm

The DNABERT clustering algorithm is an intuitive approach to applying the DNABERT model to microbiome reads. From running the reads through the DNABERT pipeline, pre-processing them using functions from their github repository, we created high-dimensional representations of each read. These representations were improved using the SBERT-WK algorithm (7), which incorporates the hidden states of a BERT model to create sentence embeddings (in this case reads are treated as 'sentences'). Once the high-dimensional embeddings are created, we employed a hierarchical clustering algorithm. We found that cosine similarity was the best performing distance metric, but metrics such as Euclidean and Manhattan distances were also considered. The pipeline is summarized in the following algorithm:

Algorithm 1: DNABERT OTU Clustering

```
Require
  DNABERT
  SBERT-WK
  READS
  C = Read count matrix
  N = Number of clusters to make
Embeddings = [ ]
i = 0
for r in READS:
  Embeddings[i] = SBERT-WK(DNABERT(r))
  i+=0
Clusters = hierarchical_cluster(Embeddings, N)
C = C.group_by(Clusters).sum()
Return C
```

2.2 Colorectal Cancer Analysis

To explore the value of the DNABERT clustering, we applied it to datasets from the MicrobiomeHD resource (8). We focused on colorectal cancer detection, since there is a documented need for both better CRC treatment and detection (9), while current non-invasive screening techniques aren't perfect (10). We applied the above DNABERT clustering to three MicrobiomeHD colorectal cancer datasets and compared the strength in downstream signals to both phylogenetic clusters and sumaclust clusters. We analyzed the signal strengths by running 5-fold cross-validation using random forest and lasso logistic regression and measuring the AUCs across the test data. Below are backgrounds on the studies the datasets we used are coming from. In all cases, there was other patient information available, but we only used the OTU data as predictors to isolate the predictive value of our clusters.

Dataset 1: *Structural Segregation of Gut Microbiota between Colorectal Cancer Patients and Healthy Volunteers* (11). The researchers collected microbiome samples from two groups. The first was 46 patients with colorectal cancer from Shanghai Cancer Hospital. The second was 56 healthy individuals having medical checkups, who never had any gastrointestinal tract disorders. All participants had been residents of Shanghai for at least 10 years. The researchers had clustered their reads into 837 distinct OTUs, so we focused on further reducing this set of clusters based on the given representative reads.

Dataset 2: *Human Intestinal Lumen and Mucosa-Associated Microbiota in Patients with Colorectal Cancer* (12). Collected stool samples from 21 CRC patients at First Affiliated Hospital, College of Medicine, Zhejiang University, China. 22 fecal samples were also collected form healthy volunteers at the same hospital. We started clustering from their given 1,617 OTUs.

Dataset 3: *Microbiota-Based Model Improves the Sensitivity of Fecal Immunochemical Test for Detecting Colonic Lesions* (13). Sampled data from 490 patients using 16S rRNA gene sequencing. 120 of the patients had CRC, 198 had adenomas, and 172 were healthy. We focused on using their microbiome data to distinguish between the healthy and the CRC groups. This dataset had too many distinct OTUs for either clustering method to run on a cpu, so only the OTUs whose abundance sum across all samples was greater than .001 were kept, leaving 13,617 OTUs. While this isn't an ideal reduction method for most analyses, it is appropriate here since we are exploring clustering methods' ability to retain the signals left in the data, and removing the least common OTUs will impact all clustering approaches in the same way.

3 Results

3.1 Evaluating Clusters

We did not find significantly better AUCs when comparing the DNABERT clustering algorithm to sumaclust. Figures 1-3 illustrate the AUCs from 5-fold cross-validation when applying random forest models to clustered CRC datasets. In general, we see both clustering methods improve the signals in our predictive models, although there is always an unsurprising drop-off once the number of clusters gets too low. The DNABERT and sumaclust clusters do outperform clusters resulting from the reads' assigned phylogenies. Dataset 3 shows relatively better results for DNABERT clustering when using more than 2000 clusters, although the overall difference is still not significant.

Figure 1: Comparing Clusters on Dataset 1 Using Random Forest AUCs

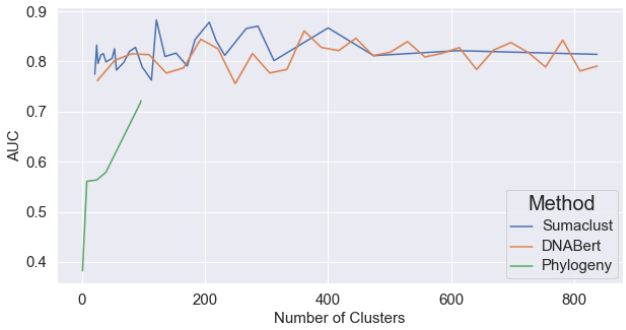


Figure 2: Comparing Clusters on Dataset 2 Using Random Forest AUCs

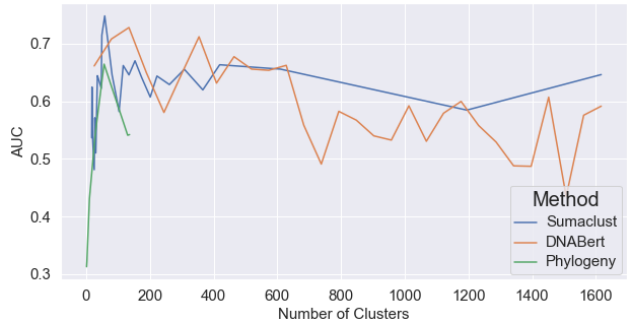
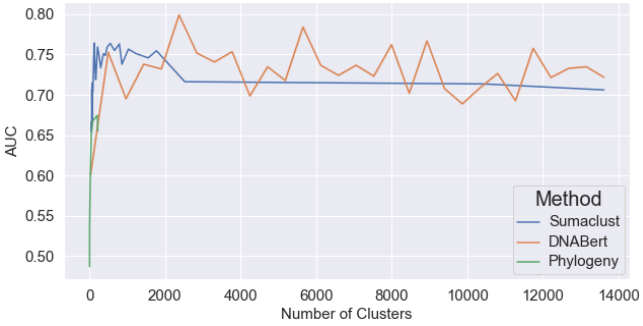


Figure 3: Comparing Clusters on Dataset 3 Using Random Forest AUCs



We sometimes see larger improvements from the clustering methods when using a logistic regression model (Figure 4). This isn't too surprising, since having fewer clusters with more elements makes it more likely that a linear signal identified in one cluster would generalize to a testing set. However, clustering to a very small number of OTUs and using a logistic regression model appears to be a rarely used analysis approach, most likely due to the consequential increased challenge in identifying a specific species associated with an increased risk for a disease/condition.

Figure 4: Comparing Clusters on Dataset 1 Using Lasso Logistic Regression AUCs

3.2 Analyzing Clusters

Table 1 is the summary of a significant DNABERT cluster in Dataset 1 (coefficient of -2.59), which had predictive value on the test set (corresponds to the logistic regression AUC for 50-cluster DNABERT grouping). We also observe a significant number of reads from each phylogeny that are excluded from the cluster. This shows there is meaningful value captured in our DNABERT model, often deviating from and outperforming BLAST phylogenies, although the DNABERT performance isn't any better than sumaclust. More importantly, we show that any clustering method can be used to observe a group of

relevant OTUs, which leads to better predictions on a test set.

Table 1: Summary of DNABERT Cluster Associated with Lower CRC Risk		
Phylogeny	#OTUs in cluster	#OTUs out of cluster
Bacteroidales Bacteroidaceae Bacteroides	72	54
Bacteroidales Prevotellaceae Prevotella	23	14
Bacteroidales Porphyromonadaceae Parabacteroides	19	12
Bacteroidales Prevotellaceae Paraprevotella	4	1
Bacteroidales Porphyromonadaceae Porphyromonas	3	4
Bacteroidales Porphyromonadaceae Odoribacter	2	2

3.3 Computational Complexity

One of the main downsides of the DNABERT clustering model is the relatively high computational costs. Using a cpu, running the embedding on a read takes on average .23 seconds (38 minutes for 10,000 reads). Furthermore, the hierarchical clustering method has a higher computational complexity than other approaches, since the DNABERT model is embedding reads into 784 dimensional vectors, making the similarity metrics much more expensive than something like sumaclust. The $\mathcal{O}(n)$ embedding time and significant $\mathcal{O}(n^2)$ hierarchical approach makes it not feasible to run DNABERT clustering onto a raw dataset with every collected read. Therefore, to use DNABERT clusters, one should first use an alternate method to group reads down to $\mathcal{O}(10,000)$ OTUs, and then run DNABERT on representative reads from those OTUs. While the run time in this situation will still be longer than most methods, it wouldn't be too significant.

4 Discussion

While the DNABERT clustering algorithm didn't outperform sumaclust, there are possible improvements to be made in this approach. The first issue is that the DNABERT model was trained on predicting masked elements within genomic sequences, which doesn't closely relate to this particular prediction task. We would expect better results from our clustering method if we were using a DNABERT model that was fine-tuned on microbiome datasets. The fact that we obtained results that are comparable to sumaclust with just the pre-trained model suggests there is potential to outperform standard OTU methods with a fine-tuned DNABERT.

In theory, a deep learning model could capture which discrepancies in kmers are more/less significant within a microbial community, which gives it greater potential for high-quality OTU clustering than most delineating methods. While in most practical situations, it would take too long to apply DNABERT clustering to a full microbiome dataset, it is possible than an optimal approach would use a fine-tuned DNABERT as a way to further cluster outputs from alternative methods, such as sumaclust. Therefore, there is still potential for significant improvements in OTU clustering through DNABERT embeddings.

We are in general seeing improvement in performance when clustering by either method. The lower dimensions (i.e. 200-400) seem to rarely be used in related analyses (11) (12) (13), although based on our results, clustering down to these lower dimensions make it easier for machine-learning models to capture significant signals in the datasets. This implies that OTU clustering is often underused in the context of predictive modelling. Of course, in other inferential analysis, there is motivation to isolate specific microbial markers, potentially making it less pratical to use larger clusters containing many phylogenices, such as Table 1.

Overall, the pre-trained DNABERT model's hierarchical clustering method did not outperform clustering algorithms. However, the DNABERT clustering could be improved through fine-tuning a model, so it is still possible that OTU methods could be enhanced with our proposed algorithm.

5 Data Availability

All of the MicrobiomeHD datasets (8) analyzed in this paper are public and can be accessed at `https://zenodo.org/record/569601#.X95hxi3Mzx4`. Dataset 1 was collected and preprocessed by Wang et al (11). The second was collected and preprocessed by Xiang et al (12), and the third by Baxter et al (13).

6 Code Availability

Our code is available at `https://github.com/gaustin15/W4671_project`. The repository uses publicly available code and pre-trained models that can be found at `https://github.com/jerryji1993/DNABERT` and `https://github.com/BinWang28/SBERT-WK-Sentence-Embedding`. Our project provides instructions for environment setups and specific items to download, as well as bash commands to download the datasets used in our analysis.

References

[1] Ji, Yanrong, Zhihan Zhou, Han Liu, and Ramana V Davuluri. 2020. "DNABERT: Pre-Trained Bidirectional Encoder Representations from Transformers Model for DNA-Language in Genome." Cold Spring Harbor Laboratory. https://doi.org/10.1101/2020.09.17.301879.

[2] Rui-xue Ding, Wei-Rui Goh, Ri-na Wu, Xi-qing Yue, Xue Luo, Wei Wei Thwe Khine, Jun-rui Wu, Yuan-Kun Lee, Revisit gut microbiota and its impact on human health and disease, Journal of Food and Drug Analysis, Volume 27, Issue 3, 2019, Pages 623-631, ISSN 1021-9498, https://doi.org/10.1016/j.jfda.2018.12.012.

[3] Schmidt TS, Matias Rodrigues JF, von Mering C. Limits to robustness and reproducibility in the demarcation of operational taxonomic units. Environ Microbiol. 2015 May;17(5):1689-706. doi: 10.1111/1462-2920.12610. Epub 2014 Sep 29. PMID: 25156547.

[4] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser and Illia Polosukhin. Attention Is All You Need, 2017; arXiv:1706.03762.

[5] Mercier, Celine, Sumaclust, https://git.metabarcoding.org/obitools/sumatra/wikis/home/

[6] Kopylova, Evguenia, Jose A. Navas-Molina, Céline Mercier, Zhenjiang Zech Xu, Frédéric Mahé, Yan He, Hong-Wei Zhou, Torbjørn Rognes, J. Gregory Caporaso, and Rob Knight. 2016. "Open-Source Sequence Clustering Methods Improve the State Of the Art." Edited by Nicola Segata. MSystems 1 (1). https://doi.org/10.1128/msystems.00003-15.

[7] Bin Wang and C. -C. Jay Kuo. SBERT-WK: A Sentence Embedding Method by Dissecting BERT-based Word Models, 2020; arXiv:2002.06652.

[8] Duvallet, Claire, Sean Gibbons, Thomas Gurry, Rafael Irizarry, and Eric Alm. 2017. "Microbiomehd: The Human Gut Microbiome In Health And Disease." Zenodo. https://doi.org/10.5281/ZENODO.797943.

[9] Cronin, Kathleen A., Andrew J. Lake, Susan Scott, Recinda L. Sherman, Anne-Michelle Noone, Nadia Howlader, S. Jane Henley, et al. 2018. "Annual Report to the Nation on the Status of Cancer, Part I: National Cancer Statistics." Cancer 124 (13): 2785–2800. https://doi.org/10.1002/cncr.31551.

[10] Thomas F. Imperiale, M.D., David F. Ransohoff, M.D., Steven H. Itzkowitz, M.D., Theodore R. Levin, M.D., Philip Lavin, Ph.D., Graham P. Lidgard, Ph.D., David A. Ahlquist, M.D., and Barry M. Berger, M.D., 2014; Multitarget Stool DNA Testing for Colorectal-Cancer Screening, https://www.nejm.org/doi/full/10.1056/nejmoa1311194

[11] Wang, T., Cai, G., Qiu, Y. et al. Structural segregation of gut microbiota between colorectal cancer patients and healthy volunteers. ISME J 6, 320–329 (2012). https://doi.org/10.1038/ismej.2011.109

[12] Chen W, Liu F, Ling Z, Tong X, Xiang C (2012) Human Intestinal Lumen and Mucosa-Associated Microbiota in Patients with Colorectal Cancer. PLoS ONE 7(6): e39743. https://doi.org/10.1371/journal.pone.0039743

[13] Baxter, Nielson T., Mack T. Ruffin IV, Mary A. M. Rogers, and Patrick D. Schloss. 2016. "Microbiota-Based Model Improves the Sensitivity of Fecal Immunochemical Test for Detecting Colonic Lesions." Genome Medicine 8 (1). https://doi.org/10.1186/s13073-016-0290-3.