Finding phylogeny-aware and biologically meaningful averages of metagenomic samples: L_2UniFrac

Wei Wei[1], Andrew Millward[2], David Koslicki[1,2,3]

[1]Huck Institutes of Life Sciences, Pennsylvania State University
[2]Department of Computer Science and Engineering, Pennsylvania State University
[3]Department of Biology, Pennsylvania State University

January 19, 2023

Abstract

Metagenomic samples have high spatiotemporal variability. Hence, it is useful to summarize and characterize the microbial makeup of a given environment in a way that is biologically reasonable and interpretable. The UniFrac metric has been a robust and widely-used metric for measuring the variability between metagenomic samples. We propose that the characterization of metagenomic environments can be achieved by finding the average, a.k.a. the barycenter, among the samples with respect to the UniFrac distance. However, it is possible that such a UniFrac-average includes negative entries, making it no longer a valid representation of a metagenomic community. To overcome this intrinsic issue, we propose a special version of the UniFrac metric, termed L_2UniFrac, which inherits the phylogenetic nature of the traditional UniFrac and with respect to which one can easily compute the average, producing biologically meaningful environment-specific "representative samples". We demonstrate the usefulness of such representative samples as well as the extended usage of L_2UniFrac in efficient clustering of metagenomic samples, and provide mathematical characterizations and proofs to the desired properties of L_2UniFrac. A prototype implementation is provided at: `https://github.com/KoslickiLab/L2-UniFrac.git`.

1 Introduction

1.1 Background and context

Microbes play a vital part in many aspects of our lives, and many diseases and conditions are related to the microbiota in and on our body. Examples of which can include various kinds of cancers [1, 18], as well as in behavioral conditions such as sociability [4] or Autism Spectrum Disorder [5]. Despite holding the key to many of the known problems, metagenomic studies are faced with multiple challenges. One of these difficulties lies in the high variability in metagenomic samples [13, 15]. Since samples for a given environment collected at different sub-environments or at different time points can be vastly different in microbial distribution, it can be difficult to characterize a specific environment by an average or representative distribution.

In this paper, we will present an approach to create such average or representative distributions of a collection of metagenomes by defining a new modification of the classic UniFrac metric [8, 10] which we call the L_2UniFrac. We will look at applications of these representative samples, as

well as further providing mathematical characterizations of the L_2UniFrac metric by giving formal mathematical proofs of its properties.

The code used to generate all data and figures in this manuscript can be found in `https://github.com/KoslickiLab/L2-UniFrac-Paper.git` and a prototype implementation for L_2UniFrac can be found in `https://github.com/KoslickiLab/L2-UniFrac.git`.

1.2 Rationale of approach

In a typical environment-related metagenomics study, clustering is often performed after sample collection and data analysis, the result of which are visualized and analyzed by methods such as Principal Coordinate Analysis (PCoA). It is therefore natural that such clusters can be used as a basis or guide in describing an environment, or in finding a representative distribution that characterizes the environment. Intuitively, it is natural that given a large sample size, the centroid (also called the average or the barycenter) of a clustering of these samples can be used as a representative of the samples in this environment. With such an average, it will be much easier to distinguish the signature microbiomes unique to the environment from random noise that come from microbes rarely in the samples. It will also make the comparison between two environments much easier, simplifying it to the comparison between two sample averages.

Finding the barycenter of a group of distribution is not as trivial as it seems. For the purpose of scalability and practicality, the method of computing such a barycenter should be relatively easy and the result easily interpretable biologically. The key to both of these properties lies in the choice of distance metric. A clustering process like this requires a distance metric such that the dissimilarity between two microbial distributions, a property also known as the beta-diversity, can be measured. There are multiple such beta-diversity metrics. In this paper, we will focus on the discussion of finding the barycenter with respect to the UniFrac metric.

There are two main reasons why the UniFrac metric was chosen among the others. The first is due to its robustness. Being one of the most widely-used phylogenetic metrics, the UniFrac metric has demonstrated its usefulness in many areas ranging across clinical studies [7], environmental studies [14], and forensic science [16]. Compared to non-phylogenetic metrics such as the Jaccard index and the Bray-Curtis dissimilarity index, the UniFrac metric takes into account the phylogenetic relationships among the organisms, giving it more biological context and hence more interpretability. In real applications, the UniFrac metric has also demonstrated its superiority over other methods, such as in sensitivity in the identification of enterotypes, which are subtypes of the same environment [7], making it an ideal candidate. The second reason is due to the unique and interesting mathematical characterization of UniFrac, which makes this biologically-motivated problem mathematically interesting at the same time, and leaves rooms for mathematical generalization that could potentially be useful for other related problems yet to be explored.

We begin by first understanding this mathematical characterization and how it is related to our goal of finding the barycenter with respect to the UniFrac distance. About seven years after the development of the original UniFrac in 2005 [8], it was shown by Evans et al. that the UniFrac metric is closely related to a mathematical problem that had been studied for centuries: the optimal transport problem [2]. The optimal transport problem uses a distance known as the Wasserstein distance, or the Earth Mover's distance. Evans showed that the UniFrac distance is equivalent to the 1-Wasserstein distance over a phylogenetic tree. Under this formulation, instead of the original computation of UniFrac as a fraction of branch lengths, it can be equivalently computed by representing metagenomic samples by probability distributions and "aggregating" the distributions up a phylogenetic tree through matrix multiplication, obtaining the aggregated vectors in what we will

call the "L_1UniFrac space", followed by taking the L_1 norm between two such aggregated vectors. This alternative characterization of UniFrac, termed the "L_1UniFrac" by Koslicki and McClelland [9], has several benefits. First, it allows an alternative representation of a metagenomic sample as an aggregated vector in the L_1UniFrac space, in which the distance between two vectors can be easily computed by taking the L_1-norm. Also, this representation of data points in the L_1UniFrac space leaves room for computation of the L_1-mean, or median, among a group of samples, giving rise to an easy-to-compute centroid, which can be viewed as the average of the samples in the L_1UniFrac space. This idea of computing the "L_1-mean UniFrac" was first mathematically conceptualized by Koslicki and McClelland [9], who saw it as an opportunity to find the representative sample of an environment with respect to the UniFrac metric.

However, the "L_1-mean sample" computed using this method does have a detrimental issue. We want the mean/average to have a readily interpretable biological meaning of representing an average sample in the environment, and not simply as an aggregated vector in the L_1UniFrac space. The former requires it being a distribution vector with each entry being non-negative and all the entries sum up to 1, such that each entry represents the relative abundance of an organism in the sample. This desired property cannot be guaranteed when the mean vector in the L_1UniFrac space is projected back to the distribution space. Both experiments using real world data and mathematical deductions have shown that the resulting projected vector could contain negative entries [12]. This makes the result biologically meaningless. Our proposed solution is an alternative characterization of UniFrac using the L_2 norm instead of L_1. Without changing the nature of being a solution to an optimal transport problem, the so-termed L_2UniFrac retains the original spirit of UniFrac of being a phylogeny-aware dissimilarity metric and offers comparable robustness, while circumventing the issue of having negative values in L_1UniFrac mean.

2 Methods

Herein, we describe the L_2UniFrac metric and how to use it to take averages. Full details and proofs are provided in the Supplementary material.

2.1 L_1 and L_2 Unifrac

Given a phylogenetic tree T with N ordered nodes, a metagenomic sample can be represented as a probability distribution $P = (P_1, P_2, ..., P_N)$ with entries in the same order as the nodes of T, such that P_i represents the relative abundance of organism/taxa i in sample P.

Define

$$w_j(i) = \begin{cases} 1 & \text{if } i \text{ is a node on the subtree of } T \text{ rooted at node } j \\ 0 & \text{otherwise.} \end{cases} \tag{1}$$

Let W be an $N \times N$ matrix with row j being w_j scaled by length of the branch connecting node j and its ancestor. The matrix $W_{\sqrt{\cdot}}$ is defined similarly, but with the rows scaled by the square root of the branch length.

We can now recall the definition of the UniFrac metric (see, for example, and of [9, 11, 2]):

Definition 1 (UniFrac)**.** For two metagenomic samples represented by the probability distributions P and Q, and with W as defined above for some fixed phylogenetic tree, and for $||\cdot||_1$ the standard L_1 norm,

$$\text{UniFrac}(P, Q) = ||W(P - Q)||_1. \tag{2}$$

By simply changing the norm from the L_1 to L_2 norm, we obtain:

Definition 2 (L_2UniFrac)**.** For two metagenomic samples represented by the probability distributions P and Q, and with $W_{\sqrt{\cdot}}$ as defined above for some fixed phylogenetic tree, and for $||\cdot||_2$ the standard L_2 norm,

$$L_2\text{UniFrac}(P,Q) = ||W_{\sqrt{\cdot}}(P-Q)||_2. \tag{3}$$

In the Supplementary material, we demonstrate that one can take meaningful averages with respect to the L_2UniFrac metric, but not with the traditional (L_1) UniFrac.

In practice though, we do not need to form the large matrix $W_{\sqrt{\cdot}}$, but rather proceed in a post-order aggregation to implement the matrix multiplications $W_{\sqrt{\cdot}}P$ and $W_{\sqrt{\cdot}}Q$ in equation (3) in Definition 2: see Algorithm 1. After applying Algorithm 1 to compute $W_{\sqrt{\cdot}}P$ and $W_{\sqrt{\cdot}}Q$, the

Algorithm 1 L_2-aggregate: an algorithm to obtain the L_2-aggregated vector given a probability vector P, resulting in a vector in L_2UniFrac space.

Input:
P, T, where P is a probability vector with entries representing relative abundances summing up to 1, T being the phylogenetic tree with the ancestor of a node i denoted by $a(i)$ and the branch length between i and $a(i)$ denoted by $l(i)$.
Initialization: $\overline{P} = P$
for i in $1, \ldots, |T|-1$ **do** ▷ Ordered from the leaves to the root (i.e. post-order)
    $\overline{P}[a(i)] += \overline{P}[i]$ ▷ aggregating mass up
    $\overline{P}[i] = \overline{P}[i] \cdot \sqrt{l(i)}$
end for
return $\overline{P}$

L_2UniFrac can then be computed by taking the simple L_2-norm of the difference of the resulting vectors.

Later, we will need to also compute the inverse of this operation (i.e. compute $W^{-1}_{\sqrt{\cdot}}P$ and $W^{-1}_{\sqrt{\cdot}}Q$), the algorithm for which is similar to the above; see Algorithm 2.

Algorithm 2 Inverse-aggregate: an algorithm that reverse L_2-aggregate to obtain a probability vector in the original space, given an aggregated vector in the L_2UniFrac space.

Input:
$\overline{P}$, T, where $\overline{P}$ is a probability vector in the L_2UniFrac space, T being the taxonomic tree with the ancestor of a node i denoted by $a(i)$ and the branch length between i and $a(i)$ denoted by $l(i)$.
Initialization: $P = \overline{P}$
for i in $1, \ldots, |T|-1$ **do** ▷ Ordered from the leaves to the root (i.e. post-order)
    $v = \overline{P}[i]$
    $P[a(i)] -= \frac{1}{\sqrt{l(i)}} * v$
end for
return P

2.2 Computing L_2UniFrac averages

In the Supplementary material, we demonstrate that one cannot form biologically meaningful (L_1) UniFrac averages of a collection of metagenomic samples represented by a collection of probability vectors. In short, this is due to the probability simplex not being closed under the operation of taking medians. However, the probability simplex is closed under means (See Supplementary material section S1.4). As such, we describe taking averages with respect to the L_2UniFrac metric only.

The definition of an average (or more precisely, a barycenter) with respect to the L_2UniFrac metric is as follows:

Definition 3 (L_2UniFrac barycenter)**.** Given a collection of probability distributions $\{P^i\}_{i=1}^M$ representing metagenomic samples, each of which is given by an N-length vector indexed by the nodes in a fixed phylogenetic tree T, the average, or barycenter, of the set $\{P^i\}_{i=1}^M$ is given by P^*:

$$P^* = \arg\min_{\mathbf{x}} \sum_{i=1}^{M} L_2\text{UniFrac}(P^i, \mathbf{x}). \tag{4}$$

Thus, in practice, to compute a "representative sample" of a collection of vectors $\{P^i\}_{i=1}^M$, we proceed as follows: First, form $\{W_{\sqrt{\cdot}}P^i\}_{i=1}^M$ by repeat application of Algorithm 1. Then, form the L_2UniFrac barycenter P^* by taking the component-wise mean of the collection $\{W_{\sqrt{\cdot}}P^i\}_{i=1}^M$. Finally, use Algorithm 2 to compute $W_{\sqrt{\cdot}}^{-1}P^*$, the result of which will be a probability distribution representing the L_2UniFrac average of the collection $\{P^i\}_{i=1}^M$. Most interestingly, we show in the Supplementary material, Claim 4 that the result of this process is as simple as taking the component-wise mean of the collection of vectors $\{P^i\}_{i=1}^M$.

For two collections of vectors (representing, say, two different environments) $\{P^i\}_{i=1}^{M_1}$ and $\{Q^i\}_{i=1}^{M_2}$, the "representative sample" vectors $W_{\sqrt{\cdot}}^{-1}P^*$ and $W_{\sqrt{\cdot}}^{-1}Q^*$ can be used as a proxy of their respective collections. Thus, instead of needing to form all pairwise L_2UniFrac distance calculations as:

$$L_2\text{UniFrac}(P_i, Q_j) \quad \text{for } i = 1 \dots M_1, j = 1 \dots M_2,$$

and averaging over the environments, we can much more efficiently do the single computation:

$$L_2\text{UniFrac}(W_{\sqrt{\cdot}}^{-1}P^*, W_{\sqrt{\cdot}}^{-1}Q^*).$$

This amounts to applying Algorithm 1 to the component-wise averages of $\{P^i\}_{i=1}^{M_1}$ and $\{Q^i\}_{i=1}^{M_2}$, then taking the L_2-norm of the difference between these two vectors.

In the following, we will refer to "L_pUniFrac." By this we mean the following: for a sample distribution P (living in the probability simplex), left-multiplying by the aforementioned W (modified depending on the L_p space under consideration), we obtain a vector WP living in "L_pUniFrac space."

3 Results

3.1 Clustering in L_2UniFrac space

We hypothesize that phylogeny-aware clustering of samples, which is a common procedure in metagenomic analyses, can be done much faster in the L_pUniFrac space than in the distribution

space. The reason is as follows: The traditional way of UniFrac-based clustering (in the distribution space) requires the pairwise UniFrac distances be computed prior to applying clustering methods. The computation of UniFrac is not trivial. Pairwise comparison aggravates the cost of computation by an exponential factor as sample size increases. On the other hand, when samples are represented as aggregated vectors in the L_pUniFrac space, the UniFrac distance can be readily computed by simply computing the L_p-norm. The bulk of the computation lies only in transforming P to WP through matrix multiplication, which is linear with respect to sample size.

We tested this hypothesis using L_2UniFrac as a representative of the L_pUniFrac general case. To demonstrate the improvement on clustering speed, we randomly selected 1,000 samples out of the 6,067 samples of 16S data obtained from `Qiita` [3] (study ID 1928), consisting of samples collected from four body sites: skin, saliva, vagina, and feces. Out of these samples, we randomly sampled with sample size ranging from 50 to 800 in step of 50. Each of these samples were clustered using two methods: the conventional matrix-based method, in which a pairwise UniFrac distance matrix is computed using the EMDUniFrac [10] algorithm, followed by k-medoids clustering, as well as our proposed method, in which samples were first aggregated to the L_2UniFrac space, followed by k-means clustering on aggregated vectors. We further computed the Fowlkes-Mallows score for each of the instances as a measure of the clustering quality. The comparisons of both the time cost and the clustering quality are shown in Figure 1.

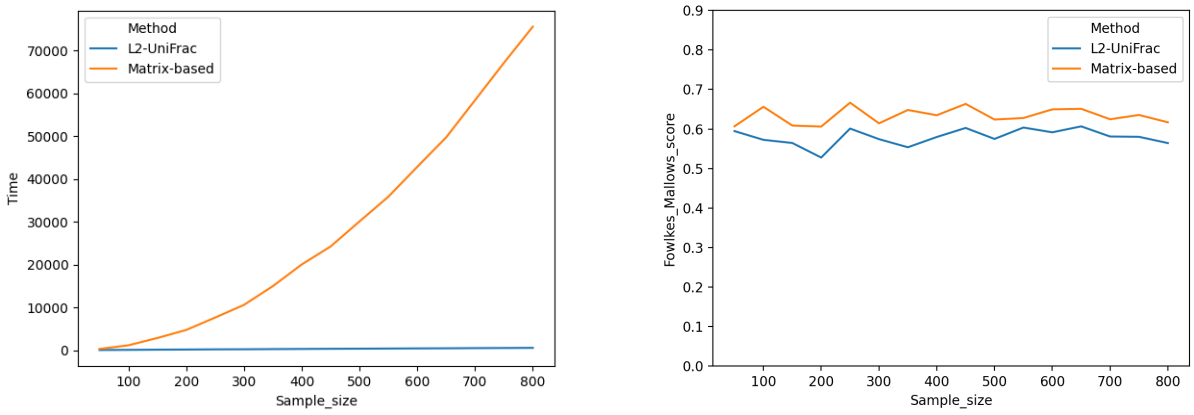

Figure 1: Comparison between the performance of the traditional matrix-based clustering, where pairwise UniFrac distance matrix was computed first and used as the basis for clustering, versus L_2-clustering, where data points were directly clustered in the L_2UniFrac space using L_2 norm as a distance metric. Left: running time. Right: clustering quality measured using Fowlkes-Mallows score, with a higher score indicating a better clustering.

As expected, for the traditional method, clustering time increased exponentially as sample size increases. According to the figure, for as few as 300 samples, the traditional matrix-based method would require approximately 3 hours to complete, whereas clustering in L_2UniFrac space took only a few minutes. This improvement on speed does come with a trade-off in clustering quality, as suggested by the figure on the right. However, the difference in clustering score is only around 0.1 on a scale of 0 to 1 and remained fairly constant in the experiments performed. In the case of large data size where the exact accuracy is not of top priority, the significant improvement in speed would deem clustering on L_2UniFrac space much more practical than the traditional 'all pairwise UniFrac" approach in real life scenarios.

3.2 Environment fingerprinting and classification

3.2.1 Finding the representative sample

In this section, we illustrate the process of finding the average sample using L_2UniFrac specific algorithms.

The illustration was performed using data with study ID 714 from `Qiita` [3], consisting of 528 samples from different environments. We first performed Algorithm 1 on all samples to obtain the aggregated vectors in the L_2UniFrac space. We then took the component-wise L_2 mean of these vectors. This mean vector was projected back to the distribution space using Algorithm 2, obtaining an average sample of the original samples. Principal Coordinate Analysis (PCoA) plots were used to show the relative relationship among these samples. Out of the five environments, we removed two that had too few samples and singled out each environment together with its representative data point to better observe their relationship for the remaining three. The results are shown in Figure 2. From the figure, a representative sample corresponds roughly to the centroid of the cluster consisting of the samples belonging to that environment. It turns out that this average sample can be equivalently computed by simply computing the component-wise mean of the original distributions, as shown in the Supplementary material Section S1.4.2. This further simplifies and speeds up the computation and is a unique advantage of L_2UniFrac that cannot be achieved using L_1UniFrac (see Supplementary material Section S1.4.1), making L_2UniFrac more than simply an alternative version of UniFrac but instead a novel metric having its own unique applications and properties.

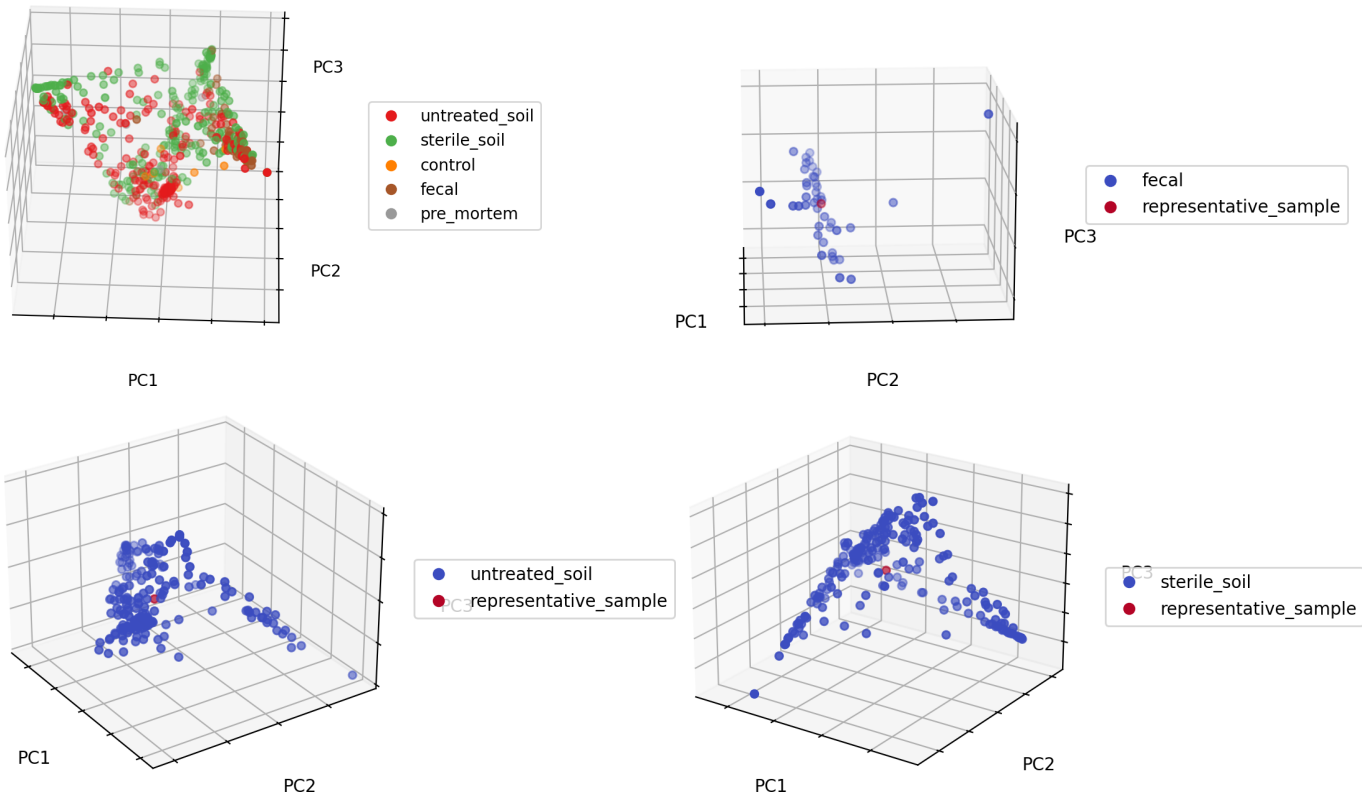

Figure 2: PCoA plots showing the representative samples with respect to all the samples from the respective environments.

3.2.2 Classification

Up to this point, we have demonstrated the process of obtaining the average sample amongst a pool of samples from the same environment. Ideally, this representative sample should be specific to the environment such that there is sufficient degree of differences between the representative samples of two different environments to significantly distinguish one from another. In this section, we test this hypothesis using a classification test using the same data set from Section 3.1 consisting of 16S samples from four body sites. Typically, when a group of unknown samples are given, clustering is first performed and classes are assigned based on the clustering result, such as based on the minimum distance to the centroid of the clusters, or by majority vote from a certain number of closest points. This type of methods can be time consuming and the unsupervised nature of this method also does not guarantee a direct correspondence between the segregation of the clusters and that of the actual traits of interest. The alternative we propose here is to first find the representative samples corresponding to each environment using a large size of known sample. For any subsequent new sample, its class can be assigned based on similarity with the representative samples. The accuracy of such assignments can therefore be used to gauge the specificity of such representative samples.

To this end, we partitioned the samples of each body site into 80 percent of training samples and 20 percent of testing samples. As representatives of current clustering-based methods, we considered k-medoids and k-means as clustering methods. For each of these clustering methods, we first performed clustering, using the pre-computed pairwise L_2UniFrac distance matrix for k-medoids and probability distribution vectors for k-means, setting the number of clusters equal to the number of body sites present. With 80 percent of the training data, the label for each cluster was assigned by majority vote. For each testing sample, the labels assigned by clustering was compared against its true label. For our L_2UniFrac based method, a representative sample was first computed using the training data. We then computed the L_2UniFrac distance between each of the testing samples and each of the representative samples and assign the test sample to the body site of which the L_2UniFrac between the representative sample produced the minimum L_2UniFrac distance. We used different scoring metrics from the `sci-kit learn` python package to evaluate the performance of all three classification methods. Figure 3 shows one such score, accuracy, used. The performance evaluated using other scores is shown in Figure (Supplementary Figure S1).

The result shows that our method out-performed the other two methods in most cases. The poor performance of k-medoids and k-means can be partially attributed to the fact that the labels assigned by majority vote after clustering were not consistent with the original classes, though the number of clusters were set to be equivalent to the true number of classes. For instance, in some cases, two clusters were assigned the same label by majority vote, while some of the original classes were missing in this process. This significantly affected the results.

Of course, this comparison is between two different kinds of approaches: supervised and unsupervised. In the case of the L_2UniFrac method, the exact classes were known and the representative samples were created from samples from a known environment. In the case of k-means and k-medoids, however, the clustering was unsupervised, significantly increasing the proportion of false positives and false negatives. On the other hand, this in turn shows the very advantage of our method of having the potential to convert a traditionally more unsupervised method into one with a more supervised nature by creating a "reference point" for each environment.

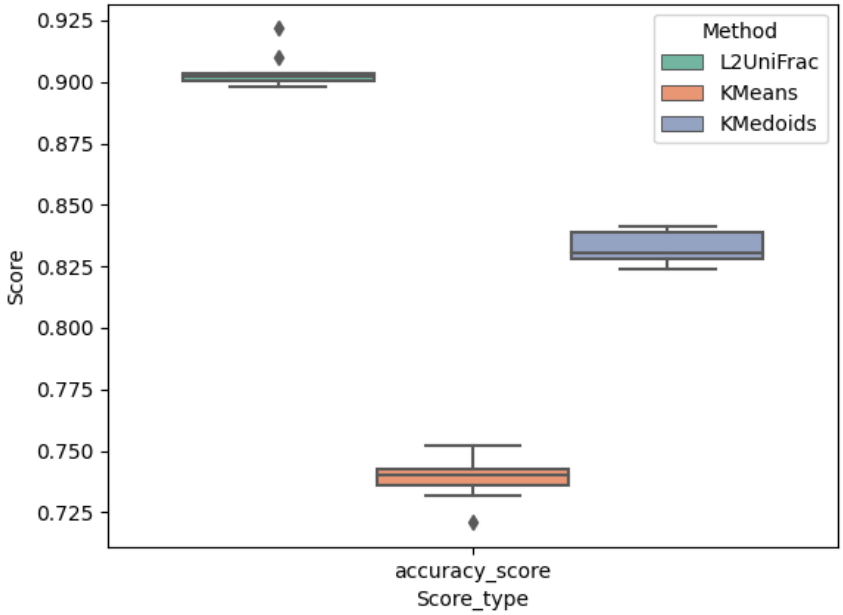

Figure 3: The accuracy scores of three classification methods. The accuracy score is defined as the proportion of correct classifications over approximately 1210 classifications.

3.2.3 Identifying differentially abundant organisms

An additional motivation of finding the average sample that characterizes an environment lies in the hope of identifying signature microbiomes that distinguish one environment significantly from another. Traditionally, in order to identify the most differentially abundant organisms, the abundances of each organism are to be compared across all samples. The results can then be visualized commonly using a box and whisker plot showing the average abundances of the organisms in each sampling environment, based on which a conclusion can be drawn on if the differences are significant, with the aid of statistical methods. We propose the alternative method based on the "flow" between average samples obtained through L_2UniFrac, giving rise to a phylogeny-aware comparison between sample groups.

The formulation of UniFrac as an optimum transport problem allows the differences in two samples to be represented as the "flow" between the two distributions. More specifically, for two samples (or averages) P and Q, the vector $W_{\sqrt{\cdot}}P - W_{\sqrt{\cdot}}Q$ represents the flow (or flux) across edges in the phylogenetic tree when moving abundance from P to overlap that of Q. This vector $W_{\sqrt{\cdot}}P - W_{\sqrt{\cdot}}Q$ is called the differential abundance vector (see [10] for the L_1 analog). The larger the absolute value of the flow between two organisms at a certain node in the phylogentic tree, the greater the difference is between the two organisms at this node/taxa. As such, we propose to obtain first the average samples from each environment and then compute the flow between pairwise average samples for the detection of differentially abundant taxa.

To test the feasibility of our method, we used the PRJEB6070 (BioProject 266076) study downloaded from HumanMetagenomeDB [6], consisting of 1261 gut samples grouped by three conditions: colorectal cancer, adenoma, and control. Unlike the previous experiments using 16S rRNA data, this dataset consists of whole genome shotgun (WGS) data. Using the principle adopted by the WGSUniFrac method [17], the average abundance with respect to each taxonomic ID is computed

for each condition, giving rise to three representative samples. Using the taxonomic tree in place of the phylogenetic tree with the reciprocal-based assignment of branch lengths as suggested by the WGSUniFrac method [17], the pairwise flow among the three representative samples were computed and illustrated in Figure 4.

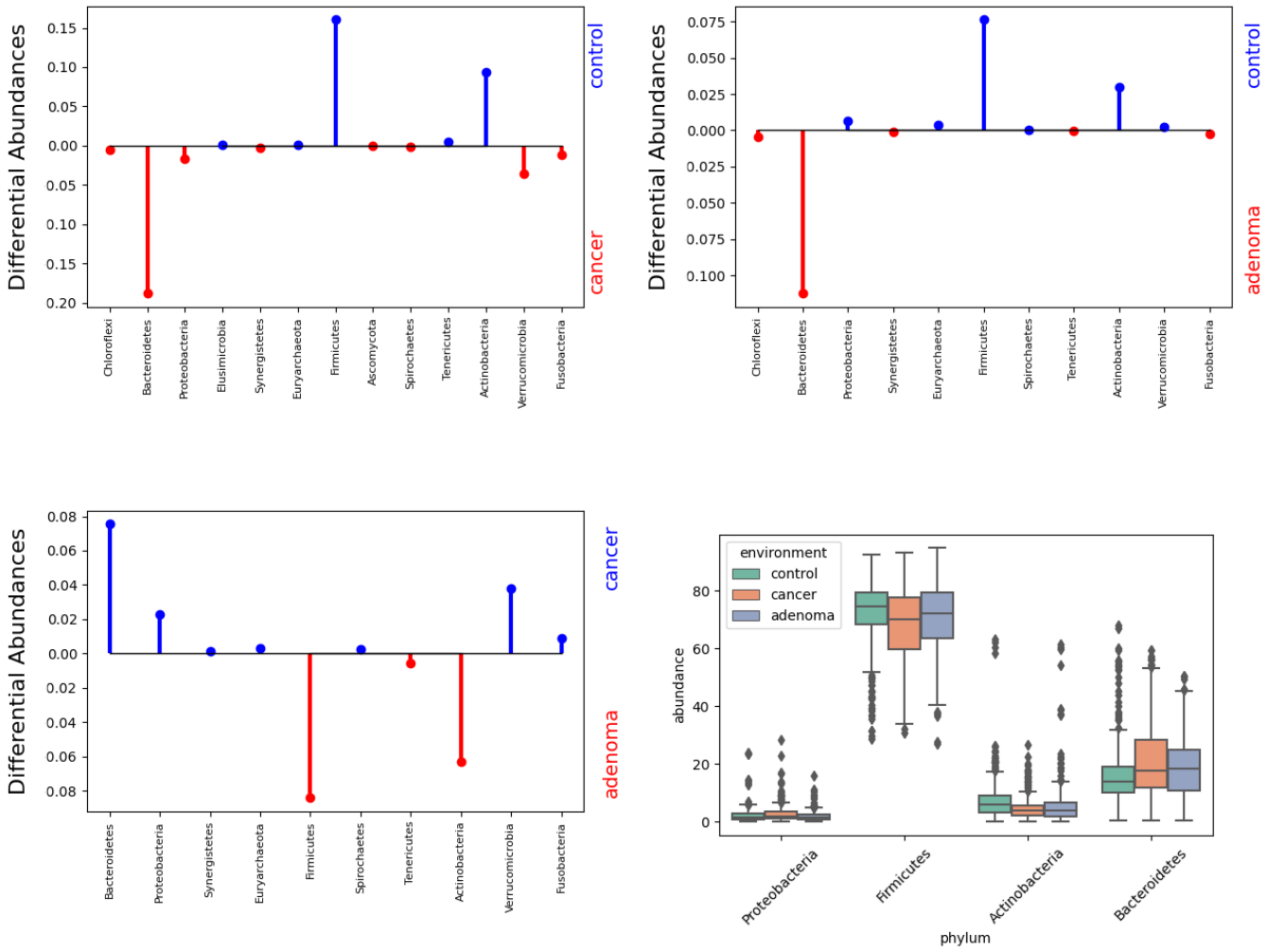

Figure 4: Top two and bottom left: top differentially abundant phyla between each of the condition pairs. The values are measured in terms of the L_2UniFrac flow. Bottom right: a traditional box plot of taxa relative abundances showing the spread of selected phyla.

Based on the differential abundance plots, we identified Bacteroidetes, Firmicutes, Actinobacteria, and Proteobacteria to be some of the phyla with the highest degree of distinguishability across the three conditions. The differential abundances plots show that compared to control, cancer and adenoma conditions tend to have a higher abundance of Bacteroidetes while having a lower abundance in Firmicutes and Actinobacteria. These agree with the box plot in the same Figure 4. Interestingly, these phyla had been observed in a previous study [18], supporting the sensitivity of our method. However, though the same phyla were identified to be correlated with the conditions, the trend of correlation does not seem to agree, with our data showing an increased abundance in Bacteroidetes and decreased abundances in Firmicutes and Actinobacteria in adenoma and cancer

conditions compared to the control group. The study carried out by Yachida et al. in [18] demonstrated elevation in all of the above-mentioned phyla. This is likely due to different data sets being compared. Perhaps more studies were to be conducted to reach a decisive conclusion. One thing we do agree, though, is the fact that the observed patterns seems to be progressive, with adenoma falling in between control and cancer, making these phyla strong candidates as indicators of the development of colorectal cancer.

3.3 Discussion

In this paper, we proposed L_2UniFrac as an alternative of the traditional L_1UniFrac. This L_2UniFrac preserves the robustness of L_1UniFrac as a phylogenetic metric for beta-diversity and allows one to compute the average distribution with respect to L_2UniFrac metric, which is not possible with the original version of UniFrac. Furthermore, we further explored the properties of the L_2UniFrac space, in which lie the aggregated vectors obtained by aggregating the distributions up the phylogenetic tree. In this space, clustering of metagenomic samples can be performed with much higher efficiency, circumventing the computation of the pairwise distance matrix, with negligible sacrifice of clustering quality. Given the invertibility of the aggregation process, as shown in Section S1.2 , the L_2-mean of the vectors in the L_2UniFrac space can be taken and projected back to the distribution space, giving rise to an average sample with respect to L_2UniFrac. We further showed that this projected mean is also a probability distribution, and is actually none other than the component-wise mean of the metagenomic distributions. This property gives the component-wise average of metagenomic distributions a phylogeny-aware biological interpretation as the L_2 norm of their differences is exactly the L_2UniFrac value.

We also demonstrated some of the potential usage of such average samples through experimentation. Most significantly, such average samples can serve as fingerprints for specific environments, allowing a biologist to better characterize the signature microbes belonging to a specific environment. Our results in Section 3.2 show that despite the between-sample variability even within the same environment, the average sample is sufficient to distinguish one environment from another to a fair degree. The average sample obtained using this method is stable to the environment in the sense that given the same environment and sufficiently large sample size, the variance between different average samples obtained using different sample pools or at different time points can be minimized. The reason is due to the nice property of its equivalence with the component-wise L_2-mean, making this process equivalent to taking a sample mean. By the Central Limit Theorem, when the sample size is sufficiently large, the sample means will be approximately normally distributed, with a standard deviation inversely proportional to the number of times samples are taken. Of course, the difficulty in defining an environment still remains, such as in the case of disease development, where different stages of disease are often artificially defined based on physiological differences. However, it can be envisioned that by pooling samples from adjacent stages, the average sample of intermediate stages can be computed. Coupled with the method to compute the flow between two distributions as demonstrated in Section 3.2.3, phylogeny-aware differential abundance profiles can be obtained, with the potential to provide a trajectory on the dynamic changes in metagenomic diversity with respect to the development of the disease to user-defined resolution.

4 Acknowledgment

This work was supported by the NIH grant 1R01GM146462-01.

References

[1] Joyita Banerjee, Neetu Mishra, and Yogita Dhas. "Metagenomics: A new horizon in cancer research". In: *Meta Gene* 5 (2015), pp. 84–89. ISSN: 2214-5400. DOI: 10.1016/j.mgene.2015.05.005.

[2] Steven N. Evans and Frederick A. Matsen. "The phylogenetic Kantorovich–Rubinstein metric for environmental sequence samples". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 74.3 (2012), pp. 569–592. ISSN: 1467-9868. DOI: 10.1111/j.1467-9868.2011.01018.x.

[3] Antonio Gonzalez et al. "Qiita: rapid, web-enabled microbiome meta-analysis". In: *Nature Methods* 15.10 (2018), pp. 796–798. ISSN: 1548-7091. DOI: 10.1038/s41592-018-0141-9.

[4] Katerina V.-A. Johnson et al. "Sociability in a non-captive macaque population is associated with beneficial gut bacteria". In: *Frontiers in Microbiology* 13 (2022), p. 1032495. ISSN: 1664-302X. DOI: 10.3389/fmicb.2022.1032495.

[5] Dae-Wook Kang et al. "Microbiota Transfer Therapy alters gut ecosystem and improves gastrointestinal and autism symptoms: an open-label study". In: *Microbiome* 5.1 (2017), p. 10. DOI: 10.1186/s40168-016-0225-7.

[6] Jonas Coelho Kasmanas et al. "HumanMetagenomeDB: a public repository of curated and standardized metadata for human metagenomes". In: *Nucleic Acids Research* 49.D1 (2020), gkaa1031–. ISSN: 0305-1048. DOI: 10.1093/nar/gkaa1031.

[7] Chao Liang et al. "Diversity and enterotype in gut bacterial community of adults in Taiwan". In: *BMC Genomics* 18.Suppl 1 (2017), p. 932. DOI: 10.1186/s12864-016-3261-6.

[8] Catherine Lozupone and Rob Knight. "UniFrac: a New Phylogenetic Method for Comparing Microbial Communities". In: *Applied and Environmental Microbiology* 71.12 (2005), pp. 8228–8235. ISSN: 0099-2240. DOI: 10.1128/aem.71.12.8228-8235.2005.

[9] Jason McClelland. "Wasserstein B-diversity metrics over graphs: Derivation, efficient computation and applications". PhD thesis. 2018.

[10] Jason McClelland and David Koslicki. "EMDUniFrac: exact linear time computation of the UniFrac metric and identification of differentially abundant organisms". In: *Journal of Mathematical Biology* 77.4 (2018), pp. 935–949. ISSN: 0303-6812. DOI: 10.1007/s00285-018-1235-9.

[11] Daniel McDonald et al. "Striped UniFrac: enabling microbiome analysis at unprecedented scale". In: *Nature methods* 15.11 (2018), pp. 847–848.

[12] Andrew Millward. "L2 UniFrac". MA thesis. the Pennsylvania State University, 2022.

[13] Stefano Mocali and Anna Benedetti. "Exploring research frontiers in microbiology: the challenge of metagenomics in soil microbiology". In: *Research in Microbiology* 161.6 (2010), pp. 497–505. ISSN: 0923-2508. DOI: 10.1016/j.resmic.2010.04.010.

[14] Vanessa Moura et al. "The influence of surface microbial diversity and succession on microbiologically influenced corrosion of steel in a simulated marine environment". In: *Archives of Microbiology* 200.10 (2018), pp. 1447–1456. ISSN: 0302-8933. DOI: 10.1007/s00203-018-1559-2.

[15] N. Nunan et al. "In Situ Spatial Patterns of Soil Bacterial Populations, Mapped at Multiple Scales, in an Arable Soil". In: *Microbial Ecology* 44.4 (2002), pp. 296–305. ISSN: 0095-3628. DOI: 10.1007/s00248-002-2021-0.

[16] Lin-Lin Wang et al. "A novel approach for the forensic diagnosis of drowning by microbiological analysis with next-generation sequencing and unweighted UniFrac-based PCoA". In: *International Journal of Legal Medicine* 134.6 (2020), pp. 2149–2159. ISSN: 0937-9827. DOI: 10.1007/s00414-020-02358-1.

[17] Wei Wei and David Koslicki. "WGSUniFrac: Applying UniFrac Metric to Whole Genome Shotgun Data". In: *22nd International Workshop on Algorithms in Bioinformatics (WABI 2022)*. Ed. by Christina Boucher and Sven Rahmann. Vol. 242. Leibniz International Proceedings in Informatics (LIPIcs). Dagstuhl, Germany: Schloss Dagstuhl – Leibniz-Zentrum für Informatik, 2022, 15:1–15:22. ISBN: 978-3-95977-243-3. DOI: 10.4230/LIPIcs.WABI.2022.15. URL: https://drops.dagstuhl.de/opus/volltexte/2022/17049.

[18] Shinichi Yachida et al. "Metagenomic and metabolomic analyses reveal distinct stage-specific phenotypes of the gut microbiota in colorectal cancer". In: *Nature Medicine* 25.6 (2019), pp. 968–976. ISSN: 1078-8956. DOI: 10.1038/s41591-019-0458-7.