Using Topological Data Analysis to find discrimination between microbial states in human microbiome data

Mehrdad Yazdani*[1,2], Larry Smarr[1,3] and Rob Knight[4]

[1] *California Institute for Telecommunications and Information Technology, University of California San Diego, USA*
[2] *Open Medicine Institute, Mountain View, California, USA*
[3] *Harry E. Gruber Professor, Department of Computer Science and Engineering, University of California San Diego, USA*
[4] *Department of Pediatrics, University of California San Diego, USA*

The vast collection of microbial cells, referred to as the human microbiome, forms an ecology of diverse microbial organisms that lives with us in symbiosis. Since the human microbiome ecology differs dramatically in different body sites and individuals, understanding how and what changes in the ecology are of crucial importance. In this study we investigate Topological Data Analysis (TDA) as an unsupervised learning and data exploration tool to identify changes in microbial states. We compare TDA with other well-established methods, such as Principle Component Analysis (PCA) and Principle Coordinate Analysis (also known as Multidimensional Scaling or MDS), using a previously published dataset of high-resolution time series of the microbiome from 3 different sites (mouth, hands, and gut) from 2 healthy (one female, one male) subjects. Since previous studies have shown that microbial communities of healthy subjects are highly stable over time (unless disturbed by an external variable), we expect to identify 6 total microbial communities corresponding to the different body site and subject combinations in our dataset. We show that PCA and MDS reveal 3 distinct clusters that correspond to the three different body sites. However, these methods do not discriminate samples based on the subjects. We find here that TDA identifies distinct groups that discriminate between the female and male gut samples and also separate between the skin and tongue body sites as well. This suggests that TDA is able to identify groups of clusters that other methods may potentially miss.

1 Introduction and background

In this study we investigate several unsupervised learning methods to identify clusters or subsets of samples in longitudinal data that correspond to different microbial states. Since the lack of ground truth on what constitutes different microbial states makes evaluating unsupervised learning methods challenging, we apply our methods on a dataset [3] (available publicly through [2]) that is a high-resolution time series of the microbiome from 2 healthy subjects sampled from 3 different body sites. Therefore, our data has at least 6 known states corresponding to the different body site and subject combinations. The female (F4) and male (M3) subjects were sampled daily for 15 and 6 months respectively at the gut (feces), the mouth (saliva), and the skin (left and right palms) body sites. We have a total of 1,433 samples for M3 and 534 samples for F4. The details of the sequencing and the data can be found in [3].

*myazdani@gmail.com

Using this dataset, we compare Topological Data Analysis (TDA) [4, 9] with several well-established unsupervised methods that are commonly used in ecology and microbiome studies such as Principle Component Analysis (PCA) and Multidimensional Scaling (MDS) [10, 11] (for example, see Figure 4 in [6, 11]). TDA can be used as a general framework for unsupervised learning to embed high dimensional datasets into graph-based representation. In PCA and MDS, however, for visualizing high dimensional data we are limited to projecting data to a small number of dimensions (typically 2-3). TDA on other hand can use an arbitrary number of projections to embed in the graph and uses the method of persistent homology so that the analysis is robust against noise and scale of data. In this work we apply TDA to embed the microbial samples of this dataset into a graph that distinguishes between the two subjects and their respective three body sites.

2 Methods

To establish baseline comparisons, we compute PCA on our dataset using the relative abundances of the taxonomic families of microbes for each sample [5, 10, 11]. As discussed in [5] (amongst others), using a non-Euclidean metric for measuring biological and ecological diversity is preferred over PCA. We therefore also compute MDS [10, 11] using the Bray-Curtis [5] distance that's commonly used in Ecology for comparison. Additionally, we also also compute MDS using the unweighted UniFrac distance [8] as another more biologically meaningful measure of biological diversity that uses domain knowledge of the phylogenetic tree of the microbes.

We compare these methods with TDA as implemented by the Ayasdi webapp software [1]. TDA is a general framework for embedding high dimensional datasets in a graph (for a general overview of TDA see [9]). To construct the graph, TDA requires a metric and a set of functions that map data points to a real valued vector space. In this study, the data that we use are the relative abundances of the taxonomic families (one can use other levels as well) of microbes for each sample, and we use the unweighted UniFrac metric. The set of functions that we use here for mapping the data is based on the Stochastic Neighbor Embedding [7] algorithm referred to as the Neighborhood Lens by Ayasdi (in two dimensions). The resolution we use for the Neighborhood Lens for both dimensions is 200 points with a gain of 80% overlap. Note, however the our results (and TDA in general) is not sensitive to the specific parameters. Single linkage clustering is performed and the heuristics for the optimization are described in [12].

3 Results

Figure 1 shows three commonly used unsupervised methods (PCA, MDS using Bray-Curtis distance, and MDS using the unweighted UniFrac distance) applied to the samples in our data. As was also shown in [3], we see that over the course of the samples there is a clear separation between the three different body sites. However, these unsupervised methods do not discriminate clearly between the microbial states corresponding to the two subjects

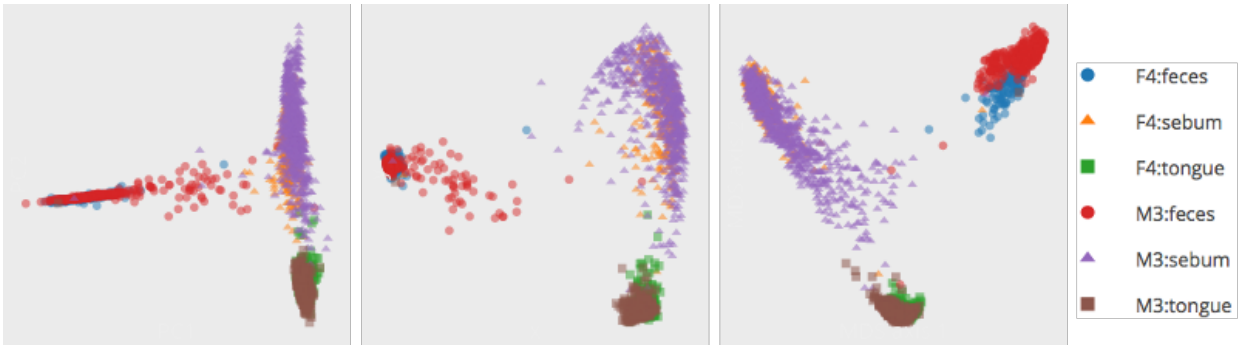
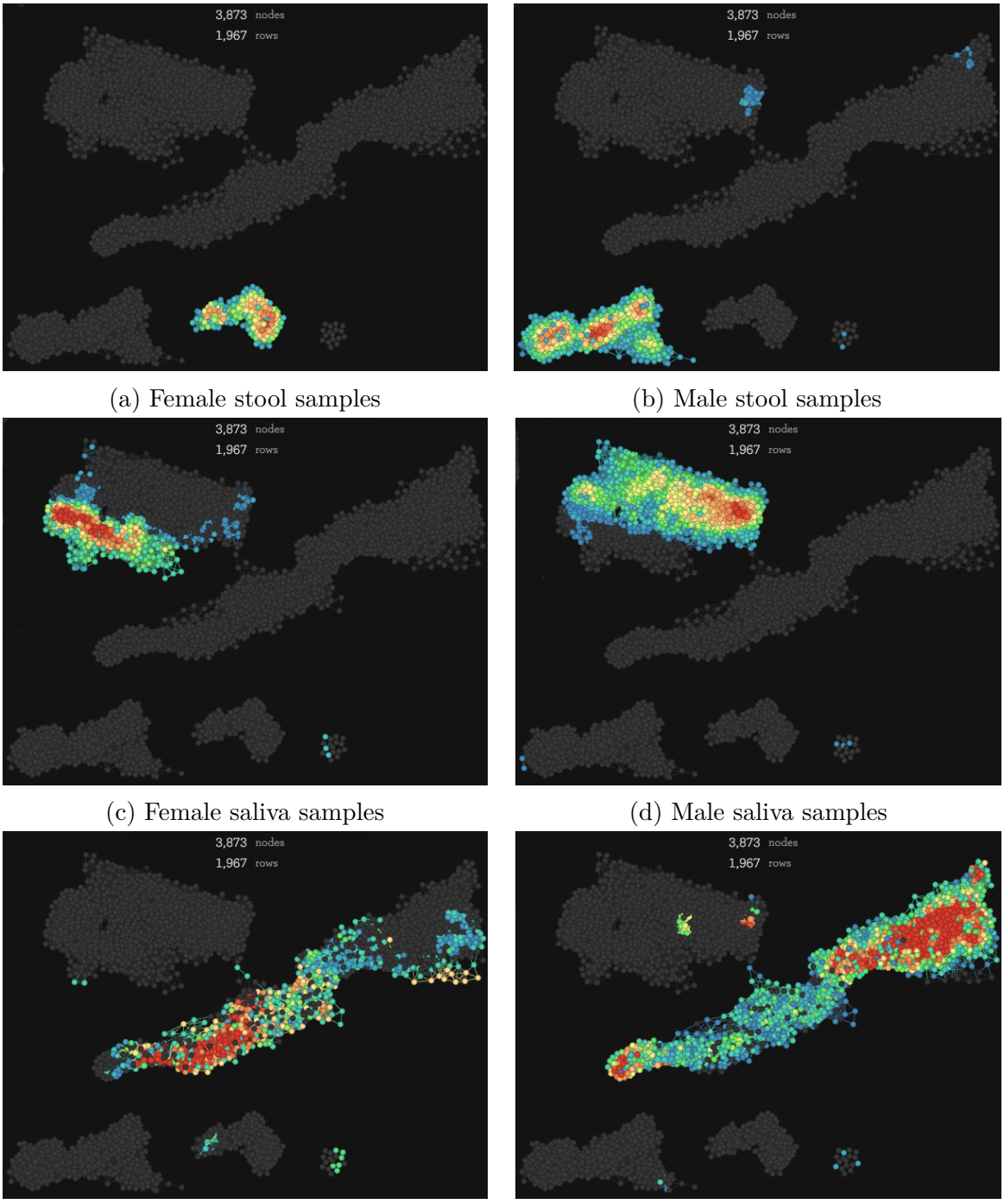


Figure 1: Left panel shows the PCA using of the family relative abundances of our data set. Middle panel shows the MDS of using the Bray-Curtis distance and the right panel is MDS using the UniFrac distance.From this analysis it appears that the samples from the male and female are not significantly different within body sites

in our data. Figure 2 on the other hand shows applying TDA using Ayasdi [1] with the Neighborhood Lens [7] and unweighted UniFrac [8] metric applied. For each panel, we color the nodes in the graph by the proportion of samples that belong to a subject and body site (red corresponds to higher proportion of samples). Note that in TDA each node in the graph corresponds to several samples and edges drawn based on overlap of samples between nodes. We see that as we shift from every subject and body site, the central density of the nodes in the graph changes accordingly. In Figures 2 (a) and 2 (b), corresponding to the female and male stool samples respectively, we see the most dramatic shift as the majority of nodes in the graph consist of two separate connected components. These two separate connected components fit as natural candidates of being distinct groups (which, given that we know the labels in the data, we know in fact correspond to two different subjects). Figures 2 (c) and 2 (d), corresponding to the saliva samples, also show a separation between the female and male subjects as the shift in color in the heatmap indicates. Finally, in Figures 2 (e) and 2 (f) corresponding to skin samples we observe discrimination, but the separation is not as strong as the other body sites.

4 Discussion and Future Work

In this paper we applied TDA to a previously published high-resolution time series of human microbiome data for discriminating between microbial states. We find that the graph embedding that TDA establishes discriminates between the 6 states that correspond to 2 subjects and the 3 different body sites in the data, while MDS only distinguishes between the body sites. We used this previously published dataset to establish benchmarks for identifying subsets of samples that are distinct (that is, finding clusters of samples that discriminate between the two subjects and three different body sites in the data). The flexibility of TDA as a general unsupervised framework provides an ideal tool for Exploratory Data Analysis of unlabelled data. Having established this method on well-studied dataset for healthy subjects, for future work we will apply TDA on subjects with disease conditions.



(a) Female stool samples

(b) Male stool samples

(c) Female saliva samples

(d) Male saliva samples

(e) Female skin samples

(f) Male skin samples

Figure 2: Unsupervised Topological Data Analysis (TDA, computed using Ayasdi [1]): the color of nodes indicates the proportion of data samples corresponding to the specific subject and body site (red means higher). In (a) and (b) there are two connected components corresponding to female and male stool samples. In (c) and (d) there is little overlap in the samples corresponding to the female and male saliva samples. In (e) and (f), the significant proportions of samples between female and male samples are in shifted in the graph.

5 Acknolwedgements

We thank Ayasdi's Devi Ramanan and Jenya Koylova and Zhenjiang Xu from the Knight Lab for helpful suggestions for this paper. We also acknowledge support from Calit2 and it's Qualcomm Institute, the San Diego Supercomputer Center, and the Amar Foundation. Part of this research was performed while Mehrdad Yazdani was visiting the Institute for Pure and Applied Mathematics (IPAM), which is supported by the National Science Foundation.

References

[1] Ayasdi Inc., menlo park, california. http://ayasdi.com. Accessed: 2016-05-01.

[2] Qiita. http://qiita.microbio.me. Accessed: 2016-03-21.

[3] J Gregory Caporaso, Christian L Lauber, Elizabeth K Costello, Donna Berg-Lyons, Antonio Gonzalez, Jesse Stombaugh, Dan Knights, Pawel Gajer, Jacques Ravel, Noah Fierer, et al. Moving pictures of the human microbiome. *Genome Biol*, 12(5):R50, 2011.

[4] Gunnar Carlsson. Topology and data. *Bulletin of the American Mathematical Society*, 46(2):255–308, 2009.

[5] Michael Greenacre and Raul Primicerio. *Multivariate analysis of ecological data*. Fundacion BBVA, 2014.

[6] Micah Hamady and Rob Knight. Microbial community profiling for human microbiome projects: Tools, techniques, and challenges. *Genome research*, 19(7):1141–1152, 2009.

[7] Geoffrey E Hinton and Sam T Roweis. Stochastic neighbor embedding. In *Advances in neural information processing systems*, pages 833–840, 2002.

[8] Catherine Lozupone and Rob Knight. Unifrac: a new phylogenetic method for comparing microbial communities. *Applied and environmental microbiology*, 71(12):8228–8235, 2005.

[9] PY Lum, G Singh, A Lehman, T Ishkanov, Mikael Vejdemo-Johansson, M Alagappan, J Carlsson, and G Carlsson. Extracting insights from the shape of complex data using topology. *Scientific reports*, 3, 2013.

[10] Kantilal Varichand Mardia, John T Kent, and John M Bibby. Multivariate analysis. 1980.

[11] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2015.

[12] Gurjeet Singh, Facundo Mémoli, and Gunnar E Carlsson. Topological methods for the analysis of high dimensional data sets and 3d object recognition. In *SPBG*, pages 91–100, 2007.