AN ABSTRACT OF THE DISSERTATION OF

Jason McClelland for the degree of Doctor of Philosophy in Mathematics presented on September 7, 2018.

Title: Wasserstein β-Diversity Metrics over Graphs: Derivation, Efficient Computation and Applications

Abstract approved: _____

David Koslicki

Microbial ecology has been transformed by metagenomics, the study of the genetic information in entire communities of organisms. In the following we develop metagenomic tools arising from the classic Wasserstein metric as applied to questions regarding the diversity between microbial communities. We provide a novel proof of the characterization of the successful UniFrac metric as the Wasserstein metric over a graph-theoretic tree, and use the proof to develop an extremely efficient computational algorithm. The analytic framework we develop is then leveraged to provide formulations for the distribution of this Wasserstein based metric. We implement these ideas and demonstrate their utility on realworld datasets. We next turn to applying the Wasserstein metric as a reference-free diversity metric by utilizing de Bruijn graphs, mathematical structures at the heart of genome assembly techniques. We show how these techniques are related to established phylogenetically-aware diversity metrics. We then implement our results using newly developed approximation techniques for the computation of the Wasserstein metric and demonstrate this novel metric's utility in comparison to established metrics.

Wasserstein β-Diversity Metrics over Graphs: Derivation, Efficient Computation and Applications

by

Jason McClelland

A DISSERTATION

submitted to

Oregon State University

in partial fulfillment of
the requirements for the
degree of

Doctor of Philosophy

Presented September 7, 2018
Commencement June 2019

Doctor of Philosophy dissertation of Jason McClelland presented on September 7, 2018

APPROVED:

Major Professor, representing Mathematics

Head of the Department of Mathematics

Dean of the Graduate School

I understand that my dissertation will become part of the permanent collection of Oregon State University libraries. My signature below authorizes release of my dissertation to any reader upon request.

Jason McClelland, Author

ACKNOWLEDGEMENTS

There are a great number of people to whom I owe much gratitude for helping to see me to the end of my education. I would first like to thank my advisor, David Koslicki. Without his patience, thoughtfulness and understanding I doubt I would have been able to see this through. His work ethic, creativity and mathematical curiosity have been, and I suspect will continue to be, sources of inspiration in my transition from math student to working mathematician.

He is the last in a long line of teachers that I have been fortunate enough to have the opportunity from which to learn. My advisor for my Master's degree, Clayton Petsche, brought a playfulness and ease of demeanor to math that I actively try to encourage in my own teaching and discussion of the subject. My first term analysis professor, David Finch, gave lectures which were sharp, meaningful and filled with an understanding of the needs of his students, I aspire to do the same. As an undergraduate student I had two mentors, Pete Goetz and Dustin Poppendieck, that gave far more time and effort to my education than they were ever compensated for. The former showed me how math could inspire real passion, the latter how important kindness and determination are in everything you do, including teaching and research. Finally, long ago when I was writing high school newspaper articles and political cartoons instead of mathematical proofs, my English teacher and newspaper advisor Dwight Evans showed me that growing up need not make a person humorless or boring.

My family and friends have been impossibly supportive of the very long road through education that I elected to take. My parents, Joyce Jackson and John McClelland, made many sacrifices in raising me to become the person I am today. There is no way I would have continued my education without the encouragement they gave. My brother Tracy McClelland, and extra brother Justin Mitchell, have all always helped to give me a home

to come home to. My grandparents, in particular Alice and Kenneth Jackson, were a huge part of my childhood. Spending summers at campouts with many people five or six times my age had a strong impact on how I learned to treat others and how I learned to value work. My friends, both here in Corvallis and back in California, believed in me. I love you all.

Finally, I want to thanks my cats, Moses and Raccoon, and my bike, a 2011 Jamis Aurora. Sometimes I needed reminders that there is more to life than math, they did the trick.

TABLE OF CONTENTS

TABLE OF CONTENTS (Continued)

TABLE OF CONTENTS (Continued)

LIST OF FIGURES

LIST OF FIGURES (Continued)

LIST OF FIGURES (Continued)

WASSERSTEIN β-DIVERSITY METRICS OVER GRAPHS: DERIVATION, EFFICIENT COMPUTATION AND APPLICATIONS

1 GENERAL INTRODUCTION

> *"No discovery of mine has made, or is likely to make, directly or indirectly, for good or ill, the least difference to the amenity of the world"* - G.H Hardy.

The Hardy-Weinberg principle states that in the absence of external evolutionary influences, gene frequencies in a population of organisms will remain constant from one generation to the next. Hardy related the mathematical demonstration of this fact in the brief 1908 paper '*Mendelian proportions in a mixed population*' [44]. As a student of the history of mathematics is already aware, Hardy was not a biologist but rather a famous English number theorist and great champion of the pursuit of mathematics for its own sake. He saw his result in genetics as a triviality, but it was certainly this author's first exposure to the work of the man. It came as some great shock later in life to learn that the Hardy of the Hardy-Weinberg principle was precisely the same Hardy that had collaborated with Ramanujan and authored '*A Mathematician's Apology*,' from which the above quote is taken. We note that an understanding of the nature of genetics has transformed both the world's understanding of biology as well as its applications. As Hardy's discovery of this principle resides in every introductory discussion of the subject, we suspect his quote to be in error.

The above anecdote highlights the power of bringing a mathematician's expertise to questions arising in biology. More specifically, in bringing that expertise to those questions related to the genetic information contained in organisms or groups of organisms. That is the precise purpose of this document. In the following we provide a mathematical foundation to usage of the Wasserstein metric, a metric on probability measures over a metric space, in the context of β-diversity metrics for microbial ecology.

Summary of Content

Supposedly, the problem underlying gene frequency was related to Hardy while in the midst of a game of cricket. We are not so fortunate as to be considering questions so succinctly expressed, and so find that understanding the analytic tools in microbial ecology requires the survey of a great deal of material related to biology, such as the context in which β-diversity metrics arise and the needs of biologists in their utilization and analysis. We therefore dedicate the bulk of Section 1.1.1 to this information, assuming no biological background on the part of our audience.

On the other hand, given that this is a work of mathematics for a mathematically-trained audience, we do assume knowledge of the foundations of probability, analysis and linear algebra used throughout. Thus we spend Section 1.2 relating the specific theory of the Wasserstein metric and the mathematical context in which is arises. This includes a survey of its use in other, nonbiological, applications, from which we draw inspiration for new uses. We include a summary of tools for the computation of the Wasserstein metric in the finite setting to which we later apply it.

In Section 1.3, in preparation for Sections 2 and 3, we relate common results in the theory of graphs, the metric spaces over which we will be considering the Wasserstein metric. This is in no small part borne out of a desire to establish a common set of notation and definitions for the field. In particular, we introduce and discuss de Bruijn graphs, a mathematical structure utilized in genomics.

In Section 2 we proceed to demonstrate our main results. We prove in Section 2.2.1 an alternate characterization for the Wasserstein metric between relative abundances assigned to a phylogenetic tree, and utilize this to demonstrate a useful invariant behind the UniFrac distance, a well-used β-diversity metric. We produce in Section 2.2.4 a constructive proof that the UniFrac distance is the 1-Wasserstein metric, and adapt this proof to a highly

efficient and rapid algorithm for computing the UniFrac metric between relative abundances. We demonstrate in Section 2.3 how casting the ordination technique known as Double Principal Coordinate Analysis as the Euclidean distance between images of the action of linear transformation allows for its efficient computation and comparison to related metrics. Utilizing the mathematical framework we have established above, in Section 2.4 we derive formulations of the probability density function of the UniFrac metric under the assumption of Dirichlet distributed relative abundances, a distributional assumption inspired by biologists use of the Dirichlet-Multinomial distribution in modeling metagenomic datasets. For the biologically-minded, we demonstrate in Section 2.5 the utility of our results on datasets, both real world and synthetic.

Inspired by the utility of the Wasserstein metric, both in the above applications as well as the field of image analysis, in Section 3 we introduce a novel metric for probability distributions defined on genomic sequence datasets utilizing the Wasserstein metric on vertex-weighted de Bruijn graphs. In Sections 3.2.2 and 3.2.3 we adapt approximation algorithms for the computation of the Wasserstein metric, one heuristic algorithm known classically as the minimum cost method and another which computes an entropically-regularized version of the Wasserstein metric. We then apply these algorithms to metagenomic datasets in Section 3.3 and compare this reference-free metric on metagenomes against others used commonly in biology. We benchmark our heursitic approximation against other solution algorithms for the Wasserstein metric in Section 3.3.1 before discussing potential sources of improvement and future research.

We conclude in Section 4 with a summary of the results we have demonstrated and a brief outline of potential future work.

1.1 Background Material for Microbial Ecology

1.1.1 Introduction to Microbial Ecology

Microbes are the dominant form of life on Earth. Constituting 350 to 550 billion tons of biomass [123], microbial life is pervasive in every ecosystem science has ever investigated. They have been found everywhere from the bottom of the sea-floor [24] to the upper reaches of the atmosphere [106]. As many as 1000 microbial species exist in the human body at any moment [117], the composition and distribution of which has been implicated in diseases as varied as cancer [57] and depression [51]. Microbial life shapes the geology and atmosphere of our planet [17] and is so ubiquitous and resilient on Earth that our best hopes for finding life elsewhere in our Solar system lie in microbes [24].

The search for new microbial diversity does not require looking so far afield though. Estimates for the number of distinct microbial species on Earth extend to 1 trillion, of which 99.999% have not been identified [66]. This lack of understanding of microbes is due in no small part to their invisibility outside of the scope of the tools of science.

Hooke and Leeuwenhoek's use of the microscope in *Micrographia* (1665) transformed our understanding of biology [36] by showing that there was a invisible world of life around us. In recent years, a new set of tools has led to a new transformation in our understanding of the microbial world, that of next-generation or high-throughput sequencing technologies [99]. Recent advances in technology have made possible the rapid sequencing of the genetic material from both individual species as well as entire microbial communities.

But these new tools come packaged with new analytic and computational challenges. Roughly speaking, the genetic information in a single human being is encoded in a string of 6.4 billion letters from the alphabet $\{A, C, T, G\}$ [58]. The first instance of determining that information ended in 2001 after more than 13 years of work and an estimated cost of 1 to 3 billion dollars [22]. At the present moment, an Illumina X10 can sequence 18,000

human genomes per year at a cost of roughly $1000 per genome [40]. Managing and interpreting that volume of information demands new mathematical and computational solutions.

In the following section we discuss fundamentals of metagenomics and phylogenetics so as to better understand the analytic tools used in these fields. We discuss the measures of community diversity used by biologists and the manner in which biologists interpret their use, with an eye toward understanding the needs of biologists so as to motivate both new analytic tools as well as useful improvements to those tools currently employed.

1.1.2 Overview of the Foundations of Metagenomics

The language of life is written in the four letters $\{A, T, C, G\}$. These represent the 4 nucleotides adenine, thymine, cytosine and guanine which form the heritable, information encoding elements of deoxyribonucleic acid (DNA) or what is known as *genetic material*. These nucleotides are known as *base pairs* (bps), as each of cytosine and adenine only pair with guanine and thymine, respectively, in the double-stranded structure of DNA. A *gene* is a sequence of these base pairs which contains information for the construction of a protein. The expression of genes, that is the construction of proteins, is mediated by ribonucleic acid (RNA), itself a molecule similar to DNA but with the information encoding alphabet $\{A, U, C, G\}$, representing adenine, uracil, cytosine and guanine, respectively.

While genes are very important, they are not the entire story of our genetic material. Or even much of the story, as it turns out. A human being has an estimated 20,000 genes constituting less than 3% of the totality of our genetic material [21]. We call those segments of the genetic material in an organism which contain genes *coding regions* and the remainder *noncoding regions*. While the exact purpose of the noncoding regions is not clear, it is clear than it is not 'junk DNA' as was once thought. Indeed, more than 80% of these noncoding regions have now been associated with a variety of biological processes,

mostly related to gene regulation and expression [21].

The totality of the genetic material in an organism is referred to as its *genome*. While there is necessarily some variation between genomes for individuals in a species, the vast majority of the information between genomes is conserved among members of a species. As an example, total genetic variation in human beings is estimated to be 0.6% of the total genome [20]. The study of the totality of the genetic information in an organism or species is known as *genomics*, in contrast to the study of the genes known as *genetics*.

The study of a microbial genome begins with the determination of its contents, that is by *sequencing* of the genome [26]. The sequencing of a single genome starts by isolating and replicating genetic material. In the case of microbial life, this involves *culturing* the organism, that is, growing the organism in a lab. The methods which follow vary and evolved in a sequence of 'generations' [105], the first of which was Sanger sequencing. Sanger sequencing is a 'chain-terminating' method which produces continuous fragments of translated DNA, known as *reads*, of length 500 to 1000 bps long. The technologies which followed are generally known as 'sequencing by synthesis' methods, and include techniques such as *pyrosequencing*. These methods produce reads 50 to 300 bps in length but at a speed much faster than that of Sanger sequencing. The current generation of sequencing technology is known as 'large fragment single molecule' sequencing, which produces very long reads, up to 30,000 to 50,000 bps long, but with higher error rates. Each of this technologies still find application today to the meet the various needs of researchers regarding cost, speed and accuracy [105].

After sequencing, a researcher is left with a large number of relatively small reads, not a genome. Thus begins the *assembly* problem [26], reassembling the reads into larger pieces, known as *contigs*, and then into yet larger pieces, known as *scaffolds*, before merging into whole genomes. Here is where the necessity of mathematical tools begins. As noted previously, genomes can be billions of base pairs long, and sequencing technologies produce

reads of lengths on the order of 100s or 1000s of base pairs long. Note that since we are not sequencing the entirety of the genome in one piece, redundancy in our gathering of genetic information is necessary. The *coverage* of a sequencing is the expected number of times any individual base pair is transcribed. The coverage required for accurate assembly varies by genome length and read length, but is generally 30 to 100 times [26]. Additional coverage also helps address the error inherent in sequencing technologies, which vary by technology but are generally on the order of 0.1-1.0% [39].

Solutions to the assembly problem are divided into two principle approaches, based on the available information. Ideally, someone has already assembled a genome for a related organism. We can then use this *reference genome* as a template with which to reassemble our genome [63]. Alternately, and in at least one case necessarily, there is no reference genome with which to guide our assembly, and thus we are stuck with the problem of *de novo* assembly. That is, of reassembling our reads without a priori knowledge of their connectivity.

Mathematical techniques with which to solve the *de novo* assembly problem fall into three categories [75], overlap-layout-consensus (OLC) methods, de Bruijn graph methods and greedy algorithms. We defer a more thorough treatment of this subject to Section 1.3.5, but briefly, OLC and de Bruijn methods represent either reads, or segments of reads, respectively, as vertices in a graph such that edges in the graph correspond to potential assemblies of those reads. A full assembly of a genome, or subsection of a genome, is then a path which traverses each edge of that graph. Greedy algorithms predate the use of de Bruijn and OLC methods and are generally less efficient [75]. If the de Bruijn and OLC methods seek a globally optimal solution for assembly, by asking for a path which best assembles all of the reads, greedy algorithms ask for locally optimal assemblies, by taking a read and looking for the best extension of the read by matching overlap between reads. This extension process continues for as long as possible, and then these assembled pieces

of genome are compared against each other for consensus and thus assembly.

We have described the beginnings of a genomics workflow so as to give context to the challenges that face the metagenomics researcher. *Metagenomics* studies the totality of the genetic material in an entire community of organisms present in an environmental sample. So now suppose we have gathered a representative environmental sample from a microbial community, be it from the soil [78], the ocean [126] or the human body [117]. Following the process involved in determining the genome of an individual microorganism, we ought to isolate an individual species' genetic material. Unfortunately, this task is generally not possible, as an estimated more than 90% of species are unculturable under lab conditions [110].

To address this complication, researchers generally proceed down one of two paths of metagenomic analysis. The first is *whole genome shotgun sequencing* in which the combined genetic material is sequenced all at once, using any of the technologies described above [96, 102]. The other common avenue for metagenomic analysis is *16S rRNA sequencing* [56, 122]. The 16S ribosomal RNA (rRNA) is a small component of ribosomal RNA, roughly 1500 base pairs in length, which has been shown to be highly conserved in structure and function, both over time and between species, but which also contains highly variable regions. These conserved sections make it a reliable 'molecular clock' [128]. At the same time, highly variable regions allow for species identification [91] which allows classification and analysis of 16S rRNA gene to be used for the reconstruction of phylogenies as similarities in 16S rRNA sequences have been shown to be positively correlated [82] to similarities in phenotypes in microbial genomes. We defer a more complete discussion of phylogenetics to Section 1.1.3, but, briefly, a phylogeny is a description of the interrelated evolutionary history of a group of organisms. This analysis is made possible by amplicon sequencing [122], in which polymerase chain reactions (PCR) are used to selectively duplicate a segment of genetic material, in this case the 16S rRNA gene, prior to

sequencing.

There are challenges and benefits to each of these methods. In whole genome shotgun analysis the assembly problem is more complicated due to a variety of factors [83, 46]. Microbial species do not occur in uniform abundances, and, as such, there arises nonuniform coverage in sequencing between species. Additionally, sections of repeated genetic information both inside a given genome and between genomes make the assembly question ill-posed. There may be many potential assemblies of sequence reads which 'jump' genomes, wherein it becomes unclear how to reassemble genomes which share conserved genetic information. Finally, many highly interrelated species may exist together. Small variations between genomes for such species make assembly more challenging. Current metagenomic assembly techniques rely on extensions of the ideas outlined above, optimized for the size and complexity of metagenomes [121]. These factors make whole genome shotgun analysis more costly and time-intensive than 16S rRNA sequencing. Whole genome shotgun sequencing does allow for the potential identification of biological function through gene identification, something not possible in 16S rRNA sequencing. As whole genome shotgun sequencing utilizes a greater amount of genetic information, it has greater resolving power and is thus better able to distinguish highly similar species.

Having briefly described the field metagenomics and the methods by which researchers explore microbial communities, we turn to a more detailed discussion of phylogenetics and the tools used to study the evolutionary history and interrelatedness of biological species.

1.1.3 Definitions and Methods Related to Phylogenetics

As noted in our discussion of 16S rRNA sequencing, one of the principal uses of metagenomic analysis is the construction and understanding of phylogenies or phylogenetic trees, a concept we define now. *Phylogenetics* is the study of the shared evolutionary history and interrelatedness of a group of organisms.

Definition 1.1.1 (Phylogenetic tree)**.** A *phylogenetic tree* or *phylogeny* is a representation of the pattern of evolution and the sequence of common ancestors for a species or group of species [130].

A variety of essentially interchangeable terms are used in describing such groupings of species in the context of microbial ecology, such a *taxon* or *operational taxonomic unit (OTU)*. Our language will reflect whatever is most common in a given application.

In a phylogenetic tree vertices represent species, either extant or inferred. Extant species are represented as leaves. A *speciation event* is a bifurcation at an internal node of the tree, representing the beginning of a new evolutionary lineage. A *clade* is the group of all species which can trace their lineage back to a single speciation event, and thus a single common ancestor. Traveling back up the tree from the leaves, we travel backwards in time. The lengths of the edges in a phylogenetic tree generally represent the expected number of substitutions at each location in the genome of a species [125].

In the case that the rate of substitutions is constant either over time or between lineages, we say that the *molecular clock* holds and in this case the number of substitutions can be used as a surrogate for measurements of time. In this case, we can use this as a means by which to infer a root or last common ancestor of a group of species. We call such trees *rooted*. Figure 1.1 depicts a model rooted phylogenetic tree for a group of organisms. Phylogenetic tree can certainly be much more complicated that this, containing far more OTUs and evolutionary relationships. Figure 1.2 depicts a phylogenetic tree containing approximately 3000 species, constructed from 16S rRNA data.

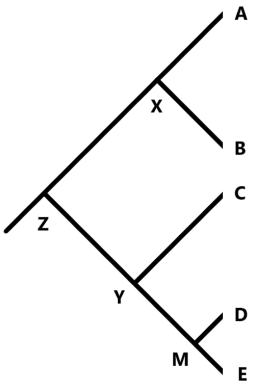


FIGURE 1.1: A model rooted phylogenetic tree depicting extant species A-E, in which A-B belong to a clade whose most recent common ancestor is X.

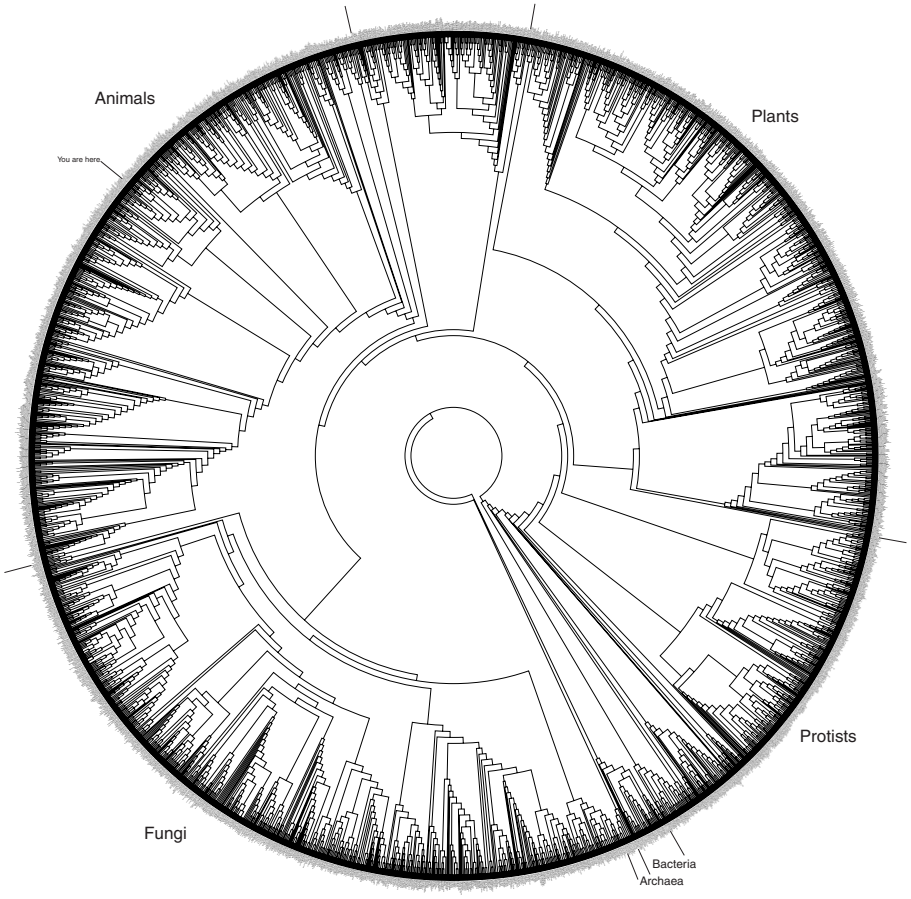


FIGURE 1.2: A phylogenetic tree depicting approximately 3000 species from the spectrum of Earth's biology. Source: David M. Hillis, Derrick Zwickl, and Robin Gutell, University of Texas

Evolutionary relationships between groups of organisms are not directly observable, and so phylogenetic trees must be inferred. The vast amount of genomic sequence data now available is a tremendous source of information for determining the evolutionary relationships between organisms. In particular, as mentioned in Section 1.1.2, 16S rRNA sequencing has been extremely successful in the inference of phylogenetic trees for microbial life [131] due to its stability and ubiquity among species, making it a reliable molecular clock. The two most common groups of methods for reconstructing phylogenetic trees from sequence data are distance-matrix methods and character-based methods [130]. Distance-matrix methods are pertinent to future discussion, so we briefly outline a few such methods here.

In distance-matrix methods, a metric on *aligned* sequences, that is, sequences in which areas of conserved function have been identified, are used to generate a matrix of all pairwise distances between genetic sequences, generally using some Markov model as a basis for the metric [86]. That distance-matrix is then used to generate an optimal tree by methods such as least squares, minimum evolution and neighbor-joining.

Letting $\mathbf{D}$ be such a matrix of observed pairwise distances between sequences, let, for any given phylogenetic tree, $\hat{\mathbf{D}}$ be a matrix of pairwise distances derived from an assumed molecular clock in the tree. Least squares seeks a tree which minimizes the sum of the squares of differences between the expected molecular clock distance in the tree and the observed distance between sequences. That is it seeks a tree such that

$$Q = \sum_i \sum_j (\mathbf{D}(i,j) - \hat{\mathbf{D}}(i,j))^2$$

is as small as possible. Minimal evolution seeks a tree in which branch lengths are minimal, again deriving distances from an expected molecular clock.

The most popular method [130] is *neighbor-joining* [34]. Neighbor-joining begins with a graph in which all taxa are joined to a common node. It then chooses a pair of taxa i and

j which minimizes the quantity

$$Q = (r-2)\mathbf{D}(i,j) - \sum_k (\mathbf{D}(i,k) + \mathbf{D}(j,k))$$

where r is the number of taxa adjacent to our initial node. The taxa i and j are then agglomerated into a common taxa descending from an inserted additional taxa u. Distances from u to each of i and j are then computed via the formula

$$\mathbf{D}(i,u) = \frac{1}{2} \cdot D(i,j) + \frac{1}{2(r-2)} \left[\sum_k (\mathbf{D}(i,k) - \mathbf{D}(j,k)) \right].$$

The taxa i and j are then removed from the distance matrix and list of taxa, being replace by u. Distances from u to the remaining taxa k are computed via the formula

$$\mathbf{D}(u,k) = \frac{1}{2}[\mathbf{D}(i,k) - \mathbf{D}(j,k)] + \frac{1}{2}[\mathbf{D}(j,k) - \mathbf{D}(j,u)].$$

The algorithm terminates when all of the taxa we began with are resolved from the initial node.

The construction of phylogenetic trees highlight the utility of metrics between genomes. We see that, by the very notion of a species or OTU in a genetic context, groups of species whose genomes are more related are defined to be more evolutionarily related. We turn from a discussion of how species are the same, that is how they are interrelated, to a discussion of the ways in which biologists define groups of species or communities of organisms to be different. That is, metrics of community diversity.

1.1.4 Survey of Methods in Community Dissimilarity Measurements

One of the chief applications of metagenomic analysis via high-throughput sequencing is the ability to conduct large-scale surveys of the spatiotemporal diversity of microbial communities [88]. The ability to reconstruct phylogenies for communities of organisms from a specific location, at a specific time, via analysis of 16S rRNA or whole genome shotgun datasets gives researchers the ability to understand the phylogenetic composition

and relative abundance of species in a community and how those quantities affect and are affected by environmental factors.

Ecologists define measures of the richness of an ecological community in terms of α-*diversity*, the site specific composition of biological communities, β-*diversity*, the variation in species composition between sites in an environment, and γ-*diversity*, the variability found in an entire ecosystem [124]. Metagenomics has given ecologists access to datasets describing the richness found in hundreds [69] and thousands [16] of environmental samples, which has led to the increased importance and application of analytic measures of such diversity.

Our chief interest lies in measures of β-diversity, but we will briefly touch on measures of α-diversity for context. Before continuing our discussion, we define an important term for casting ecological questions in a mathematical context, that of *relative abundance*.

Definition 1.1.2 (Relative abundance)**.** Given a environmental sample A, let S_A denote the set of species or OTUs present in A. For each $i \in S_A$ let n_i denote the number of specimens of i in A. Then the *relative abundance* $\mathbf{p}_A(i)$ of i in A is

$$\mathbf{p}_A(i) = \frac{n_i}{\sum_{j \in S_A} n_j}$$

and the *relative abundance of* A is the vector $\mathbf{p}_A$, indexed in some order. When comparing samples A and B, we will generally take the index set to be the species or OTUs found in the union of A and B.

The simplest measure of α-diversity is *species richness* [71], which is merely the number of species or OTUs (see Section 1.1.3) present in a community. In Whittaker's initial definition of the term α-diversity he takes this as "the most generally appropriate" [124] formulation, though it takes into account no information on species abundance. Measures of α-diversity which account for species abundances generally take the form of a reciprocal weighted mean, or some function of that quantity, as follows.

Definition 1.1.3 (order q diversity)**.** For a community of S species, where species $i \in S$ has relative abundance $\mathbf{p}(i)$, the *order q diversity* ${}^qD$ is

$${}^qD = 1 \Big/ \sqrt[q-1]{\sum_{i=1} \mathbf{p}^q(i)}.$$

Taking $q = 0$ we get species richness, $q = 2$ yields what is known as the *Simpson index*, taking $q = 1$ and applying the natural logarithm yields the *Shannon index*. Each correlates with an effective number of species [115].

On the other hand, β-diversity measures are far more varied [88, 7], given the wide variety of ways one might construct for comparing things which are not the same. Measures of β-diversity can be described as quantitative or qualitative, and phylogenetic or nonphylogenetic [80]. Phylogenetic measures account for the interrelatedness of species in defining differences between communities, inferred from some phylogeny, while nonphylogenetic do not. Quantitative measures account for the difference in relative abundance in species or OTUs between communities, while qualitative measure account only for the absence or presence of species or OTUs. In many cases [88], the same analytic tools can be considered in each context by recasting a dataset of abundances as binary absence-presence values or a dataset of absence-presence data as proportions of the total species accounted. We follow the literature in describing the most common applications of these tools.

Nonphylogenetic examples of β-diversity measures include Bray-Curtis dissimilarity [11], Jaccard Index [98], Sørenson index, modified Gower measure [3], Hellinger Distance (see Definition 1.2.5) and χ^2 distance (see Defintion 1.2.6). We describe each briefly.

Definition 1.1.4 (Bray-Curtis)**.** The *Bray-Curtis dissimilarity* $BC_{A,B}$ between sample specimen counts A and B is defined as

$$BC_{A,B} = 1 - \frac{2|A \cap B|}{|A| + |B|},$$

where the intersection $|A \cap B|$ above is defined as the sum over all species present of the minimum of the number of specimens counted from each sample.

Definition 1.1.5 (Jaccard)**.** The *Jaccard index* $J_{A,B}$ between samples A and B is

$$J_{A,B} = \frac{|S_A \cap S_B|}{|S_A| + |S_B| - |S_A \cap S_B|}$$

where S_A, S_B denote the set of species or OTUs recorded in each sample.

Definition 1.1.6 (Sørenson)**.** The *Sørenson index* $S_{A,B}$ between samples A and B is

$$S_{A,B} = \frac{|S_A \cap S_B|}{|S_A| + |S_B|}$$

using the notation of Definition 1.1.5.

Note that the Jaccard and Sørenson indices are qualitative measures of β-diversity which adhere to the relationship $S_{A,B} < J_{A,B}$, while Bray-Curtis is the quantitative version of the Sørenson index.

Definition 1.1.7 (Modified Gower)**.** Define l_+ by

$$l_+(x) = \begin{cases} \log_{10}(x) + 1 & x \neq 0 \\ 0 & x = 0. \end{cases}$$

Let $n_{A,B}$ denote the number of species or OTUs found in either of a pair of samples A and B. Then the Modified Gower $MG_{A,B}$ measure between samples A and B with relative abundances $\mathbf{p}_A$ and $\mathbf{p}_B$ is

$$MG_{A,B} = \frac{\sum_i^{n_{A,B}} |l_+(\mathbf{p}_A(i)) - l_+(\mathbf{p}_B(i))|}{n_{A,B}}$$

Hellinger and χ^2 are described in Section 1.2.4 and are computed on vectors of relative abundances.

Phylogenetic examples of β-diversity measures include community distance, community distance-nearest taxon distance [112], PhyloSor [13] and UniFrac [68, 67].

In the following we assume familiarity with the material related to graph theory detailed in Section 1.3. For microbial community samples A and B let $T = (V, E, \rho)$ be a rooted

phylogenetic tree which has been constructed for the combined communities (see Section 1.1.3). Let l be a weight function for the edges of T. Let d_T be the induced path length metric in T between species or OTUs in A and B. Let $\mathbf{p}_A, \mathbf{p}_B$ be the vectors of relative abundance. Let S_A, S_B be the species or OTUs present in each of A and B, let $S_{A,B}$ the species or OTUs present in their union, and let n_A, n_B and $n_{A,B}$ the number of elements in each set.

Definition 1.1.8 (Community distance)**.** The *community distance* $CD_{A,B}$ between samples A and B is defined as

$$CD_{A,B} = \frac{1}{n_A \cdot n_B} \sum_{i \in S_A} \sum_{j \in S_B} \mathrm{d}_T(i,j).$$

Definition 1.1.9 (Community Distance-Nearest Taxon)**.** The *community distance-nearest taxon* $CDNT_{A,B}$ between samples A and B is defined as

$$CDNT_{A,B} = \frac{1}{n_A} \sum_{i \in S_A} \min\{j \in S_B | \mathrm{d}_T(i,j)\}.$$

Each of the above is qualitative as described but can be adapted to a quantitative measure by weighting the summands involved by their corresponding relative abundances.

Our next measure of β-diversity was given by Rao [97] in an attempt to give a unifying mathematical framework similar to that of Defintion 1.1.3 for measures of diversity between communities.

Definition 1.1.10 (Diversity Index)**.** Let d be any symmetric measure of difference between species or OTUs. Then for samples X and Y containing species S_X and S_Y with relative abundance vectors $\mathbf{p}_X$ and $\mathbf{p}_Y$ we define the *diversity index* $h_{X,Y}$ to be

$$h_{X,Y} = \sum_{i \in S_X} \sum_{j \in S_Y} \mathbf{p}_X(i)\mathbf{p}_Y(j)\mathrm{d}(i,j).$$

Definition 1.1.11 (Dissimilarity Index)**.** The *dissimilarity index* D_h between samples A and B for a given measure of difference d is given by

$$D_h(A,B) = h_{A,B} - \frac{1}{2}\left(h_{A,A} + h_{B,B}\right).$$

The following terminology will be useful in the definition of the following phylogenetic β-diversity measures. We say an edge e in T *belongs* to sample A in the case that $H \cap S_A$ is nonempty while $H \cap S_B$ is empty for the branch H of T (Definition 1.3.8) defined by the deletion of e. We say an edge belongs to both if $H \cap S_{A,B}$ is nonempty and say it belongs neither otherwise.

Definition 1.1.12 (PhyloSor)**.** Let E_A,E_B and $E_{A,B}$ be the set of edges belonging to A,B and both, respectively. Then, given our weight function l, the *PhylSor* $PS_{A,B}$ diversity measure between A and B is given by

$$PS_{A,B} = \frac{2 \cdot \sum_{e \in E_{A,B}} l(e)}{\sum_{e \in E_A} l(e) + \sum_{e \in E_B} l(e)}.$$

Note that PhyloSor is qualitative. It is the phylogenetic application of the idea behind the Sørenson index, where species or OTU absence-presence is weighted by evolutionary distance.

Finally, we discuss UniFrac, a phylogenetic β-diversity measure which has formulations which are qualitative, *unweighted UniFrac*, as well as quantitative, *weighted UniFrac*. We retain the notation used in the definition of PhyloSor in our description of each.

Definition 1.1.13 (Unweighted UniFrac)**.** Let E_A and E_B be the set of edges belonging to A and B, respectively. Then the *unweighted UniFrac* metric $UF^u_{A,B}$ between A and B is given by

$$UF^u_{A,B} = \frac{\sum_{e \in E_A} l(e) + \sum_{e \in E_B} l(e)}{\sum_{e \in E_A \cup E_B} l(e)}.$$

That is, UniFrac measures the *fraction* of a phylogeny which is *unique* to each of the communities. Figure 1.3 shows the identification of edges belonging uniquely to each of two samples. We define *weighted UniFrac* similarly. For an edge e let H_e be the branch defined by e. Let $\mathbf{p}_A, \mathbf{p}_B$ be the relative abundances for samples A and B.

Definition 1.1.14 (Weighted UniFrac)**.** The *weighted UniFrac* metric $UF_{A,B}$ between samples A and B is given by

$$UF_{A,B} = \sum_{e \in E} l(e) \cdot | \sum_{v \in H_e} (\mathbf{p}_A(v) - \mathbf{p}_B(v))|.$$

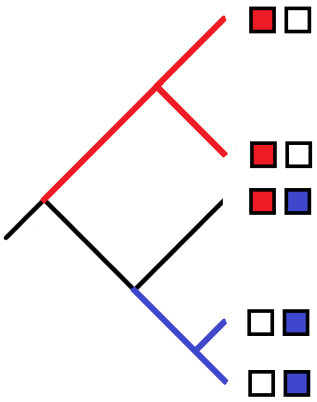


FIGURE 1.3: A depiction of edge identification on a phylogenetic tree T used in the computation of UniFrac between samples A and B. The presence of a species or OTU in sample A is indicated by a red box, that of sample B in blue. Edges identified with each of A and B are colored correspondingly.

In the example depicted, let the tree be ultrametric (Definition 1.3.5) of depth 3. Suppose the upper two OTUs terminate in edges of length 1, and thus the length of edges which belong to sample A is 4. Suppose that the bottom two edges are of length 0.5 and that these edges belong to a clade which arises from a edge of length 1, so that the edges which belong to sample B is 2. The edge which contains OTUs from both is then necessarily 1.5 and thus the total length of edges belonging to either is 9. In this case $UF^u_{A,B} = 6/9$.

While the property of being ultrametric is not necessary in the defintion of UniFrac, in either its quantitative or qualitative forms, the existence of disparities in branch lengths yield over-weighting of quickly evolving taxa, that is, those with longer edge lengths [68]. In such cases it may be advisable to compute *normalized* versions of UniFrac as follows.

Let d_T be the distance function in T, and define D by

$$D = \sum_{v \in T} d_T(v, \rho) \cdot (\mathbf{p}_A(v) + \mathbf{p}_B(v)).$$

Thus D is the weighted average distance between OTUs from A and B to the root ρ.

Definition 1.1.15 (Normalized UniFrac)**.** Given samples A and B, with UniFrac distances $UF^u_{A,B}$ and $UF_{A,B}$, define the *normalized UniFrac distances* $\overline{UF}^u_{A,B}$ and $\overline{UF}_{A,B}$ by $\overline{UF}^u_{A,B} = UF^u_{A,B}/D$ and $\overline{UF}_{A,B} = UF_{A,B}/D$.

In this section we have addressed a variety of ways in which biologists compare groups of organisms, in particular the formulation of β-diversity metrics. Each of the metrics discussed thus far require the construction of an underlying phylogenetic tree in comparing groups of organisms. We turn next to metrics that do not have this requirement, so-called reference-free metrics in metagenomics.

1.1.5 Survey of Methods in Reference-free Metagenomic Comparison

The methods for comparing communities of microbial organisms outlined in Section 1.1.4 share one common constraint, they require the determination of precisely which species or OTUs are present in a sample before making comparisons between communities. A dataset of sequence reads needs to be transformed into a list of species or OTUs and, hopefully, relative abundances. This generally requires that those sequences be assembled and aligned. These are generally referred to as binning methods [73], wherein community composition and relative abundance are derived from placing contigs into 'bins' which have been assigned to species or OTUs via the use of reference databases. Such methods are powerful but suffer from the difficulties related to sequence alignment and assembly described in Section 1.1.2. Here we describe *reference-free methods* for genomic and metagenomic comparison, which do not rely on sequence alignment and assignment to species or OTUs.

The first of such methods are variations on the notation of *factor frequencies*. In these methods an information-theoretic framework is adapted wherein k-mer frequencies (see Section 1.3.4 for notation and definitions related to theory of words) from sequence reads are analyzed directly by application of a variety of tools. In the description that follows, let A be B metagenomic samples and let W_A and W_B be collections of sequence reads from each, respectively.

The simplest such tool is the *dinucleotide odds ratio* [54], derived for genomic comparison.

Definition 1.1.16 (Dinucleotide Odds Ratio)**.** Given symbols X and Y, let XY denote their concatenation. Define the *dinucleotide odds ratio* ρ_{XY} to be

$$\rho_{XY} = \frac{freq_{W_A}(XY)}{freq_{W_A}(X) \cdot freq_{W_A}(Y)}$$

for each ordered pair (X, Y) from $\{A, C, T, G\}^2$.

The 16 dinucleotide odds ratios measure the deviation from uniformly random expectation of the occurrence of 2-mers, strings of symbols from $\{A, C, G, T\}$ of length 2, in a genome based on letter frequency. This discriminant, when applied to genes from a variety of organisms, has been shown [54, 50] to be a sort of 'genomic signature' capable of both identifying taxa and characterizing the evolutionary distance between taxa.

Applying more robust tools from information theory has led to the use of the Jensen-Shannon divergence d_{JS} (Definition 1.2.4) based upon the Kullback-Leibler divergence. These methods have been applied successfully to factor frequencies of k-mers as a distance metric for phylogenetic tree construction in mammals [104] and Hepatitis viruses by neighbor-joining (See Section 1.1.3 for information related to phylogenetic trees and their construction).

Sims et al. [104] also explored the ideal range of values for k in such metrics, giving lower and upper bounds as follows, albeit in differing notation. Let G be a collection

of sequence reads from a genome. They establish $l_{\min} = \max_{n\in\mathbb{N}} f_G(n)$, where f_G is the complexity function for the set G (Defintion 1.3.11), as the minimum value to achieve maximum discrimination power for Jensen-Shannon based factor frequency methods. They empirically approximate the value for a genome of length n as $l_{\min} = \log_4(n)$, and give explicit computations for rat mitochondrial genomes (n=16 kBps) and human chromosome 1 (n=230 MBps) as $l_{\min} = 7$ and $l_{\min} = 14$, respectively. This lower bound is somewhat obvious, the discriminating power of a metric only improves as it is able to consider more distinct features.

Their upper bound is derived as follows. We first extend the idea behind the dinucleotide odds ratios to more general k-mers. Let $i = w_2 \dots w_{(k-1)}$ be a word of length $(k-2)$ in the genetic alphabet. Let $p = w_2 \dots w_k$ and let $s = w_1 \dots w_{(k-1)}$ be words of length $(k-1)$ containing i as a prefix and suffix, respectively. Then the expected frequency $\widehat{freq}_w$ of the k-mer $w = w_1 w_2 \dots w_k$, given the observed $(k-2)$-mer and $(k-1)$-mer frequencies, is

$$\widehat{freq}_w = \frac{freq_s \cdot freq_p}{freq_i}.$$

Let $\widehat{freq}^k(G)$ be the vector of expected k-mer frequencies for a genome G, given the observed frequencies $freq^{(k-1)}(G)$ and $freq^{(k-2)}(G)$. Then $d_{KL}(\hat{freq}^k(G), freq^k(G))$ is a measure of the additional information gained by considering k-mer frequencies, relative to the information already contained in the distributions of $(k-1)$ and $(k-2)$-mers. Let ϵ be small and positive, and set $l_{\max} = \min_{n\in\mathbb{N}}\{n \,\big|\, d_{KL}(\widehat{freq}^n(G), freq^n(G)) > \epsilon\}$. That is, a measure based in analyzing k-mers stops gaining discriminating power when considering larger values of k no longer garners new information regarding factor frequencies. They empirically approximate this in their application to phylogenetic tree construction via neighbor-joining to be the least k such that the phylogenetic tree generated ceases to change. For rat mitochondrial genomes (n=16 kBps) they determined $l_{\max} = 14$.

An alternate, purely experimental approach, was taken to determining optimal k-mer sizes

in [129]. Wu et al. analyzed synthetic sequences in which a fixed 'mother' sequence was generated under the assumption of uniform and independent nucleotide distribution and 'son' sequences were derived by mutating the mother sequence at each base with fixed probability $p = 0.01, 0.02, \ldots, 0.99, 1.00$. Here the goal was to determine which size k-mer best balanced capturing the similarity transmitted from mother to son with the noise introduced by mutation. They experimentally derived optimal k values of 7, 8 and 9 for comparison of sequences of length approximately 700-2500, 2500-5000 and 5000-6100, respectively.

Having discussed a variety of measures of diversity used in microbial ecology, including their formulation, scope and parameter values, we turn briefly to their analysis. Each of the above returns a number or collection of numbers when used to describe the diverity observed in differing microbial communities. When used to compare many such samples, ecologists are faced with many, many such numbers. We next discuss how ecologist give significance to individual measurements and interpret large collections of pairwise measurements for the purpose of generating hypotheses.

1.1.6 Survey of Techniques in Ecological Data Analysis

As we have stated, the revolution in high-throughput metagenomic sequencing has led to the generation of tremendous amounts of data containing new insights into how microbial communities are composed, interrelated and varying in both time and space. Uncovering those insights, detecting the biological signals inside large datasets, requires not only analytic tools for interpreting metagenomic datasets directly, but also tools for interpreting and understanding the measurements, models and inferences built from those datasets. Here we briefly describe the means by which microbial ecologists answer the following two questions. Given a model or metric, how do I understand the significance or sensitivity to error of the results? Given a large collection of measurements relating communities of

organisms, how do I detect structure or order in a dataset so as to infer structure or order in biological communities? Understanding and addressing the ways in which biologists use data motivates the construction and improvement of the tools which generate data.

While the field of metagenomics is new, the desire to understand the distribution and evolution of species abundance is not and has led to many theories [84] regarding the the relative abundance of species in communities. Under the recent 'Unified Neutral Theory of Biodiversity' [10, 84] relative abundances of species are described by *Dirichlet-multinomial* distributions. This, in addition to the analytic tractability and simplicity of the Dirichlet-multinomial distribution, has led to its frequent use in modeling metagenomic datasets [85, 6]. We first define the multinomial distribution.

Definition 1.1.17 (Multinomial Distribution)**.** Let $X_1, X_2, \ldots, X_n$ be a sequence of n independent trials, each with k mutually exclusive and exhaustive possible outcomes. For each $i \in \{1, \ldots, k\}$, say outcome i occurs with fixed probability p_i. Then the number of occurrences for each of the k possible outcomes after our n trials is a random variable given by the *multinomial distribution.*

The multinomial distribution is a discrete probability distribution with parameters $n > 0$ and $\{p_1, \ldots, p_k\}$ such that $\sum_{i=1}^{k} p_i = 1$ supported on the set $(x_1, x_2, \ldots, x_k)$ such that each $x_i \in \{0, \ldots, n\}$ and $\sum_i x_i = n$. The probability mass function for the multinomial distribution is given by

$$f(x_1, \ldots, x_k; n, p_1, \ldots, p_k) = \frac{n!}{x_1! \ldots x_n!} p_1^{x_1} \cdot \cdots p_k^{x_k}$$

for the parameters described above.

In applications to metagenomics, the p_i represent the assumed fraction of OTU i in the communities genomic information, as sampled in a collection of sequence reads. This information is not generally known beforehand, thus the vector of p_i is more often assumed to be a random variable itself. If we assume that the p_i are *Dirichlet* distributed, we

arrive at the Dirichlet-multinomial distribution. We proceed with a few relevant definitions related to the Dirichlet-multinomial distribution and its formulation. We first recall the definition of the *Gamma function* $\Gamma(z)$ and the *Beta function* $B(\alpha)$.

Definition 1.1.18 (Gamma function)**.** For a $z \in \mathbb{C}$ with nonnegative real part, we define the *Gamma function* $\Gamma(z)$ by

$$\Gamma(z) = \int_0^\infty x^{z-1} e^{-x} \mathrm{d}x.$$

Definition 1.1.19 (Multivariate Beta function)**.** For $\alpha = (\alpha_1, \ldots, \alpha_n)$, with each $\alpha_i > 0$, the *multivariate beta function* $B(\alpha)$ is given by

$$B(\alpha) = \frac{\prod_{i=1}^n \Gamma(\alpha_i)}{\Gamma\left(\sum_{i=1}^n \alpha_i\right)}.$$

We are now prepared to give a formal defintion of the Dirichlet distribution and the related Dirichlet-multinomial distribution.

Definition 1.1.20 (Dirichlet Distribution)**.** The *Dirichlet distribution* is a probability distribution with parameters $\alpha = (\alpha_1, \ldots, \alpha_n)$ defined on the open $(n-1)$-dimensional simplex in $\mathbb{R}^n$ such that the probability density funciton is

$$f(x_1, \ldots, x_n; \alpha_1, \ldots, \alpha_n) = \frac{1}{B(\alpha)} \prod_{i=1}^n x_i^{\alpha_i - 1}.$$

Definition 1.1.21 (Dirichlet-Multinomial Distribution)**.** The *Dirichlet-multinomial* distribution is a compound probability distribution describing a random variable $x = (x_1, \ldots, x_k)$ such that x is distributed by the multinomial distribution with parameters $p_1, \ldots, p_k$ drawn from the Dirichlet distribution.

In addition to comparison against such probabilistic models describing metagenomic datasets, *Monte Carlo permutation tests* [41] are often [114, 132, 116] used in measuring the significance of metrics in metagenomic studies. The idea is very straightforward.

Given a measurement between two communities or samples A and B of sizes m and n, respectively, we wish to know whether the measured difference between A and B is greater than would be expected due to chance. So we repeatedly select random samples or communities S_1 and S_2 of sizes m and n, generally 1000s or 10,000s of such pairs, and measure the difference between them. This estimates the distribution of the measurement, and the fraction of simulated pairs which fall above the distance between A and B gives an indication of the significance.

We next address the ways in which analytic measures are used in metagenomics, particularly with respect to exploratory data analysis. Here researchers are interested in seeing large scale structure in a dataset so as to formulate scientific hypotheses. The broad term for such tools in statistics are *ordination techniques*. The idea is to organize a set of objects or observations such that object which lie close together with respect to some easily observed distance, such as the Euclidean metric in the plane, are more related. Two of the most utilized such tools, particularly in microbiology, is that of *principal component analysis (PCA)* and *principal coordinate analysis (PCoA)* [95, 136].

PCA takes as input a dataset of m observations $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_m$ of a multivariate random variable with n quantitative components, that is $\mathbf{x}_j = (x_{1,j}, x_{2,j}, \ldots, x_{n,j})$ for each $1 \leq j \leq n$. PCA seeks to produce the coefficients $\mathbf{c}_i = (c_{i,1}, c_{i,2}, \ldots, c_{i,n})$, for each $1 \leq i \leq n$, of a set of n uncorrelated linear combinations of the components of the $\mathbf{x}_i$, known as *principal components*, such that

$$\mathbf{c}_1^t \cdot \mathbf{x}_j = \sum_i c_{1,i} \cdot x_{i,j}$$

has maximum variance, and that each subsequent $\mathbf{c}_i$ captures as much of the remaining variance as possible while remaining uncorrelated. We follow [52] in deriving these components. We first recall a minor result necessary for our work.

Proposition 1.1.1 (Covariance of a linear transformation of a multivariate random variable)**.** *Say $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)$ is a multivariate random variable such that each x_i has*

finite variance and $\boldsymbol{T}$ *is a linear transformation, then*

$$\operatorname{cov}(\boldsymbol{T}\boldsymbol{x}) = \boldsymbol{T}^t \Sigma \boldsymbol{T}$$

where $\operatorname{cov}(\boldsymbol{x}) = \Sigma$.

Proof. Let $\mathbb{E}(\mathbf{x}) = m$. Then the covariance matrix of $\mathbf{x}$ is defined as $\operatorname{cov}(\mathbf{x}) = \mathbb{E}((\mathbf{x} - m)(\mathbf{x} - m)^t)$. Thus we have, by the linearity of the expectation, that

$$\begin{aligned}
\operatorname{cov}(\mathbf{T}\mathbf{x}) &= \mathbb{E}((\mathbf{T}\mathbf{x} - \mathbb{E}(\mathbf{T}\mathbf{x}))(\mathbf{T}\mathbf{x} - \mathbb{E}(\mathbf{T}\mathbf{x}))^t) \\
&= \mathbb{E}((\mathbf{T}\mathbf{x} - \mathbf{T}\mathbb{E}(\mathbf{x}))(\mathbf{T}\mathbf{x} - \mathbf{T}\mathbb{E}(\mathbf{x}))^t) \\
&= \mathbb{E}((\mathbf{T}(\mathbf{x} - m))(\mathbf{T}(\mathbf{x} - m))^t) \\
&= \mathbb{E}(\mathbf{T}(\mathbf{x} - m)(\mathbf{x} - m)^t)\mathbf{T}^t) \\
&= \mathbf{T}\mathbb{E}((\mathbf{x} - m)(\mathbf{x} - m)^t)\mathbf{T}^t \\
&= \mathbf{T}\Sigma\mathbf{T}^t
\end{aligned}$$

as required. □

Returning to our consideration of PCA, let Σ be the known covariance matrix, or a sample covariance matrix S which estimates Σ, for the n components of our multivariate random variable. Let $\mathbf{c}_1$ be the vector which maximizes $\operatorname{var}(\mathbf{c}_1^t \cdot \mathbf{x}) = \mathbf{c}_1^t \Sigma \mathbf{c}_1$. As written, the value is not finite for nonconstant $\mathbf{x}$, thus we constrain $\mathbf{c}_1$ to have unit norm, that is $\mathbf{c}_1^t \mathbf{c}_1 = 1$. The multivariate optimization problem of maximizing $\mathbf{c}_1^t \Sigma \mathbf{c}_1$ subject to $\mathbf{c}_1^t \mathbf{c}_1 = 1$ can be solved by use of Lagrange multipliers. That is, maximize

$$\mathbf{c}_1^t \Sigma \mathbf{c}_1 - \lambda(\mathbf{c}_1^t \mathbf{c}_1 - 1)$$

for Lagrange multiplier λ.

Taking the derivative with respect to $\mathbf{c}_1$ and setting this equal to zero we see that

$$2\Sigma\mathbf{c}_1 - 2\lambda\mathbf{c}_1 = 0$$

so that

$$(\Sigma - \lambda \mathbb{I})\mathbf{c}_1 = 0.$$

Thus $\mathbf{c}_1$ is a eigenvector with corresponding eigenvalue λ. To determine our choice of eigenvector, recall that we wish to maximize

$$\mathbf{c}_1^t \Sigma \mathbf{c}_1 = \mathbf{c}_1^t \lambda \mathbf{c}_1 = \lambda$$

and thus we take the largest possible eigenvalue for Σ. Note that since the covariance matrix is always positive semi-definite, the eigenvalues are all non-negative. It can be shown [52] that $\mathbf{c}_2$ is the unit norm eigenvector corresponding to the second largest eigenvalue, and so on.

Alternately [31], we can consider our m multivariate quantities as an $m \times n$ matrix $\mathbf{M}$ whose columns are the observations, and we may perform the same eigenvalue decomposition on a related matrix to yield our principal components.

Definition 1.1.22. Given $\mathbf{M}$ above, define $\mathbf{M}_c$ by subtracting from each column its mean. We then define the *Gram matrix* $G = \mathbf{M}_c \mathbf{M}_c^t$

The Gram matrix of inner products defined above differs from our covariance matrix by transposition and scaling, and yields the same principal components as above [31].

PCA is frequently used in biology for ordination of taxa or OTU distribution by selecting the first few, say 2 or 3, principal components and plotting the dataset in these transformed coordinates. In these applications it is less than ideal [95], as we are maximizing the retained variation in the Euclidean distance given by embedding the dataset in lower dimensions. If the Euclidean distance between taxa or OTU distributions is not meaningful, then there may be little meaning in preserving it. Our next ordination technique, PCoA, is similar but seeks to maximize the retained variance given by some hopefully more meaningful metric.

PCoA, also known as *metric multidimensional scaling* takes as input an $n \times n$ matrix $\mathbf{D}$ of metric similarity, or dissimilarity, measures between n objects of interest and seeks to embed the those n objects into low dimensional Euclidean space, typically of dimensions 2 or 3, in such a way as to preserve relationships between objections. We follow [61] in giving a brief description of the method. Given a matrix of pairwise distances $\mathbf{D}$, we first transform $\mathbf{D}$ into the related matrix $\mathbf{A}$ by

$$\mathbf{A}(i,j) = -\frac{1}{2}\mathbf{D}^2(i,j).$$

Defining $\overline{\mathbf{A}_{i,\cdot}}, \overline{\mathbf{A}_{\cdot,j}}$ and $\overline{\mathbf{A}}$ as, respectively, the row, column and overall means for the elements of $\mathbf{A}$, we then we then define the matrix $\mathbf{\Delta}$ so that

$$\mathbf{\Delta}(i,j) = \mathbf{A}(i,j) - \overline{\mathbf{A}_{i,\cdot}} - \overline{\mathbf{A}_{\cdot,j}} + \overline{\mathbf{A}}.$$

It can be shown [61] that these transformations preserve the encoded metric information in D. The principal coordinates are then the eigenvectors of $\mathbf{\Delta}$, the first cooridinate corresponding to the largest eigenvalue and so on.

PCoA is frequently used in examining the sets of pairwise distances generated from the community or genomic metrics discussed in Sections 1.1.4 and 1.1.5, such as UniFrac. An example of PCoA as applied to the Human Microbiome Project data utilizing the Bray-Curtis metric for pairwise distances is given in Figure 1.4 [65].

One framework for the application of PCoA is an adaption of Rao's Dissimilarity Index, Definition 1.1.11, for the generation of the underlying distance matrix. This application of the Dissimilarity Index, utilizing a measure of OTU or species difference given by the metric distance in a phylogenetic tree, was developed into an ordination method given in [89] as *Double Principal Coordinate Analysis*. The 'double' in the title refers to the inclusion of two sorts of data, relative abundance and OTU dissimilarity. In the language we have adapted this is a phylogenetically-aware β-diversity metric packaged together with PCoA for ordination.

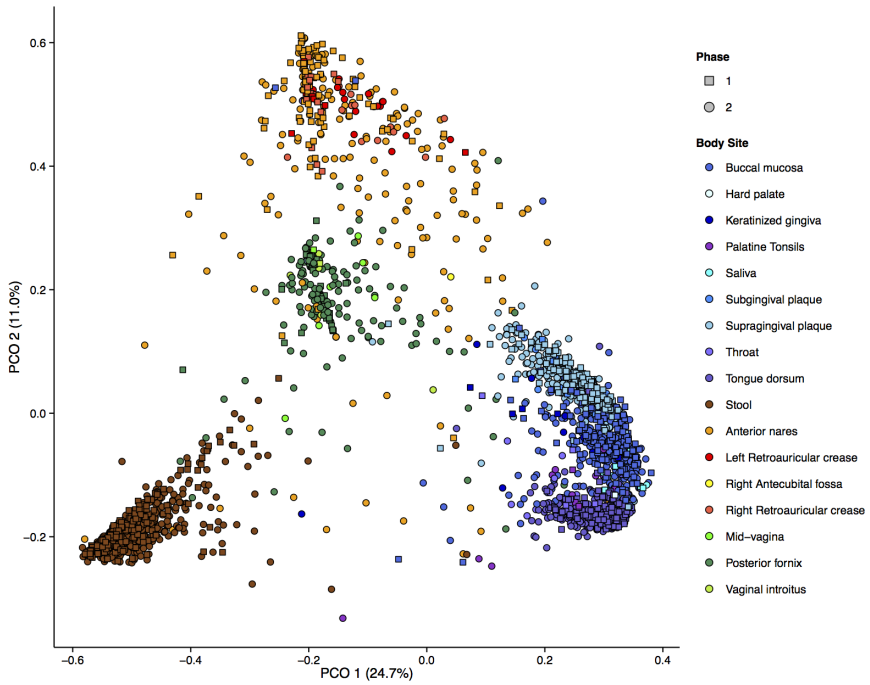


FIGURE 1.4: An example of Principal Coordinate Analysis (PCoA) for the purpose of dataset ordination in microbial ecology. In this plot pairwise distances between microbiome samples from a variety of body locations were generated utilizing Bray-Curtis and then the dataset was projected onto the first two principal coordinates. Such plots are used for exploratory data analysis [65].

Definition 1.1.23 (Double Principal Coordinate Analysis)**.** Let samples A and B assigned to a phylogenetic tree T be given. Let $\mathbf{p}_A$ and $\mathbf{p}_B$. Let d_T be the induced path-length metric in T and let D_h be the Dissimilarity Index defined from pairwise distances given by d_T. Then DPCoA is ordination using PCoA given a distance matrix M generated pairwise distances using D_h.

There is a direct connection between PCoA and PCA in the case where our D matrix is given by the L_2 distances between points. Let $\mathbf{M}$ an $n \times m$ matrix whose columns are $\mathbf{x}_k$ for $1 \leq k \leq n$. Let $\mathbf{D}$ be the $m \times m$ matrix of L_2 distances between columns of M.

Abusing notation, let $\mathbf{D}^2$ denote the component-wise squared matrix of distances and let $\overline{\mathbf{D}}$ denote the mean value of $\mathbf{D}$. Then, applying the parallelogram law, we have that [31]

$$\begin{aligned}
\mathbf{D}^2(i,j) &= \|\mathbf{x}_i - \mathbf{x}_j\|_2^2 \\
&= \langle \mathbf{x}_i, \mathbf{x}_i \rangle + \langle \mathbf{x}_j, \mathbf{x}_j \rangle - 2 \cdot \langle \mathbf{x}_i, \mathbf{x}_j \rangle \\
&= \langle \mathbf{x}_i, \mathbf{x}_i \rangle + \langle \mathbf{x}_j, \mathbf{x}_j \rangle - 2 \cdot \langle \mathbf{x}_i, \mathbf{x}_j \rangle \\
&= \langle \mathbf{x}_i - \overline{\mathbf{D}}, \mathbf{x}_i - \overline{\mathbf{D}} \rangle + \langle \mathbf{x}_j - \overline{\mathbf{D}}, \mathbf{x}_j - \overline{\mathbf{D}} \rangle - 2 \cdot \langle \mathbf{x}_i - \overline{\mathbf{D}}, \mathbf{x}_j - \overline{\mathbf{D}} \rangle
\end{aligned}$$

Letting $\mathbf{G}$ denote the Gram matrix given above, we see that

$$\mathbf{G}(i,j) = \langle \mathbf{x}_i - \overline{\mathbf{D}}, \mathbf{x}_j - \overline{\mathbf{D}} \rangle$$

so that

$$\mathbf{G}(i,j) = -\frac{1}{2} \cdot \left(\mathbf{D}^2(i,j) - \langle \mathbf{x}_i - \overline{\mathbf{D}}, x_i - \overline{\mathbf{D}} \rangle + \langle \mathbf{x}_i - \overline{\mathbf{D}}, \mathbf{x}_i - \overline{\mathbf{D}} \rangle \right).$$

That is

$$\mathbf{G} = -\left(\mathbf{I} - \frac{\mathbf{1}_n}{n} \right) \frac{\mathbf{D}^2}{2} \left(\mathbf{I} - \frac{\mathbf{1}_n}{n} \right).$$

Note that this is precisely the transformation, though in compacted form, utilized in PCoA to produce the Gram matrix whose eigenvectors form the principal coordinates. That is, PCoA is PCA when the distance is L_2.

In the above we have covered some of the mathematical tools and techniques used by biologists in interpreting metrics between biological communities. We next turn to a purely mathematical discussion regarding metrics in probability spaces that will ultimately help in understanding the mathematical foundations to the biological analyses we have discussed.

1.2 Introduction to the Wasserstein Metric

1.2.1 Introduction to the Wasserstein Metric

One of the first moments in the education of a mathematics student in which they see the real power of their field is optimization. In calculus you discover that you don't need to guess endlessly to determine how large a pasture you can build adjacent to your barn with 200 feet of fencing, a simple application of the derivative does the trick.

The utility of optimization, finding minima and maxima for a given function on a given domain, is endless, and, unfortunately, often far more difficult than encountered fencing in that pasture. Consider the following example, an example which will motivate much of our remaining discussion. Imagine that an otherwise level field has been excavated, holes dug and dirt piled randomly. How much work is required to fill the holes back in? Is there a plan which describes how to go about filling in the holes most efficiently? This is an example of an important class of problems [120] known broadly as *optimal transport* and the measure of the optimal amount of work required, however work might be defined, is known broadly as the *transport metric*. This measure of the minimal amount of work required to move all that dirt becomes a very useful measure of distance with a multitude of applications.

The theory of optimal transport began in 1781 with the work of Gaspard Monge [120], a French mathematician who formalized the problem in precisely the same soil-moving context we described above. He called his problem 'Les de'blais et les remblais' or 'Excavation and embankments', and was concerned with the optimal transport of soil for construction of forts and roads.

Pursuit of this theory was continued by the Soviet mathematician and economist Leonid Vital'evich Kantorovich in the early twentieth century [119]. He developed the tools of linear programming to tackle this and other optimization problems arising in economic

models. Much as linear programming was rediscovered in the West, in particular in the work of George Dantzig during World War II [119], the theory behind optimal transport was rediscovered in many guises throughout the years.

The transport metric devised by Kantorovich in the field of economics was also described by Vaserstein [118] (transliterated as Wasserstein) in the field of probability, Mallows [72] in the field of statistics, and Rubner [100] in the field of computer science. As such, it has collected a variety of names; Kantorovich-Rubinstein metric, Wasserstein metric, Mallows distance and the Earth mover's distance. In referring to the transport metric we will primarily use the names Wasserstein metric and Earth mover's distance, as these are the names most common in the literature related to mathematics and computer science. We continue our discussion with a more formal definition of the Wasserstein metric.

1.2.2 Definitions Related to the Wasserstein Metric

We begin our more rigorous discussion of the Wasserstein metric by recalling a few standard definitions and results related to probability.

Let (X, d) denote a complete metric space. Let $\mathcal{B}$ be the Borel σ-algebra of sets from X generated by d. We say the pair $(X, \mathcal{B}$ is a *Polish space* when (X, d) is separable.

We say a probability measure is *locally finite* if for every $x \in X$ there exists a $U \in \mathcal{B}$ of finite measure such that $x \in U$. We say that a probability measure is *inner regular* if for every $U \in \mathcal{B}$ we have that the measure of U is equal to the supremum over measures of compact subsets of U. We say a measure defined on $\mathcal{B}$ is a *Radon measure* if it is inner regular and locally finite. Let $M(X)$ denote the set of all Radon probability measures on X.

We say that a measure has *finite pth moment* for $1 \leq p \leq \infty$ if for some $x_0 \in X$ we have that

$$\int_X d(x, x_0)^p \mathrm{d}\mu(x) < \infty.$$

Note that this definition is independent of the choice of x_0, since for another choice x_1 we have that $d(x, x_1)^p \leq 2^{p-1}(d(x, x_0)^p + d(x_0, x_1)^p)$.

Let $M_p(X)$ denote the set of all Radon measures on (X, d) with finite pth moments.

Now let μ and ν be elements of $M_p(X)$ for some p and define $\Gamma(\mu, \nu)$ to be the set of all measures γ on $X \times X$ such that for all measurable sets $A \in \mathcal{B}$ we have that $\gamma(A, X) = \mu(A)$ and $\gamma(X, A) = \nu(A)$. Notice that $\Gamma(\mu, \nu)$ is nonempty, as we may always take the product measure of μ and ν.

For a fixed measure γ we refer to the related measures μ and ν as defined above as its *marginals*. We will refer to γ as a *coupling* or *flow* between μ and ν. Allowing for greater generality, we can extend this definition of a coupling to the case where (X, μ) and (Y, ν) are a pair of probability spaces and γ is a measure on $X \times Y$ with appropriate marginals.

We are now equipped to define one of our principal objects of study, the p-Wasserstein metric.

Definition 1.2.1. (p-Wasserstein distance) The *p-Wasserstein distance* $W_p(\mu, \nu)$ on $M_p(X)$ is defined as

$$W_p(\mu, \nu) = \left(\inf_{\gamma \in \Gamma(\mu, \nu)} \int_{X \times X} \mathrm{d}(x, y)^p \gamma(dx, dy) \right)^{1/p}.$$

Notice that by applying Minkowski's inequality and the triangle inequality, we see that the integral above is bounded by the sum of the pth moments of μ and ν and is thus finite, by hypothesis.

Our definition is quite general, but our interest lies in the particular case in which X is the vertex set of some graph G endowed with a metric induced by path length (see Proposition 1.3.1). Thus we will ultimately restrict our discussion of the Wasserstein metric to this more finite setting. In our next section we proceed to related some of the standard theory of the Wasserstein metric, in particular alternate formulations.

1.2.3 Summary of Standard Results Related to the Wasserstein Metric

We now discuss some properties of the Wasserstein distance. While it is clear that the Wasserstein distance is well-defined, that does not guarantee the existence of a coupling which realizes the distance. There does happen to exist such a coupling, and so we start with demonstrating that fact. We follow the treatment in [120] to outline the proof. In all that follows X is a Polish space equipped with its Borel σ-algebra.

Recall that a sequence of probability measures $\{\mu_n\}_{n=1}^{\infty}$ *converges weakly* on X to a probability measure μ if $\mathbb{E}_n[f]$ converges to $\mathbb{E}[f]$ for all bounded, continuous functions f. Further, we say a set of probability measures U on X is *tight* if for every $\epsilon > 0$ there exists a compact subset X_ϵ of X such that for all $\mu \in U$ we have that $\mu(X \backslash X_\epsilon) < \epsilon$.

For completeness we state the following useful result related to the compactness of sets of measures.

Theorem 1.2.1 (Prokhorov 1956)**.** *Let X be a Polish space and $\mathcal{P}(X)$ the set of all probability measures on X. Then there is a complete metric on $\mathcal{P}(X)$ equivalent to the topology of weak convergence and $K \subset \mathcal{P}(X)$ has compact closure with respect to this metric if and only if K is tight [120].*

We now prove that there does exist a coupling realizing the Wasserstein distance between probability measures $\mu, \nu \in M_p(X)$.

Theorem 1.2.2 (Existence of a Coupling Realizing the Wasserstein Distance)**.** *For $\mu, \nu \in M_p(X)$ of a Polish space X there exists [120] $\gamma \in \Gamma(\mu, \nu)$ which minimizes*

$$\int_{X \times X} \mathrm{d}(x, y)^p \gamma(dx, dy).$$

Proof. Using the notation of the statement of the proof, we first show that $\Gamma(\mu, \nu)$ is tight. First note that $\{\mu\}$ and $\{\nu\}$ are tight subsets of X, as X is a Polish space. Let $\epsilon > 0$.

Then there exists $U_\mu, U_\nu \subset X$ such that $\mu(X\backslash U_\mu) < \epsilon/2$ and $\nu(X\backslash U_\nu) < \epsilon/2$. Thus for any $\gamma \in \Gamma(\mu,\nu)$ we have that

$$\gamma\Big((X\times X)\backslash(U_\mu \times U_\mu)\Big) \leq \mu(X\backslash U_\mu) + \mu(X\backslash U_\nu) \leq \epsilon.$$

Thus $\Gamma(\mu,\nu)$ is tight. It follows from Prokhorov's theorem that $\Gamma(\mu,\nu)$ has compact closure. In fact, $\Gamma(\mu,\nu)$ is closed and so is compact. To see this, let $\{\gamma_n\}$ converge weakly to γ in $X \times X$. Let A be compact in X and let f_k be a sequence of continuous functions converging to the indicator function of A in the first component. Then, by dominated convergence, we see that

$$\begin{aligned}
\mu(A) &= \lim_{n\to\infty} \gamma_n(A\times X) \\
&= \lim_{n\to\infty} \int_{X\times X} \left(\lim_{k\to\infty} f_k\right) d\gamma_n \\
&= \lim_{k\to\infty} \left(\lim_{n\to\infty} \int_{X\times X} f_k d\gamma_n\right) \\
&= \lim_{k\to\infty} \int_{X\times X} f_k d\gamma \\
&= \int_{X\times X} \left(\lim_{k\to\infty} f_k\right) d\gamma \\
&= \gamma(A\times X)
\end{aligned}$$

As our space is Polish, it is Radon. Hence our measures are inner regular, and so it follows by approximation from within by compact sets that $\gamma(U\times X) = \mu(U)$ for all measurable sets U. A symmetric argument in the second components shows that $\gamma \in \Gamma(\mu,\nu)$ and thus $\Gamma(\mu,\nu)$ is compact.

Now suppose $\{\gamma_n\}$ is a sequence of probability measures such that $\lim_{n\to\infty} \int_{X\times X} \mathrm{d}(x,y)^p \gamma_n = W_p(\mu,\nu)^p$. By passing to some subsequence if necessary, we can assume that $\{\gamma_n\}$ converges to some $\gamma \in \Gamma(\mu,\nu)$. Then by, passing to a sequence of bounded and continuous approximations of the distance function and utilizing

monotone convergence, we have that

$$\begin{aligned} W_p(\mu,\nu)^p &= \lim_{n\to\infty} \int_{X\times X} \mathrm{d}(x,y)^p d\gamma_n \\ &= \lim_{n\to\infty} \int_{X\times X} \left(\lim_{k\to\infty} \min(k, \mathrm{d}(x,y)^p) \right) d\gamma_n \\ &= \lim_{k\to\infty} \lim_{n\to\infty} \int_{X\times X} \min(k, \mathrm{d}(x,y)^p) d\gamma_n \\ &= \int_{X\times X} \left(\lim_{k\to\infty} \min(k, \mathrm{d}(x,y)^p) \right) d\gamma \\ &= \int_{X\times X} \mathrm{d}(x,y)^p d\gamma \end{aligned}$$

Thus γ is a minimizer for the Wasserstein distance. □

We now proceed to show that the Wasserstein distance indeed defines a metric on $M_p(X)$.

Recall that for separable metric spaces X, Y we say that a map $f : X \to Y$ is *Borel* if the inverse image of any open subset of Y is a Borel subset of X. We say that a map which assigns to each $x \in X$ a measure μ_x of Y is Borel if for all Borel subsets $U \subset Y$ the map which sends each x to the quantity $\mu_x(U)$ is Borel.

For a probability measure μ of X and Borel map $f : X \to Y$ we define the *push-forward measure* $f_{\#}\mu$ *of* Y by $f_{\#}\mu(U) = \mu(f^{-1}(U))$ for all measurable subsets U of Y. It is a straightforward exercise to show that this defines a measure on the Borel subsets of Y. Given a product space $X \times Y$ let $\pi^x : X \times Y \to X$ be the canonical projection onto the first component.

To begin, we state without proof a technical lemma on the existence of couplings with fixed marginals in a product of three probability spaces.

Lemma 1.2.1 (Gluing lemma)**.** *Let* $(X, \mu_x), (Y, \mu_y)$ *and* (Z, μ_z) *be Polish probability spaces. If* $\gamma_{x,y}$ *is a coupling of* (μ_x, μ_y) *and* $\gamma_{y,z}$ *is a coupling of* (μ_y, μ_z) *then one can con-*

struct a measure $\gamma_{x,y,z}$ *on* $X \times Y \times Z$ *such that* $\pi^{X\times Y}_{\#}\gamma_{x,y,z} = \gamma_{x\times y}$ *and* $\pi^{Y\times Z}_{\#}\gamma_{x,y,z} = \gamma_{y,z}$ *[120].*

Clearly gluing lemma is an appropriate title, as we 'glue' the couplings $\gamma_{x,y}$ and $\gamma_{y,z}$ together along their common marginal μ_y. We can leverage this fact to show that the Wasserstein distance is indeed a metric.

Theorem 1.2.3 (Wasserstein Metric)**.** *For all* $p \geq 1$ *the* p*-Wasserstein distance given in Definition 1.2.1 satisfies the mathematical defintion of a metric for the finite* p*th moment measures of a Polish space* X *[120].*

Proof. Let X be a Polish space, let $p \geq 1$ be fixed and let μ, ν, ω be probability measures on X with finite pth moments. By symmetry of the integral defining W_p it is clear that $W_p(\mu, \nu) = W_p(\nu, \mu)$.

Now suppose $W_p(\mu, \nu) = 0$. Then there exists a coupling γ for μ and ν which is concentrated on the diagonal in $X \times X$, that is, it is assigns to any subset not lying on the line $y = x$ measure zero. Letting π^1, π^2 denote the projections into the first and second components of $X \times X$, respectively, we see that $\pi^1_{\#}\gamma = \pi^2_{\#}\gamma$ since $y = x$ on the support of γ and thus $\mu = \nu$.

Now suppose $\gamma_{\mu,\nu}$ is an optimal coupling for μ and ν and $\gamma_{\nu,\omega}$ is an optimal coupling for ν and ω. By the gluing lemma there exists a measure $\gamma_{\mu,\nu,\omega}$ on $X \times X \times X$ such that the push-forward measure into the first two components is $\gamma_{\mu,\nu}$ and the push-forward measure into the second two components is $\gamma_{\nu,\omega}$. It follows that the push-forward measure of $\gamma_{\mu,\nu,\omega}$ by projection into the first component is μ and the push-forward measure of $\gamma_{\mu,\nu,\omega}$ by projection into the third component is ω. So we have that

$$W_p(\mu, \omega) \leq \left(\int_{X\times X} \mathrm{d}(x, z)^p \gamma_{\mu,\nu,\omega}(dx, dz) \right)^{1/p}$$

$$
\begin{aligned}
&= \left(\int_{X\times X\times X} \mathrm{d}(x,z)^p \gamma_{\mu,\nu,\omega}(dx,dy,dz) \right)^{1/p} \\
&\leq \left(\int_{X\times X\times X} (\mathrm{d}(x,y)+\mathrm{d}(y,z))^p \gamma_{\mu,\nu,\omega}(dx,dy,dz) \right)^{1/p} \\
&\leq \left(\int_{X\times X\times X} \mathrm{d}(x,y)^p \gamma_{\mu,\nu,\omega}(dx,dy,dz) \right)^{1/p} \dots \\
&\dots + \left(\int_{X\times X\times X} \mathrm{d}(y,z)^p \gamma_{\mu,\nu,\omega}(dx,dy,dz) \right)^{1/p} \\
&= \left(\int_{X\times X} \mathrm{d}(x,y)^p \gamma_{\mu,\nu}(dx,dy) \right)^{1/p} + \left(\int_{X\times X} \mathrm{d}(y,z)^p \gamma_{\nu,\omega}(dy,dz) \right)^{1/p} \\
&= W_p(\mu,\nu) + W_p(\nu,\omega)
\end{aligned}
$$

The third and fourth equalities above are justified by the triangle inequality in X and $L^p(\gamma_{\mu,\nu,\omega})$. It follows that W_p satisfies the triangle inequality and thus is a metric. □

We now proceed to state an important duality formula for the Wasserstein metric. Our principal interest is in the W_p metric for $p = 1$ and $p = 2$ and so we will restrict ourselves to those cases for the remainder of our discussion.

Recall that a function is *Lipschitz* if the difference quotient is uniformly bounded. More precisely, for a function $f : X \to \mathbb{R}$ let

$$Lip(f) = \sup \left\{ \frac{|f(x) - f(y)|}{d(x,y)} \,\middle|\, x, y \in X, x \neq y \right\}.$$

When $Lip(f)$ is finite, f is Lipschitz and $Lip(f)$ is its *Lipschitz constant*.

Let $Lip_1(X)$ denote the set of all $f : X \to \mathbb{R}$ such that $Lip(f) \leq 1$. Note that for $f \in Lip_1(X)$, $|f(x)| \leq |f(x_0)| + d(x, x_0)$. Hence for every $f \in Lip_1(X)$, f is integrable with respect to every measure in $M_1(X)$. We shall see that in seeking the value of the W_1 between measures μ and ν we may, instead of asking for a coupling which minimizes, ask for a Lip_1 function which maximizes. We state, without proof, this famous theorem.

Theorem 1.2.4 (Kantorovich, Rubinstein 1958)**.** *Let (X, d) be a Polish space and let $\mu, \nu \in M_1(X)$. Then [120]*

$$W_1(\mu, \nu) = \sup \left\{ \int_X f \mathrm{d}\mu - \int_X f \mathrm{d}\nu, \middle| \ f \in Lip_1(X) \right\}$$

To summarize our discussion so far, the p-Wasserstein distance over a Polish space X between a pair of measures is realized by some optimal coupling of those measures. The W_p distance satisfies the definition of a metric on the Borel measures of X with finite pth moments. In the case of $p = 1$, there is a dual formulation of the metric in terms of Lipschitz-1 functions. We next turn to a comparison of the metric W_p to other senses of distance on probability spaces, with applications to both mathematics and biology.

1.2.4 Comparison of the Wasserstein Metric to Other Metrics in Probability Spaces

We continue our discussion of the Wasserstein metric by discussing other measures of difference in probability spaces and their relationship to the Wasserstein metric. The metrics below were chosen for discussion due to their theoretical significance, relationship to the Wasserstein metric [37] or use in applications. We follow the treatment in [37] in the following discussion.

We begin with the definition of a few metrics, and less formal senses of distance, on probability measures.

Definition 1.2.2 (Discrepancy)**.** Let X be a metric space. We define the *discrepancy metric* d_D on the probability measures of X by

$$\mathrm{d}_D(\mu, \nu) = \sup_{\text{closed balls C}} |\mu(C) - \nu(C)|.$$

Recall that we say a measure μ *dominates* a measure ν if, for any measurable set A, we have that $\mu(A) = 0$ implies that $\nu(A) = 0$.

Definition 1.2.3 (Relative Entropy (Kullback-Leibler Divergence))**.** Let Ω be any measurable space. Let f, g be densities of μ, ν, respectively, with respect to a dominating measure ω and let S_μ be the support of μ in Ω. We define the *relative entropy* or *Kullback-Leibler divergence* d_{KL} of μ and ν as

$$\mathrm{d}_{KL}(\mu, \nu) = \int_{S_\mu} f \log(f/g) d\omega.$$

As a consequence of the Radon-Nikodym theorem, this definition is independent of the dominating measure. Further, note that the relative entropy is not a metric. It is neither symmetric nor does it satisfy the triangle inequality. It is, however, non-negative and zero precisely when $\mu = \nu$. It was defined by Kullback and Leibler in 1951 as a generalization of the Shannon's information theoretic definition of entropy. The symmetric version of the Kullback-Leibler divergence is the *Jensen-Shannon divergence*.

Definition 1.2.4 (Jensen-Shannon Divergence)**.** Let Ω be any measurable space. Let f, g be densities of μ, ν, respectively, with respect to a dominating measure ω. We define the *Jensen-Shannon Divergence* d_{JS} of μ and ν as

$$\mathrm{d}_{JS}(\mu, \nu) = \frac{1}{2} \cdot \mathrm{d}_{KL}(\mu, \nu) + \frac{1}{2} \cdot \mathrm{d}_{KL}(\nu, \mu).$$

Definition 1.2.5 (Hellinger)**.** Let Ω be any measurable space. For measures μ and ν on Ω, having densities f and g, respectively, with respect to some dominating measure λ we define the *Hellinger distance* d_H to be

$$\mathrm{d}_H(\mu, \nu) = \left[\int_\Omega (\sqrt{f} - \sqrt{g})^2 d\lambda \right]^{1/2}.$$

In the case that Ω is countable we can express the Hellinger distance as $\mathrm{d}_H(\mu, \nu) = \left[\sum_{\omega \in \Omega} (\sqrt{\mu(\omega)} - \sqrt{\nu(\omega)})^2 \right]^{1/2}$.

Definition 1.2.6 (χ^2-distance)**.** Let Ω be any measurable space. For measures μ and ν on Ω, having densities f and g, respectively, with respect to some dominating measure λ,

let the support of each be S_μ and S_ν. We then define the χ^2-*distance* d_{χ^2} to be

$$\mathrm{d}_{\chi^2}(\mu,\nu) = \int_{S_\mu \cup S_\nu} \frac{(f-g)^2}{g} d\lambda.$$

Note that the χ^2-distance is not a metric, nor even symmetric in the arguments.

Definition 1.2.7 (Prokhorov)**.** Let X be any metric space. For any Borel set U and $\epsilon > 0$, let $U_\epsilon = \{x | \inf_{y \in U} \mathrm{d}(x,y) \leq \epsilon\}$. We then define the *Prokhorov metric* d_P by

$$\mathrm{d}_P(\mu,\nu) = \inf\{\epsilon > 0 | \mu(U) \leq \nu(U_\epsilon) + \epsilon \text{ for all Borel } U\}.$$

The Prokhorov metric does satisfy the definition of a metric [47] and is of theoretical importance as it metrizes weak convergence of measures on a separable metric space.

Definition 1.2.8 (Total Variational Distance)**.** Let X be any measurable space. We then define the *Total Variational Distance* d_{TV} by

$$\mathrm{d}_{TV}(\mu,\nu) = \sup_{A \subset X} |\mu(A) - \nu(A)|.$$

Letting $D = \{(x,y) \in X \times X | x \neq y\}$, an alternate characterization of the total variational distance in terms of couplings is $\mathrm{d}_{TV}(\mu,\nu) = \inf\{\gamma(D) | \gamma \in \Gamma(\mu,\nu)\}$.

Having defined several metrics or other measures of distance related to probability measures, we now state and recount proofs of several relationships between these and the 1-Wasserstein metric W_1. We begin with an important relationship between the Wasserstein and Prokhorov metrics on the probability measures of a bounded metric spaces.

Theorem 1.2.5 (Prokhorov and Wasserstein)**.** *For a bounded metric space X, the Prokhorov metric d_P and the 1-Wasserstein metric W_1 satisfy the following relationship*

$$(\mathrm{d}_P)^2 \leq w_1 \leq (\mathrm{diam}(X) + 1) \cdot \mathrm{d}_P$$

for $\mathrm{diam}(X) = \sup\{\mathrm{d}(x,y) | x,y \in X\}$ *[37].*

Proof. Let $\epsilon > 0$. Let $D_\epsilon = \{(x,y) \in X \times X | d(x,y) \leq \epsilon\}$. Then for any coupling $\gamma \in \Gamma(\mu,\nu)$ we have

$$\begin{aligned}\int_{X\times X} \mathrm{d}(x,y)\gamma(dx,dy) &= \int_{D_\epsilon} \mathrm{d}(x,y)\gamma(dx,dy) + \int_{(X\times X)\backslash D_\epsilon} \mathrm{d}(x,y)\gamma(dx,dy) \\ &\leq \epsilon \cdot \gamma(D_\epsilon) + \mathrm{diam}(X) \cdot \gamma\Big((X\times X)\backslash D_\epsilon\Big) \\ &= \epsilon + (\mathrm{diam}(X) - \epsilon) \cdot \gamma\Big((X\times X)\backslash D_\epsilon\Big)\end{aligned}$$

By Strassen's theorem [111] we see that if $\mathrm{d}_P(\mu,\nu) \leq \epsilon$ then there exists a $\gamma \in \Gamma(\mu,\nu)$ such that $\gamma((X\times X)\backslash D_\epsilon) \leq \epsilon$.

Therefore we have that

$$\int_{X\times X} \mathrm{d}(x,y)\gamma(dx,dy) \leq \epsilon + (\mathrm{diam}(X) - \epsilon)\cdot\epsilon \leq (\mathrm{diam}(X)+1)\cdot\epsilon.$$

Thus, by taking the infimum over all couplings and setting $\epsilon = \mathrm{d}_P(\mu,\nu)$, we see that $W_1 \leq (\mathrm{diam}(X)+1)\mathrm{d}_P$.

To prove the other bound, we set $\epsilon = \sqrt{W_1(\mu,\nu)}$ and use Chebyshev's inequality's to deduce that

$$\int_{D_\epsilon} \gamma(dx,dy) \leq \frac{1}{\sqrt{W_1(\mu,\nu)}} \int_{X\times X} \mathrm{d}(x,y)\gamma(dx,dy) \leq \sqrt{W_1(\mu,\nu)}.$$

Finally, by using Strassen's theorem in the other direction and recalling the notation of Definition 1.2.7, we note that $\int_{D_\epsilon} \gamma(dx,dy) \leq \epsilon$ implies that for all Borel sets B we have that $\mu(B) \leq \nu(B) + \epsilon$ so that $\mathrm{d}_P \leq \sqrt{W_1(\mu,\nu)}$, as required. □

As we have already stated, the d_P metrizes weak convergence in separable metric spaces. We now see that W_1 generates the same topology, and so we garner the following result.

Corollary 1.2.1 (Wasserstein Metrizes Weak Convergence of Measures on Bounded Metric Spaces)**.** *Let X be a bounded metric space. Then the 1-Wasserstein metric W_1 metrizes the weak topology on the probability measures of X [37].*

We now state and prove a relationship between W_1 and the discrepancy metric d_D over the probability measures of a finite metric space [37].

Theorem 1.2.6 (Wasserstein and Discrepancy)**.** *Let X be a finite metric space. Let $\mathrm{d}_{min} = \min_{x \neq y} \mathrm{d}(x, y)$. Then [37] we have that*

$$\mathrm{d}_{min} \cdot \mathrm{d}_D \leq W_1.$$

Proof. Recalling Theorem 1.2.4, we define for a closed ball B in X the function

$$h(x) = \begin{cases} \mathrm{d}_{min} & x \in B \\ 0 & \text{else} \end{cases}$$

Clearly $Lip(h) \leq 1$. We then see that for any pair of probability measures μ and ν that

$$\begin{aligned} \mathrm{d}_{min} \cdot |\mu(B) - \nu(B)| &= |\int_X h d\mu - \int_X h d\nu| \\ &\leq W_1(\mu, \nu) \end{aligned}$$

Taking that B which maximizes the left hand side yields the desired result. □

In the last of our analytic comparisons between metrics on probability measures, we compare the 1-Wasserstein metric with the Total Variation metric d_{TV} on bounded or finite metric spaces.

Theorem 1.2.7 (Wasserstein and Total Variation)**.** *Let X be a bounded metric space. Then [37]*

$$W_1 \leq \mathrm{diam}(X) \cdot \mathrm{d}_{TV}.$$

Now supposed X is a finite metric space. Then setting $\mathrm{d}_{min} = \min_{x \neq y} \mathrm{d}(x, y)$ we have that

$$\mathrm{d}_{min} \cdot \mathrm{d}_{TV} \leq W_1.$$

Proof. Recalling the coupling characterization of d_{TV} we see that for $D = \{(x,y)|x \neq y\}$ and any coupling $\gamma \in \Gamma(\mu,\nu)$ we have that

$$\begin{aligned}\int_{X\times X} \mathrm{d}(x,y)d\gamma &= \mathrm{diam}(X) \cdot \int_{X\times X} \mathrm{d}(x,y)d\gamma \\ &\leq \mathrm{diam}(X) \cdot \gamma(D)\end{aligned}$$

By taking the infimum over all couplings we yield our first result.

Now suppose X is finite. We then get that

$$\begin{aligned}\int_{X\times X} \mathrm{d}(x,y)d\gamma &= \mathrm{diam}(X) \cdot \int_{X\times X} \mathrm{d}(x,y)d\gamma \\ &\geq \mathrm{d}_{min} \cdot \gamma(D)\end{aligned}$$

and thus our second result. □

We now have a sense of the 1-Wasserstein metric's relationship to other notions of distance between probability measures in metric spaces. We have also shown that for bounded metric spaces the 1-Wasserstein metric metrizes the topology of weak convergence of measures. We now turn to discussing the applications of the Wasserstein metric in mathematics, science and engineering.

1.2.5 Survey of Applications of the Wasserstein Metric

The Wasserstein metric is a natural and powerful sense of distance between probability measures which has found applications in a variety of fields. It has been independently discovered in various branches of both pure and applied mathematics [119]. Here we will discuss a few examples of the application and formulation of the Wasserstein metric used in the study optimization, statistics and computer science.

Optimization

As mentioned in our introduction to this subject, Gaspard Monge first formulated the optimal transport problem in 1781 [120] in the context of the following very applied problem. Suppose you wish to construct an earthen structure at a fixed location by excavating soil from a predetermined set of locations. What is the optimal method of going about moving the dirt?

Formalizing the problem slightly, let R be a bounded region in $\mathbb{R}^2$. Let $c(r_1, r_2) : R^2 \to \mathbb{R}_{\geq 0}$ be the cost of transporting a unit of material $r_1 \in R$ to $r_2 \in R$. Let $S(r) : R \to \mathbb{R}_{\geq 0}$ be the amount soil required to build our structure at location $r \in R$ and let $E(r) : R \to \mathbb{R}_{\geq 0}$ be the amount of soil to be excavated at location $r \in R$.

Since we would rather not dig up any soil that is not going to be put to use in our structure, we will require that $\int_R S(r)\, dr = \int_R E(r)\, dr$.

Now let a *transport plan* T between E and S be a function $T : R^2 \to \mathbb{R}_{\geq 0}$ such that for each $r_0 \in R$ we have that $\int_R T(r_0, r)\, dr = E(r_0)$ and $\int_R T(r, r_0)\, dr = S(r_0)$. This is merely a description of the ultimate destination of our excavated soil for each point $r \in R$, our first condition, and an assurance that each location $r \in R$ received precisely enough material for construction, our second condition.

To determine how much it might cost to build our structure, we do the following: pick a transport plan, look at every pair of points in R, see how much dirt was moved from here to there for that transport plan, multiply that by the cost of moving dirt from here to there and then total those costs. Invoking a little calculus, the cost of building our bit of earthworks is then

$$C_{E,S} = \int_{R^2} c(r_1, r_2) \cdot T(r_1, r_2)\, dr_1 dr_2$$

for a given transport plan T. Letting $T_{E,S}$ be the set of all transport plans between E and S we see that finding the minimal cost corresponds to the $T \in T_{E,S}$ which minimizes the

integral.

We now simplify and abstract the problem. First, we choose units such that $\int_R S(r)\,dr = \int_R E(r)\,dr = 1$. Next, we take the cost function c of moving the soil to be proportional to the work, in the physics sense of force times distance. For a material of uniform density, this means the cost is proportional to the distance traveled. Simplifying, we take that constant of proportionality to be 1.

Notice that after our simplification, our functions E and S are now probability measures on R and that $T_{E,S}$ is the set of all couplings of those measures. That is, the minimal cost of construction under these assumptions is

$$\inf_{T \in T_{E,S}} \int_{R^2} \mathrm{d}(r_1, r_2)\, T(dr_1, dr_2)$$

or the 1-Wasserstein distance between E and S.

Monge attacked this problem using the tools of descriptive geometry [120], while some 150 years later Kantorovich developed *linear programming*, that is optimization of linear objective functions subject to linear constraints, to address the problem in the context of the mathematical theory of economics. The problem presented above in terms of soil and structures can just as easily be recast in terms of goods and consumers. Before continuing, we define a few terms related to finite metric spaces that will be useful for this problem and, indeed, the rest of our discussion.

Definition 1.2.9 (Distance Matrix)**.** Let (X, d) be a finite metric space and let $n = |X|$. The *distance matrix* $\mathbf{D}$ for X with respect to d is the $n \times n$ matrix, indexed by the elements of X, such that $\mathbf{D}(i, j) = \mathrm{d}(i, j)$ for all $i, j \in X$.

Definition 1.2.10 (Marginals of a Matrix)**.** Let $\mathbf{M}$ be an n by m matrix. Let $\mathbf{1}_n$ and $\mathbf{1}_m$ be the column vectors of length n and m, respectively, whose entries are identically 1. We say the pair of vectors μ and ν are the *marginals* of M if $\mathbf{1}_n^t \cdot \mathbf{M} = \mu$ and $\mathbf{M} \cdot \mathbf{1}_m = \nu$. We may also say μ is the *column sum* of M and ν is the *row sum* of M.

Returning to our economic problem, we pass to a finite metric space (X, d), say $|X| = n$, let $\mu(x)$ to be the number of goods ready for delivery at each $x \in X$, and let $\nu(x)$ denote the number of goods required at each location $x \in X$. Suppose that k goods are required, and that this quantity is equal to the number available. Let $\mathbf{D}$ be the distance matrix for X. Let $T_{\mu,\nu}$ be the set of all matrices $\mathbf{T}$ which have μ and ν as marginals. If we again assume that the cost of transporting good is proportional to the distance traveled, and, by normalizing if necessary, take the constant of propotionality to be 1, we see that the minimum cost of associated with this economic allocation problem is

$$\min_{\mathbf{T} \in T_{\mu,\nu}} \sum_i \sum_j \mathbf{D}(i,j) \cdot \mathbf{T}(i,j).$$

Normalizing each of μ and ν by k, we see that this again corresponds to the Wasserstein metric. Further, it is now a linear programming problem, where multiplication against the distance matrix forms our linear objective function and satisfying the marginals form our set of linear constraints. It was in this context that Kantorovich developed the theory of both linear programming and optimal transport [119].

Dynamical Systems and Partial Differential Equations

In 1970, Dobrushin [25] coined the term 'Vasershtein metric' in a paper regarding the existence and uniqueness of random fields. This was in reference to Vasershtein's 1969 paper [118, 101] which used the metric between distributions P and Q given by $\inf[\mathbb{E}\mathrm{d}(X,Y)]$ where this infimum is taken with respect to all random variables X, Y with distributions P, Q in studying dynamical systems.

Statistics

In a statistical context, studying the asymptotic distribution of a sequence of jointly distributed random variables, Mallows [72] constructed the metric

$$\rho^2(F,G) = \int_0^1 (f(w) - g(w))^2 dw$$

for F, G distributions of finite variance and zero mean, and f the essentially unique monotone function such that $f(F(x)) = x$ almost everywhere with respect to F.

Mallows demonstrated that ρ metrizes convergence of distributions in the Lévy metric, of which the Prokhorov metric from Definition 1.2.7 is a generalization, and that ρ has as an equivalent formulation as

$$\rho^2(F_1,F_2) = \min_{\lambda \in \Lambda(F_1,F_2)} \int (x-y)^2 d\lambda(x,y)$$

where $\Lambda(F_l, F_2)$ is the set of bivariate distribution functions on $\mathbb{R} \times \mathbb{R}$ with marginals equal to F_1 and F_2, respectively. That is, the 2-Wasserstein distance between the distributions.

Computer Science

In 1999 Rubner et al. [100] defined a distance between distributions they called the *Earth mover's distance* for the purpose of content based computer image retrieval. Here the problem is to identify a given image by comparison against a set of previously identified reference images. Clearly some sort of metric or other means of comparison on the set of images is necessary.

As a way of constructing such a metric, they first define a *histogram* as a mapping from a set of d-dimensional integer vectors in i into $\mathbb{R}_{\geq 0}$. Here the vectors i are called *bins* and represent the a range of values in the spectrum of some image feature, such as color content or intensity, and the value of h_i is the number of pixels in an image which fall into the range defined by i.

They next define a *ground distance* as a measure of dissimilarity, metric of otherwise, between the bins in a histogram. Using this ground distance they define the *signature* $\{s_j = (m_j, w_j)\}$ of an image as a set of clusters of image features, or bins, where each cluster s_j is represented by some measure of central tendency m_j for the cluster and then weighted by the number or fraction w_j of pixels belonging to that cluster. Here the number of clusters in a signature may vary, depending on the complexity of an image.

Given signatures $P = \{(p_1, w_{p_1}), \ldots, (p_m, w_{p_m})\}$ and $Q = \{(q_1, w_{q_1}), \ldots, (q_m, w_{q_n})\}$ of m and n clusters, respectively, and an $m \times n$ matrix $\mathbf{D}$ such that $\mathbf{D}(i,j)$ is the distance between clusters p_i and q_j they solve the following transport problem. Determine the $m \times n$ matrix $\mathbf{F}$ which minimizes

$$WORK(P, Q, F) = \sum_{i=1}^{m} \sum_{j=1}^{n} \mathbf{D}(i,j)\mathbf{F}(i,j)$$

subject to the constraints

$$\begin{aligned}
\mathbf{F}(i,j) &\geq 0, \quad \forall i,j \\
\sum_{j=1}^{n} \mathbf{F}(i,j) &\leq w_{p_i}, \quad \forall i \\
\sum_{i=1}^{n} \mathbf{F}(i,j) &\leq w_{q_j}, \quad \forall j \\
\sum_{i=1}^{m} \sum_{j=1}^{n} \mathbf{F}(i,j) &= \min(\textstyle\sum_{i=1}^{m} w_{p_i}, \sum_{j=1}^{m} w_{q_j}).
\end{aligned}$$

Having solved for an optimal matrix $\mathbf{F}$, they finally define the Earth mover's distance EMD by

$$EMD(P, Q) = \frac{\sum_{i=1}^{m} \sum_{j=1}^{n} \mathbf{D}(i,j)\mathbf{F}(i,j)}{\sum_{i=1}^{m} \sum_{j=1}^{n} \mathbf{F}(i,j)}.$$

The normalization factor here helps to avoid skewing toward signatures with fewer clusters. In the case that P and Q given by proportions of pixels of a common set of clusters and the ground distance is a mathematical metric, we see that EMD is the 1-Wasserstein metric over a finite metric space.

They demonstrate that EMD is an effective metric for image recognition which is insensitive to noise and which allows for partial matches between images, that is, matches between regions of images.

The Wasserstein metric, under the name EMD or otherwise, has since appeared in a variety of computer science applications, such as image tracking [135], machine learning [62, 4] and text analysis [59].

In this section we have discussed the ways in which the Wasserstein metric has been applied in a variety of contexts, reaching from its first formulation more than two hundred years ago to its recent applications in computer science. In particular, we have discussed the use of the Wasserstein metric in image analysis as a valuable way to lift notions of distance between small components of a set, in this case pixels or regions of pixels, to comparisons of the large scale structure of sets, entire images. This provides valuable motivation in some of the work which follows. We note that we have yet to determine how to compute the value of the metric in any setting, and so this is the subject of our next discussion.

1.2.6 Survey of Computational Methods for the Wasserstein Metric

A variety of numerical methods have be devised over the years with which to compute or approximate the Wasserstein metric in various settings. In particular, its utilization in computer science as the Earth mover's distance have lead to a number of novel solutions and approximations. In their initial formulation of the Earth mover's distance, Rubner et al [100] utilized the classic Transportation Simplex algorithm as a solution method. The Transportation Simplex algorithm is an adaption of Dantzig's original Simplex algorithm for linear programming [45] to the solution of the optimal transport problem. We begin with a brief discussion of linear programming and the Simplex algorithm.

The traditional form of a linear programming problem is, for fixed vectors $\mathbf{b}, \mathbf{c} \in \mathbb{R}^n$ and matrix $\mathbf{M} \in \mathbb{R}^{n \times n}$, to find $x \in \mathbb{R}^n$ which maximizes the *objective function* $Z = \mathbf{c}^t \cdot \mathbf{x}$

subject to the constraints that $\mathbf{Ax} \leq \mathbf{b}$ and that the entries of $\mathbf{x}$ are nonnegative. It can be shown [81] that problems in which we seek to minimize Z or satisfy constraints which are equalities can be solved by the Simplex method either through merely negating our objective function or introducing additional slack variables to enforce equality.

We call the set of potential solutions to $\mathbf{Ax} \leq \mathbf{b}$ for $\mathbf{x}$ in the nonnegative orthant of $\mathbb{R}^n$ the *feasible region*. It can be shown [45] that the feasible region is a potentially unbounded convex polytope in $\mathbb{R}^n$. We state the following very useful theorem related to the existence and location of extreme values for a linear programming problem.

Theorem 1.2.8. *Given a linear programming problem of the form* $\max Z = \boldsymbol{c}^t \cdot \boldsymbol{x}$ *subject to the constraints that* $\boldsymbol{Ax} \leq \boldsymbol{b}$ *defined above, if an extreme value for* Z *occurs in the feasible region, then it occur on one or more of the vertices of the convex polytope defined by the feasible region [45].*

We call these vertices *basic feasible solutions*.

We now demonstrate how to cast the Wasserstein metric as a linear programming problem, known generally as the Transportation or Network Simplex problem. Given a finite metric space (X, d), with $|X| = n$, we can view the distance matrix $\mathbf{D}$ as

$$\mathbf{D} = \begin{bmatrix} d(x_1, x_1) & d(x_1, x_2) & \dots & d(x_1, x_n) \\ d(x_2, x_1) & d(x_2, x_2) & \dots & d(x_2, x_1) \\ \vdots & \vdots & \ddots & \vdots \\ d(x_n, x_1) & d(x_n, x_2) & \dots & d(x_n, x_n) \end{bmatrix}$$

as the vector

$$\mathbf{c}_D = (-d(x_1, x_1), -d(x_1, x_2), \dots, -d(x_1, x_n), -d(x_2, x_n), \dots \\ \dots, -d(x_{n-1}, x_n), -d(x_n, x_1), \dots, -d(x_n, x_n))$$

in $\mathbb{R}^{n^2}$.

Define the $2n \times n^2$ matrix $\mathbf{A}$

$$\mathbf{A} = \begin{bmatrix} 1 & \dots & 1 & 0 & \dots & \dots & \dots & \dots & 0 \\ 0 & \dots & 0 & 1 & \dots & 1 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & \dots & \dots & \dots & \dots & 0 & 1 & \dots & 1 \\ 1 & \dots & 1 & 0 & \dots & \dots & \dots & \dots & 0 \\ 0 & \dots & 0 & 1 & \dots & 1 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & \dots & \dots & \dots & \dots & 0 & 1 & \dots & 1 \end{bmatrix}$$

where each row contains n consecutive entries of 1 and for row i, the first nonzero entry appears in the $n \cdot (i-1) + 1$ column.

Given measures μ, ν on X we can define the vector

$$\mathbf{b} = (\mu(x_1), \mu(x_2), \dots, \mu(x_n), \nu(x_1), \nu(x_2), \dots, \nu(x_n))$$

in $\mathbb{R}^{2n}$.

For these definitions of $\mathbf{A}, \mathbf{b}, \mathbf{c}_D$ the Wasserstein metric is a linear programming problem. The Wasserstein metric between μ and ν is given by the vector x in $\mathbb{R}^{n^2}$ which minimizes $c_D^t \cdot x$.

Simplex algorithms solve a linear programming problem iteratively. It can be shown [45] that if a vertex of the feasible region is not a maximizer for the objective function Z, then it is adjacent to an edge on which Z is strictly increasing. Thus, intuitively, Simplex algorithms select a vertex of the polytope and at each iteration traverse an edge which increases Z. This choice of an edge is known as a *pivot rule* and generally involves solving a system of linear equations to determine a new choice of edge [81]. If this edge is infinite in length, then Z is unbounded and thus the linear programming problem has no solution.

Otherwise, this edge terminates in another vertex. Either this vertex maximizes Z, and thus we terminate, or we select another edge and continue.

We pause from our discussion on solutions to the optimal transportation problem to define a few useful terms related to the study of algorithms.

Definition 1.2.11 (Big O Notation)**.** For functions $f, g : \mathbb{R} \to \mathbb{R}$, such that $g(x)$ is eventually strictly positive, we say $f(x)$ has order of growth $O(g(x))$ if there exists an $N \in \mathbb{R}_{>0}$ such that $f(x) \leq N \cdot g(x)$ for all sufficiently large x.

Definition 1.2.12 (Algorithmic Time Complexity)**.** Given an algorithm $\mathcal{A}$, let n be a measure of the size of the input to $\mathcal{A}$. Typical examples of such a measure of size are the number of unknown variables to be determined or the number values to be processed. We say that the *time complexity* of $\mathcal{A}$ is $g(n)$ or $O(g(n))$, for a function $g(n) : \mathbb{N} \to \mathbb{R}$, if, for an input of size n, the number of elementary mathematical operations required for the algorithm to terminate has order of growth $O(g(n))$.

The time complexity of an algorithm is frequently not uniform accross all inputs of a given size. In such cases, we may refer to the *worst-case time complexity*, *average time complexity* or *best-case time complexity*, that is, the rate of growth of the maximum, average or minimum number of elementary operations required for the algorithm to terminate.

Although Dantzig's original Simplex algorithm has enjoyed a tremendous amount of success in practical application due to an observed number of iterations which is linear in the size of the linear programming problem, the Simplex algorithm has demonstrated exponential worst-case time complexity [55, 81]. The Transportation Simplex problem described above has additional structure to the objective function and constraints which allows for more efficient solution techniques.

The Transportation Simplex method [1] interprets a vertex of the feasible region as a spanning tree for the underlying graph. The algorithm, through a variety of pivot rules,

selects an edge to add to the spanning tree. This necessarily introduces a cycle to the spanning tree. Flows are reduced around this cycle until the the algorithm returns to a new spanning tree, and thus a new vertex of the feasible region. This continues until no choice of new edge reduces our flow.

Let $G = (V, E)$ be a graph with n vertices, m edges and where C is a bound for the length of each edge. The best known Simplex algorithm for optimal transport over a graph [113] has time complexity $O((nm\log(n))\min(\log(nC), m\log(n))$ for a graph with. That algorithm improves, through the use of more efficient data structures, the algorithm presented in [87], which is in turn a variation on the classic *Network Simplex* method.

There are several methods for the selection of a starting point from which to iterate for the Transportation Simplex algorithm [45]. The most simplistic is that of the *Northwest corner rule*. Letting $\mathbf{M}$ be our initial basic feasible solution, we build $\mathbf{M}$ iteratively. We increase $\mathbf{M}(1,1)$ until we meet one of the two marginal contraints and then stop. In the case that we have satisfied the marginal constraint for the row, we proceed to the $\mathbf{M}(2,1)$ element and repeat. Else we proceed to the $\mathbf{M}(1,2)$ element and repeat. We iterate this process until we have satisfied all of the marginal contraints and generated a basic feasible solution.

While this method is easy to implement, it is ignorant of the costs involved in transport and varies given a reordering of the variables. The *minimum cost method* is an alternate approach to generating a basic feasible solution which improves on this method by considering cost in the selection of elements of the initial flow to saturate. In this method we begin by selecting the minimum cost between between distinct elements for the transport problem, and sets the flow between these elements to a value which satisfies one or other of the constraints defined by the marginals. This process iterates, satisfying all low cost pairs before advancing to higher cost pairs.

In the above we have discussed methods for producing the exact minimizing value to

the problem of optimal transport. We turn now to a discussion of a set of methods for approximating the optimal transport metric via *regularization*. Regularization is the inclusion of additional constraints on an objective function so as to make the minimization problem more computationally tractable [107], and is common in optimization. A useful form of regularization for optimal transport is *entropic-regularization*. We now define the *entropy* of a probability measure in a finite setting.

Definition 1.2.13 (Entropy)**.** Let μ be a measure on a finite set X such that $\mu(x) > 0$ for all $i \in X$. Then the *entropy of* μ is given by

$$H(\mu) = -\sum_i \mu(i)\log(\mu(i)).$$

The notion of entropically-regularized transport is to include some fraction of the entropy of a coupling γ between measures as an additional constraint in the minimization problem underlying optimal transport. We follow the treatment in [107] in the derivation of the following.

Definition 1.2.14. Let measures μ and ν on a finite metric space X be given, say $|X| = m$, and recall that $\Gamma(/mu, /nu)$ is the set of all flows or couplings between μ and ν. Let $\mathbf{D}$ be the distance matrix in X and let $\alpha > 0$ be fixed. We then define the *entropically-regularized transport problem* by

$$W_{1,\alpha}(\mu,\nu) = \min_{\gamma\in\Gamma(\mu,\nu)} -\alpha H(\gamma) + \sum_{i,j}\mathbf{D}(i,j)\gamma(i,j).$$

We can express the function being minimized in the above as follows

$$\begin{aligned}
-\alpha H(\gamma) + \sum_{i,j}\mathbf{D}(i,j)\gamma(i,j) &= \sum_{i,j}\alpha\gamma(i,j)\log(\gamma(i,j)) + \mathbf{D}(i,j)\gamma(i,j) \\
&= \alpha\sum_{i,j}\gamma(i,j)\left(\log(\gamma(i,j)) + \mathbf{D}(i,j)/\alpha\right) \\
&= \alpha\sum_{i,j}\gamma(i,j)\left(\log(\frac{\gamma(i,j)}{e^{-\mathbf{D}(i,j)/\alpha}})\right)
\end{aligned}$$

Letting $\mathbf{K}_\alpha$ the matrix such that $\mathbf{K}_\alpha(i,j) = e^{-\mathbf{D}(i,j)/\alpha}$ we note that

$$\alpha \sum_{i,j} \gamma(i,j) \left(\log \frac{\gamma(i,j)}{e^{-\mathbf{D}(i,j)/\alpha}} \right) = \alpha \mathrm{d}_{KL}(\gamma, \mathbf{K}_\alpha)$$

where d_{KL} is the Kullback-Leibler divergence from Definition 1.2.3.

The above is a multivariate calculus problem, the optimization of a differentiable function subjection to set of equality constraints, and thus we can apply the method of Lagrange multipliers. Let $\boldsymbol{\lambda}_\mu$ and $\boldsymbol{\lambda}_\nu$ be vectors of Lagrange multipliers for each of the constraints defined by the marginals of our coupling γ. We then have a Lagrange multiplier function $\Lambda(\gamma, \boldsymbol{\lambda}_\mu, \boldsymbol{\lambda}_\nu)$ of the form

$$\begin{aligned}\Lambda(\gamma, \boldsymbol{\lambda}_\mu, \boldsymbol{\lambda}_\nu) = &\sum_{i,j} \mathbf{D}(i,j)\gamma(i,j) + \alpha \sum_{i,j} \gamma(i,j)\mathbf{D}(i,j) \\ &+ \sum_i \boldsymbol{\lambda}_\mu(i) \left(\mu - \sum_j \gamma(i,j) \right) + \sum_j \boldsymbol{\lambda}_\nu(j) \left(\nu - \sum_a \gamma(i,j) \right)\end{aligned}$$

Now let $\mathbf{1}$ be a column vector of ones and let the $\log(\gamma)$ be given as the component-wise logarithm. Taking the gradient of the above and setting it equal to $\mathbf{0}$ yields

$$0 = \gamma + \alpha \mathbf{1}\mathbf{1}^t + \alpha \log(\gamma) - \boldsymbol{\lambda}_\mu \mathbf{1}^t - \mathbf{1}\boldsymbol{\lambda}_\nu$$

so that

$$\log(T) = \frac{(\boldsymbol{\lambda_\mu} - \alpha \mathbf{1})}{\alpha} + \frac{\mathbf{1}\boldsymbol{\lambda_\nu}^t}{\alpha} + \log(\mathbf{K}_\alpha).$$

Letting $\mathbf{p} = \exp[\frac{\boldsymbol{\lambda_\mu} - \alpha \mathbf{1}}{\alpha}]$, $\mathbf{q} = \exp[\frac{\boldsymbol{\lambda_\nu}}{\alpha}]$ and, for a vector $\mathbf{v}$, letting $\mathrm{diag}(v)$ be the matrix whose diagonal elements are v, we then have that

$$\gamma = \mathrm{diag}(\mathbf{p}) K_\alpha \mathrm{diag}(\mathbf{q}).$$

Hence by a change of variables we now seek the $2n$ components of $\mathbf{p}$ and $\mathbf{q}$ instead of the n^2 components of γ. Further, recalling the constraints imposed by our marginals, we have, after a bit of reassociation of matrix products, that

$$\mathbf{p} \otimes (K_\alpha \mathbf{q}) = \mu$$

$$\mathbf{q} \otimes (K_\alpha^t \mathbf{p}) = \nu$$

where $\otimes$ denotes the element-wise product of vectors, also known as the *Hadamard product*. This also inspires an extremely succinct algorithm for the computation of entropically-regularized transport, the *Sinkhorn algorithm*. Let $\oslash$ denote component-wise division.

Definition 1.2.15 (Sinkhorn Algorithm)**.** Using the notation above, let p^0, q^0 be arbitrary probability distributions on X such that both are strictly positive. Then the *Sinkhorn algorithm* for entropically-regularized transport is given by iteration of

$$\mathbf{p}^{k+1} = \mu \oslash (K_\alpha \mathbf{q}^k)$$
$$\mathbf{q}^{k+1} = \nu \oslash (K_\alpha \mathbf{p}^{k+1})$$

until the quantity

$$\mathrm{diag}(\mathbf{p}^{k+1}) K_\alpha \mathrm{diag}(\mathbf{q}^{k+1})$$

converges.

It can be shown [8] that the above algorithm converges asymptotically and efficiently to the optimal γ. Thus we have a tool for computing an approximation of the optimal transport metric which circumvents some of the time-complexity shortcomings of thr Transportation Simplex algorithm.

We have discussed a classic solution technique for optimal transport problems, the Simplex algorithm, and its refinements in the case of optimal transport. This discussion also highlighted facets of the geometry of the solution set to optimal transport problems, and thus to the Wasserstein metric. We have also discussed a method to approximate the optimal transport metric via entropic regularization as well as described an algorithm which converges to the optimal value for the entropically-regularized Wasserstein metric. We next shift emphasis from a discussion of the ways of measuring distance between

probability measures defined on metric spaces to metric spaces themselves, in particular to those of graphs.

1.3 Introduction to Graph Theory and its Applications to Biology

1.3.1 Introduction to Graph Theory and Combinatorics

A classic problem in mathematics is that of attempting to pass from local information to that of global information. In what ways can information about the immediate vicinity of a point tell us about the large scale structure of a mathematical object? Indeed many basic notions in mathematics, such as that of compactness or continuity, are valued for precisely this reason. They describe mathematical settings in which we are able to make this leap. The birth of the theory of graphs can be cast as precisely such a question and an introduction to the subject would be incomplete without the obligatory mention of the story.

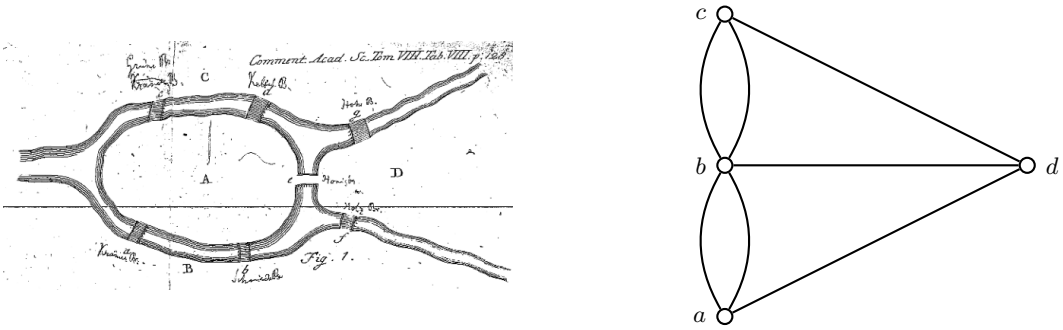


FIGURE 1.5: Figure from Euler's 1735 paper 'Solution problematis as geometriam situs pertinentis' on the solution to the Seven Bridges problem [MAA Euler Archive] and a more modern presentation of the same graph.

The city of Königsberg, present day Kalingrad, is bisected by the Pregel River. In addition to the components of the city on each of the banks, two large islands in the river are also inhabited. There are seven bridges which interconnect the various components of the city. Leonhard Euler considered the problem of whether one might be able to wander the city in

such a way as to cross each bridge precisely once, abstracted by him in Figure 1.5. In 1735 he presented the paper 'Solution problematis as geometriam situs pertinentis' in which he proves that such a tour of the town is not possible. His solution relates the local structure of the graph, the number of bridges departing from a given landmass in this case, to a global graph property, that of a particular sort of path through the graph.

The application of graph theory in this often recounted story of the origins of the field is actually quite close to one of the chief applications of graphs in our further discussion. We will be concerned with two classes of graphs, those of *trees* (Defintion 1.3.4) and *de Bruijn* graphs (Defintion 1.3.9). While trees are very common, de Bruijns graphs are a bit more exotic and have their origin in the study by Nicolaas Govert de Bruijn [23] of the following problem.

Let $\mathcal{A}$ be a set of n symbols, and let $k \in \mathbb{N}$ be fixed. There are clearly n^k distinct *words* or ordered tuples of length k we can construct from a set n things. If we took all n^k words of length k and concatenated them together we would have a word made from $k \cdot n^k$ symbols which contained within it each of our n^k. It is easy to imagine we can do better and find a shorter word satisfying the same constraint. At a minimum such a word must have n^k starting positions, one for each of our n^k distinct words, and the final starting position must end with an additional $(n-1)$ symbols, to finish forming that last word. Is there a word that achieves this minimum length of $n^k + (n-1)$? Take $(A) = \{0, 1, 2, 3\}$, so that $k = 4$, and $n = 3$. Then $n^k + (n-1) = 4^3 + 2 = 66$. Consider this string of 66 symbols

$$000100200301101201302102202303103203311121131221231321332223233300.$$

This one does the trick and we shall see that this is not a special property of the numbers 3 and 4, but rather a general property of a certain graph upon which Euler would have been able to take his tour.

We recount both of these problems not only before their inherent aesthetic appeal, which

we see as self-evident, but also because they underlie a very useful application of graphs to computational biology. Building up big words from little ones in an efficient manner, the process at the heart of our string of 66 symbols above, is more than a novelty when applied to genomes. It is also a useful motivation for a novel sense of distance on strings of symbols which we hope to show has applications to metagenomics. We next establish some useful definitions for the theory of graphs.

1.3.2 Definitions Related to Graph Theory and Combinatorics

We begin with the definition of a multigraph G. Let V be a nonempty, finite set and let E be a multiset of two-element multisets containing elements from V. The pair (V, E) form a *multigraph* and we write $G = (V, E)$. We refer to the elements of V as the *vertices* or *nodes* of G. We refer to the elements of E as the *edges* of G and say an element $\{u, v\} \in E$ is a edge between u and v.

Definition 1.3.1 (Graph)**.** A *graph* is a multigraph such that each element of E is distinct and for each $\{u, v\} \in E$ we have that $u \neq v$. That is there are not multiples edges between vertices and there are no edges between a vertex and itself.

A *digraph* is a graph whose edge set consists of ordered pairs of elements from V, that is, the edges have a fixed orientation. For brevity's sake we will restrict our references from this point forward to graphs, though these definitions apply to each of these objects, given the appropriate modifications.

Let $v \in V$ and let E_v be the subset of E whose elements contain v. The *degree* of v is the cardinality of E_v. For $u \in V$, we say u and v are *adjacent* if $\{u, v\} \in E$. The *order* of G is the cardinality of V. We now define a very useful way of encoding the connectivity of a graph, that of the *adjacency matrix*.

Definition 1.3.2 (Adjacency Matrix)**.** The *adjacency matrix* of a graph $G = (V, E)$ is a

square matrix A, indexed by $V \times V$ such that $A(i,j) = 1$ if $(i,j) \in E$ and $A(i,j) = 0$ otherwise.

We say a graph H is a *subgraph* of G if the vertices and edges of H are subsets of V and E, respectively. For a subset U of V, the subgraph *induced by* U is the subgraph of G whose vertices are precisely U and whose edge set is maximal, with respect to set inclusion, as a subset of E. Similarly for a subset F of E, the subgraph *induced by* F is the subgraph of G whose edges are precisely U and whose vertex set is minimal, with respect to set inclusion, as a subset set of V.

We say a graph is *edge-weighted* when considering a strictly positive function $l : E \to \mathbb{R}$. We call l a *weight function* for G or that it defines a *length* for the edges of G. Note that we can always consider a graph to be edge-weighted by taking the weight which is identically 1. We similarly define a graph to be *vertex-weighted* when considering a nonnegative function $p : V \to \mathbb{R}$.

A *path* P from v to u in G is a sequence of $n+1$ elements $\{x_0, x_2, ..., x_n\}$ from V such that $u = x_0$, $v = x_n$ and x_i is adjacent to x_{i+1} for all $i \in \{0, ..., n-1\}$. A *cycle* is a path from v to v in G.

For a graph without explicit edge-weights the *length* of the path $\{x_0, x_2, ..., x_n\}$ is n, the number of edges involved. For an edge-weighted graph with weight function l, we define the length to be $\sum_{i=0}^{n-1} l(x_i, x_{i+1})$, the sum of edge-weights along the path.

We say a graph is *connected* if, for all u and v in V, there exists a path from u to v. For $v \in V$ we define the *connected component containing* v as the maximal connected subgraph containing v as a vertex. Similarly, for $e \in E$ we define the *connected component containing* e as the maximal connected subgraph containing e as a edge.

Definition 1.3.3 (Bridge)**.** We say an edge $e \in E$ is a *bridge* if the connected component containing e in G fails to remain connected in the subgraph of G induced by $E \backslash \{e\}$.

Definition 1.3.4 (Tree)**.** A *tree* is a connected graph T containing no cycles.

In a tree T we may distinguish a single vertex $\rho \in V$ as the *root* of T. When drawing T we typically place ρ at the top of the page. A tree with a root is called *rooted*. A vertex adjacent to precisely one additional vertex in a tree is called a *leaf*. We call a vertex *internal* if it is not a leaf.

Definition 1.3.5 (Ultrametric Tree)**.** If the length of the path from the root ρ to any leaf v_l in T is precisely d, for some fixed d we say T is an *ultrametric tree* of *depth* d.

We say a rooted tree T is *binary* if the root ρ is adjacent to precisely two vertices, which we call *daughters* and, inductively, each internal vertex has in turn precisely two daughters of its own.

Definition 1.3.6 (Perfect Binary Tree)**.** We a binary tree is *perfect* if, for the trivial edge-weighting, it is ultrametric of depth d, for some $d \geq 1$.

Having established a bit of the language of graph theory, we turn to a few simple results in the field which will be useful for our future discussion.

1.3.3 Summary of Standard Results Related to Graph Theory

We first consider some standard results related to the mathematical theory of graphs in general before more specifically discussing the properties of trees. In the following, let $G = (V, E)$ be a graph.

Proposition 1.3.1 (Classic)**.** *Given a connected graph $G = (V, E)$, there is a natural metric space structure on the set of vertices V. That is, given $u, v \in V$ let l be the minimal length of a path from u to v. Then $d(u, v) = l$ defines a metric on V.*

Proof. We need to prove that the metric defined above is well-defined, symmetric, positive-definite and satisfies the triangle inequality.

As G is connected, there exists a path between any pair of vertices u and v in V. The lengths of paths from u to v is thus a non-empty, finite set of non-negative numbers, and thus contains a minimal element. So $d(u, v)$ is well-defined.

The path from u to v of minimal length can be traversed backwards, defining a path from v to u, which is clearly of minimal length. Hence $d(u, v) = d(v, u)$.

As we are counting, $d \geq 0$, and a path is of length 0 precisely when it contains no edges, that is it contains only one vertex. Hence d is positive-definite.

Finally, say $u, v, w \in V$ and $d(u, v) = l$, $d(v, w) = m$. By concatenating our paths, from u to v and then from v to w, we see that there exists a path from u to w of length bounded by $l + m$. That is, $d(u, w) \leq l + m$.

It follows that d defines a metric on V. □

Definition 1.3.7 (Path metric)**.** The metric described above, given by the length of the minimal path between vertices in graph G is known as the *induced path metric* in a graph G.

We recount a famous result [27] related to the origins of graph theory recounted in our introduction to the subject and which is related to an important application of graph theory in computational biology. We say a graph G is *Eulerian* if there exists a cycle in G which traverses each edge of G precisely once.

Theorem 1.3.1 (Euler 1735)**.** *A connected graph G is Eulerian if and only if the degree of each vertex is even.*

Having described a few a few result related the to Graph theory in general, we specialize for a moment and describe results explicitly related trees. In the following discussion, let $T = (V, E)$ be a tree.

Proposition 1.3.2 (Classic)**.** *In a tree T, for each pair $u, v \in V$ there exist a unique path from u to v.*

Proof. Since T is connected, by our definition of a tree, it suffices to show uniqueness.

To produce a contradiction, suppose not, and let $\{x_0, x_1, ..., x_n\}$ and $\{y_o, y_1, ..., y_m\}$ be a pair of distinct paths from u to v, where $u = x_0 = y_0$ and $v = x_n = y_m$. Without loss of generality, assume $n \leq m$. Now let i be the first index such that $x_i \neq y_i$ and let $\{j, k\}$ be the first pair of indices such that each is greater than i and that $x_j = y_k$.

By construction, $i \neq 0$ and $\{j, k\}$ exist, as $\{n, m\}$ satisfy all but the minimality.

By the definition of $\{j, k\}$, the elements $\{x_i, x_{i+1}, ..., x_{j-1}\}$ and $\{y_i, y_{i+1}, ..., y_{k-1}\}$ are all pairwise distinct. Further, consecutive elements from each set are adjacent in T, as they are consecutive elements in a path. As $x_j = y_k$ and $x_i = y_i$ it follows that $\{x_i, x_{i+1}, ..., x_j, y_{k-1}, y_{k-2}, ..., y_i\}$ is a cycle. This is a contradiction, as T is a tree, and hence the statement is proved. □

Corollary 1.3.1 (Classic)**.** *In a tree T, each edge is a bridge.*

Proof. Let $e \in E$ be arbitrary. Say $e = \{u, v\}$. Then e is the single edge on the path from u to v. By uniqueness, this is the only path from u to v in T, and thus the graph induced by removing e is disconnected. □

Thus the deletion of an edge e in a graph leaves the graph having two connected components. As these are connected acyclic graphs, they are in turn trees, *subtrees* of T.

Definition 1.3.8 (Branch)**.** Let T be a rooted tree, with root ρ and let $v \neq \rho$ be a vertex in T. Let B_v and B_ρ be the subtrees formed by the deletion of an edge adjacent to v in the path from v to ρ, such that $v \in B_v$ and $\rho in B_\rho$. Then we say that B_v is the *branch* of T defined by v.

In this section we have established the natural metric structure on graphs. Additionally, we have noted a few of the useful topological properties of trees, properties which will be useful in our following discussion. We next turn to the second collection of graphs of interest, that of de Bruijn graphs.

1.3.4 Summary of Definitions and Results Related to de Bruijn Graphs

We begin with some definitions related to the theory of words. An *alphabet* $\mathcal{A}$ is a finite set of symbols. Examples include $\mathcal{A} = \{0, 1\}$ the binary alphabet, and what we will refer to as the *genetic alphabet* $\mathcal{A} = \{A, C, G, T\}$, the alphabet representing the four genetic nucleotides.

A *word* is a finite tuple of elements from $\mathcal{A}$. We will denote a word by the concatenation of the symbols comprising it, that is a word w is $w_1 w_2 ... w_k$ for $w_i \in \mathcal{A}$.

The *length* of a word w is the number of symbols in the tuple comprising it. We denote the length of a word w by $|w|$. We will often refer to a word of length k as a *k-mer*, particularly in the context of the alphabet $\{A, C, G, T\}$. Let the *empty word* be the word of length 0 containing no symbols.

We denote the set of all words of length k generated by $\mathcal{A}$ as $\mathcal{A}^k$ and the set of all finite words generated by $\mathcal{A}$ as $\mathcal{A}^* = \bigcup_{k \in \mathbb{N}} \mathcal{A}^k$. There is a natural algebraic structure on $\mathcal{A}^*$ given by juxtaposition. That is, given words $v = \{v_1 \dots v_n\}$ of length n and $w = \{w_1 \dots w_m\}$ of length $\in \mathcal{A}^*$ define their product vw as the word $vw = \{v_1 \dots v_n w_1 \dots w_m\}$ of length $n + m$.

We say a word v is a *factor* of a word w if there exist words w_p and w_s such that $w = w_p v w_s$. We say v is a *prefix* of w in the case that v is a factor and w_p is the empty word. We say v is a *suffix* of w in the case that v is a factor and w_s is the empty word. We say a word w is a *right-extension* of v if v is a prefix of w and $|v| = |w| - 1$, and say w is a *left-extension* of v if v is a suffix of w and $|v| = |w| - 1$.

Let $\leq$ be a total ordering on the symbols in an alphabet $\mathcal{A}$. We may then lift this ordering to a total ordering on $\mathcal{A}^k$ by defining $v = v_1v_2...v_k \leq w = w_1w_2...w_k$ if either $v = w$ or $v_i \leq w_i$ for i the first index in which the letters comprising v and w differ. We refer to this as the *lexicographical order* on $\mathcal{A}^k$.

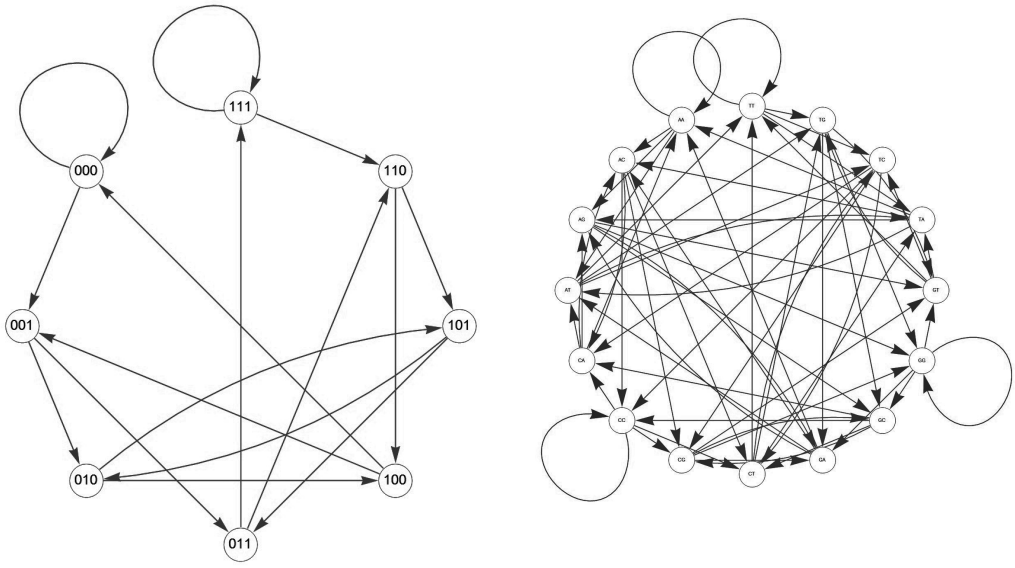
Definition 1.3.9 (*de Bruijn graph).* The *k-dimensional de Bruijn graph* $B_k(\mathcal{A})$ is the directed graph with vertex set

$$V = \mathcal{A}^k$$

and edge set

$$E = \{(v, w) \in V \times V \mid v_2v_3...v_k = w_1w_2...w_{k-1}\}.$$

That is, for words v and w there is a directed edge between them in $B_k(\mathcal{A})$ if the suffix of length $k-1$ of v agrees with the prefix of length $k-1$ of w. Figures 1.6a and 1.6b show representations of the de Bruijn graphs for words of length 3 on a binary alphabet and words of lengths 2 on the genetic alphabet, respectively.



(a) $B_3(\{0,1\})$

(b) $B_2(\{A,C,G,T\})$

FIGURE 1.6: Depiction of $B_3(\{0,1\})$ and $B_2(\{A,C,G,T\})$, the 3-dimensional de Bruijn graph from a binary alphabet and the de Bruijn graph of 2−mers from the genetic alphabet.

The *k-dimensional symmetric de Bruijn graph* $B_k^*(\mathcal{A})$ is the undirected graph with vertex set

$$V = \mathcal{A}^k$$

and edge set

$$E = \{(v,w) \in V \times V \mid v_2v_3...v_k = w_1w_2...w_{k-1} \text{ or } v_1v_2...v_{k-1} = w_2w_3...w_k\}.$$

Figure 1.7 shows a representation of the symmetric de Bruijn for 2−mers from the genetic alphabet.

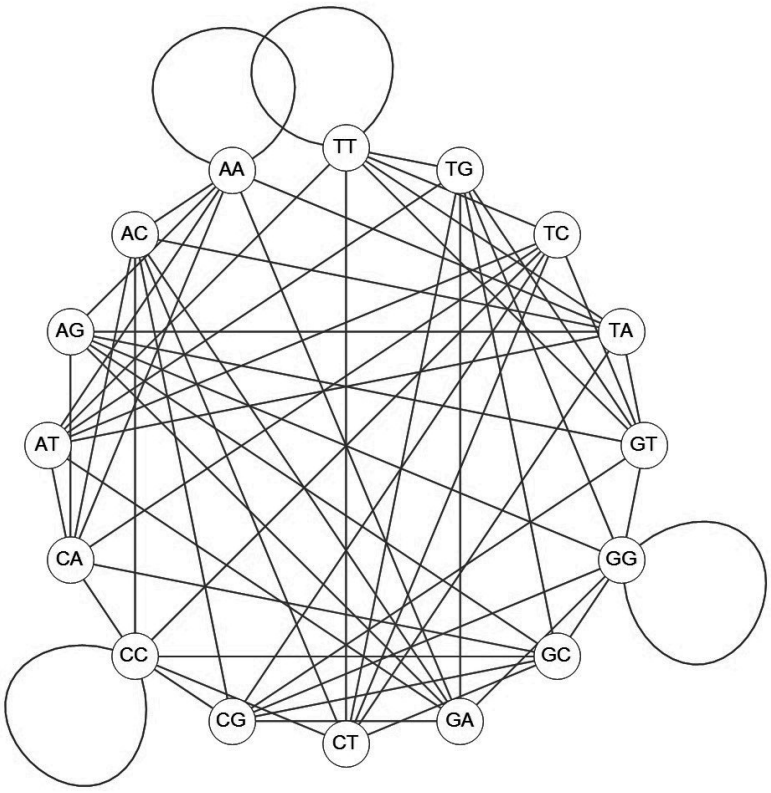


FIGURE 1.7: Depiction of $B_2^*(\{A, C, G, T\})$, the symmetric de Bruijn graph of 2−mers from the genetic alphabet.

Our interest in de Bruijn graphs is related to genome assembly, as we shall see in the following section, but we would be remiss to not address the problem we described briefly in our introduction. While this is mostly for the enjoyment of the author, it does highlight the manner in which de Bruijn graphs are used in genomics.

Theorem 1.3.2 (de Bruijn 1946)**.** *Given an alphabet $\mathcal{A}$ such $|\mathcal{A}| = n$ and let $k \in \mathbb{N}$ be fixed. Then there exists a word of length $n^k + (n-1)$ containing every element of $\mathcal{A}^k$ as a factor.*

Proof. Consider the de Bruijn graph $B^*_{k-1}(\mathcal{A})$. First note that the degree of each vertex w is $B^*_k(\mathcal{A})$ is $2n$, as we form n edges by appending each of our n symbols to the end of w and another n edges by appending our n symbols to the beginning of w. Thus, by Theorem 1.3.1 our graph is Eulerian. Further note that each edge corresponds to a unique element of $\mathcal{A}^k$, as a word is uniquely defined by simultaneously knowing its maximal prefix and suffix. Hence any Eulerian cycle in $B^*_{k-1}(\mathcal{A})$ describes a cyclic string of $|B^*_k(\mathcal{A})| = n^k$ symbols containing each element of $\mathcal{A}^k$. To generate the noncyclic version of our solution we need to traverse the first $(n-1)$ edges of our cycle twice, once to begin our first factor and a second time to complete our last factor. □

We continue with a simple way of defining a measure on de Bruijn graphs.

Definition 1.3.10 (Occurrence)**.** For $v, w \in \mathcal{A}^*$ the *number of occurrences* of v in w is

$$occ_v(w) = |\{i | w_i w_{i+1} ... w_{i+|v|-1} = v\}|.$$

Clearly $occ_v(w) = 0$ when $|v| > |w|$.

This is the merely the number of instances in which v appears as a factor of w.

For W a finite subset of $\mathcal{A}^*$ and $v \in \mathcal{A}^*$, the *factor frequency of v in W* is

$$freq_v(W) = \frac{\sum_{w \in W} occ_v(w)}{\sum_{w \in W} \max\{0, |w| - |v| + 1\}}.$$

Define $freq^k(W)$ to be the vector, in lexicographical order, of $freq_v(W)$ for $v \in \mathcal{A}^k$.

Proposition 1.3.3. *Using the notation defined above, $freq^k(W)$ defines a measure on $\mathcal{A}^k$.*

Proof. Since $\mathcal{A}^k$ is a finite set, we need only demonstrate that $freq^k(W)$ integrates to 1 over $\mathcal{A}^k$. That is

$$\sum_{v \in \mathcal{A}^k} freq^k(W)(v) = \sum_{v \in \mathcal{A}^k} freq_v(W)$$

$$= \sum_{v \in \mathcal{A}^k} \frac{\sum_{w \in W} occ_v(w)}{\sum_{w \in W} \max\{0, |w| - |v| + 1\}}$$
$$= \frac{\sum_{v \in \mathcal{A}^k} \sum_{w \in W} occ_v(w)}{\sum_{w \in W} \max\{0, |w| - |v| + 1\}}$$

Notice that the numerator and the denominator here both count the same quantity, the number of factors, counting with repetition, of length k appearing in the words in W. The numerator does so by counting the occurrences over all possible factors of length k, the denominator does so by counting the number of consecutive sequences of length k in each $w \in W$. Thus the sum is 1, as required. □

We next describe a useful, and related, function when considering factor frequencies and de Bruijn graphs, that of the *complexity function* of a word or collection of word.

Definition 1.3.11 (Complexity Function)**.** For a word w over an alphabet $\mathcal{A}$ define the *complexity function* $f_w : \mathbb{N} \to \mathbb{N}$ by

$$f_w(n) = |\{v \,\big|\, v \text{ is a factor of } w \text{ and } |v| = n\}|.$$

We extend the complexity function to a set of words W by counting the number of factors of length n in any of the words $w \in W$.

We now relate a useful formula for the computation of the path distance in each of the above de Bruijn graphs.

Theorem 1.3.3 (Graph Distance in $B_k(\mathcal{A})$)**.** *Let $v = v_1 \dots v_k, w = w_1 \dots w_k \in \mathcal{A}^k$. Let d_{dB} be the path length distance in the de Bruijn graph. Then [64]*

$$\mathrm{d}_{dB}(v, w) = k - \max\{i | \, 1 \leq i \leq k, v_{k-s+1} \dots v_k = w_1 \dots w_i\}.$$

Theorem 1.3.4 (Graph Distance in $B_k^*(\mathcal{A})$)**.** *Let $v = v_1 \dots v_k, w = w_1 \dots w_k \in \mathcal{A}^k$. Let d_{dB^*} be the path length distance in the symmetric de Bruijn graph. Define*

$$l_{i,j}(v, w) = \max\{s | \, s \leq j, s \leq k - i + 1, v_i \dots v_{i+s-1} = w_{j-s+1} \dots w_j\}$$

$$r_{i,j}(v,w) = \max\{s |\, s \leq i, s \leq k-j+1, v_{i-s+1} \dots v_i = w_j \dots w_{j-s+1}\}$$

Then [64]

$$\mathrm{d}_{dB^*}(v,w) = 2k - 1 + \min_{1 \leq i,j \leq k} \{i - j - \max\{l_{i,j}(v,w), r_{i,j}(v,w)\}\}.$$

Note that in the case of a directed graph, our distance on vertices is not a metric as it is not symmetric. On the other hand, utilizing the path distance on the symmetric de Bruijn graph does yield a metric on $\mathcal{A}^k$, which we will make use of later in our work. We break from our discussion of de Bruijn graphs to introduce an alternate pair of metrics on the space of finite words [77]. We begin by defining a set of operations on words, that of *insertion, deletion* and *substitution*. In the following let v be a word of length k from the alphabet $\mathcal{A}$.

Definition 1.3.12 (Insertion)**.** We say w is obtained from v by an *insertion* if there exists $\alpha \in \mathcal{A}$ and factors x and y of v such that $v = xy$ and $w = x\alpha y$.

Definition 1.3.13 (Deletion)**.** We say w is obtained from v by an *deletion* if there exists $\alpha \in \mathcal{A}$ and factors x and y of v such that $v = x\alpha y$ and $w = xy$.

Definition 1.3.14 (Substitution)**.** We say w is obtained from v by an *substitution* if there exists $\alpha, \beta \in \mathcal{A}$ and factors x and y of v such that $v = x\alpha y$ and $w = x\beta y$.

We now define the *edit distance* and the related *longest common subsequence distance (LCS)* on elements of $\mathcal{A}^*$.

Definition 1.3.15 (Edit Distance)**.** Let $v, w \in \mathcal{A}^*$. Define $\mathrm{d}_E(v,w)$ to be the minimum number of applications of insertion, deletion or substitution needed to obtain w from v.

Definition 1.3.16 (Longest Common Subsequence (LCS) Distance)**.** Let $v, w \in \mathcal{A}^*$. Define $\mathrm{d}_{LCS}(v,w)$ to be the minimum number of applications of insertion or deletion needed to obtain w from v.

It is clear that the above are both metrics, given the symmetry of the operations involved. The LCS distance is so named because of the following alternate characterization.

Proposition 1.3.4 (Alternate Characterization of LCS). *Let $v = v_1 \dots v_n$ and $w = w_1 \dots w_m$ be words. Let I be the set of strictly increasing functions $\rho : \mathbb{N} \to \mathbb{N}$. Then [64]*

$$\mathrm{d}_{LCS}(v, w) = (n + m) - 2 \cdot \max\{s | \exists\, \rho_v, \rho_w \in I \text{ s.t. } v_{\rho_v(1)} \dots v_{\rho_v(s)} = w_{\rho_w(1)} \dots w_{\rho_w(s)}\}$$

where we take the maximum over the empty set to be 0.

That is, we may take the LCS to be the number of symbols left unmatched in a pairing between symbols comprising words which respects order.

In this section we have defined the language used in understanding the theory of words and the de Bruijn graph, a natural mathematical object which encodes information regarding the factors of words. We have also defined a few natural metric structures on the space of words, those arising from de Bruijn graphs, as well as those derived from of a set of operations on words. We now turn to a discussion of the applications of graphs, particularly the manner in which graphs have been used in biology.

1.3.5 Survey of the Applications of Graph Theory to Genomic Assembly

Applications of the mathematical formalism of graphs appear in a variety of contexts in the biological sciences, including networks describing biomolecular interaction and structure [48, 9], ecosystem ecology [14], epidemiology and phylogenetics [90] (see Section 1.1.3). In molecular biology, of chief importance is the understanding of the structure and function of the molecular building blocks of life. Graphs have been implemented in predicting the shape of molecules, such as predicting RNA secondary structures [9] and protein shape [38]. In the field of genomics, graph theory has been particularly useful for giving a computational framework to the problem of *de novo genome assembly*. Our chief purpose

here will to be to describe the application of graphs to this problem. See Section 1.1.2 for information related to genomics which is assumed in the following.

Given a collection of short sequence reads, the assembly problem is that of finding the most parsimonious sequence, that is, the simplest genome, from which this given collection might have arrived as factors. As we have seen, de Bruijn graphs are a useful object when considering the manner in which longer words can be built up from a set of factors. Indeed, de Bruijn graphs have been at the heart of solving this problem [19] by solving an interpretation of the same Eulerian path problem described in Theorem 1.3.2, but in a genomic context. There are many current de Bruijn graph based genomic assembly packages, such as Velvet [134], SOAPdenovo2 [70] and SPAdes [5], which vary in their optimization for particular genomes sizes, sequencing technologies or methods of error correction. We describe the general method utilized by following the algorithm for EULER [92], one of the original such assemblers.

Given that errors occur in the process of genomic sequencing, error correction is made before assembly begins. By leveraging the relatively high coverage rates supplied by sequencing technologies, error correction can be done by consensus among sequence reads. Additional processing to reduce error can be made by applying dynamic programming techniques to determine the minimum number of edits to a set of sequences for an assembly to exist and then thresholding 'low-quality' reads having little agreement with the rest of the sequences.

We begin assembly itself by selecting a positive integer k and deriving from our collection W of sequence reads a multiset W_k of all k-mers appearing in W. The optimal values for k vary by assembler technology, sequence read length and genome length [18] and are often chosen experimentally, by comparing the quality of assemblies under each. Typical values are from $k = 20$ to $k = 70$.

Letting $\mathcal{A} = \{A, C, T, G\}$ we select and build the vertex set for the de Bruijn of $(k-1)$-

mers, $B_{(k-1)}(\mathcal{A})$. Note that each directed edge in $B_{(k-1)}(\mathcal{A})$ corresponds uniquely to a k-mer, given by the word whose first proper prefix is given by the originating vertex and whose first proper suffix is the terminal vertex. Thus we then take our set of k-mers W_k and populate the edges, with multiplicity, of $B_{(k-1)}(\mathcal{A})$.

A path of length n in this graph is thus a word of length $n+k-1$, with overlapping factors of length k appearing among the known set of k-mers in the multiset of sequence reads. By producing an Eulerian path which traverses each edge of the graph, we produce an assembly of the genome. There are linear time algorithms for solving the Eulerian path problem [53], given that the degree requirements of the graph are satisfied so that a solution exists, and so, in theory, we have reduced the problem of genome assembly to that of the Eulerian path problem. Complexities arise due to unaddressed errors in sequence data, uneven or incomplete genome coverage and boundary conditions for sequences. These constraints are addressed otherwise, or are used to place additional constraints on the choice of an Eulerian path.

In the above we have described the application of graphs to questions arising in genomics, particularly to of sequence assembly. We have highlighted how adjacency in the de Bruijn graph can be used to find sequences which contain a given set of k-mers as consecutive factors. This material is important motivation for the work in Section 3. The treatment here regarding graphs in a biological context is complementary to our discussion in Section 1.1.3 regarding the application of trees to Phylogenetics. We are now prepared to begin the discussion of our main results, the Wasserstein metric in the context of biologically relevant graphs.

2 COMPUTATION AND FOUNDATIONS OF THE UNIFRAC METRIC FOR MICROBIAL COMMUNITY ECOLOGY

2.1 Introduction

The UniFrac metric (Definitions 1.1.13 and 1.1.14) is a robust and well-used β-diversity metric in metagenomics. Evans and Matsen demonstrated in [28], using the dual formulation of the 1-Wasserstein metric (Definition 1.2.4), that the UniFrac metric is the 1-Wasserstein metric (Definition 1.2.1) for path distance (Definition 1.3.7) in a phylogenetic tree (Defintion 1.1.1) and is given by integration over subtrees of the tree. They utilized that observation to generalize the UniFrac metric, by considering alternate expressions for weighting the differences between relative abundances over a phylogenetic tree, and to develop the use of Monte Carlo permutation tests for significance, in both UniFrac and their generalized metrics.

In the following we provide an alternate and elementary proof of the same fact, that UniFrac is the 1-Wasserstein metric given by integration over the subtrees of a phylogenetic tree. The constructive proof builds a minimizing coupling between samples and highlights a useful invariant behind the UniFrac metric, that of a *weighted differential abundance vector* between relative abundances. We utilize the structure of the integral to construct an efficient linear time algorithm, EMDUniFrac, which computes UniFrac orders of magnitude faster than previous implementations while returning more information, that of the weighted differential abundance vector and a minimizing coupling between samples. These results were published as *EMDUniFrac: exact linear time computation of the UniFrac metric and identification of differentially abundant organisms* with David Koslicki in J. Math. Biol. (2018) https://doi.org/10.1007/s00285-018-1235-9. The ideas contained within Algorithm 2.2.2 were incorporated into a collaborative work which is in

revision, to be published as *Striped UniFrac: enabling microbiome analysis at unprecedented scale* with Daniel McDonald, Yoshiki VÃąazquez-Baeza, David Koslicki, Nicolai Reeve, Zhenjiang Xu, Antonio Gonzalez, Rob Knight, Nature Methods (2018).

We next show how modifying the linear structure underlying the UniFrac metric allows for the computation of a related biological ordination technique known as Double Principal Coordinate Analyis (DPCoA, Definition 1.1.23). We show how combining this realization with a mathematical understanding of the principles behind DPCoA allows for an efficient algorithm which circumvents the need to compute pairwise distances between relative abundances in the generation of DPCoA.

Finally, we demonstrate how considering the UniFrac metric between relative abundances as the L_1 norm of the image of their difference under a linear transformation allows for the formulation for the expected UniFrac distance between Dirichlet distributed sample relative abundances. We show how this has utility to the Dirichlet-Multinomial (Definition 1.1.21) model frequently used for sequence read data.

We conclude by demonstrating the effectiveness of these tools on both real-world and synthetic metagenomic datasets, before discussing potential work for the future.

2.2 Efficient Computation of the UniFrac Metric as the Wasserstein Metric

2.2.1 Alternate Characterization of the 1-Wasserstein Metric over a Tree

In the following we begin our demonstration that the 1-Wasserstein metric $W_1(P, Q)$ between probability distributions P and Q over a tree T is given by an edge-weighted integration over all subtrees of the absolute value of the difference between P and Q by producing an alternate characterization for the minimizing flow realizing the Wasserstein metric. We first require some definitions.

Let P and Q be probability distributions on a tree T with distance matrix $\mathbf{D}$ and edge set E. Recall that $\Gamma(P,Q)$ is the set of all flows or couplings between P and Q in T. By an abuse of notation, we write $i \in T$ to denote a vertex of our tree. For such a vertex $i \in T$ we will say i is a *source* if $P(i) \geq Q(i)$ and say i is a *sink* otherwise. Let T_{source} and T_{sink} denote the sets of sources and sinks, respectively.

Next, we select an arbitrary vertex and distinguish it as the root ρ of T. While the choice of a root in a phylogenetic tree is biologically motivated, it is, for our current mathematical purposes, merely a convenience of notation. For each $i \in T$ let $a(i)$ be the unique neighbor of i in T which lies on the path from i to ρ in T. Thus the edges of T are determined by the set of ordered pairs $(i, a(i))$ for $i \in T$. Let e_i denote the edge $(i, a(i))$. As T is a tree, each edge $e \in E$ is a bridge. Thus the removal of an edge partitions the vertices into two disjoint subsets. We denote the subset containing ρ by T_e and the other by T'_e. Let $l : E \to \mathbb{R}_{\geq 0}$ define a set of edge weights or lengths on E. For $i, j \in T$, define $\pi(i,j)$ to be the set of edges comprising the unique minimal path from i to j in T, so that $\mathbf{D}(i,j) = \sum_{e \in \pi(i,j)} l(e)$ is the distance from i to j in T.

Lemma 2.2.1 (McClelland 2018)**.** *We have that*

$$W_1(P,Q) = \min_{\boldsymbol{M} \in \Gamma(P,Q)} \sum_{e \in E} \sum_{i \in T_e} \sum_{j \in T'_e} l(e) \left(\boldsymbol{M}(i,j) + \boldsymbol{M}(j,i)\right).$$

Proof. Let $1_{\pi(i,j)}(e) : E \to \{0,1\}$ be the indicator function of the path from i to j in T. That is, $1_{\pi(i,j)}(e) = 1$ if e is an edge in the path from i to j and is 0 otherwise. We then have that for any flow $M \in \Gamma(P,Q)$

$$\sum_{i,j \in T} \mathbf{D}(i,j)\mathbf{M}(i,j) = \sum_{i \in T} \sum_{j \in T} \left(\sum_{e \in E} l(e) 1_{\pi(i,j)}(e) \right) \mathbf{M}(i,j) \tag{2.2.1}$$

$$= \sum_{e \in E} \sum_{i \in T} \sum_{j \in T} l(e) 1_{\pi(i,j)}(e) \mathbf{M}(i,j) \tag{2.2.2}$$

$$= \sum_{e \in E} \sum_{\substack{i \in \\ T_e \cup T'_e}} \sum_{\substack{j \in \\ T_e \cup T'_e}} l(e) 1_{\pi(i,j)}(e) \mathbf{M}(i,j) \tag{2.2.3}$$

$$= \sum_{e \in E} \left(\sum_{i \in T_e} \sum_{j \in T'_e} l(e)\mathbf{M}(i,j) + \sum_{i \in T'_e} \sum_{j \in T_e} l(e)\mathbf{M}(i,j) \right) \tag{2.2.4}$$

$$= \sum_{e \in E} \sum_{i \in T_e} \sum_{j \in T'_e} l(e) \left(\mathbf{M}(i,j) + \mathbf{M}(j,i) \right). \tag{2.2.5}$$

The above equalities are justified as follows. To begin, (2.2.1) follows from the definition of the distance function and the use of the characteristic function of the path between vertices to expand the summation over all edges of the graph. Next, (2.2.2) and (2.2.3) reorder the summation and express the vertex set in terms of the partitions defined above by edge deletion. We have that $1_{\pi(i,j)}(e) = 1$ if and only if the vertices i and j belong to distinct partitions T_e and T'_e, from which (2.2.4) follows. Finally, in (2.2.5) we condense the summation notation by reordering the last sum and grouping terms. Taking the minimum over all $\mathbf{M} \in \Gamma(P,Q)$ yields the 1-Wasserstein metric on the left hand side, and thus the desired result is obtained. □

Next, we prove a lower bound on the summands involved in the above definition of the 1-Wasserstein metric.

Lemma 2.2.2 (McClelland 2018)**.** *For any flow $\boldsymbol{M} \in \Gamma(P.Q)$ and any $e \in E$ we have that*

$$\sum_{i \in T_e} \sum_{j \in T'_e} l(e)(\boldsymbol{M}(i,j) + \boldsymbol{M}(j,i)) \geq l(e) \left| \sum_{i \in T_e} P(i) - Q(i) \right|.$$

Further, the vector $\boldsymbol{d}_a$ indexed by the edges of T and having entries $\boldsymbol{d}_a(e) = l(e) \sum_{i \in T_e} \sum_{j \in T_{e'}} \boldsymbol{M}(i,j) - \boldsymbol{M}(j,i)$ is unique, regardless of the minimizing flow M.

Proof. We have that

$$l(e) \left| \sum_{i \in T_e} P(i) - Q(i) \right| = \left| l(e) \sum_{i \in T_e} \left(\sum_{j \in T} \mathbf{M}(i,j) - \sum_{j \in T} \mathbf{M}(j,i) \right) \right| \tag{2.2.6}$$

$$= \left| \sum_{i \in T_e} l(e) \sum_{j \in T} \mathbf{M}(i,j) - \mathbf{M}(j,i) \right| \tag{2.2.7}$$

$$= \left| \sum_{i \in T_e} \left(\sum_{j \in T_e} l(e)(\mathbf{M}(i,j) - \mathbf{M}(j,i)) + \sum_{j \in T'_e} l(e)(\mathbf{M}(i,j) - \mathbf{M}(j,i)) \right) \right| \tag{2.2.8}$$

$$= \left| \sum_{i \in T_e} \sum_{j \in T_e} l(e)(\mathbf{M}(i,j) - \mathbf{M}(j,i)) + \sum_{i \in T_e} \sum_{j \in T'_e} l(e)(\mathbf{M}(i,j) - \mathbf{M}(j,i)) \right| \tag{2.2.9}$$

$$= \left| \sum_{i \in T_e} \sum_{j \in T'_e} l(e)(\mathbf{M}(i,j) - \mathbf{M}(j,i)) \right| \tag{2.2.10}$$

$$\leq \sum_{i \in T_e} \sum_{j \in T'_e} l(e)(\mathbf{M}(i,j) + \mathbf{M}(j,i)). \tag{2.2.11}$$

Equations (2.2.6) and (2.2.7) above follow from expanding $P(i)$ and $Q(i)$ in terms of the row and column sums of M. Equations (2.2.8) and (2.2.9) reorganize the inner sums by way of the partitions T_e and T'_e and then group terms. Next we note that $\sum_{i \in T_e} \sum_{j \in T_e} l(e)(\mathbf{M}(i,j) - \mathbf{M}_(j,i)) = 0$ as each term $\mathbf{M}(i,j)$ occurs precisely twice, once with each sign, which is reflected in (2.2.10) above. This line also demonstrates the uniqueness of $\mathbf{d}_a(e)$, as the quantity is here shown to be equal to $\sum_{i \in T_e} P(i) - Q(i)$, which depends only on the distributions P and Q. Finally, we apply the triangle inequality to yield our result. □

By the lemmas above, it suffices to demonstrate that there exists a flow $\mathbf{M}$ which, for every edge e, satisfies $\sum_{i \in T_e} \sum_{j \in T'_e} l(e)(\mathbf{M}(i,j) + \mathbf{M}(j,i)) = l(e) \left| \sum_{i \in T_e} P(i) - Q(i) \right|$. Further note that the expression on the right is precisely the summands involved in Definition 1.1.14. In our next section we present EMDUniFrac, which produces such a flow M.

2.2.2 EMDUniFrac: Description

The pseudocode for EMDUniFrac is contained in Algorithm 2.2.1. Intuitively, the algorithm begins at the leaves of the tree and 'pushes' mass toward the root; satisfying the

sources and sinks for each subtree encountered in the progression. The matrix $\mathbf{G}$ tracks the mass still needed to be moved to or from each vertex by the algorithm, while the vector w tracks the length of paths traversed by mass at each step.

To implement EMDUniFrac, we first choose an ordering of the set of vertices of T such that for $i, j \in T$, i is an element of the path from j to ρ only if $i \geq j$. A natural such ordering is defined by partitioning the vertices of T by the disjoint circles of radius $r \in \mathbb{N}$ centered at ρ, and then ordering vertices such that increasing indices correspond to partitions defined by descreasing radii.

We now establish a bit of notation for the following algorithm. We then let $\mathbf{G}$ and $\mathbf{M}$ be a pair of matrices whose rows and columns are indexed by the vertices of T with respect to an ordering as above. Let $\mathbf{G}_{i,\cdot}$ denote the ith row of the matrix G. Initialize both $\mathbf{G}$ and $\mathbf{M}$ to be the zero matrix. Let $\mathbf{w}$ be a vector indexed by the vertices of T, initialized to be the zero vector. For any vector $\mathbf{u}$, define skel($\mathbf{u}$) to be the binary vector of the same dimension as $\mathbf{u}$ such for all i, skel($\mathbf{u}(i)$) $= 1$ if $\mathbf{u}(i) \neq 0$ and skel($\mathbf{u}(i)$) $= 0$ otherwise.

2.2.3 EMDUniFrac: Algorithm

Algorithm 2.2.1. *EMDUniFrac (McClelland 2018)*

Input:

$P, Q, \rho, T, E = \{i, a(i)\}$ *For* $i \in T, l$

Initialization:

$\boldsymbol{M}, \boldsymbol{G} = \mathbf{0}$

EMDUniFrac$(P, Q) = 0$

DiffAbund $= \boldsymbol{0}$

$\boldsymbol{w} = \boldsymbol{0}$

Iterations:

1: ***for*** $i = 1, ..., |T|$ ***do***

2: $\quad \boldsymbol{M}(i,i) = \min(P(i), Q(i))$

3: $\quad G(i,i) = P(i) - Q(i)$

4: $\quad w = w + l(i, a(i))\text{skel}(G_{i,\cdot})$

5: $\quad$ ***for*** j *such that* $\boldsymbol{G}(i,j) > 0$ ***do***

6: $\quad\quad$ ***for*** k *such that* $\boldsymbol{G}(i,k) < 0$ ***do***

7: $\quad\quad\quad \boldsymbol{M}(j,k) = \min(\boldsymbol{G}(i,j), -\boldsymbol{G}(i,k))$

8: $\quad\quad\quad \boldsymbol{G}(i,j) = \boldsymbol{G}(i,j) - \boldsymbol{M}(j,k)$

9: $\quad\quad\quad \boldsymbol{G}(i,k) = \boldsymbol{G}(i,k) + \boldsymbol{M}(j,k)$

10: $\quad\quad\quad \text{EMDUniFrac}(P,Q) = \text{EMDUniFrac}(P,Q) + (\boldsymbol{w}(j) + \boldsymbol{w}(k)) \cdot \boldsymbol{M}(j,k)$

11: $\quad\quad$ ***end for***

12: $\quad$ ***end for***

13: $\quad \boldsymbol{G}_{a(i),\cdot} = \boldsymbol{G}_{a(i),\cdot} + \boldsymbol{G}_{i,\cdot}$

14: $\quad \text{DiffAbund}((i, a(i))) = l(i, a(i)) \sum_{t \in T} \boldsymbol{G}(i,t)$

15: $\quad G_{i,\cdot} = \boldsymbol{0}$

16: ***end for***

Output:

$\boldsymbol{M}$, EMDUniFrac(P,Q) *and* DiffAbund

2.2.4 EMDUniFrac: Proof of Correctness, Speed and Space Requirements

What follows is a brief technical lemma used to prove that the matrix M produced by EMDUniFrac is indeed a flow. between distribution P and Q.

Lemma 2.1 (McClelland 2018). Let $m \in T$ be arbitrary. Then for all $n \in T$ such that n is a vertex along the path from m to ρ, when $i = n$ in the loop beginning at line 1 of Algorithm 2.2.1 we have that one of the following hold:

If m is a source, then at the beginning of line 4 of Algorithm 2.2.1 we have that

$$P(m) = G(n,m) + \sum_{k \in T} M(m,k)$$
$$Q(m) = \sum_{k \in T} M(k,m).$$

Alternately, if m is a sink, then at the beginning of line 4 of Algorithm 2.2.1 we have that

$$P(m) = \sum_{k \in T} M(m,k)$$
$$Q(m) = -G(n,m) + \sum_{k \in T} M(k,m).$$

Proof. This follows by induction. Suppose m is a source and let $i = m$ in the loop at line 1 of Algorithm 2.2.1. Then $\min(P(m), Q(m)) = Q(m)$ and hence, by construction, $M(m,m) = Q(m), G(m, m = P(m) - Q(m)$. Further, before beginning the loop at line 4 of Algorithm 2.2.1, every other entry of the mth row of M and G are zero. This is because the elements of these rows are first potentially assigned nonzero values for $i = m$ in the midst of lines 6, 7 or 8. Thus at the beginning of line 4 of Algorithm 2.2.1, we have

$$P(m) = G(m,m) + \sum_{k \in T} M(m,k),$$
$$Q(m) = \sum_{k \in T} M(k,m).$$

Thus the claim holds for $i = m$.

Now suppose inductively that the above equalities holds when $i = j$ for some vertex $j \geq m$ on the path from m to ρ in T. We shall show the equalities holds for $i = a(j)$. As Algorithm 2.2.1 proceeds in the loop at line 1 to the vertex for $i = a(j)$, we have that $G(a(j), m \geq 0$ and thus by line 5 of Algorithm 2.2.1, the m-th column of M is left unchanged. Hence the sum $\sum_{k \in T} M(k,m)$ remains unchanged.

Additionally, any change to $G(a(j), m)$ during the loop at line 5 is compensated by a change to $\sum_{k \in T} M(m, k)$, thus

$$G(a(j), m) + \sum_{k \in T} M(m, k) = G(j, m) + \sum_{k \in T} M(m, k) = P_m.$$

Thus, inductively, the claims holds for all vertices along the path from m to ρ in T and m a source. Symmetric reasoning holds for the case of m a sink. □

We now prove our main result.

Theorem 2.1 (McClelland 2018). The EMDUniFrac algorithm in Algorithm 2.2.1 produces the 1-Wasserstein distance $W_1(P, Q)$ and a corresponding minimizing flow M.

Proof. We first show that M is indeed a flow. Upon the algorithm reaching the root ρ, that is when $i = |T|$ in line 4 of Algorithm 2.2.1, we have traversed every vertex of T, so that

$$\begin{aligned}
0 &= 1 - 1 && (2.2.12)\\
&= \sum_{k \in T} P(k) - Q(k) && (2.2.13)\\
&= \sum_{k \in T_{source}} P(k) - Q(k) + \sum_{k \in T_{sink}} P(k) - Q(k) && (2.2.14)\\
&= \sum_{k \in T_{source}} \left(G(\rho, k) + \sum_{l \in T} M(k, l) - \sum_{l \in T} M(l, k) \right) \dots \\
&\dots + \sum_{k \in T_{sink}} \left(\sum_{l \in T} M(k, l) - \left(-G(\rho, k) + \sum_{l \in T} M(l, k) \right) \right) && (2.2.15)\\
&= \sum_{k \in T} \sum_{l \in T} M(l, k) - \sum_{k \in T} \sum_{l \in T} M(k, l) + \sum_{k \in T_{source}} G(\rho, k) + \sum_{k \in T_{source}} G(\rho, k) && (2.2.16)\\
&= \sum_{k \in T} G(\rho, k). && (2.2.17)
\end{aligned}$$

The above equalities are justified as follows. In (2.2.15) we expand the terms $P(k)$ and $Q(k)$ in terms of the matrices G and M, as shown in Lemma 3, since ρ is an element of

the path from any vertex to ρ. We then group terms in (2.2.16) and (2.2.17) by repeatedly using that $T_{source} \cup T_{sink} = T$, before canceling the symmetric summations of the elements of M.

It then follows that the sum of the positive elements of $G(\rho, \cdot$ is equal to the sum of the negative elements of $G(\rho, \cdot$, and thus, by construction of the loops at lines 4 and 5 of Algorithm 2.2.1, the algorithm must terminate with $G(\rho, \cdot$ identically zero. As we still have that for each $i \in T$, $P(i) = \sum_{k \in T} M(j,k), Q(i) = \sum_{k \in T} M(k,j)$, up to the addition or subtraction of $G(\rho, i) = 0$, M must be a flow.

Now we show that M minimizes the sum defining the 1-Wasserstein distance. By Lemmas 1 and 2, it suffices to show that $\sum_{i \in T_e} \sum_{j \in T'_e} l(e)(M(i,j) + M(j,i)) = |\sum_{i \in T_e} P(i) - Q(i)|$ for all $e \in E$. Given the ordering of the vertices chosen for the algorithm above, let $n \in T - \{\rho\}$ be arbitrary. To begin, we make some observations regarding the structure of the matrix G and its relationship to M in the algorithm. Note, that by construction, at the termination of the loop at line 4 of Algorithm 2.2.1 for $i = n$, the entries of $G(n, \cdot)$ all have the same sign, as the the loops at lines 4 and 5 have the effect of pairwise choosing elements of opposite signs and using one to reduce the magnitude of the other. This process terminates when elements of one or the other sign are exhausted. Second, note that for $k \in T'_{e_n}$ and $m > n$, either $G(m,k) = 0$ or has the same sign as $G(n,k)$, as any change to the entries of $G(\cdot, k)$ is made to move the value toward zero by a quantity bounded by the magnitude of the entry. This again follows from examination of the inner most loop of the algorithm, as well as the evolution of rows of G.

Finally, note that across all $i \in T'_{e_n}, j \in T_{e_n}$ either $M(j,i) = 0$ or $M(j,i) = 0$. This follows since $M(i,j)$, respectively $M(j,i)$, is only assigned a nonzero value in the case of $G(m, i > 0$, respectively $G(m, i < 0$. By the above observation regarding the signs of the elements of $G(n, \cdot)$, only one of these conditions holds across i, j.

Now, without loss of generality, assume

$$\left|\sum_{i\in T_{e_n}} P(i) - Q(i)\right| = \sum_{i\in T_{e_n}} P(i) - Q(i)$$

as the argument for the alternate case is analogous. We then have that

$$\left|\sum_{i\in T_{e_n}} P(i) - Q(i)\right| = \sum_{i\in T_{e_n}} P(i) - Q(i) \tag{2.2.18}$$

$$= \sum_{i\in T_{e_n}} \sum_{j\in T} M(i,j) - M(j,i) \tag{2.2.19}$$

$$= \sum_{i\in T_{e_n}} \sum_{j\in T'_{e_n}} M(i,j) - M(j,i) \tag{2.2.20}$$

$$= \sum_{i\in T_{e_n}} \sum_{j\in T'_{e_n}} M(i,j) + M(j,i). \tag{2.2.21}$$

The change of sign in moving from (2.2.20) to (2.2.21) follows from the above observation that at least one of $M(i,j)$ or $M(j,i)$ must be identically zero, and that the sum must be non-negative. Hence $-M(j,i) = 0 = M(j,i)$. Scaling the above equality by $l(e_n)$ yields

$$\left|\sum_{i\in T_{e_n}} P(i) - Q(i)\right| = \sum_{i\in T_{e_n}} \sum_{j\in T'_{e_n}} M(i,j) + M(j,i).$$

Having achieved the lower bound established in Lemma 2, we must have that the flow M is a minimizer for the sum defining $W_1(P,Q)$. □

Theorem 2.2 (McClelland 2018). Let $|\text{supp } P|, |\text{supp } Q|$ denote the number of elements in the support of the probability distributions P and Q, respectively. Let $s = |\text{supp } P| + |\text{supp } Q|$. Then the EMDUniFrac algorithm has time and space complexity $O(s)$.

Proof. We first consider the time complexity of EMDUniFrac. Note that each iteration of the loop at line 5 of Algorithm 2.2.1 has the effect of satisfying a source i or sink j, that is, establishing the appropriate row sum i or column sum j of the matrix M. Further, the loop at line 5 only visits a pair of vertices (i,j) in the case that both source i and sink j

have not been satisfied, that is, that both $P(i) \neq \sum_{k \in T} M(i,k)$ and $Q(i) \neq \sum_{k \in T} M(k,i)$. As there are s such row or column sums to satisfy, the loop at line 5 is evaluated at most s times. Hence the time complexity of the algorithm is, in total, linear in s.

Now we examine the space requirements of EMDUniFrac. By the above, the matrix M is sparse. That is, there are most s evaluations of the loop at line 5 of Algorithm 2.2.1 and thus, including the assignment of values to M at line 2 of the algorithm, at most $2s$ non-zero entries in M. Additionally, line 3 of the algorithm assigns a nonzero entry to G at most n times, while line 12 has the effect of passing non-zero entries of G from one row to another prior to being removed in line 13. Thus the number of nonzero entries of G is bounded by s. Finally, the vector w in Algorithm 2.2.1 is one dimensional, having at most s nonzero entries. Hence the total space requirements of the algorithm are also linear in s. □

In this section we have demonstrated the correctness and efficiency of an algorithm which computes the UniFrac metric while producing an optimal flow. In our next section we relate a proof which demonstrates a method for the computation of the UniFrac metric which does not produce an optimizing flow.

2.2.5 EMDUniFrac: Linear Algebra Proof of Correctness

We present another proof of our previous result, that the 1-Wasserstein metric is given by integration over all subtrees of the absolute value of the difference between distributions, in the spirit of [28]. This does not produce a minimizing flow M, but it does allow us to characterize the W_1 as the L_1 norm of a readily constructed linear transformation W.

Consider a rooted tree T with root ρ. Identify the subtrees of T with the nodes of T, so that subtree i is the subtree which does not contain ρ formed by deletion of the edge $(i, a(i))$ from the path from node i to the root ρ. The subtree corresponding to ρ is T. Let

the vectors $\{w_i | \ 1 \leq i \leq n\}$ be such that w_i is the indicator function for subtree w, that is $v_i(j) = 1$ for those nodes j in subtree i and zero otherwise

Let W be the $n \times n$ matrix whose rows correspond to the vectors w_i scaled by the corresponding edge weight $l(i, a(i))$. Let P and Q br probability distributions on T, given as column vectors ordered such that entry i corresponds to the root of subtree i.

Theorem 2.2.1. *Using the above definitions, the 1-Wasserstein metric between distributions P and Q is given by [28]*

$$\|W(P - Q)\|_{L_1}.$$

Proof. Recall that by Theorem 1.2.4 we may express the $W_1(P, Q)$ distance between distributions P and Q as

$$W_1(P, Q) = \max_f \sum_{t \in T} f(t)(P(t) - Q(t))$$

for Lipschitz $f \in Lip_1(T)$. It follows from standard facts from analysis, which are perhaps more trivial in our current discrete setting, that f can be expressed as an indefinite integral for a function bounded in absolute value by the Lipshitz constant. That is, we may write any such f as

$$f(t) = \sum_{s \in \pi(t,\rho)} g(s) \cdot l(s, a(s))$$

for some $g : T \to [-1, 1]$, up to the value of $f(\rho)$, which does not alter the value of the maximization.

For a fixed f we then have that

$$\sum_{t \in T} f(t)(P(t) - Q(t)) = \sum_{t \in T} \left(\sum_{s \in \pi(t,\rho)} g(s) \cdot l(s, a(s)) \right) (P(t) - Q(t)) \tag{2.2.22}$$

$$= \sum_{t \in T} \sum_{s \in T} 1_{\pi(t,\rho)}(s) \cdot g(s) \cdot l(s, a(s)) \cdot (P(t) - Q(t)) \tag{2.2.23}$$

$$= \sum_{s \in T} \sum_{t \in T} 1_{\pi(t,\rho)}(s) \cdot g(s) \cdot l(s, a(s)) \cdot (P(t) - Q(t)) \tag{2.2.24}$$

$$= \sum_{s \in T} g(s) \cdot l(s, a(s)) \sum_{t \in T'_{(s,a(s))}} (P(t) - Q(t)) \tag{2.2.25}$$

Where we have again in (2.2.22) expressed summation over a path as summation against an indicator function over T and noted in (2.2.25) that, after interchanging the order of integration, for a fixed node s, the set of vertices t for which s is an element of the path $\pi(t, \rho)$ is precisely the subtree defined by s.

Letting $u(i)$ be the ith component of the vector $u = W(P-Q)$ we see that, by construction, $u(i) = l(i, a(i)) \sum_{t \in T'_{(i,a(i))}} (P(i) - Q(i))$. It follows that $W_1(P, Q) = \max_g \sum_{i \in T} g(i)u(i)$ for $|g| \leq 1$. Clearly we achieve a maximum when $g(i) = 1$ for $u(i) \geq 0$ and $u(i) = -1$ for $u(i) < 0$, that is, when $g(i)u(i) = |w(i)|$. Thus $W_1(P, Q) = \|W(P - Q)\|_{L_1}$ as required.

We note that the above formulation allows for an more efficient implementation of the UniFrac metric in those instances in which we are uninterested in capturing a minimizing flow. By expressing the action of the matrix $\mathbf{W}$ implicitly we are able to recover the UniFrac metric in time linear in the number of OTUs in a pair of samples, without having to interact with the matrix which contains the elements of a minimizing flow. We present pseudocode for this simplified version of the algorithm now.

2.2.6 EMDUniFrac: Algorithm without Flow

In the following algorithm let $T = (V, E)$ have root ρ. Let m be the maximum number of edges in a path from ρ to any vertex in T. Let $S_k = \{v \in T | \, \mathrm{d}(v, \rho) = k\}$ for d the metric which merely counts unweighted edges, for each $1 \leq k \leq m$. For each vertex $v \in T$ let $Dau(v)$ be the set of daughters of v in T, that these are the vertices adjacent to v in the branch which has v as its base.

Algorithm 2.2.2. *EMDUniFrac: Without Flow (McClelland 2018)*

Input:

$P, Q, \rho, T, E = \{i, a(i)\}$ *For* $i \in T, l$

Initialization:

EMDUniFrac$(P, Q) = 0$

$\boldsymbol{w = 0}$

$D = P - Q$

Iterations:

1: ***for*** $i = 1, ..., m$ ***do***

2: ***for*** $v \in S_i$ ***do***

3: ***for*** $w \in Dau(v)$ ***do***

4: *D(v) = D(v) + D(w)*

5: ***end for***

6: ***end for***

7: ***end for***

8: ***for*** $i = 1, \ldots, |T|$ ***do***

9: EMDUniFrac$(P, Q) =$ EMDUniFrac$(P, Q) + (l(i, a(i))) \cdot |D(i)|$

10: ***end for***

Output:

EMDUniFrac(P, Q)

2.3 Efficient Computation of a PCoA Motivated, UniFrac-Related Metric for Ordination

2.3.1 Introduction to the Rapid Computation of DPCoA

One of the chief applications of the UniFrac metric is as measure of dissimilarity for use in ordination techniques, such as Principle Coordinate Analysis (PCoA) (see Section 1.1.6), for the purpose of exploratory data analysis. The pairwise distances between 1,000s or 10,000s of samples are carefully computed, and then embedded on 2 or 3 dimensions in

such a manner as maximize retained variance in the measured distances between low dimensional points. This embedding necessarily just approximates the relationships between our samples, and thus we are throwing away information which we have labored to produce.

But what if we knew before hand which information we wanted to keep afterwards? What if we only computed the aspects of the metric we were actually interested in observing? As an aside to our work regarding the UniFrac metric, in the following we outline a solution to precisely that question that utilizes a UniFrac-related and biologically significant metric based in the L_2 distance between weighted differential abundance vectors. Let P and Q be relative abundances assigned to a phylogenetic tree T and let $\mathbf{W}$ be the matrix defined in Section 2.2.5 such that $\|W(P-Q)\|_{L_1}$ is the UniFrac metric between P and Q. Let $\mathbf{W}_{\sqrt{\cdot}}$ denote the matrix formed by scaling the rows of $\mathbf{W}$ by the reciprocal of the length of the corresponding edge length, so that $\mathbf{W}_{\sqrt{\cdot}}$ has rows which are indicator functions for subtrees scaled by the square root of the length of the edge defining the subtree.

In [28] Evans and Matsen noted that there was a biological significance to the quantity $\|\mathbf{W}_{\sqrt{\cdot}}(P-Q)\|_{L_2}$ albeit in the form of a integral over subtrees of a component-wise squared difference of relative abundances, and thus, as written, not a linear function of the abundances themselves. We modify the expression to suit our purposes and notation and denote it $\mathrm{d}_{UFL2}(P,Q)$, but the result remains the same. We relate the derivation of that significance in the following. Consider

$$\mathrm{d}_{UFL2}(P,Q)^2 = \|\mathbf{W}_{\sqrt{\cdot}}(P-Q)\|_{L_2}^2 \tag{2.3.1}$$

$$= \sum_{i\in T}(\mathbf{w}_i^t P - \mathbf{w}_i^t Q)^2 \tag{2.3.2}$$

$$= \sum_{i\in T}(\mathbf{w}_i^t P)^2 + (\mathbf{w}_i^t Q)^2 - 2(\mathbf{w}_i^t P)(\mathbf{w}_i^t Q). \tag{2.3.3}$$

For the sake clarity, we expand one of the above terms as an example

$$\sum_{i\in T}(\mathbf{w}_i^t P)^2 = \sum_{i\in T}\left(\sum_{j\in T}\mathbf{1}_{\pi(j,\rho)}(i)\sqrt{l(i)}P(j)\right)^2$$

$$= \sum_{i \in T} \left[\left(\sum_{j \in T} \mathbf{1}_{\pi(j,\rho)}(i) \sqrt{l(i)} P(j) \right) \left(\sum_{k \in T} \mathbf{1}_{\pi(k,\rho)}(i) \sqrt{l(i)} P(k) \right) \right]$$
$$= \sum_{i \in T} \sum_{j,k \in T} l(i) \mathbf{1}_{\pi(j,\rho) \cap \pi(k,\rho)}(i) P(j) P(k)$$

Note that if we let $a_{j,k}$ denote the last common ancestor of j and k, then $\mathbf{1}_{\pi(j,\rho)\cap\pi(k,\rho)} = \mathbf{1}_{\pi(a_{j,k},\rho)}$. Further, we have that $\mathrm{d}(a_{j,k},\rho) = (1/2) \cdot (\mathrm{d}(j,\rho) + d(k,\rho) - \mathrm{d}(j,k))$ which we utilize in continuing the above expansion

$$\begin{aligned} \sum_{i \in T} (\mathbf{w}_i^t P)^2 &= \sum_{j,k \in T} P(j) P(k) \sum_{i \in T} l(i) (\mathbf{1}_{\pi(a_{j,k},\rho)}(i) \\ &= \sum_{j,k \in T} P(j) P(k) \cdot \mathrm{d}(a_{j,k},\rho) \\ &= \frac{1}{2} \sum_{j,k \in T} P(j) P(k) \cdot (\mathrm{d}(j,\rho) + \mathrm{d}(k,\rho) - \mathrm{d}(j,k)). \end{aligned}$$

Returning to our previous work and applying the above expansion to each of the terms in Equation 2.3.3 yields

$$\begin{aligned} \mathrm{d}_{UFL2}^2 = \frac{1}{2} \sum_{j,k \in T} P(j) P(k) \cdot (\mathrm{d}(j,\rho) + \mathrm{d}(k,\rho) - \mathrm{d}(j,k)) \ldots \\ \ldots + \frac{1}{2} \sum_{j,k \in T} Q(j) Q(k) \cdot (\mathrm{d}(j,\rho) + \mathrm{d}(k,\rho) - \mathrm{d}(j,k)) \ldots \\ \ldots - \sum_{j,k \in T} P(j) Q(k) \cdot (\mathrm{d}(j,\rho) + \mathrm{d}(k,\rho) - \mathrm{d}(j,k)) \end{aligned}$$

Isolating the terms which depend upon ρ and thus only one of j or k, yields

$$\begin{aligned} \sum_{j,k \in T} \frac{1}{2} P(j) P(k) \cdot (\mathrm{d}(j,\rho) + \mathrm{d}(k,\rho)) \ldots \\ \ldots + \frac{1}{2} Q(j) Q(k) \cdot (\mathrm{d}(j,\rho) + \mathrm{d}(k,\rho)) - P(j) Q(k) \cdot (\mathrm{d}(j,\rho) + \mathrm{d}(k,\rho)) \\ = \sum_{j,k \in T} \frac{1}{2} P(j) (P(k) - Q(k)) \cdot (\mathrm{d}(j,\rho) + \mathrm{d}(k,\rho)) \\ \ldots - \frac{1}{2} Q(k) (P(j) - Q(j)) \cdot (\mathrm{d}(j,\rho) + \mathrm{d}(k,\rho)) \end{aligned}$$

$$
\begin{aligned}
&= \sum_{j\in T} \frac{1}{2} P(j) \sum_{k\in T} (P(k) - Q(k)) \cdot \mathrm{d}(k,\rho) \dots \\
&\quad \dots - \sum_{k\in T} \frac{1}{2} Q(k) \sum_{j\in T} (P(j) - Q(j)) \cdot \mathrm{d}(j,\rho) \\
&= \quad 0.
\end{aligned}
$$

Thus we have that

$$
\mathrm{d}^2_{UFL2} = \sum_{j,k\in T} P(j)Q(k) \cdot \mathrm{d}(j,k) - \frac{1}{2} \left(\sum_{j,k\in T} P(j)P(k) \cdot \mathrm{d}(j,k) + \sum_{j,k\in T} Q(j)Q(k) \cdot \mathrm{d}(j,k) \right).
$$

The treatment in [28] ends with the above statement, which has a clear biological signifigance. We weight the sum of all pairwise distances between community members by their relative abundances and then compare the average of such measurements among a pair of communities to the measurement between communities. Thus we are comparing the overall biological 'spread' in an evolutionary sense between of a pair of communities, as encoded by pairwise distances in a phylogenetic tree, against the average 'spread' of the communities themselves. It was noted in [33] that this is precisely Double Principle Coordinate Analysis (DPCoA), see Section 1.1.6. It is currently implemented as described in [33] in the *'phyloseq' R* package, a component of the Bioconductor initiative [35] for bioinformatic tools development.In [33] they note the computational inefficiency of implementing DPCoA compared with UniFrac, while using the formulation for DPCoA given by [28]. They note in 2012 a run time of approximately 40 minutes of a 32 core Linux cluster for a tree containing approximately 2500 OTUs.

We note that casting DPCoA as the L_2 distance between images under a linear transformation has computational benefits. As shown in Section 1.1.6, in the case of an L_2 distance matrix for PCoA, PCoA on a set of pairwise distances is precisely PCA on the data points themselves. Thus we can determine the principle coordinates for a matrix of pair-wise distances d_{UFL2} from the images $\mathbf{W}_{\sqrt{\cdot}}(P)$, without constructing the the quadratic in time

and space distance matrix. But, as we noted in Section 1.1.6, we can further produce the covariance matrix for the set of images $\mathbf{W}_{\sqrt{\cdot}}(P)$ from the covariance matrix for the relative abundances themselves. In the case that relative abundances are supported only on the leaves of T, this yields a matrix which is at least twice as sparse as the matrix of images and does not require computation of the action of the $\mathbf{W}_{\sqrt{\cdot}}$.

Further, given the first few principle coordinates, say $\mathbf{c}_1$ and $\mathbf{c}_2$, the products $\mathbf{W}^t_{\sqrt{\cdot}}\mathbf{c}_1$ yield the projections of the action of $\mathbf{W}^t_{\sqrt{\cdot}}$ into only the desired coordinates, thus we can compute, via a pair of inner products, precisely the PCoA transformed dataset and and no additional components. Note that any additional work in performing the eigenvalue decomposition of the matrix of relative abundances would have been embedded in the construction of a PCoA plot anyways, we have merely done the work upfront so as to avoid the construction of the full set of images and their pairwise distances.

Finally, this allows DPCoA to be cast in a similar theoretical framework as other linear transformation used for ordination, such those described in [60] by Legendre and Gallagher. They noted the use by biologists of PCA to transform raw abundance dataset, and thus the ordination of communities with respect to the L_2 distance, and presented a variety of linear transformations which allowed for the use of PCA to produce more meaningful relationships. This included the χ^2 and Hellinger distances discribed in Section 1.2.4. Regardless of the computational benefit, the above observation allows for Rao's diversity index to be included in a list of diversity metrics which are given by 'ecologically meaning transformations' of relative abundance datasets, as described defined in [60].

2.3.2 DPCoA via PCA: Description

In the following we outline the description of the algorithm to compute *DPCoA* via PCA. As we are not interested in capturing the minimizing flow underlying the metric, we utilize the action of the matrix $\mathbf{W}_{\sqrt{\cdot}}$ described in our introduction implicitly in a pair of subrou-

tines. For ease of notation we assume in the following that our underlying phylogenetic tree T is a perfect binary tree with root ρ and depth b, and thus 2^b leaves and $2^{b+1}-1$ vertices, and let S be a set of k probability distributions on T. Index the vertices of our tree, and hence the probability distributions in S, in the following manner. Let the root be 1, and for each vertex indexed by x, iteratively index each daughter vertex as $2 \cdot x$ and $2 \cdot x + 1$.

We begin by computing Σ, the covariance matrix for the elements of S. Recall that if a random vector x has covariance matrix C, then for a matrix M of appropriate dimension

$$\mathrm{cov}(Mx) = M^t C M.$$

We use Σ in this way to compute of the covariance matrix for the vectors $\mathbf{W}_{\sqrt{\cdot}}(p_i)$. We note that since the rows of $\mathbf{W}_{\sqrt{\cdot}}$ are given as scaled indicator functions for subtrees, the columns are then scaled indicator functions for paths from a given vertex to the root. Thus in computing the action of $\mathbf{W}^t_{\sqrt{\cdot}}$ on a vector, we may inductive express the value of $\mathbf{W}^t_{\sqrt{\cdot}}(i)$ in terms of the value of the vertex adjacent to i in a path to the root. By our indexing system for a binary tree, this vertex is $\lfloor (i/2) \rfloor$ for any $i > 2$.

Having performed the above computation, we then compute the singular value decomposition of $\mathrm{cov}(\mathbf{W}_{\sqrt{\cdot}}(p_i))$ to determine the principal coordinates of interest, say $\mathbf{c}_1, \ldots, \mathbf{c}_n$, of the pairwise distance under Rao's diversity index. The projection of the set of images onto those coordinates is then given by

$$\begin{aligned} c_k^t \mathbf{W}_{\sqrt{\cdot}} p_j &= (c_k^t \mathbf{W}_{\sqrt{\cdot}}) p_j \\ &= (\mathbf{W}^t_{\sqrt{\cdot}} c_k)^t p_j. \end{aligned}$$

Thus we may construct the our set of PCoA plots via inner products with the set of vectors $\mathbf{W}^t_{\sqrt{\cdot}} c_k$. In the following section we present pseudocode which expresses the above.

2.3.3 DPCoA via PCA: Algorithm

Algorithm 2.3.1. *DPCoA via PCA*

Input:

S, l ▷ *S a set of relative abundances, l the length function for the phylogenetic tree*

n ▷ *n the desired number of principle coordinates*

Iterations:

1: $\Sigma_S = \mathrm{cov}(S)$

2: $\Sigma_{WS} = \Sigma_S$

3: ***for*** $j = 2, \ldots, d$ ***do***

4: ***for*** $i = 2, \ldots, d$ ***do***

5: $\Sigma_{WS}(j,i) = \sqrt{l(i,a(i))} \cdot \Sigma_{WS}(j,i) + \Sigma_{WS}(j, \lfloor (i/2) \rfloor)$ ▷ *The action of* $\Sigma \boldsymbol{W}_{\sqrt{\cdot}}$

6: ***end for***

7: ***end for***

8: ***for*** $i = 2, \ldots, d$ ***do***

9: ***for*** $j = 2, \ldots, d$ ***do***

10: $\Sigma_{WS}(i,j) = \sqrt{l(i,a(i))} \cdot \Sigma_{WS}(i,j) + \Sigma_{WS}(i, \lfloor (j/2) \rfloor)$ ▷ *The action of* $\boldsymbol{W}^t_{\sqrt{\cdot}}(\Sigma \boldsymbol{W}_{\sqrt{\cdot}})$

11: ***end for***

12: ***end for***

13: $(c_1, \ldots, c_n) =$ *First n principal components generated from* $\mathrm{SVD}(\Sigma_{WS})$

14: ***for*** $j = 1, \ldots, n$ ***do***

15: ***for*** $i = 2, \ldots, d$ ***do***

16: $c_{w,j}(i) = \sqrt{l(i,a(i))} \cdot c_{w,j}(i) + c_{w,j}(\lfloor (i/2) \rfloor)$ ▷ *The action of* $\boldsymbol{W}^t_{\sqrt{\cdot}} c_j$

17: ***end for***

18: ***end for***

19: ***for*** $j = 1, \ldots, n$ ***do***

20: ***for*** $s \in S$ ***do***

21: $PCoA(i, s) = \langle s, c_{w,j} \rangle$ ▷ *Transformation of the dataset into the principal coordinates*

22: ***end for***

23: ***end for***

Output:

$PCoA$

2.4 Expectation of the UniFrac Metric

Having alternate characterizations and efficient computation of the UniFrac metric, as well as noted the application of our results to the related ordination technique DPCoA, we we change our focus to understanding the expected value of the UniFrac metric. In this section we generate expressions for the probability density function of the UniFrac metric under the assumption of a frequently employed distribution for the modeling of metagenomic datasets. We determine $\mathbb{E}(\text{UniFrac}(P, Q))$ when P and Q are drawn from a Dirichlet distribution (Definition 1.1.20) using previous work on the relationship between the difference of Beta distributed random variables (Defintion 2.4.1).

2.4.1 Application of the Dirichlet Distribution in UniFrac for Dirichlet-Multinomial Distributed Sequence Data

While our efforts to speed computation of the UniFrac metric make analysis of significance in measured UniFrac distances via Monte Carlo methods more tractable, it is tempting to see if we can derive exact expressions for the expected value of the UniFrac metric given a model of the underlying distribution of OTUs on a phylogenetic tree.

As we noted in Section 1.1.6, the Dirichlet-Multinomial (Definition 1.1.21) is a common probabilistic model for the distribution of sequence read data. Given a fixed phylogenetic tree T, and a set S of sequence reads, we select a subset T_S of the nodes of T, thus a collection of taxa or OTUs for the phylogenetic tree, upon which the sample generating the sequence data is to be supported. This is typically some fixed taxonomic depth in the tree to which the use of 16S rRNA analysis will assign sequences. We then draw a Dirichlet distributed probability that each sequence read s is assigned to some node $v \in T_S$. The assignment of the set of sequences to nodes is then a Multinomial distributed random variable.

What we have at the end of this model is a distribution for sequence assignments, not for the vector of relative abundances which forms the basis for UniFrac distances. Supposing that $T_S = n$ and letting $p = (p_1, \ldots, p_n)$ be the probability distribution which forms the parameter for the Multinomial and $t = (t_1, \ldots, t_n)$ be the random vector of sequence counts, we note that the marginal distributions for each component t_i is the Binomial with parameter p_i. Considering our Binomially distributed marginals as a sequence of independent Bernoulli trials, a standard application of the central limit theorem yields that t_i/n is normally distributed with expectation $\mathbb{E}(t_i/n) = (n \cdot p_i)/n = p_i$ and variance $Var(t_i/n) = (p_i \cdot (1 - p_i))/n$. In the case of sequence read data, n stretches into the 100,000s for the metagenomic coverage generated by modern sequencing techniques. In these circumstances the Dirichlet prior for the Dirichlet-Multinomial is an excellent model for the UniFrac metric, one which is compatible with methods being utilized currently by researchers.

In the following section we utilize properties of the Dirichlet distribution and its marginal distribution, the Beta distribution, to generate formulas for the expectation of Dirichlet distributed relative abundance datasets.

2.4.2 Derivation of Expected Values for the UniFrac Metric

By utilizing our characterization of the UniFrac metric between relative abundance vectors P and Q as $\|W(P-Q)\|_{L_1}$, we can generate formulations for the expectation of the UniFrac metric between P and Q by first exploring the distribution of $P - Q$ when P and Q are Dirichlet distributed random variables.

For ease of use, we recall that we say a random variable $X = (X_1, \ldots, X_n)$ is Dirichlet distributed, $X \sim \mathrm{Dir}(X, \alpha)$, if it is supported on the interior of the unit simplex in $\mathbb{R}^n$ and has probability density function given by

$$f(x_1, \ldots, x_{n-1}; \alpha_1, \ldots, \alpha_n) = \frac{1}{\mathrm{B}(\alpha)} \prod_{i=1}^{n} x_i^{\alpha_i - 1}$$

for shape parameters $\alpha = (\alpha_1, \ldots, \alpha_n) > 0$. Note that the x_i satisfy $x_n = 1 - \sum_{i=1}^{n-1} x_i$, $x_1, ..., x_{n-1} > 0$ and $\sum_{i=1}^{n-1} x_i < 1$, where the normalizing constant $\mathrm{B}(\alpha)$ is given by the multivariate Beta function (Definition 1.1.19).

We first define the *Beta distribution* [32].

Definition 2.4.1 (Beta distribution)**.** The *Beta distribution* $\mathrm{Beta}(\alpha, \beta)$ is a probability distribution supported on the unit interval and has probability density function

$$f(x; \alpha, \beta) = \frac{1}{\mathrm{B}(\alpha, \beta)} x^{\alpha - 1}(1 - x)^{\beta - 1}$$

for shape parameters $\alpha, \beta > 0$.

The Beta distribution bears more than a passing similarity to the Dirichlet distribution, a connection we formalize now. We now describe the marginal distributions of the components of a Dirichlet distributed random variable [32].

Proposition 2.4.1 (Marginals of Dirichlet distribution [32])**.** *Suppose a* $(X_1, \ldots, X_n)$ *is Dirichlet distributed with shape parameters* $\alpha_1, \ldots, \alpha_n$*. Let* $\alpha_0 = \sum_{i=1}^{n} \alpha_i$*. Then the marginal distribution of* X_j *is* $\mathrm{Beta}(\alpha_j, \alpha_0 - \alpha_j)$.

As we are interested in the image of our random variable under a linear transformation, we recount the following very useful *aggregation property* of the Dirichlet distribution [32].

Proposition 2.4.2 (Dirichlet aggregation property [32]). *Say $X = (X_1, \ldots, X_n)$ is Dirichlet distributed with shape parameters $(\alpha_1, \ldots, \alpha_n)$. If the random variable X' is constructed by omitting x_i, x_j from X and replacing with $x_i + x_j$, that is $X' = (x_1, ..., x_i + x_j, ..., x_n)$ then X' is Dirichlet distributed with shape parameters $(\alpha_1, ..., \alpha_i + \alpha_j, ..., \alpha_n)$.*

Now suppose that we are considering the UniFrac metric over tree T with $n + 1$ edges, thus n nodes and the trivial edge-weight. Thus the matrix W defined in Section 2.2.5 has $n + 1$ rows, each row being the indicator function for a given subtree in T. Let w_k, for $1 \leq k \leq n+1$ be the rows of W. Further, say that the relative abundance vector P is drawn from a Dirichlet distribution with shape parameters $\alpha = (\alpha_1, ..., \alpha_n)$. Let $\alpha_0 = \sum_{i=1}^{n} \alpha_i$. The following proposition follows directly from the aggregation property for the Dirichlet distribution

Proposition 2.4.3 (McClelland 2018). *$WP(k)$, the kth component of the vector WP, is distributed as*

$$WP(k) \sim \text{Beta}(\langle \alpha, w_k \rangle, \alpha_0 - \langle \alpha, w_k \rangle).$$

given the above definitions.

Thus determining the distribution of the elements of the differential abundance vector $W(P-Q)$ relies on the determination of the distribution of the difference of beta variables. We follow the treatment in [93] to determine the distribution. As a step toward that goal we first define the *Pochhammer symbol* $(a)_m$.

Definition 2.4.2 (Pochhammer symbol). Define the *Pochhammer symbol* $(a)_m$ by

$$(a)_m = \frac{\Gamma(a+m)}{\Gamma(a)}$$

for $m > 0$ and $(a, m) = 1$ for $m = 0$.

We now define *Appell's First Hypergeometric* function F_1.

Definition 2.4.3. *Appell's first hypergeometric function* F_1 is given by

$$F_1(a, b_1, b_2; c; x, y) = \sum_{n=0}^{\infty} \sum_{m=0}^{\infty} \frac{(a)_{m+n}}{(c)_{m+n}} (b_1)_m (b_2)_n \frac{x_1^m}{m!} \frac{x_2^n}{n!}$$

It can be shown [29] that F_1 converges for $|x_1|, |x_2| < 1$. Picard [94] derived the following integral expression for F_1, which is useful for our purposes.

Theorem 2.4.1 (Picard 1881)**.** *Let* a, b_1, b_2, c *be complex numbers. If* $Re(a), Re(c-a) > 0$ *and* $F_1(a, b_1, b_2; c; x_1, x_2)$ *converges then [94]*

$$F_1(a, b_1, b_2; c; x_1, x_2) = \frac{\Gamma(c)}{\Gamma(a)\Gamma(c-a)} \int_0^1 u^{a-1}(1-u)^{c-a-1}(1-ux_1)^{-b_1}(1-ux_2)^{-b_2} du$$

The following identities [29] related to F_1 will also be required. We have that

$$\begin{aligned} F_1(a, b_1, b_2; c; x, y) &= (1-x)^{c-(a+b_1)}(1-y)^{-b_2} \\ &\qquad \cdot F_1(c-a, c-(b_1+b_2), b_2; c; x(y-x)/(y-1)) \qquad (2.4.1) \\ &= (1-x)^{-b_1}(1-y)^{c-(a+b_2)} \\ &\qquad \cdot F_1(c-a, b_1, c-(b_1+b_2); c; (x-y)/(x-1), y). \qquad (2.4.2) \end{aligned}$$

We are now prepared for a description of the density of a difference of Beta distributed random variables.

Theorem 2.3 (Pham-Gia 1993). Say X, Y are independent random variable with $X \sim \text{Beta}(\alpha_X, \beta_X)$ and $Y \sim \text{Beta}(\alpha_Y, \beta_Y)$. Let $A = B(\alpha_X, \beta_X) \cdot B(\alpha_Y, \beta_Y)$. Then [93] $D = X - Y$ has probability density function

$$f(d) = \begin{cases} \frac{1}{A} B(\alpha_X, \beta_Y) d^{\beta_X + \beta_Y - 1}(1-d)^{\alpha_Y + \beta_X - 1}. & \text{for } 0 < d \leq 1 \\ \quad F_1(\beta_X, \alpha_X + \beta_X + \alpha_Y + \beta_Y - 2, 1-\alpha_X; \beta_X + \alpha_Y; (1-d), 1-d^2) & \\ \frac{1}{A} B(\alpha_X, \beta_Y)(-d)^{\beta_X + \beta_Y - 1}(1-d)^{\alpha_Y + \beta_X - 1}(1-d)^{\alpha_X + \beta_Y - 1}. & \text{for } -1 \leq d < 0 \\ \quad F_1(\beta_Y, 1-\alpha_Y, \alpha_X + \beta_X + \alpha_Y + \beta_Y - 2; \alpha_X + \beta_Y; 1-d^2, 1+d) & \end{cases}$$

In the case that $\alpha_X + \alpha_y > 1$ and $\beta_X + \beta_Y > 1$ then

$$f(0) = \frac{1}{A} B(\alpha_X + \alpha_Y - 1, \beta_X + \beta_Y - 1)$$

Proof. The density of $D = X - Y$ is given by the convolution of the densities for each of X and Y. Thus for $0 < d \leq 1$ we have that

$$\begin{aligned} f(d) &= \frac{1}{A} \int_0^{1-d} (d+x)^{\alpha_X - 1}(1-d-x)^{\beta_X - 1} x^{\alpha_Y - 1}(1-x)^{\beta_2 - 1} \mathrm{d}x \qquad (2.4.3) \\ &= \frac{1}{A} d^{\alpha_X - 1}(1-d)^{\beta_X - 1} \int_0^{1-d} x^{\alpha_Y - 1}(1-x)^{\beta_Y - 1}(1 + \frac{x}{d})^{\alpha_X - 1}(1 - \frac{x}{1-d})^{\beta_X - 1} \mathrm{d}x \end{aligned}$$

Changing variables so that $w = x/(1-d)$ yields

$$\begin{aligned} f(d) = \frac{1}{A} d^{\alpha_X - 1}(1-d)^{\beta_X - 1}(1-d)^{\alpha_Y - 2} \cdot \\ \int_0^1 w^{\alpha_Y - 1}(1-(1-d)w)^{\beta_Y - 1}(1 - \frac{w(d-1)}{d})^{\alpha_X - 1}(1-w)^{\beta_X - 1} \mathrm{d}w. \end{aligned}$$

After simplyfing and applying Picard's theorem, we can express the integral in terms of F_1 so that

$$\begin{aligned} f(d) = \frac{1}{A} \cdot d^{\alpha_X - 1}(1-d)^{\alpha_Y + \beta_X - 1} B(\alpha_Y, \beta_X) \cdot \\ F_1(\alpha_Y, 1-\beta_Y, 1-\alpha_X; \beta_X + \alpha_Y; 1-d, 1-\frac{1}{d}). \end{aligned}$$

We now apply Equation 2.4.1 above to yield

$$\begin{aligned} f(d) = \frac{1}{A} B(\alpha_X, \beta_Y) d^{\beta_X + \beta_Y - 1}(1-d)^{\alpha_Y + \beta_X - 1} \\ F_1(\beta_X, \alpha_X + \beta_X + \alpha_Y + \beta_Y - 2, 1-\alpha_X; \beta_X + \alpha_Y; (1-d), 1-d^2), \end{aligned}$$

our desired result. Note that $|1-d|, |1-d^2| < 0$ so that F_1 converges. The proof for $-1 \leq d < 0$ is analogous.

In the case that $\alpha_X + \alpha_Y > 1$ and $\beta_X + \beta_Y > 1$, setting $d = 0$ in Equation 2.4.3 we have that

$$f(0) = \frac{1}{A} \int_0^1 x^{\alpha_X - 1}(1-x)^{\beta_X - 1} x^{\alpha_Y - 1}(1-x)^{\beta_2 - 1} \mathrm{d}x$$

$$= \frac{1}{A}\int_0^1 x^{(\alpha_X+\alpha_Y-1)-1}(1-x)^{(\beta_X+\beta_Y-1)-1}$$
$$= \frac{1}{A}B(\alpha_X+\alpha_Y-1, \beta_X+\beta_Y-1)$$

□

As an example of utilizing the above, consider a perfect binary tree of depth 2, and thus 4 leaves. Let P, Q be Dirichlet distributed relative abundance vectors, supported only on the leaves and with shape parameter uniformly 1. This choice of parameters represents a uniform random distribution for OTUs at the leaves. Letting $X_i = P_i - Q_i$, for $1 \leq i \leq 4$, X_i has density

$$f(x_i) = \begin{cases} \frac{3}{10}(1+x_i)^3(6-3x_i+x_i^2) & \text{for } -1 \leq x < 0 \\ \frac{3}{10}(1-x_i)^3(6+3x_i+x_i^2) & \text{for } 0 \leq x < 1 \end{cases}$$

and thus $\mathbb{E}[|X_i|] = \frac{3}{14}$.

Letting $Y_{i,i+1} = (P_i + P_{i+1}) - (Q_i + Q_{i+1})$, $Y_{i,i+1}$ has density

$$f(y_{i,i+1}) = \begin{cases} \frac{6}{5}(1+y_{i,i+1})^3(y_{i,i+1}^2-3y_{i,i+1}+1) & \text{for } -1 \leq x < 0 \\ \frac{6}{5}(1-y_{i,i+1})^3(y_{i,i+1}^2+3y_{i,i+1}+1) & \text{for } 0 \leq x < 1 \end{cases}$$

and thus $\mathbb{E}[|Y_{i,i+1}|] = \frac{9}{35}$.

Thus we have that

$$\mathbb{E}[\text{UniFrac}(P,Q)] = \mathbb{E}[|X_1| + ... + |X_4| + |Y_{1,2}| + |Y_{3,4}|] = \frac{48}{35}$$

Utilizing the implementation of Appell's First Hypergeometric function F_1 in Wolfram Mathematica, we applied the above to perfect binary trees of depths 2 through 6. The results are shown in Table 2.1. As consequence of Theorem 1.2.7, an upper bound for the 1-Wasserstein metric, and thus UniFrac, is the diameter of the tree. The case of our binary

trees, that equates to twice the depth. So as to compare values lying in the range of 0 to 1 for a variety of trees, we have also included the expectation normalized by the diameter of the tree in Table 2.1.

TABLE 2.1: Expected values for the UniFrac metric and the UniFrac metric normalized by the tree diameter between a pair of Dirichlet distributed relative abundance vectors, with shape parameters chosen so as to represent a uniform random distribution on the leaves of perfect binary trees of depths 2-6.

Depth	Expectation	Normalized Expectation
2	1.37143	0.342857
3	1.96022	0.326703
4	2.41646	0.302057
5	2.75824	0.275824
6	3.00946	0.250788

In the above we has presented a solution to the expectation for a collection of particularly simple examples. The application to other, more complicated examples is no more complicated. Note that differences in branch length merely scale the summands in the formula for the expectation above and so are easily incorporated. Having in this section determined an expression for the expectation of the UniFrac metric in the case of Dirichlet distributed relative abundances, we now turn to applying our efficient algorithm for the the computation of the UniFrac metric itself to actual datasets.

2.5 Applications

In the following we demonstrate the utility of applying our results related to the Wasserstein metric and UniFrac to datasets, both real-world and synthetic.

2.5.1 Application of EMDUniFrac to Data

To demonstrate the utility of EMDUniFrac, we utilized it to analyze a real world 16S rRNA dataset from a previous study [127]. The original dataset consists of 454 pyrosequenced fecal samples from a cohort of 40 twin pairs. We utilized phylogenetic tree classifications from QIIME/QIITA [15]. For simplicity, we focused on the phylum level, and so summed classifications to this level. From the dataset of 454 samples we selected a subset consisting of 49 healthy samples and 16 ulcerative colitis samples and used the silva taxonomic tree [133] for the EMDUniFrac computation.

We evaluated the EMDUniFrac algorithm on all $\binom{65}{2} = 2,080$ pairs of samples and performed principle coordinate analysis (PCoA) on the resulting distance matrix. The result of this is contained in part (A) of Figure 2.1. Next, we combined all the healthy samples and combined all the ulcerative colitis samples and evaluated EMDUniFrac on these two combined samples. The returned minimizing flow is depicted in part (B) of Figure 2.1. The corresponding weighted differential abundance vector is shown in part (C).

2.5.2 Comparison of EMDUniFrac to Alternate Solution Methods for UniFrac

As modern comparative metagenomics studies often perform all pairwise UniFrac distance computations for datasets consisting of tens to thousands of samples, it is important to compute such distances in an efficient manner. As we showed in Section 2.2.4, Algorithm 2.2.1 used to compute EMDUniFrac runs in space and time complexity linear in the total support of the input vectors, and thus less than or equal to the number of vertices in the tree.

To assess practical performance of Algorithm 2.2.1, we compared it to the fastest previous implementation of UniFrac, called FastUniFrac [43]. We randomly generated trees (using the ete2 toolkit [49]) with the number of leaf vertices ranging from 10 to 90,000. We then randomly produced pairs of distributions on the leaves using an exponential distribution

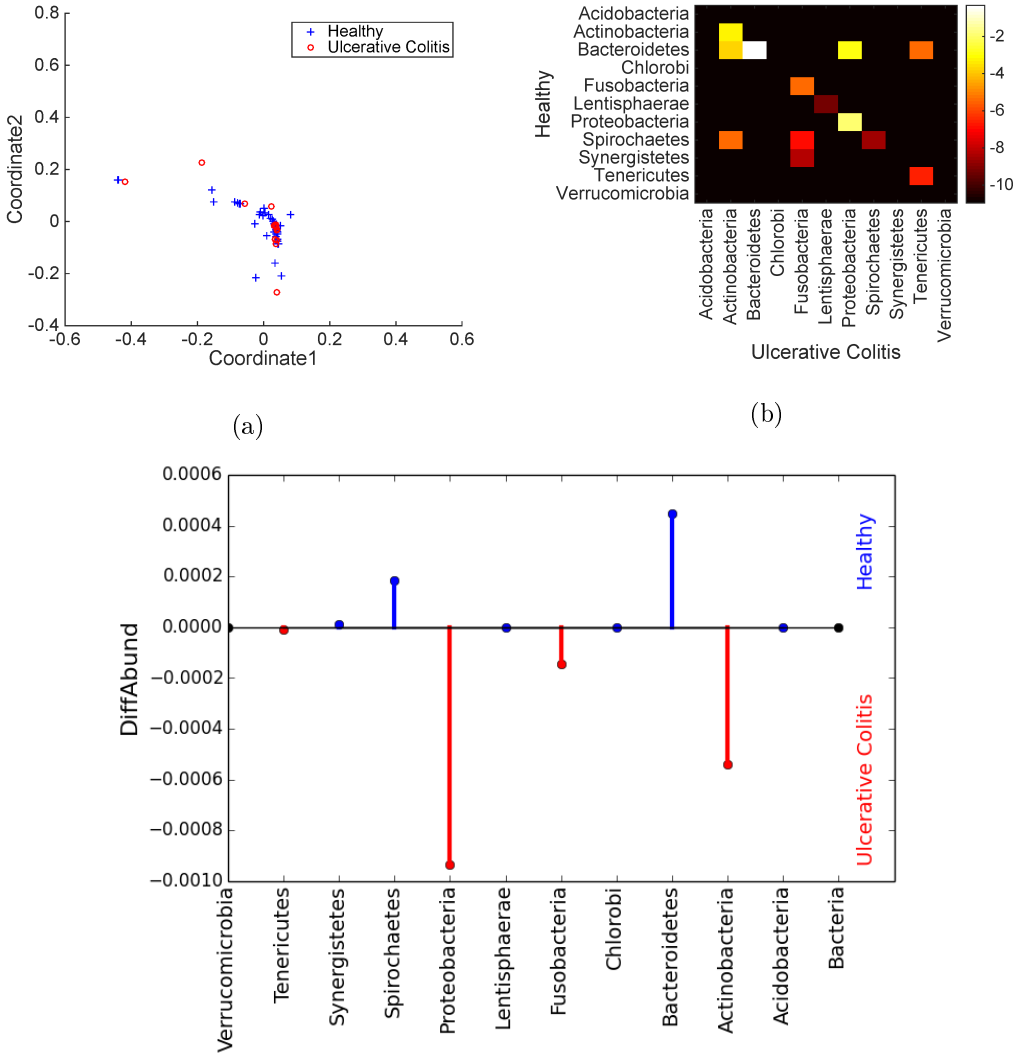


FIGURE 2.1: Results of the application of EMDUniFrac on real-world data. Part (A) is the PCoA plot of the EMDUniFrac distance matrix between all pairs of samples analyzed. Compare to the similar plot in Figure 2 of [127]. Part (B) contains a heat map of the minimizing flow for the combined healthy and ulcerative colitis samples. This heat map is scaled logarithmically for visualization purposes. Part (C) depicts the differential abundance vector between the combined healthy and Ulcerative Colitis samples and indicate which organisms are differentially abundant in the samples, demonstrating usefulness over the PCoA plot in part (A).

with scale parameter 1. Importantly, EMDUniFrac can handle distributions with weights on leaf vertices as well as internal vertices while FastUniFrac only allows distributions with weights on the leaf vertices. We performed 10 replicates for each number of tree leaves and 10 replicates for each tree topology.

Using the same fixed computational resources, we then ran FastUniFrac, EMDUniFrac in a mode that computes and returns the optimal flow as given in Algorithm 2.2.1, and EMDUniFrac in a mode that compute only the distance as given in Algorithm 2.2.2, and thus not an optimal flow, so as to return output identical to that of FastUniFrac. The average timings (over each number of tree leaves) are depicted in Figure 2.2.

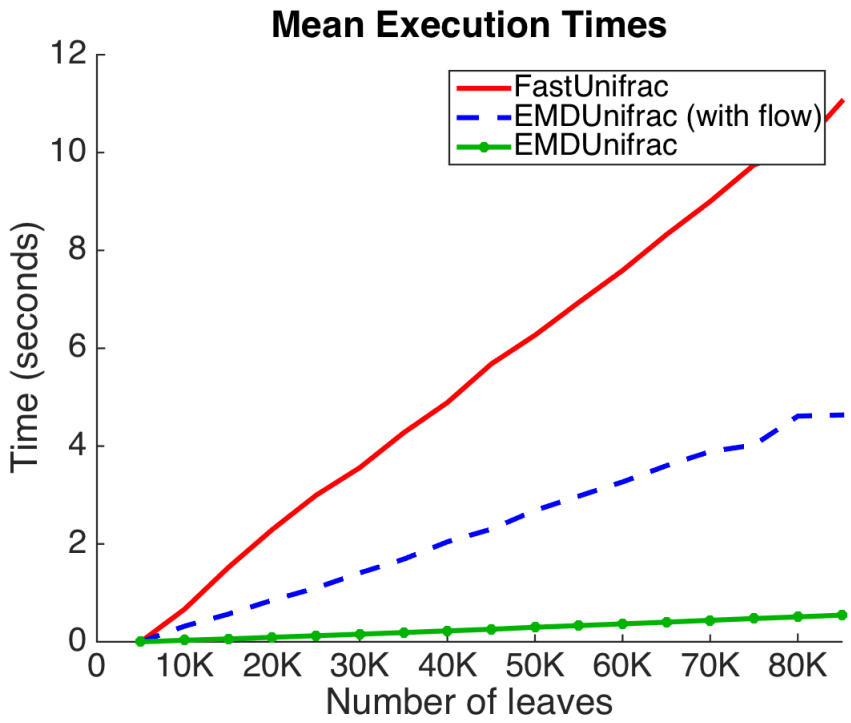


FIGURE 2.2: Speed comparison of FastUniFrac to EMDUniFrac (while also returning the minimizing flow) and EMDUniFrac (while returning just the distance). Trees are generated with random topology and abundances are random realizations of an exponential distribution and are supported on the leaves.

2.6 Discussion

2.6.1 Results

Having applied EMDUniFrac to real-worlds datasets and compared the efficiency of the algorithm against the fastest previous implementation of the UniFrac metric, we now interpret the results of our applications to real and synthetic datasets.

Even though upon visual inspection, the PCoA plot in part (A) of Figure 2.1 does not show much distinction between healthy and ulcerative colitis samples (compare to the similar plot contained in Figure 2 of [127]), the differential abundance vector leads to the immediate conclusion that the ulcerative colitis samples are primarily enriched for Actinobacteria and Proteobacteria, while being deficient in Bacteroidetes. This observation is consistent with other studies where the same trend was observed in irritable bowel disease subjects, but using alternate analysis techniques [30, 109, 74], and demonstrates how utilizing the minimizing flow results in more information than simply using an ordination technique (here PCoA) on the pairwise UniFrac distances.

These results from our comparison of indicate that in either mode, EMDUniFrac is more computationally efficient than FastUniFrac, and when just the resulting distance is desired and thus Algorithm 2.2.2 is utilized, EMDUniFrac takes less than half a second to run, even on trees with 90,000 leaves. Note that our implementations of EMDUniFrac are non-optimized, Python implementations.

2.6.2 Future Work

There are multiple avenues for continued research in understanding the UniFrac metric from a mathematical perspective. While faster direct computation seems unlikely, ways to rapidly estimate the metric seem feasible. Our observation regarding computation of DPCoA was borne out of attempts to perform the same mathematical sleight of hand,

via PCA on a dataset of transformed relative abundances, to avoid generating billions of pairwise distances when seeking PCoA plots for tens of thousands of samples. There do exists techniques for generating L_1 versions of PCA, such as described in [12, 79], which seek coordinate transformations which embed a dataset in lower dimensions in such a way as to preserve the L_1 distances between datapoints. Understanding the relationship between such coordinate transformations and the UniFrac metric is a source for future work, in addition to implementation and exploration of our observations regarding DPCoA itself.

Additionally, and perhaps as more of a mathematical curiosity than anything, while considering the difference of Dirichlet distributed random variable under the UniFrac metric, definite patterns were observed in the structure of the rational functions which give the densities. Attempts to determine a more succinct expression for the densities cost the author more than a few moments of thought and is likely to be something which will be consider when less pressing matters are at hand. In the same vein, but perhaps more biologically signifigant, would be the application of our results related to the structure of the UniFrac metric and its computation to other models for sequence count datasets.

Finally, there exists alternate mathematical frameworks in which to consider relative abundances, particularly that of *compositional data analysis (CDA)* [2]. We have not touched on these tools for understanding relative abundances in our treatment of metagenomics or the UniFrac metric, as our results and considerations have not utilized them. They present an alternate approach to the consideration of relative abundances as probability distributions. Extremely briefly, one can consider CDA as 'projective geometry for geologists', in which we view our relative abundances as equivalence classes of proportions. From this framework the probability distributions we have utilized are just one possible normalization. The techniques were born out of the statistical analysis of geologic datasets and give a means by which to consider the interior of the unit simplex as a Hilbert space. While

these techniques have definite drawbacks, the requirement of nonnegativity on the part of the components being a particularly biologically egregious one, they offer theoretical benefits and an alternate mathematical structure for the problem of comparing proportions arising in metagenomics. There has been recent work on this subject [103]. Finding ways to incorporate the framework of CDA into the mathematics behind the UniFrac metric, or the highlight the ways in which the techniques are complementary, is a goal for future research.

With that, we conclude our discussion of the UniFrac metric and turn to the formulation of a novel β-diversity metric, based in another application of the Wasserstein metric.

3 REFERENCE-FREE METAGENOMIC COMPARISON USING THE WASSERSTEIN METRIC

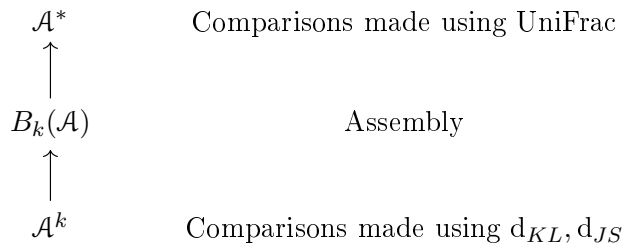
3.1 Introduction

In the work previous to this section we have considered UniFrac, a β-diversity metric between microbial communities defined by relative abundances assigned to some common phylogenetic tree. In Section 1.1.2 we noted that species or OTU identification in a metagenomic sample is hampered by challenges related to the culturing of microbial species, the lack of reference databases and the overlap between related species. These challenges motivate reference-free methods of comparison between metagenomic sample, like those described in Sections 1.1.4 and 1.2.4.

These reference-free methods act on probability distributions defined on k-mers generated from sets of sequence reads, as described in Section 1.1.2, including examples such as Jensen-Shannon (d_{JS}, Definition 1.2.4) divergence. As we noted in Section 1.3.2, there is another utility for k-mers, that of genome assembly. Distributions of k-mers are assigned to a de Bruijn graph in such a way that a solution to the Eulerian path problem corresponds to an assembly of a component of a genome or, in the case of much smaller assemblies, reads are fitted together via overlap-layout-consensus (OLC) methods by finding agreement between read ends. Such assemblies are the source of OTU identification and relative abundance estimation, and thus the input to an analysis via UniFrac.

Recalling the notation of Section 1.3.4, we can therefore consider the methods described thus far as making comparisons at either the top or the bottom of the diagram in Figure 3.1.

FIGURE 3.1: Depiction of relationships between β-diversity metrics and genomic assembly.



At the bottom, we compare distributions on $k-$mers, elements of $\mathcal{A}^k$ for the genetic alphabet $\mathcal{A} = \{A, C, G, T\}$, directly, by methods which are ignorant of their origins as factors of a genome. Alternately, we lift collections of $k-$mers to assemblies, elements of $\mathcal{A}^*$, by way of the de Bruijn graph $B_k(\mathcal{A})$ or the OLC methods described in Section 1.1.2, at the top of the diagram, which we then compare via β-diversity metrics on relative abundances such as UniFrac. Clearly each method has benefits, UniFrac in informed by the biological context of proximity in a phylogenetic tree and the reference-free methods allow for comparisons between datasets for which less background information is known.

In this section we prove that for k sufficient to allow unambiguous assembly of a metagenomic dataset via de Bruijn graphs, convergence of a set of measures defined on assemblies derived from that dataset is equivalent to the convergence of the set of derived measure with respect to the $k-$mer occurrence defined on the set of metagenomic sequence reads. We then use this to motivate $EMDeBruijn$, a reference-free metric between metagenomic datasets which makes β-diversity comparisons by utilizing the structures used in assembly, but without preforming the assembly itself. $EMDeBruijn$ utilizes the Wasserstein metric between $k-$mer occurrence distributions defined on metagenomic samples, in which the underlying ground metric between k-mers is derived from the de Bruijn graph. Due to the infeasibility of exact computation of the Wasserstein metric in this setting, we introduce a pair of approximations, one heuristic and another borrowed from the burgeoning field

of image processing via the Wasserstein metric. We then apply these approximations to *EMDeBruijn* to real-world datasets. The results of these applications are then compared against alternate methods for reference-free comparison. Finally, we explore alternate ground metrics for use in Wasserstein-based reference-free comparison of metagenomic sample.

3.1.1 Motivation for a Reference-free Wasserstein Metric on Metagenomic Datasets

As we have shown, the UniFrac metric is the 1-Wasserstein metric when we take a ground distance between OTUs given by path length in a common phylogenetic tree. As we commented in Section 1.1.3, the edge-lengths of a phylogenetic tree are given in terms of an expected number of substitutions per location in the genomes between organisms. Formalizing this, given an OTU with genome $u \in \mathcal{A}^*$, let u_a be the most recent ancestor of u. Then the length of an edge between u and u_a is given by

$$l(u, u_a) = \frac{s(u, u_a)}{|u|}$$

where $s(u, u_a)$ is the expected or estimated number of substitutions between the genomes of u and u_a. This over simplifies the situation somewhat, but, in microbial ecology in particular, agreement between genomes or genomic regions such as the 16s rRNA gene as described in Section 1.1.2 is taken as the very definition of an OTU and is the source for the information used in the construction of phylogenetic trees.

Thus the edge-length in a phylogenetic tree can be interpreted as a normalized version of an edit distance (Definition 1.3.15) between genomes. In this light we can interpret the evolutionary distance given by path length in a phylogenetic tree as a constrained edit distance between genomes, constrained in that we are obligated to compute the evolutionary distance with respect to a sequence of substitutions that pass through the most recent common ancestor of a pair of OTUs. It follows that we can view the UniFrac metric as

measuring a quantity related to the 1-Wasserstein distance between genomes for a ground edit distance.

This motivates considering the following abstraction which relates convergence of measures defined on a finite subset of $S \subset \mathcal{A}^*$ with respect to the Wasserstein metric and the convergence of the projection of those measures into $\mathcal{A}^k$, when k is chosen so that assembly via de Bruijn graphs is unambiguous. The spirit of the statement below is summarized as follows. If we are interested in comparing probability distributions on sets of genomes via the Wasserstein distance and we are able to unambiguously assemble the sets of genomes given a collection of k-mers, then we can recover a notion of closeness from comparing the distributions given by the k-mers themselves, without actually performing the assembly.

Recall that for words v and w the juxtaposition vw denotes concatenation of the symbols involved. Further recall that we say a word w is a *right-extension* of v if $w = v\alpha$ for some $\alpha \in \mathcal{A}$ and w is a *left-extension* of v if $w = \alpha v$ for some $\alpha \in \mathcal{A}$.

Theorem 3.1.1 (McClelland 2018)**.** *Let $S \subset \mathcal{A}^*$ be finite. Let $m = \min_{s \in S} |s|$, and suppose that there exists $k \leq m$ such that each factor of length k in S has a unique, possibly empty, left and right extension to factors of length $k+1$ in S.*

Let S_k be the set of k-mers occurring as factors in the elements of S. Let $M(S)$ and $M(S_k)$ be the set of measures defined on S and S_k, respectively. Let $\pi_k : M(S) \to M(S_k)$ be the projection of measures such that $\pi_k(\mu) = \nu$ for

$$\nu(v) = \sum_{w \in S} \frac{\mu(w) \cdot occ_w(v)}{|w| - k + 1}$$

where $occ_w(v)$ is the occurrence function (Defintion 1.3.10) which counts the number of instances of v in w. Let $W_{1,*}$ and $W_{1,k}$ denote the 1-Wasserstein metric on $M(S)$ and $M(S_k)$, respectively, with respect to an arbitrary pair of ground distances.

Then a sequence of measures $\{\mu_i\}_{i=1}^{\infty}$ converges in $M(S)$ to a measure μ with respect to $W_{1,*}$ if and only if the sequence of projections $\{\pi_k(\mu_i)\}_{i=1}^{\infty}$ converges in $M(S_k)$ to $\pi_k(\mu)$

with respect to $W_{k,*}$.

We first require an observation regarding uniqueness of factors which mirrors the use of de Bruijns graphs in assembly.

Lemma 3.1.1 (McClelland 2018)**.** *Given the above hypotheses and notation, for each $a \in S_k$ there exists a unique $b \in S$ such that a is a factor of b.*

Proof. This follows from our property regarding uniqueness of extensions as follows. Let $a = a_1 \dots a_k \in S_k$ be arbitrary. Then either a has a unique right-extension to a factor $a\alpha_1$ for some $\alpha_1 \in \mathcal{A}$ or a is a suffix to some word in S. If a is not a suffix, then $a_2 \dots a_k\alpha_1 \in S_k$ and so has either has a unique right-extension or is a suffix. Inductively, we may keep appending the unique elements extending our word, to say $a_1a_2 \dots a_k\alpha_1 \dots \alpha_l$, until we necessarily terminate in a suffix of some word in S. Let $\alpha = \alpha_1 \dots \alpha_l$.

On the other hand, a is either the prefix of some word in S or there exists a unique left-extension of a by some $\beta_1 \in \mathcal{A}$, so that $\beta_1 a$ is a factor to some word in S. Thus we inductively append $\beta_m\beta_{m-1} \dots \beta_1$ to a before terminating in a prefix. Letting $\beta = \beta_m\beta_{m-1} \dots \beta_1$, we thus have that $\beta a\alpha$ is both a suffix and prefix to some word in S, that is $\beta a\alpha \in S$. Thus $\beta a\alpha$ is the unique element of S containing a as a factor. □

We now prove our result.

Proof. Let d_* be the ground metric for $W_{1,*}$ and let

$$d_{\min,*} = \min_{s,t \in S, s \neq t} \mathrm{d}_*(s,t).$$

Note that as our metric spaces are finite, convergence with respect to $W_{1,*}$ implies the pointwise convergence of measures in $\mathbb{R}^{|S|}$ as the distance between elements of S is bounded below by $d_{\min,*}$.

Further, the function π_k is clearly continuous from $\mathbb{R}^{|S|}$ to $\mathbb{R}^{|S_k|}$, as in each component it is a fixed linear combination of the measures of the elements of S. Thus convergence of $\{\mu_i\}_{i=1}^{\infty}$ in $M(S)$ to a measure μ implies convergence of $\{\pi_k(\mu_i)\}_{i=1}^{\infty}$ to $\pi_k(\mu)$ in $M(S_k)$.

We prove the alternate implication by considering the contrapositive. Suppose that we have that $W_{1,*}(\mu_n, \mu) > \delta$ for some $\delta > 0$. By the same considerations as above regarding the finite nature of out metric space, this implies that there exists $v \in S$ such that

$$|\mu(v) - \mu_n(v)| \geq \frac{\delta}{d_{\min,*} \cdot |S|}.$$

This follows from considering the Wasserstein metric as a sum $|S|$ differences in measure scaled by the distanced required for transport in some minimizing flow. Now let $\pi_k(\mu_n) = \nu_n$ and $\pi_k(\mu) = \nu$.

We claim that the above implies that there exists $w \in S_k$ such that

$$|\nu_n(w) - \nu(w)| > \frac{\delta}{(m-k+1) \cdot d_{\min,*} \cdot |S|},$$

recalling that $m = \min_{s \in S} |s|$. As this quantity depends only on δ, our result follows.

From our lemma we see that for each factor w of v we have that

$$\nu_n(w) = \frac{occ_v(w)\mu_n(v)}{|v| - k + 1}$$

as w occurs as a factor in no other v. Thus we have that

$$\begin{aligned}
|\nu_n(w) - \nu(w)| &= \left| \frac{occ_v(w)\mu_n(v)}{|v| - k + 1} - \frac{occ_v(w)\mu(v)}{|v| - k + 1} \right| \\
&= \frac{occ_v(w)}{|v| - k + 1} |\mu_n(v) - \mu(v)| \\
&> \frac{occ_v(w)}{|v| - k + 1} \frac{\delta}{d_{\min,*} \cdot |S|} \\
&\geq \frac{\delta}{(m-k+1) \cdot d_{\min,*} \cdot |S|}
\end{aligned}$$

where we have used that $occ_v(w) \geq 1$ and that $|v| \leq m$ in the last inequality. Thus the claim, and so the statement, is proved. $\square$

While the above gives weight to the notion of Wasserstein-based comparisons of k-mers distributions arising from metagenomic datasets by relating the notion to the very successful UniFrac metric, it does not address why such a Wasserstein metric is useful for these sorts of datasets to begin with. A strong analog exists between these applications in metagenomics and the application of the Wasserstein metric, generally under the name of the Earth mover's distance, to image analysis, as described in Section 1.2.5. As noted in that section, the Earth mover's distance has been useful in those contexts in part because it is capable of taking large, noisy datasets, images, and determining the relative similarity of large scale structures given comparisons of purely local features, such as the color composition of individual pixels or small groups of pixels.

These are precisely the features which unify the various phylogenetically-aware β-diversity metrics described in Section 1.1.4, they lift comparisons of the constituent parts of communities to comparisions of the communities themselves. With this in mind, we now consider the relative benefit of various ground metrics on words with respect to the problem of reference-free comparison of metagenomic datasets.

3.1.2 Comparison of Ground Metrics for Wasserstein Distance between Metagenomic Datasets

We made no mention of any particular metric on words in the our discussion relating Wasserstein convergence on finite words and the k-mer distributions arising from them, and relied solely on the necessity of the pointwise convergence of the measures. In actual applications we are more interested in a definition of distance that is useful when the distributions are not so similar. In particular, as we have noted in Section 1.1.2, sequencing produces potentially nonuniform coverage of genetic material and includes errors, and so even inside a single genome the distribution of k-mers derived from a set of sequence reads will not perfectly reflect the actual distribution of k-mers in the genome.

For these reasons we are interested, at a minimum, in versions of distance in which k-mers

are closest when they arise from the same OTU. This is the benefit of utilizing a distance derived from de Bruijn graphs as opposed to the edit distance. While the edit distance is useful in comparing alignments, that is in identifying when portions of an assembled genomes are most similar, it does not reflect the actual process of assembly well. The sequences $AAACCCCC$ and $ACCCTCCG$ are only three edits apart but a minimal assembly of these sequences is their concatenation, so in this sense they are maximally different. On the other hand, the sequences $GTTTGA$ and $TTTGAC$ are two edits away but clearly adjacent in the assembly $GTTTGAC$, and so have a distance in the de Bruijn graph of 1. Note that as we desire a true metric on words, we are obligated to take the metric d_{dB*} given by distances in the symmetric de Bruijn graph.

There are some simple inequalities between the edit distance d_E and the path distance d_{dB*} in the symmetric de Bruijn graph which can be observed. In particular, $2 \cdot \mathrm{d}_E \leq \mathrm{d}_{dB*}$ as we may view each transition in the de Bruijn graph as a pair of edits, one in which we delete a terminal symbol in a word and another when we insert a symbol to the alternate end. As our example above showed, when a pair of k-mers are adjacent in the de Bruijn graph, we have that $\mathrm{d}_{dB*} \leq \mathrm{d}_E$. The example $AAAAA$ and $ATAAA$, in which the edit distance is one but the distance in the symmetric deBruijn graph is two shows that this inequality cannot be extended to nonadjacent elements of the symmetric deBruijn graph.

Another benefit of d_{dB*} over the edit distance is the smaller size of a neighborhood about each point, which diminishes the probability that a randomly selected pair of words are adjacent. Note that the degree of any vertex in the symmetric de Bruijn graph for an alphabet of size n is at most $2n$, one for each potential left and right-extension, and thus the volume of a ball of radius 1 with respect to the distance defined by the symmetric de Bruijn graphs is independent of k. On the other hand, the number words within a ball of radius 1 with respect to the edit distance is at least $k \cdot (n-1)$, as there are this many nontrivial substitutions, and so grows at least linearly with k.

A more constrained version of the edit distance, that of the Longest Common Substring (LCS, Definition 1.3.16) distance offers some of the benefits of both of the above metrics on words. It achieves its minimum value of two for distinct k-mers when they are adjacent in an assembly, though this condition is not necessary to achieve that minimum. On the other hand, it does allow for a degree of error correction not allowed by the distance in the symmetric de Bruijn graph, as k-mers which differ by one substitution also achieve the minimum nonzero value.

In this section we have noted some of the relationships between various metrics on words, in particular highlighting the manner in which proximity of k-mers with respect to the path distance in the symmetric de Bruijn graph implies adjacency of factors in assembled genomes. We have noted some relationships between natural metrics on words and the potential benefits of each when considering the comparison of k-mers arising as factors in metagenomic sequence reads.

We now define a reference-free metric on metagenomic datasets which utilizes the 1-Wasserstein metric for a ground distance defined by path length in the symmetric de Bruijn graph.

3.2 Application of the Wasserstein Metric to Metagenomic Datasets

3.2.1 EMDeBruijn: Description

Let U and V be datasets consisting of sequence reads from a pair of metagenomic samples. That is, abstractly, for the alphabet $\mathcal{A} = \{A, C, T, G\}$, we have that $U, V \subset \mathcal{A}^*$ such that $|U|, |V|$ are finite. Choose $k \in \mathbb{N}$ such that $k \leq \min_{x \in U,V} |x|$ and let $u_k = freq_k(U)$ and $v_k = freq_k(V)$ be the vectors, indexed by the elements of $\mathcal{A}^k$, of the normalized frequency of occurrence of $k-$mers in the elements of U and V, respectively, as described in 1.3.4.

EMDeBruijn computes a approximation to the 1-Wasserstein metric between normalized frequency vectors for a ground distance between k-mers given by path distance in the symmetric de Bruijn graph. Such an approximation is necessary as the number of k-mers grows exponentially with k, making an exact solution computationally impractical. To demonstrate the proof of concept, a simple heuristic based in the minimum cost method was first implemented, as described in Section 1.2.6 as a means by which to generate an initial basic feasible solution for the transportation Simplex algorithm. This heuristic algorithm approximates the optimal transport metric by iteratively building a flow between relative abundances by first maximizing the transport between adjacent vertices in the symmetric de Bruijn graph before proceeding to maximize the remaining transport between vertices which are distance $2, 3, \ldots$, up to the diameter of the graph, thereby constructing a flow which satisfies both marginals. The pseudocode for the implementation of this algorithm is shown in Section 3.2.2.

As a way of producing a more mathematically rigorous approximation to the Wasserstein metric, the method of entropically-regularizing the optimal transport problem was then implemented. The theory behind this recently developed method is described in Section 1.2.6, while Definition 1.2.14 gives the specific form of the regularized optimization problem. The pseudocode for the implementation of this algorithm is shown in Section 3.2.3.

3.2.2 EMDeBruijn: Minimum Cost Heuristic Algorithm

Algorithm 3.2.1. *EMDeBruijn: Minimum Cost Heuristic*

Input:

$k, freq_k(V), freq_k(W)$ ▷ $freq_k(V), freq_k(W)$ *distributions on k-mers, D a distance matrix between k-mers*

Initialization:

$\gamma = 0 \in \mathbb{R}^{\mathcal{A}^k \times \mathcal{A}^k}$

$d = 0$

Iterations:

1: ***while*** $\gamma \notin \Gamma(V, W)$ ***do***

2: *Sort* $\mathcal{A}^k$ *so that* $freq_k(V)(v_i) - \sum \gamma(v_i, \cdot) \geq freq_k(V)(v_j) - \sum \gamma(v_j, \cdot)$ *for* $i \leq j$

3: ***for*** v_i ***do***

4: ***while*** $\sum \gamma(v_i, \cdot) < freq_k(V)(v_i)$ ***do***

5: *Choose* w' *with* $D(v_i, w') = d, freq_k(W)(w') \geq freq_k(W)(w') \forall w$

6: *Maximize* $\gamma(v_i, w')$ *subject to* $\sum \gamma(v_i, \cdot) < freq_k(V)(v_i)$ *and*

7: $\sum \gamma(\cdot, w') < freq_k(W)(w')$

8: ***end while***

9: ***end for***

10: $d = d + 1$

11: ***end while***

Output:

$\gamma, EMDeBruijn(freq_k(V), freq_k(W))$

3.2.3 EMDeBruijn: Entropic-Regularization Algorithm

In the following we describe our implementation of an entropically-regularized approximation to the 1-Wasserstein metric via Sinkhorn iteration as described in Section 1.2.6. In what follows, the let exp and log denote the elementwise computation of this functions, let $\otimes$ denote the elementwise, or Hadamard, product of matrices, and $\oslash$ the elementwise division of matrices.

Algorithm 3.2.2. *EMDDeBruijn: Entropically-Regularized Approximation*

Input:

$freq_k(V), freq_k(W), D$ ▷ $freq_k(V), freq_k(W)$ *distributions on k-mers,* D *a distance*

matrix between k-mers

α, rep ▷ *α the regularization parameter, rep the number of iterations for Sinkhorn projection*

Initialization:

$K_\alpha = \exp[(-1/\alpha) \otimes D]$

$v = w = \mathbf{1}$

Iterations:

1: ***for*** $i = 1, \ldots, rep$ ***do***

2: $v = A \oslash (K_\alpha \cdot w)$

3: $w = B \oslash (K_\alpha \cdot v$

4: ***end for***

5: $\gamma = diag(v) \cdot K_\alpha \cdot diag(w)$

6: $EMDeBruijn(freq_k(V), freq_k(W)) = \sum_{i,j} \gamma(i,j) \cdot \log(\gamma(i,j)/K_\alpha)$

Output:

$\gamma, EMDeBruijn(freq_k(V), freq_k(W))$

3.3 Results

3.3.1 Empirical Estimation of Error in the Minimum Cost Heuristic Approximation to the Wasserstein Metric

To gauge the accuracy of the minimum cost heuristic approximation to the Wasserstein metric for ground distances given by path length in the symmetric de Bruijn graph, the minimum cost heuristic was compared against a solution to the linear programming formulation of the optimal transport problem. 10,000 pairs of random synthetic measures were generated for $\mathcal{A}^k$ for $k = 4$ and $|\mathcal{A}| = 4$. For each pair, the minimum cost heuristic approx-

imation to the Wasserstein metric was computed. Additionally, the linear programming formulation of the optimal transport problem, as described in Section 1.2.6, was solved iteratively via non-negative least squares implementation in MATLAB to determine a minimizing flow. Iterative methods were chosen to generate solutions due to the exponential growth of the problem with respect to k. The computed values for the first 100 such pairs are displayed in Figure 3.2. The distribution of the relative error of the minimum cost heuristic approximation is included in Figure 3.3. The mean relative error between the minimum cost heuristic and the linear programming formulation was determined to be 0.010, the median such error was determined to be 0.098.

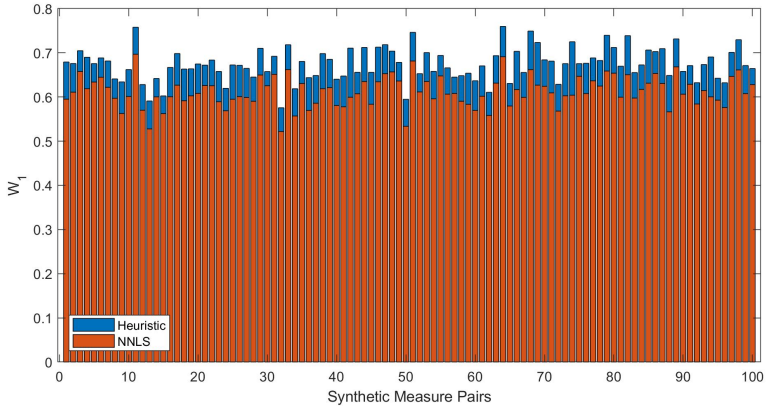


FIGURE 3.2: Comparison of the computation of the Wasserstein metric for ground distances given by path length in the symmetric de Bruijn graph for $k = 4$ and $|\mathcal{A}| = 4$ via the minimum cost heuristic and non-negative least squares solution to the linear programming formulation for 100 randomly generated synthetic sample pairs.

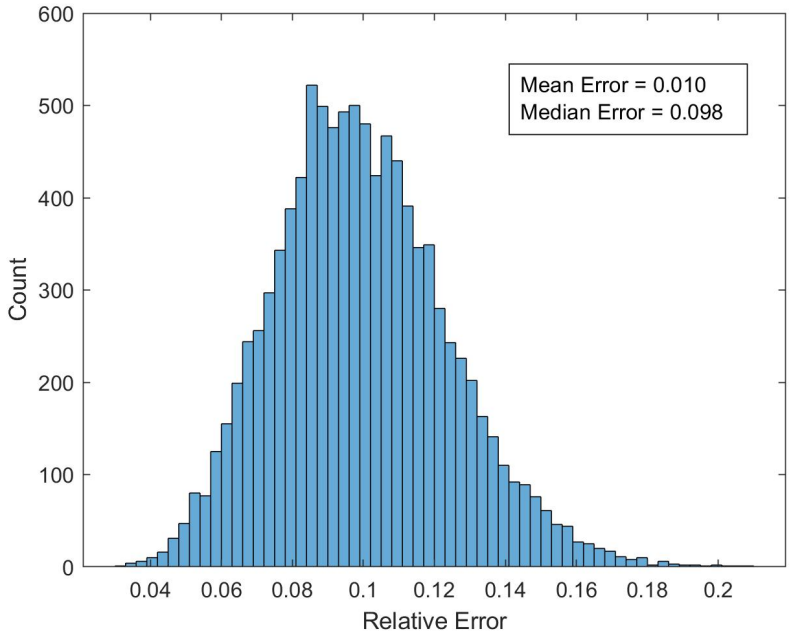
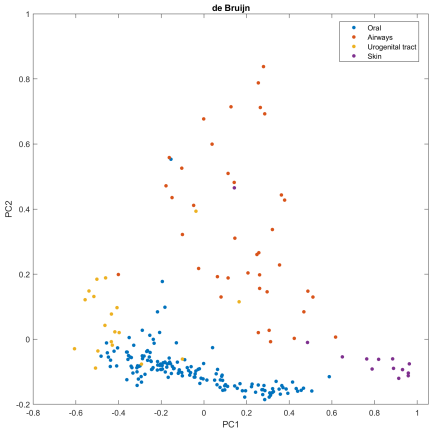


FIGURE 3.3: Distribution of the relative error in the computation of the Wasserstein metric for ground distances given by path length in the symmetric de Bruijn graph for $k = 4$ and $|\mathcal{A}| = 4$ via the minimum cost heuristic and non-negative least squares solution to the linear programming formulation for 10,000 randomly generated synthetic sample pairs.

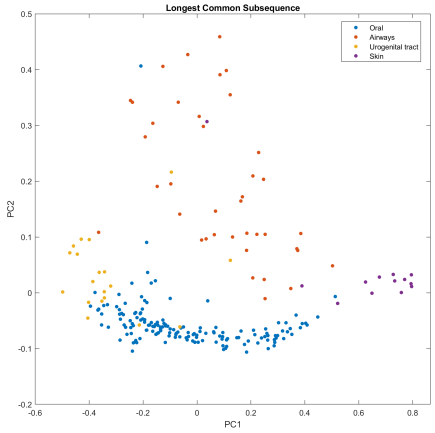
3.3.2 Application of EMDeBruijn to Real-world Datasets

To evaluate the utility of Wasserstein-based reference-free metagenomic comparisons, we first applied the minimum cost heuristic algorithm to a real-world dataset consisting of 223 samples from the Human Microbiome Project [117]. These samples, originating from body locations designated as oral, airways, urogenital tract or skin, were processed by the Broad Institute via whole genome shotgun sequencing. The downloaded datasets were processed using the FASTX-Toolkit package [42] into FASTA format sequence-read files. These were processed into k-mer counts via the dna-utils package [76].

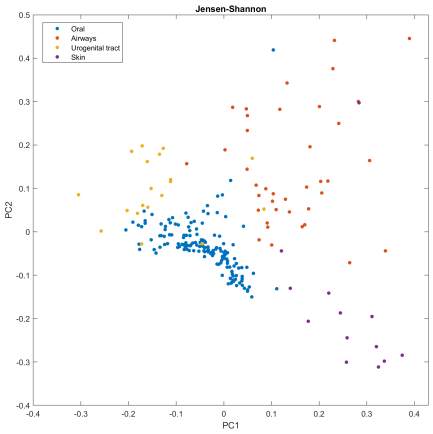
Pairwise comparisons for $k = 6$ were made for each of the following metrics: $EMDeBruijn$ using the minimum cost heursitic, L_1, Jensen-Shannon divergence and a 1-Wasserstein metric with a ground distance given by the longest common subsequence metric approximated utilizing the minimum cost heuristic. The resulting matrices of pairwise distances were then used to perform PCoA for each of the given metrics. The results are presented in Figure 3.4.



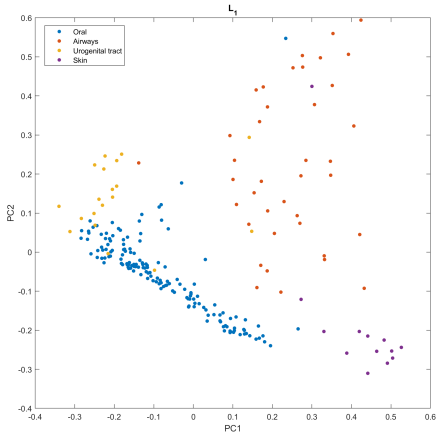
(a) EMDeBruijn: Minimum Cost



(b) LCS: Minimum Cost



(c) Jensen-Shannon

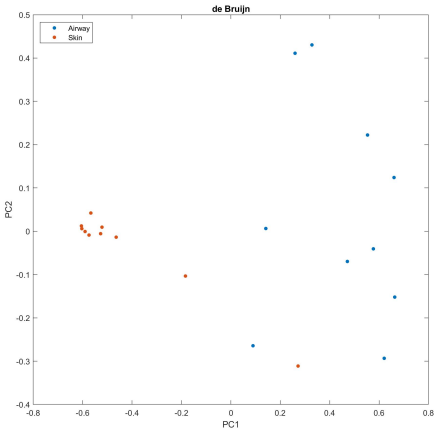


(d) L_1

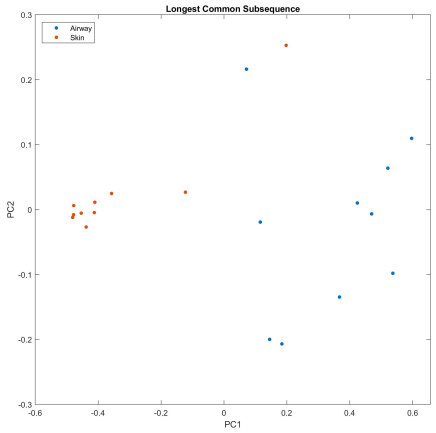
FIGURE 3.4: Principle Coordinate Analysis via minimum cost heuristic approximation of the EMDeBruijn metric for $k = 6$, 1-Wasserstein for the LCS metric and $k = 6$, Jensen-Shannon divergence, and L_1 metric of 223 metagenomic microbiome samples from the Human Microbiome Project. Samples are labeled as originating from body locations designated as oral, airways, urogenital tract or skin.

After further research, the entropically-regularized formulation of the optimal transport problem was selected as a means by which to better approximate the Wasserstein metric.

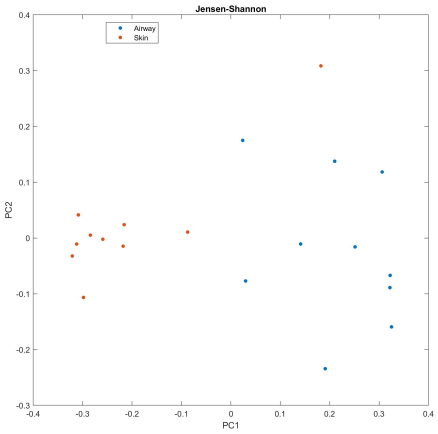
Implementation of these ideas is ongoing, but as a further proof of concept a subset of the Human Microbiome dataset used above was analyzed. Ten samples designated as originating from skin and ten samples designated as originating from airways were selected at random. The same selection of metrics were again applied, but utilizing the entropically-regularized approximation to the Wasserstein metric for both the LCS and symmetric de Bruijn distances. Parameter choices for the regularization were set uniformly at $\alpha = 0.01$ with a fixed 500 Sinkhorn iterations. The resulting pairwise distance matrices were then used to perform PCoA. The results are presented in Figure 3.5.



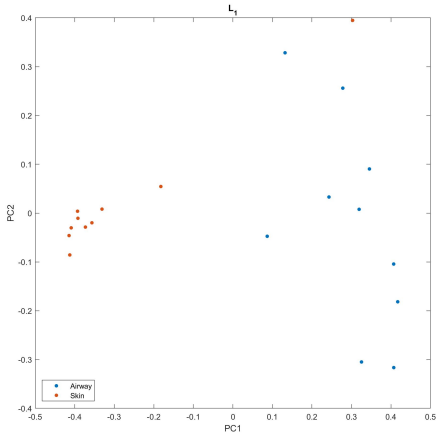
(a) EMDeBruijn: Entropic Regularization



(b) LCS: Entropic Regularization



(c) Jensen-Shannon



(d) L_1

FIGURE 3.5: Principle Coordinate Analysis via entropically-regularized approximation of the EMDeBruijn metric for $k = 6$, 1-Wasserstein distance using the longest common subsequence (LCS) ground metric for $k = 6$, Jensen-Shannon divergence, and L_1 metric of 20 metagenomic microbiome samples from the Human Microbiome Project. Samples are labeled as originating from body locations designated as airways or skin.

3.4 Discussion

3.4.1 Results

As observed in Figure 3.4, the *EMDeBruijn* metric implemented with the minimum cost heuristic appeared to show a greater degree of discrimination in separating the samples designated as originating from the the oral and urogenital tract locations from those designated as airways and skin than the Jensen-Shannon divergence, a commonly applied reference-free metric in microbiology. A similar apparent pattern was observed in comparing the minimum cost approximated Wasserstein metric for the LCS ground metric. Strikingly, ordination via the L_1 metric, the least theoretically justifiable of the metrics used, appeared to perform as well or better than any of metrics tested.

These results suggested that the use of Wasserstein metrics in reference-free comparisons had merit. This is in spite of the somewhat lackluster performance of the minimum cost heuristic in approximating the Wasserstein metric, as observed in the 10% average relative error shown in Figure 3.3. The heuristic does generate a basic feasible solution for the Wasserstein metric, and so provides a lower bound for the true value, as demonstrated visibly in Figure 3.2. Regardless, the broad distribution of relative errors demonstrate that the heuristic did not provide a consistent or precise approximation to the true value.

Adaption of the entropically-regularized approximation is a work in progress, and so our current results utilizing the method are preliminary. As such, benchmarking the accuracy of the approximation against known solution techniques has not yet been performed. The result themselves are promising though, and again Figure 3.5 demonstrates that *EMDeBruijn* appears to outperform the Jensen-Shannon divergences in demonstrating the similarity of microbial communities sample from skin.

As noted in Section 1.1.5, the heuristic arguments for the lower bound of optimal k values for a single genome comparison using reference-free techniques is on the order of $\log_4$ of the

length of the genome. With some microbial genomes being on the order of millions of base pairs long, our choice of $k = 6$ is far from optimal. Further optimization and refinement of the algorithms employed to allow for computation of these metrics with larger word lengths should only increase the resolving power of the technique.

3.4.2 Future Work

At the moment, when utilizing an Intel i7 2.2 gHz processor, computation of the entropically-regularized Wasserstein metric for $k = 6$ and a ground metric derived from the symmetric de Bruijn graph requires approximately 40 seconds. If this method of utilizing the Wasserstein metric to make β-diversity comparisons is to be truly useful, faster computation for larger values of k will be necessary. While the current implementation of the algorithm has much room for optimization, it seems unlikely to scale to the necessary values in its current form. This suggests that if these ideas are to be followed through on, another approximation technique will need to be developed.

By its very nature, a graph is defined by purely local connectivity information. From this information we have extracted a metric, but there exists a potentially more efficient algorithm which uses the local connectivity information given by the graph adjacency matrix (Defintion 1.3.2) to approximate the Wasserstein metric. This technique is known as the *convolutional Wasserstein metric* [108]. It is related to the entropically-regularized Wasserstein metric we have adapted for these computations, and further approximates the Wasserstein metric by utilizing convolution against solutions to the heat equation over a geometric domain to approximate the action of the matrix K_α utilized in the solution to the entropically-regularized transport algorithm described in Section 1.2.6. A goal for future work is to explore whether the high degree of regularity found in the de Bruijn graph can be leveraged to use these convolutional techniques to generate sufficiently efficient and accurate algorithms for optimal transport approximation to make the ideas developed in

this chapter practical.

This concludes our discussion of the application of the Wasserstein metric to reference-free β-diversity comparisons between metagenomic datasets. We now turn to a brief summary of the work and result we have developed.

4 CONCLUSION

In this dissertation we have developed the mathematical foundations for two instances of the Wasserstein metric in microbial ecology. This work has been inspired by the explosion of data arising from the sequencing of metagenomes, the aggregate genetic information of entire communities of organisms. Our research has been focused in building analytic tools to infer structure between these communities, and so better understand the vast diversity of microbial life around us.

We began our work by conveying the biological foundations of microbial ecology in Section 1.1, with an emphasis on phylogenetics, the metrics used to quantify the diversity found between microbial communities and the tools used in the analysis of ecological data. We sought to show how the methods of comparison adopted by biologists, such as the UniFrac metric, benefited from further mathematical analysis. We highlighted the ways in which the graph theoretic or combinatorial form of some of the objects of biological study, such as phylogenetic trees and genomes, allowed for mathematical means of understanding.

In Section 1.2 we explored a particular metric between measures defined on a metric space, that of the Wasserstein distance. We noted its basic theory and alternate formulations, as well applications to a variety of other fields. Our treatment gave context to the metric in Section 1.2.4, which the choice of comparisons made motivated by mathematical as well as biological significance. Of particular interest were the recently developed methods related to the Wasserstein metric, arising under the name of the Earth mover's distance, in image analysis. These applications were a source of inspiration in our own uses of the Wasserstein metric to understand difference in microbial communities. We concluded our survey of the Wasserstein metric in Section 1.2.6 with solution and approximation techniques.

Our next goal was to better understand the graph theoretic structures underlying two

important ideas in metagenomics, that of the trees used to encode evolutionary relationships and the de Bruijn used in genomes assembly. Our interest in Section 1.3 was in determining the properties of the metric spaces upon which we would be considering the Wasserstein metric, as well as to establish a common language for the various objects of study under consideration. This included giving in Section 1.3.4 a mathematical formalism, finite sequences from a fixed alphabet, with which to consider genomes and sequence reads.

Our primary results are related to the UniFrac metric and are conveyed in Section 2. We proved an alternate characterizations for the Wasserstein metric when applied to a tree and provided a novel proof that this metric, when comparing relative abundances assigned to a common phylogenetic tree, was equivalent to the successful UniFrac metric. We developed this proof into an efficient solution technique for the UniFrac in Section 2.2.4 which simultaneously computes the UniFrac metric faster than previous implementations while providing additional information, that of a minimizing flow between relative abundances and the weighted differential abundance vector. In Section 2.3, our research noted how the relationship between the ordination techniques of PCA and PCoA might allow for a more efficient application of DPCoA, and how adapting the linear transformation used in the computation of UniFrac casts DPCoA in the same light as other biologically signifigant metrics, such as the χ^2 and Hellinger distances. We proceeded in SectionUniFracErr to use the framework we developed for considering the UniFrac metric as a means to produce the expectation of UniFrac for Dirichlet distributed relative abundances, a model appropriate for the datasets in question. We then applied this method to a sequence of examples. In Section 2.5 we demonstrated the application of our work on datasets. We concluded with a description of potential for future work inspired by our research, particularly the development of ideas related to the relationship between alternate L_1 PCA formulations and the UniFrac metric and the potential for the use of these ideas in other conceptual

frameworks for relative abundances, particularly that of compositional data analysis.

Given the success of the Wasserstein metric in making phylogenetically aware β-diversity comparisons, in Section 3 we developed a framework for making reference-free metagenomic comparisons via the Wasserstein metric. We noted in Section 3.1 how the means by which phylogenetic trees are constructed can be used to interpret distances between OTUs in these trees as distances between the words representing their genomes. We showed in Section 3.1.1 how utilizing the structure given by de Bruijn graphs allowed for convergence with respect to the Wasserstein metric between finite words to be cast in terms of the k-mers comprising those word. This work was inspired not only by the success of the UniFrac, but also by the recent development of the Wasserstein metric in image analysis. We next described two approximation algorithms for the Wasserstein metric when applied to k-mers, one a heuristic used classically as a seed for the Simplex algorithm and another developed recently for the purposes of image analysis which itertively solves an entropically-regularized form of the Wasserstein metric. In Section 3.3 we demonstrated the proof-of-concept of our ideas, and showed how utilizing Wasserstein metrics performed against metrics commonly used in the reference-free comparison of sequence dataset. We ended our discussion in Section 3.4.2 by considering the computational limits of our methods and the need to increase resolving power by considering factors of greater length. We noted how using more of the local graph structure to apply convolutional solution techniques might help address these issues and make this line of analysis computationally tractable.

With that, we conclude our work.

BIBLIOGRAPHY

1. Ravindra K Ahyja, James B Orlin, and Thomas L Magnanti. Network flows: theory, algorithms, and applications. 1993.

2. J Aitchison, C Barceló-Vidal, JJ Egozcue, and V Pawlowsky-Glahn. A concise guide for the algebraic-geometric structure of the simplex, the sample space for compositional data analysis. In *Proceedings of IAMG*, volume 2, pages 387–392, 2002.

3. Marti J Anderson, Kari E Ellingsen, and Brian H McArdle. Multivariate dispersion as a measure of beta diversity. *Ecology letters*, 9(6):683–693, 2006.

4. Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein gan. *arXiv preprint arXiv:1701.07875*, 2017.

5. Anton Bankevich, Sergey Nurk, Dmitry Antipov, Alexey A Gurevich, Mikhail Dvorkin, Alexander S Kulikov, Valery M Lesin, Sergey I Nikolenko, Son Pham, Andrey D Prjibelski, et al. Spades: a new genome assembly algorithm and its applications to single-cell sequencing. *Journal of computational biology*, 19(5):455–477, 2012.

6. Yael Baran and Eran Halperin. Joint analysis of multiple metagenomic samples. *PLoS computational biology*, 8(2):e1002373, 2012.

7. Louise J Barwell, Nick JB Isaac, and William E Kunin. Measuring β-diversity with species abundance data. *Journal of Animal Ecology*, 84(4):1112–1122, 2015.

8. Jean-David Benamou, Guillaume Carlier, Marco Cuturi, Luca Nenna, and Gabriel Peyré. Iterative bregman projections for regularized transportation problems. *SIAM Journal on Scientific Computing*, 37(2):A1111–A1138, 2015.

9. Giorgio Benedetti and Stefano Morosetti. A graph-topological approach to recognition of pattern and similarity in rna secondary structures. *Biophysical chemistry*, 59(1-2):179–184, 1996.

10. Karl A Beres, Robert L Wallace, and Hendrik H Segers. Rotifers and hubbell's unified neutral theory of biodiversity and biogeography. *Natural Resource Modeling*, 18(3):363–376, 2005.

11. J Roger Bray and John T Curtis. An ordination of the upland forest communities of southern wisconsin. *Ecological monographs*, 27(4):325–349, 1957.

12. J Paul Brooks, JoseâĂŹH Dulá, and Edward L Boone. A pure l1-norm principal component analysis. *Computational statistics & data analysis*, 61:83–98, 2013.

13. Jessica A Bryant, Christine Lamanna, Hélene Morlon, Andrew J Kerkhoff, Brian J Enquist, and Jessica L Green. Microbes on mountainsides: contrasting elevational patterns of bacterial and plant diversity. *Proceedings of the National Academy of Sciences*, 2008.

14. Andrew G Bunn, DL Urban, and TH Keitt. Landscape connectivity: a conservation application of graph theory. *Journal of environmental management*, 59(4):265–278, 2000.

15. J Gregory Caporaso, Justin Kuczynski, Jesse Stombaugh, Kyle Bittinger, Frederic D Bushman, Elizabeth K Costello, Noah Fierer, Antonio Gonzalez Pena, Julia K Goodrich, Jeffrey I Gordon, et al. Qiime allows analysis of high-throughput community sequencing data. *Nature methods*, 7(5):335–336, 2010.

16. 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 biology*, 12(5):R50, 2011.

17. Thomas Cavalier-Smith, Martin Brasier, and T Martin Embley. Introduction: how and when did microbes change the world? *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, 361(1470):845–850, 2006.

18. Rayan Chikhi and Paul Medvedev. Informed and automated k-mer size selection for genome assembly. *Bioinformatics*, 30(1):31–37, 2013.

19. Phillip EC Compeau, Pavel A Pevzner, and Glenn Tesler. Why are de bruijn graphs useful for genome assembly? *Nature biotechnology*, 29(11):987, 2011.

20. 1000 Genomes Project Consortium et al. A global reference for human genetic variation. *Nature*, 526(7571):68, 2015.

21. ENCODE Project Consortium et al. An integrated encyclopedia of dna elements in the human genome. *Nature*, 489(7414):57, 2012.

22. International Human Genome Sequencing Consortium et al. Initial sequencing and analysis of the human genome. *Nature*, 409(6822):860, 2001.

23. Nicolaas Govert De Bruijn. A combinatorial problem. *Koninklijke Nederlandse Akademie v. Wetenschappen*, 49(49):758–764, 1946.

24. Gregory J Dick, Karthik Anantharaman, Brett J Baker, Meng Li, Daniel C Reed, and Cody S Sheik. The microbiology of deep-sea hydrothermal vent plumes: ecological and biogeographic linkages to seafloor and water column habitats. *Frontiers in Microbiology*, 4:124, 2013.

25. Roland L Dobrushin. Prescribing a system of random variables by conditional distributions. *Theory of Probability & Its Applications*, 15(3):458–486, 1970.

26. Robert Ekblom and Jochen BW Wolf. A field guide to whole-genome sequencing, assembly and annotation. *Evolutionary applications*, 7(9):1026–1042, 2014.

27. Leonhard Euler. Solutio problematis ad geometriam situs pertinensis. *Comm. Acad. Sci. Imper. Petropol.*, 8:128–140, 1736.

28. Steven N Evans and Frederick A Matsen. The phylogenetic kantorovich–rubinstein metric for environmental sequence samples. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 74(3):569–592, 2012.

29. Harold Exton. Multiple hypergeometric functions and applications. 1976.

30. Daniel N Frank, Allison L St Amand, Robert A Feldman, Edgar C Boedeker, Noam Harpaz, and Norman R Pace. Molecular-phylogenetic characterization of microbial community imbalances in human inflammatory bowel diseases. *Proceedings of the National Academy of Sciences*, 104(34):13780–13785, 2007.

31. Jerome Friedman, Trevor Hastie, and Robert Tibshirani. *The elements of statistical learning*, volume 1. Springer series in statistics New York, NY, USA:, 2001.

32. Bela A Frigyik, Amol Kapila, and Maya R Gupta. Introduction to the dirichlet distribution and related processes. *Department of Electrical Engineering, University of Washignton, UWEETR-2010-0006*, 2010.

33. JULIA Fukuyama, Paul J McMurdie, Les Dethlefsen, David A Relman, and Susan Holmes. Comparisons of distance methods for combining covariates and abundances in microbiome studies. In *Biocomputing 2012*, pages 213–224. World Scientific, 2012.

34. Olivier Gascuel and Mike Steel. Neighbor-joining revealed. *Molecular biology and evolution*, 23(11):1997–2000, 2006.

35. Robert C Gentleman, Vincent J Carey, Douglas M Bates, Ben Bolstad, Marcel Dettling, Sandrine Dudoit, Byron Ellis, Laurent Gautier, Yongchao Ge, Jeff Gentry, et al. Bioconductor: open software development for computational biology and bioinformatics. *Genome biology*, 5(10):R80, 2004.

36. Howard Gest. The discovery of microorganisms by robert hooke and antoni van leeuwenhoek, fellows of the royal society. *Notes and Records of the Royal Society*, 58(2):187–201, 2004.

37. Alison L Gibbs and Francis Edward Su. On choosing and bounding probability metrics. *International statistical review*, 70(3):419–435, 2002.

38. Alessandro Giuliani, Arun Krishnan, Joseph P Zbilut, and Masaru Tomita. Proteins as networks: usefulness of graph theory in protein science. *Current Protein and Peptide Science*, 9(1):28–38, 2008.

39. Travis C Glenn. Field guide to next-generation dna sequencers. *Molecular ecology resources*, 11(5):759–769, 2011.

40. Rachel L Goldfeder, Dennis P Wall, Muin J Khoury, John PA Ioannidis, and Euan A Ashley. Human genome sequencing at the population scale: a primer on high-throughput dna sequencing and analysis. *American journal of epidemiology*, 186(8):1000–1009, 2017.

41. Phillip Good. *Permutation tests: a practical guide to resampling methods for testing hypotheses*. Springer Science & Business Media, 2013.

42. A Gordon and GJ Hannon. Fastx-toolkit. http://hannonlab.cshl.edu/fastx_toolkit, 2010.

43. Micah Hamady, Catherine Lozupone, and Rob Knight. Fast unifrac: facilitating high-throughput phylogenetic analyses of microbial communities including analysis of pyrosequencing and phylochip data. *The ISME journal*, 4(1):17–27, 2010.

44. Godfrey Harold Hardy et al. Mendelian proportions in a mixed population. *Classic papers in genetics. Prentice-Hall, Inc.: Englewood Cliffs, NJ*, pages 60–62, 1908.

45. Frederick Hillier and G Lieberman. Introduction to mathematical programming, 2nd edition. 1995.

46. Adina Howe and Patrick SG Chain. Challenges and opportunities in understanding microbial communities with metagenome assembly (accompanied by ipython notebook tutorial). *Frontiers in microbiology*, 6:678, 2015.

47. Peter J Huber. Robust statistics. In *International Encyclopedia of Statistical Science*, pages 1248–1251. Springer, 2011.

48. Wolfgang Huber, Vincent J Carey, Li Long, Seth Falcon, and Robert Gentleman. Graphs in molecular biology. *BMC bioinformatics*, 8(6):S8, 2007.

49. Jaime Huerta-Cepas, François Serra, and Peer Bork. Ete 3: reconstruction, analysis, and visualization of phylogenomic data. *Molecular biology and evolution*, 33(6):1635–1638, 2016.

50. Robert W Jernigan and Robert H Baran. Pervasive properties of the genomic signature. *BMC genomics*, 3(1):23, 2002.

51. Haiyin Jiang, Zongxin Ling, Yonghua Zhang, Hongjin Mao, Zhanping Ma, Yan Yin, Weihong Wang, Wenxin Tang, Zhonglin Tan, Jianfei Shi, et al. Altered fecal microbiota composition in patients with major depressive disorder. *Brain, behavior, and immunity*, 48:186–194, 2015.

52. Ian Jolliffe. Principal component analysis. In *International encyclopedia of statistical science*, pages 1094–1096. Springer, 2011.

53. Dieter Jungnickel and D Jungnickel. *Graphs, networks and algorithms*. Springer, 2005.

54. Samuel Kariin and Chris Burge. Dinucleotide relative abundance extremes: a genomic signature. *Trends in genetics*, 11(7):283–290, 1995.

55. Victor Klee, George J Minty, and Oved Shisha. Inequalities, iii. *How Good is the Simplex Algorithm*, pages 159–175, 1972.

56. Christopher P Kolbert and David H Persing. Ribosomal dna sequencing as a tool for identification of bacterial pathogens. *Current opinion in microbiology*, 2(3):299–305, 1999.

57. Aleksandar D Kostic, Eunyoung Chun, Lauren Robertson, Jonathan N Glickman, Carey Ann Gallini, Monia Michaud, Thomas E Clancy, Daniel C Chung, Paul Lochhead, Georgina L Hold, et al. Fusobacterium nucleatum potentiates intestinal tumorigenesis and modulates the tumor-immune microenvironment. *Cell host & microbe*, 14(2):207–215, 2013.

58. Pranav Kulkarni and Peter Frommolt. Challenges in the setup of large-scale next-generation sequencing analysis workflows. *Computational and Structural Biotechnology Journal*, 15:471–477, 2017.

59. Matt Kusner, Yu Sun, Nicholas Kolkin, and Kilian Weinberger. From word embeddings to document distances. In *International Conference on Machine Learning*, pages 957–966, 2015.

60. Pierre Legendre and Eugene D Gallagher. Ecologically meaningful transformations for ordination of species data. *Oecologia*, 129(2):271–280, 2001.

61. Pierre Legendre and Louis Legendre. Numerical ecology, volume 24, (developments in environmental modelling), 1998.

62. Na Lei, Kehua Su, Li Cui, Shing-Tung Yau, and David Xianfeng Gu. A geometric view of optimal transportation and generative model. *arXiv preprint arXiv:1710.05488*, 2017.

63. Heidi EL Lischer and Kentaro K Shimizu. Reference-guided de novo assembly approach improves genome reconstruction for related species. *BMC bioinformatics*, 18(1):474, 2017.

64. Zhen Liu. Optimal routing in the de bruijn networks. In *Distributed Computing Systems, 1990. Proceedings., 10th International Conference on*, pages 537–544. IEEE, 1990.

65. Jason Lloyd-Price, Anup Mahurkar, Gholamali Rahnavard, Jonathan Crabtree, Joshua Orvis, A Brantley Hall, Arthur Brady, Heather H Creasy, Carrie McCracken, Michelle G Giglio, et al. Strains, functions and dynamics in the expanded human microbiome project. *Nature*, 550(7674):61, 2017.

66. Kenneth J Locey and Jay T Lennon. Scaling laws predict global microbial diversity. *Proceedings of the National Academy of Sciences*, 113(21):5970–5975, 2016.

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

68. Catherine A Lozupone, Micah Hamady, Scott T Kelley, and Rob Knight. Quantitative and qualitative β diversity measures lead to different insights into factors that structure microbial communities. *Applied and environmental microbiology*, 73(5):1576–1585, 2007.

69. Catherine A Lozupone and Rob Knight. Global patterns in bacterial diversity. *Proceedings of the National Academy of Sciences*, 104(27):11436–11440, 2007.

70. Ruibang Luo, Binghang Liu, Yinlong Xie, Zhenyu Li, Weihua Huang, Jianying Yuan, Guangzhu He, Yanxiang Chen, Qi Pan, Yunjie Liu, et al. Soapdenovo2: an empirically improved memory-efficient short-read de novo assembler. *Gigascience*, 1(1):18, 2012.

71. Anne E Magurran. Why diversity? In *Ecological diversity and its measurement*, pages 1–5. Springer, 1988.

72. CL Mallows. A note on asymptotic joint normality. *The Annals of Mathematical Statistics*, pages 508–515, 1972.

73. Sharmila S Mande, Monzoorul Haque Mohammed, and Tarini Shankar Ghosh. Classification of metagenomic sequences: methods and challenges. *Briefings in bioinformatics*, 13(6):669–681, 2012.

74. Chaysavanh Manichanh, Natalia Borruel, Francesc Casellas, and Francisco Guarner. The gut microbiota in ibd. *Nature Reviews Gastroenterology and Hepatology*, 9(10):599–608, 2012.

75. Jason R Miller, Sergey Koren, and Granger Sutton. Assembly algorithms for next-generation sequencing data. *Genomics*, 95(6):315–327, 2010.

76. Calvin Morrison. dna-utils. www.github.com/EESI/dna-utils/, 2014.

77. Gonzalo Navarro. A guided tour to approximate string matching. *ACM computing surveys (CSUR)*, 33(1):31–88, 2001.

78. Joseph Nesme, Wafa Achouak, Spiros N Agathos, Mark Bailey, Petr Baldrian, Dominique Brunel, Åsa Frostegård, Thierry Heulin, Janet K Jansson, Edouard Jurkevitch, et al. Back to the future of soil metagenomics. *Frontiers in microbiology*, 7:73, 2016.

79. Feiping Nie, Heng Huang, Chris Ding, Dijun Luo, and Hua Wang. Robust principal component analysis with non-greedy l1-norm maximization. In *IJCAI Proceedings-International Joint Conference on Artificial Intelligence*, volume 22, page 1433, 2011.

80. Jie Ning and Robert G Beiko. Phylogenetic approaches to microbial community classification. *Microbiome*, 3(1):47, 2015.

81. Jorge Nocedal and Stephen J Wright. Numerical optimization 2nd, 2006.

82. Ulrich Nübel, Ferran Garcia-Pichel, Michael Kühl, and Gerard Muyzer. Quantifying microbial diversity: morphotypes, 16s rrna genes, and carotenoids of oxygenic phototrophs in microbial mats. *Applied and Environmental Microbiology*, 65(2):422–430, 1999.

83. Sergey Nurk, Dmitry Meleshko, Anton Korobeynikov, and Pavel A Pevzner. metaspades: a new versatile metagenomic assembler. *Genome research*, pages gr–213959, 2017.

84. John D O'Brien and Nicolas Record. The power and pitfalls of dirichlet-multinomial mixture models for ecological count data. *bioRxiv*, page 045468, 2016.

85. Vera Odintsova, Alexander Tyakht, and Dmitry Alexeev. Guidelines to statistical analysis of microbial composition data inferred from metagenomic sequencing. *Current issues in molecular biology*, 24:17–36, 2017.

86. Brian C O'Meara. Evolutionary inferences from phylogenies: a review of methods. *Annual Review of Ecology, Evolution, and Systematics*, 43:267–285, 2012.

87. James B Orlin. A polynomial time primal network simplex algorithm for minimum cost flows. *Mathematical Programming*, 78(2):109–129, 1997.

88. Donovan H Parks and Robert G Beiko. Measures of phylogenetic differentiation provide robust and complementary insights into microbial communities. *The ISME journal*, 7(1):173, 2013.

89. Sandrine Pavoine, Anne-Béatrice Dufour, and Daniel Chessel. From dissimilarities among species to dissimilarities among communities: a double principal coordinate analysis. *Journal of theoretical biology*, 228(4):523–537, 2004.

90. DAVID PENNY. Graph theory, evolutionary trees and classification. *Zoological Journal of the Linnean Society*, 74(3):277–292, 1982.

91. Filipe Pereira, Joao Carneiro, Rune Matthiesen, Barbara van Asch, Nadia Pinto, Leonor Gusmao, and Antonio Amorim. Identification of species by multiplex analysis of variable-length sequences. *Nucleic acids research*, 38(22):e203–e203, 2010.

92. Pavel A Pevzner, Haixu Tang, and Michael S Waterman. An eulerian path approach to dna fragment assembly. *Proceedings of the National Academy of Sciences*, 98(17):9748–9753, 2001.

93. Thu Pham-Gia, Noyan Turkkan, and P Eng. Bayesian analysis of the difference of two proportions. *Communications in Statistics-Theory and Methods*, 22(6):1755–1771, 1993.

94. Émile Picard. Sur une extension aux fonctions de deux variables du probleme de riemann relatif aux fonctions hypergéométriques. In *Annales scientifiques de l'École Normale Supérieure*, volume 10, pages 305–322. Elsevier, 1881.

95. Alban Ramette. Multivariate analyses in microbial ecology. *FEMS microbiology ecology*, 62(2):142–160, 2007.

96. Ravi Ranjan, Asha Rani, Ahmed Metwally, Halvor S McGee, and David L Perkins. Analysis of the microbiome: advantages of whole genome shotgun versus 16s amplicon sequencing. *Biochemical and biophysical research communications*, 469(4):967–977, 2016.

97. C Radhakrishna Rao. Diversity and dissimilarity coefficients: a unified approach. *Theoretical population biology*, 21(1):24–43, 1982.

98. Raimundo Real and Juan M Vargas. The probabilistic basis of jaccard's index of similarity. *Systematic biology*, 45(3):380–385, 1996.

99. Jason A Reuter, Damek V Spacek, and Michael P Snyder. High-throughput sequencing technologies. *Molecular cell*, 58(4):586–597, 2015.

100. Yossi Rubner, Carlo Tomasi, and Leonidas J Guibas. The earth mover's distance as a metric for image retrieval. *International journal of computer vision*, 40(2):99–121, 2000.

101. Ludger Rüschendorf. Wasserstein metric. *Hazewinkel, Michiel, Encyclopaedia of Mathematics, Springer*, 2001.

102. Thomas J Sharpton. An introduction to the analysis of shotgun metagenomic data. *Frontiers in plant science*, 5:209, 2014.

103. Justin D Silverman, Alex D Washburne, Sayan Mukherjee, and Lawrence A David. A phylogenetic transform enhances analysis of compositional microbiota data. *Elife*, 6:e21887, 2017.

104. Gregory E Sims, Se-Ran Jun, Guohong A Wu, and Sung-Hou Kim. Alignment-free genome comparison with feature frequency profiles (ffp) and optimal resolutions. *Proceedings of the National Academy of Sciences*, pages pnas–0813249106, 2009.

105. Barton E Slatko, Andrew F Gardner, and Frederick M Ausubel. Overview of next-generation sequencing technologies. *Current protocols in molecular biology*, 122(1):e59, 2018.

106. Wenke Smets, Serena Moretti, Siegfried Denys, and Sarah Lebeer. Airborne bacteria in the atmosphere: presence, purpose, and potential. *Atmospheric Environment*, 139:214–221, 2016.

107. Justin Solomon. Optimal transport on discrete domains. *AMS Short Course on Discrete Differential Geometry*, 2018.

108. Justin Solomon, Fernando De Goes, Gabriel Peyré, Marco Cuturi, Adrian Butscher, Andy Nguyen, Tao Du, and Leonidas Guibas. Convolutional wasserstein distances: Efficient optimal transportation on geometric domains. *ACM Transactions on Graphics (TOG)*, 34(4):66, 2015.

109. Aymé Spor, Omry Koren, and Ruth Ley. Unravelling the effects of the environment and host genotype on the gut microbiome. *Nature Reviews Microbiology*, 9(4):279–290, 2011.

110. Eric J Stewart. Growing unculturable bacteria. *Journal of bacteriology*, pages JB–00345, 2012.

111. Volker Strassen et al. The existence of probability measures with given marginals. *The Annals of Mathematical Statistics*, 36(2):423–439, 1965.

112. Nathan G Swenson. Phylogenetic beta diversity metrics, trait evolution and inferring the functional beta diversity of communities. *PloS one*, 6(6):e21264, 2011.

113. Robert E Tarjan. Dynamic trees as search trees via euler tours, applied to the network simplex algorithm. *Mathematical Programming*, 78(2):169–177, 1997.

114. Atsuki Tsuruya, Akika Kuwahara, Yuta Saito, Haruhiko Yamaguchi, Takahisa Tsubo, Shogo Suga, Makoto Inai, Yuichi Aoki, Seiji Takahashi, Eri Tsutsumi, et al. Ecophysiological consequences of alcoholism on human gut microbiota: implications for ethanol-related pathogenesis of colon cancer. *Scientific reports*, 6:27923, 2016.

115. Hanna Tuomisto. A consistent terminology for quantifying species diversity? yes, it does exist. *Oecologia*, 164(4):853–860, 2010.

116. Peter J Turnbaugh, Micah Hamady, Tanya Yatsunenko, Brandi L Cantarel, Alexis Duncan, Ruth E Ley, Mitchell L Sogin, William J Jones, Bruce A Roe, Jason P Affourtit, et al. A core gut microbiome in obese and lean twins. *nature*, 457(7228):480, 2009.

117. Peter J Turnbaugh, Ruth E Ley, Micah Hamady, Claire M Fraser-Liggett, Rob Knight, and Jeffrey I Gordon. The human microbiome project. *Nature*, 449(7164):804, 2007.

118. Leonid Nisonovich Vaserstein. Markov processes over denumerable products of spaces, describing large systems of automata. *Problemy Peredachi Informatsii*, 5(3):64–72, 1969.

119. Anatoly Moiseevich Vershik. Long history of the monge-kantorovich transportation problem. *The Mathematical Intelligencer*, 35(4):1–9, 2013.

120. C. Villani. *Optimal Transport: Old and New*. Grundlehren der mathematischen Wissenschaften. Springer Berlin Heidelberg, 2008.

121. John Vollmers, Sandra Wiegand, and Anne-Kristin Kaster. Comparing and evaluating metagenome assembly tools from a microbiologistâĂŹs perspective-not only size matters! *PloS one*, 12(1):e0169662, 2017.

122. William G Weisburg, Susan M Barns, Dale A Pelletier, and David J Lane. 16s ribosomal dna amplification for phylogenetic study. *Journal of bacteriology*, 173(2):697–703, 1991.

123. William B Whitman, David C Coleman, and William J Wiebe. Prokaryotes: the unseen majority. *Proceedings of the National Academy of Sciences*, 95(12):6578–6583, 1998.

124. Robert H Whittaker. Evolution and measurement of species diversity. *Taxon*, pages 213–251, 1972.

125. Edward Orlando Wiley and Bruce S Lieberman. *Phylogenetics: theory and practice of phylogenetic systematics*. John Wiley & Sons, 2011.

126. Shannon J Williamson, Douglas B Rusch, Shibu Yooseph, Aaron L Halpern, Karla B Heidelberg, John I Glass, Cynthia Andrews-Pfannkoch, Douglas Fadrosh, Christopher S Miller, Granger Sutton, et al. The sorcerer ii global ocean sampling expedition: metagenomic characterization of viruses within aquatic microbial samples. *PloS one*, 3(1):e1456, 2008.

127. Ben P Willing, Johan Dicksved, Jonas Halfvarson, Anders F Andersson, Marianna Lucio, Zongli Zheng, Gunnar Järnerot, Curt Tysk, Janet K Jansson, and Lars Engstrand. A pyrosequencing study in twins shows that gastrointestinal microbial profiles vary with inflammatory bowel disease phenotypes. *Gastroenterology*, 139(6):1844–1854, 2010.

128. Carl R Woese. Bacterial evolution. *Microbiological reviews*, 51(2):221, 1987.

129. Tiee-Jian Wu, Ying-Hsueh Huang, and Lung-An Li. Optimal word sizes for dissimilarity measures and estimation of the degree of dissimilarity between dna sequences. *Bioinformatics*, 21(22):4125–4132, 2005.

130. Ziheng Yang and Bruce Rannala. Molecular phylogenetics: principles and practice. *Nature Reviews Genetics*, 13(5):303, 2012.

131. Pablo Yarza, Michael Richter, Jörg Peplies, Jean Euzeby, Rudolf Amann, Karl-Heinz Schleifer, Wolfgang Ludwig, Frank Oliver Glöckner, and Ramon Rosselló-Móra. The all-species living tree project: a 16s rrna-based phylogenetic tree of all sequenced type strains. *Systematic and applied microbiology*, 31(4):241–250, 2008.

132. Tanya Yatsunenko, Federico E Rey, Mark J Manary, Indi Trehan, Maria Gloria Dominguez-Bello, Monica Contreras, Magda Magris, Glida Hidalgo, Robert N Baldassano, Andrey P Anokhin, et al. Human gut microbiome viewed across age and geography. *nature*, 486(7402):222, 2012.

133. Pelin Yilmaz, Laura Wegener Parfrey, Pablo Yarza, Jan Gerken, Elmar Pruesse, Christian Quast, Timmy Schweer, Jörg Peplies, Wolfgang Ludwig, and Frank Oliver Glöckner. The silva and all-species living tree project (ltp) taxonomic frameworks. *Nucleic acids research*, 42(D1):D643–D648, 2013.

134. Daniel Zerbino and Ewan Birney. Velvet: algorithms for de novo short read assembly using de bruijn graphs. *Genome research*, pages gr–074492, 2008.

135. Qi Zhao, Zhi Yang, and Hai Tao. Differential earth mover's distance with its applications to visual tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(2):274–287, 2010.

136. Alain Zuur, Elena N Ieno, and Graham M Smith. *Analyzing ecological data*. Springer Science & Business Media, 2007.