THE PHYLOGENETIC KANTOROVICH-RUBINSTEIN METRIC FOR ENVIRONMENTAL SEQUENCE SAMPLES

STEVEN N. EVANS AND FREDERICK A. MATSEN

1. SUMMARY

We employ the Kantorovich-Rubinstein (KR) metric and L^p generalizations to compare probability distributions on a given phylogenetic tree. Such distributions arise in the context of metagenomics, where a sample of environmental sequences may be treated as a collection of weighted points on a reference phylogenetic tree of known sequences. In contrast to many applications of Kantorovich-Rubinstein ideas, the phylogenetic KR metric can be written in a closed form and calculated in linear time. Using Monte Carlo resampling of the data, we assign a statistical significance level to the observed distance between two distributions under a null hypothesis of no clustering. We also approximate the significance level using a functional of a suitable Gaussian process; in the L^2 generalized case this functional is distributed as a linear combination of χ_1^2 random variables weighted by the eigenvalues of an associated matrix. We conclude with an example application using our software implementation of the KR metric and its generalizations.

2. INTRODUCTION

2.1. Phylogenetic placement and probability distributions on a phylogenetic tree.

Next-generation sequencing technology enables sequencing of hundreds of thousands to millions of short DNA sequences in a single experiment. This has led to a new methodology for characterizing the collection of microbes in a sample: rather than using observed morphology or the results of culturing experiments, it is possible to directly sequence genetic material extracted in bulk from the sample. This technology has revolutionized the possibilities for unbiased surveys of environmental microbial diversity, ranging from the human gut (Gill et al., 2006) to acid mine drainages (Baker and Banfield, 2003).

Recently, a number of groups have proposed *phylogenetic placement* algorithms that place each environmental sequences on a fixed *reference phylogenetic tree* according to the phylogenetic maximum-likelihood criterion (Von Mering et al., 2007; Monier et al., 2008; Berger and Stamatakis, 2009; Matsen et al., 2010). The reference phylogenetic tree is constructed from previously-characterized DNA sequences.

The output of a placement algorithm is an assignment of the *query sequences* to their most likely location on the reference tree. By fixing a reference tree rather than attempting to build a phylogenetic tree for the sample from scratch, recent algorithms of this type are able to place tens of thousands of query sequences per hour per processor on a reference tree of one thousand taxa (e.g. species), with linear performance scaling in the number of reference taxa.

The emergence of these algorithms raises the question of how to make comparisons between the sets of placements that they output for two samples. One approach would be to use taxonomic annotation on the reference tree to assign taxonomic information to the query sequences. However, this is very difficult to automate: there is no canonical way to assign taxonomic information to the internal nodes, and the process can be further complicated by gene trees deviating from the species tree. Resolving such differences is an active research area, with algorithms and implementations such as Notung (Chen et al., 2000), Softparsmap (Berglund-Sonnhammer et al., 2006), and PrIME (Åkerborg et al., 2009).

Instead, we consider directly if there is a significant difference between two collections of placements on a reference tree. In doing so, we sidestep the errors introduced by the naming process and preserve the detailed information that phylogeny offers, including branch length information and information about paralogous genes. To do so, environmental sequence samples are considered as probability distributions on their corresponding reference tree as follows. Given n query sequences, each sequence fragment is assigned probability mass $1/n$. In the "point" version, the mass for a given query sequence is assigned to the location in the reference tree which best fits the location in the reference tree for that query sequence. In the "spread" version the $1/n$ mass is distributed in a continuous or discrete fashion across the tree according to the level of optimality of that location for the given query sequence.

These probability distributions can be found via likelihood-based phylogenetics. The likelihood of a phylogenetic tree under a given model of sequence evolution for a given aligned set of sequences is the probability of having those observed sequences evolve on the phylogenetic tree under the specified model (see, e.g., Felsenstein (2004)). In the likelihood setting, we define a phylogenetic placement to be a choice of attachment edge on the reference tree, position of attachment point along that edge, and "pendant" branch length (i.e. the length from the attachment point to the query node). The likelihood of a placement is the likelihood of the reference tree with the query sequence attached as specified. A maximum-likelihood point placement for a given query sequence is the maximum-likelihood attachment of the sequence to the tree.

By assuming prior probabilities on the edges, the attachment locations on those edges, and the pendant branch lengths, one can also calculate a posterior probability value for the placements. For example, one might take a uniform prior on edges and attachment locations, then an exponential or truncated uniform prior on pendant branch lengths. By integrating out pendant branch lengths, one obtains a posterior probability measure μ_i on the tree for each query sequence i; the resulting probability distribution for the sample would in that case be $\sum_i \mu_i/n$. For large data sets, it is not practical to record detailed information about the posterior probability distribution. Thus, in our implementation, the posterior probability is computed on an edge-by-edge basis by integrating out the attachment location and

the pendant branch length, and the corresponding mass is assigned to the maximum *a posteriori* location along the edge (Matsen et al., 2010). We note that the implementation of phylogenetic placement by Berger and Stamatakis (2009) allows for more flexibility in the tree, with the consequence that placements are only assigned to a given edge, rather than to a location along an edge.

2.2. **The phylogenetic Kantorovich-Rubinstein metric.** The Kantorovich-Rubinstein (KR) metric is a classical means for comparing probability distributions on a metric space. It is defined rigorously below, but it can be described intuitively in physical terms as follows. Picture each of the two probability distributions on the metric space as a collection of piles of sand with unit mass; the amount of sand at a given point is equal to the probability mass at that point. Suppose that the amount of "work" required to transport an amount of sand from one place to another is proportional to the mass of the sand moved times the distance it has to travel. Then the KR distance between two probability distributions P and Q is simply the minimum amount of work required to move sand in a configuration corresponding to P into a configuration corresponding to Q. It will require little effort to move sand between the configurations corresponding to two similar probability distributions, while more will be needed for two distributions that place most of their respective masses on disjoint regions of the metric space.

In the phylogenetic case, the optimization implicit in the definition of the KR metric can be done analytically, resulting in a closed form expression that can be evaluated in linear time, thereby enabling analysis of the volume of data produced by large-scale sequencing studies. Indeed, as shown in Section 3, the metric can be represented as a single integral over the tree, and the integral reduces to a summation with a number of terms on the order of the number of placements. In contrast, computing the KR metric in Euclidean spaces of dimension greater than one requires a linear programming optimization step.

It is remarkable that the point version of this closed-form expression for the phylogenetic KR distance (although apparently not the optimal mass transport justification for the distance) was intuited by microbial ecologists and called "weighted UniFrac." The original UniFrac distance (Lozupone and Knight, 2005) measured the fraction of branch length shared between two samples when a tree is constructed on both samples together. When the branches are weighted by the number of samples below them (Lozupone et al., 2007), one obtains a formula identical to the point version of the KR distance – see (2) below.

By demonstrating a connection between the classical KR formulation and a construct that has already proved useful in microbial ecology, we are able to leverage the statistical foundations of the KR perspective. Specifically, we introduce L^p generalizations of the KR metric that are analogous to ones on the real line due to Zolotarev (Rachev, 1991; Rachev and Rüschendorf, 1998) – the KR metric corresponds to the case $p = 1$. The generalizations do not arise from optimal mass transport considerations, but we remark in Section 4 that the square of the $p = 2$ version does have an appealing interpretation as the amount of variability in a pooling of the two samples that is not accounted for by the variability in each of them.

A natural approach to assessing the significance of an observed distance between the probability distributions associated with two samples is the following procedure analogous to a permutation test from classical non-parametric statistics that

we expand upon in Section 6. Suppose that we have two samples with m and n placements, respectively. Imagine creating a new pair of samples by taking some other subset of size m and its complement from the set of all $m+n$ placements and then computing the distance between the resulting two probability distributions. The proportion of the $\binom{m+n}{m}$ choices of such pairs of samples that result in a distance larger than the one observed in the data is an indication of the significance of the observed distance. Of course, we can rephrase this procedure as taking a uniform random subset of placements of size m and its complement and asking for the probability that the distance between the resulting probability distributions is greater than the observed one. Consequently, it is possible to approximate the proportion/probability by taking repeated independent choices of the random subset and recording the proportion of times there is a distance greater than the observed one. We describe the distribution of the distance between the two random probability distributions produced from a uniform random subset of placements of size m and its complement of size n as the *distribution of the distance under the null hypothesis of no clustering*.

We show in Section 7 that the distribution of the distance under the null hypothesis of no clustering is approximately that of a readily-computable functional of a Gaussian process indexed by the tree and that this Gaussian process is relatively simple to simulate. Moreover, we observe that when $p = 2$ this approximate distribution is that of the square root of a weighted sum of χ_1^2 random variables.

Also, we recall in Section 5 the well-known fact that any tree is a *Hadamard space*: a Hadamard space is a simply connected complete metric space in which there is a suitable notion of the length of a path in the space, the distance between two points is the infimum of the lengths of the paths joining the points, and the space has nonpositive curvature in an appropriate sense – see (Burago et al., 2001). Equivalently, a Hadamard space is a complete CAT(0) space in the sense of (Bridson and Haefliger, 1999). CAT(0) spaces have already made an appearance in phylogenetics in the construction of spaces of phylogenetic trees (Billera et al., 2001). Any compactly supported probability distribution on a Hadamard space has an appropriately defined *barycenter* that shares some of the properties of the expectation of a probability distribution on a Euclidean space and acts as a sensible "single point summary" of the distribution. We describe a simple procedure for calculating the barycenter of a probability distribution on a tree.

To illustrate the usefulness of probability metrics for distinguishing between collections of placements, we perform an example analysis using a dataset on a photosystem gene that was collected to investigate the difference in populations when organic sulfur is added to seawater. A significant difference is seen when comparing the two conditions using the KR metric and its generalizations, and the sampling distribution of the various distances under the null hypothesis is quite consistent as the exponent p varies. It is noteworthy that for this data set the barycenters of the two samples are quite close together, and so the significant value of the observed distance between the two sets of placements is not simply due to them occupying essentially non-overlapping regions of the reference tree. Also, it is not immediately clear from plots of the distributions of distances between pairs of placements for the two samples and a pooling of them that there is clustering present.

There are a number of ways to compare microbial samples in a phylogenetic context besides the method presented here. One popular means of comparing samples is the "parsimony test," by which the most parsimonious assignment of internal nodes of the phylogenetic tree to communities is found; the resulting parsimony score is interpreted as a measure of difference between communities (Slatkin and Maddison, 1989; Schloss and Handelsman, 2006). Another interesting approach is to consider a "generalized principal components analysis" whereby the tree structure is incorporated into the process of finding principal components of the species abundances (Bik et al., 2006; Purdom, 2008). The Kantorovich-Rubinstein metric complements these methods by providing a means of comparing samples that leverages established statistical methodology, that takes into account uncertainty in read location, and can be visualized directly on the tree.

3. Probability metrics on trees

In this section we introduce the phylogenetic Kantorovich-Rubinstein metric, which is a particular case of the family of Wasserstein metrics. We use a dual formulation of the KR metric to show that it can be calculated in linear time via a simple integral over the tree.

Let T be a phylogenetic tree with assigned branch lengths. Write d for the path distance on T. We assume that probability distributions have been given on the tree via collections of either "point" or "spread" placements as described in the introduction.

For a metric space (S, r), the Kantorovich-Rubinstein distance $Z(P, Q)$ between two Borel probability distributions P and Q on S is defined as follows. Let $\mathcal{R}(P, Q)$ denote the set of probability distributions R on the product space $S \times S$ with the property $R(A \times S) = P(A)$ and $R(S \times B) = Q(B)$ for all Borel sets A and B (that is, the two marginal distributions of R are P and Q). Then,

$$Z(P, Q) := \inf \left\{ \int_{S \times S} r(x, y) \, R(dx, dy) : R \in \mathcal{R}(P, Q) \right\};$$

see, for example, (Rachev, 1991; Rachev and Rüschendorf, 1998; Villani, 2003; Ambrosio et al., 2008; Villani, 2009).

There is an alternative formula for $Z(P, Q)$ that comes from convex duality. Write $\mathcal{L}$ for the set of functions $f : S \to \mathbb{R}$ with the Lipschitz property $|f(x) - f(y)| \le r(x, y)$ for all $x, y \in S$. Then,

$$Z(P, Q) = \sup \left\{ \int_S f(x) \, P(dx) - \int_S f(y) \, Q(dy) : f \in \mathcal{L} \right\}.$$

We can use this expression to get a simple explicit formula for $Z(P, Q)$ when $(S, r) = (T, d)$.

Given any two points $x, y \in T$, let $[x, y]$ be the arc between them. There is a unique Borel measure λ on T such that $\lambda([x, y]) = d(x, y)$ for all $x, y \in T$. We call λ the *length measure*. It is analogous to Lebesgue measure on the real line. Fix a distinguished point $\rho \in T$, which we think of as the "root" of the tree. For any $f \in \mathcal{L}$ with $f(\rho) = 0$, there is an λ-a.e. unique Borel function $g : T \to [-1, 1]$ such that $f(x) = \int_{[\rho, x]} g(y) \, \lambda(dy)$ (this follows easily from the analogous fact for the real line).

Given $x \in T$, put $\tau(x) := \{y \in T : x \in [\rho, y]\}$; that is, if we draw the tree with the root ρ at the top of the page, then $\tau(x)$ is the sub-tree below x. Observe that

if $h : T \to \mathbb{R}$ is a bounded Borel function and μ is a Borel probability distribution on T, then we have the integration-by-parts formula

$$\int_T \left(\int_{[\rho,x]} h(y) \, \lambda(dy) \right) \mu(dx) = \int_{T \times T} 1_{[\rho,x]}(y) h(y) \, (\mu \otimes \lambda)(dx, dy)$$

$$= \int_{T \times T} 1_{\tau(y)}(x) h(y) \, (\mu \otimes \lambda)(dx, dy)$$

$$= \int_T h(y) \left(\int_{\tau(y)} \mu(dx) \right) \lambda(dy)$$

$$= \int_T h(y) \mu(\tau(y)) \, \lambda(dy).$$

Thus, if P and Q are two Borel probability distributions on T we have

$$Z(P, Q) = \sup \left\{ \int_T g(y) \left[P(\tau(y)) - Q(\tau(y)) \right] \lambda(dy) : -1 \leq g \leq +1 \right\}.$$

It is clear that the integral is maximized by taking $g(y) = +1$ (resp. $g(y) = -1$) when $P(\tau(y)) > Q(\tau(y))$ (resp. $P(\tau(y)) < Q(\tau(y))$), so that

(1) $$Z(P, Q) = \int_T |P(\tau(y)) - Q(\tau(y))| \, \lambda(dy).$$

Interestingly, the quantity on the LHS does not depend on the choice of root, whereas *a priori* the quantity on the RHS appears to.

3.1. **Zolotarev-type L^p generalization.** The metric defined in (1) generalizes to an L^p Zolotarev-type version: for $0 < p < \infty$ we have the distances

$$Z_p(P, Q) = \left[\int_T |P(\tau(y)) - Q(\tau(y))|^p \, \lambda(dy) \right]^{\frac{1}{p} \wedge 1}$$

– see (Rachev, 1991; Rachev and Rüschendorf, 1998) for a discussion of analogous metrics for probability distributions on the real line. Intuitively, large p gives more weight in the distance to parts of the tree which are maximally different in terms of P and Q, while small p gives more weight to differences which require lots of transport.

The position of the root ρ also does not matter for this generalized version. Indeed, if ρ' and ρ'' are two choices of the root and $[\rho', \rho'']$ is the segment in the tree T joining them, then (in the obvious notation)

$$P(\tau'(y)) - Q(\tau'(y)) = P(\tau''(y)) - Q(\tau''(y)), \quad y \notin [\rho', \rho''],$$

and

$$P(\tau'(y)) - Q(\tau'(y)) = - \left[P(\tau''(y)) - Q(\tau''(y)) \right], \quad \lambda - \text{a.e. } y \in [\rho', \rho''];$$

(there is equality provided y is not an atom of P or Q), so

$$\int_T |P(\tau'(y)) - Q(\tau'(y))|^p \, \lambda(dy) = \int_T |P(\tau''(y)) - Q(\tau''(y))|^p \, \lambda(dy)$$

and the choice of the "root" is irrelevant.

3.2. **Connection with the "weighted UniFrac" metric.** Assume there are m sequences from sample A and n sequences in sample B. The "weighted UniFrac" formula is

$$(2) \qquad u = \sum_{i=1}^{n} \ell_i \left| \frac{A_i}{m} - \frac{B_i}{n} \right|$$

where ℓ_i is the length of branch i, A_i and B_i are the number of descendants of branch i from communities A and B respectively, and m and n are the total number of sequences from communities A and B, respectively (Lozupone et al., 2007). This is the same as (1) when P assigns point mass $1/m$ to each of the leaves in community A, and Q assigns point mass $1/n$ to each of the leaves in community B.

4. $Z_2^2(P,Q)$ AND ANOVA

In this section we demonstrate how $Z_2^2(P,Q)$ can be interpreted as a difference between the pooled average of pairwise distances and the average for each sample individually.

Let $\pi_1, \ldots, \pi_m$ (resp. $\pi_{m+1}, \ldots, \pi_{m+n}$) be the placements in the first (resp. second) sample, so that each π_k is a probability distribution on the tree T, $P = \frac{1}{m} \sum_{k=1}^{m} \pi_k$, and $Q = \frac{1}{n} \sum_{k=m+1}^{m+n} \pi_k$.

Observe that

$$
\begin{aligned}
Z_2^2(P,Q) &= \int_T |P(\tau(u)) - Q(\tau(u))|^2 \, \lambda(du) \\
&= \int_T P(\tau(u))^2 \, \lambda(du) - 2 \int_T P(\tau(u)) Q(\tau(u)) \, \lambda(du) \\
&\quad + \int_T Q(\tau(u))^2 \, \lambda(du) \\
&= \int_T \left[\int_T \int_T 1_{[\rho,v]}(u) 1_{[\rho,w]}(u) \, P(dv) \, P(dw) \right] \lambda(du) \\
&\quad - 2 \int_T \left[\int_T \int_T 1_{[\rho,v]}(u) 1_{[\rho,w]}(u) \, P(dv) \, Q(dw) \right] \lambda(du) \\
&\quad + \int_T \left[\int_T \int_T 1_{[\rho,v]}(u) 1_{[\rho,w]}(u) \, Q(dv) \, Q(dw) \right] \lambda(du)
\end{aligned}
$$

Now, $[\rho, v] \cap [\rho, w]$ is an arc of the form $[\rho, v \wedge w]$, where $v \wedge w$ is the "most recent common ancestor" of v and w relative to the root ρ. Hence,

$$
\begin{aligned}
\int_T 1_{[\rho,v]}(u) 1_{[\rho,w]}(u) \, \lambda(du) &= \lambda([\rho, v \wedge w]) \\
&= d(\rho, v \wedge w) \\
&= \frac{1}{2} \left[d(\rho, v) + d(\rho, w) - d(v, w) \right],
\end{aligned}
$$

and

$$
\begin{aligned}
Z_2^2(P, Q) = &\int_T \int_T \frac{1}{2} \left[d(\rho, v) + d(\rho, w) - d(v, w) \right] P(dv) P(dw) \\
&- 2 \int_T \int_T \frac{1}{2} \left[d(\rho, v) + d(\rho, w) - d(v, w) \right] P(dv) Q(dw) \\
&+ \int_T \int_T \frac{1}{2} \left[d(\rho, v) + d(\rho, w) - d(v, w) \right] Q(dv) Q(dw) \\
= &\frac{1}{2} \left[2 \int_T \int_T d(v, w) P(dv) Q(dw) \right. \\
&- \int_T \int_T d(v, w) P(dv) P(dw) \\
&\left. - \int_T \int_T d(v, w) Q(dv) Q(dw) \right].
\end{aligned}
$$

Therefore, if we set

$$
R := \frac{m}{m+n} P + \frac{n}{m+n} Q = \frac{1}{m+n} \sum_k \pi_k,
$$

then

$$
\begin{aligned}
2mn \int_T \int_T d(v, w) P(dv) Q(dw) = &(m+n)^2 \int_T \int_T d(v, w) R(dv) R(dw) \\
&- m^2 \int_T \int_T d(v, w) P(dv) P(dw) \\
&- n^2 \int_T \int_T d(v, w) Q(dv) Q(dw)
\end{aligned}
$$

and

$$
\begin{aligned}
Z_2^2(P, Q) = &\frac{1}{2} \left[\frac{(m+n)^2}{mn} \int_T \int_T d(v, w) R(dv) R(dw) \right. \\
&- \frac{(mn + m^2)}{mn} \int_T \int_T d(v, w) P(dv) P(dw) \\
&\left. - \frac{(mn + n^2)}{mn} \int_T \int_T d(v, w) Q(dv) Q(dw) \right] \\
= &\frac{1}{2} \frac{(m+n)^2}{mn} \left[\int_T \int_T d(v, w) R(dv) R(dw) \right. \\
&- \left\{ \frac{m}{m+n} \int_T \int_T d(v, w) P(dv) P(dw) \right. \\
&\left. \left. + \frac{n}{m+n} \int_T \int_T d(v, w) Q(dv) Q(dw) \right\} \right].
\end{aligned}
$$

Thus, $Z_2^2(P, Q)$ gives an indication of the "variability" present in the pooled collection π_i, $1 \leq i \leq m+n$, that is over and above the "variability" in the two collections π_i, $1 \leq i \leq m$, and π_i, $m+1 \leq i \leq m+n$.

5. CALCULATING THE BARYCENTER

It can be useful to compare probability distributions on a metric space by calculating a suitably defined "center of mass" that provides a single point summary for

each distribution. Recall the standard fact that if P is a probability distribution on a Euclidean space such that $\int |y - x|^2 P(dy)$ is finite for some (and hence all) x, then the function $x \mapsto \int |y - x|^2 P(dy)$ has a unique minimum at $x_0 = \int y\, P(dy)$. It therefore makes sense to say that a probability distribution P on an arbitrary metric space (S, r) has a *barycenter* at x_0 if $\int r(x, y)^2 P(dy)$ is finite for some (and hence all x) and the function $x \mapsto \int r(x, y)^2 P(dy)$ has a unique minimum at x_0. In general, a barycenter need not exist. However, it is well-known that barycenters do exist for probability distributions on Hadamard spaces – see Chapter 9 of (Burago et al., 2001) for the formal definition of a Hadamard space and a proof of this fact. It is a straightforward exercise to check that a tree is a Hadamard space – see Example II.1.15(4) of (Bridson and Haefliger, 1999) and note the remark after Definition II.1.1 of (Bridson and Haefliger, 1999) that a Hadamard space is the same thing as a complete CAT(0) space. Thus, a probability distribution on a tree possesses a barycenter in the above sense.

This fact may also be established directly as follows. As a continuous function on a compact metric space, the function $f : T \to \mathbb{R}_+$ defined by $f(x) := \int_T d(x, y)^2 P(dy)$ achieves its infimum. Suppose that the infimum is achieved at two points x' and x''. Define a function $\gamma : [0, d(x', x'')] \to [x', x'']$, where $[x', x''] \subseteq T$ is the arc between x' and x'', by the requirement that $\gamma(t)$ is the unique point in $[x', x'']$ that is distance t from x'. It is straightforward to check that the composition $f \circ \gamma$ is *strongly convex*; that is,

$$(f \circ \gamma)(\alpha r + (1 - \alpha)s) < \alpha(f \circ \gamma)(r) + (1 - \alpha)(f \circ \gamma)(s)$$

for $0 < \alpha < 1$ and $r, s \in [0, d(x', x'')]$. In particular, $f(\gamma(d(x', x'')/2)) = (f \circ \gamma)(d(x', x'')/2) < (f(x') + f(x''))/2$, contradicting the definitions of x' and x''.

We next consider how to compute the barycenter of a probability distribution P on a tree (T, d). For each point $u \in T$ there is the associated set of directions in which it is possible to proceed when leaving u. There is one direction for every connected component of $T \setminus \{u\}$. Thus, there is just one direction associated with a leaf, two directions associated with a point in the interior of an edge, and k associated with a vertex of degree k. Given a point u and a direction δ, write $T(u, \delta)$ for the subset of T consisting of points $v \neq u$ such that the unique path connecting u and v departs u in the direction δ, set

$$D(u, \delta) := -\int_{T(u,\delta)} d(u, y)\, P(dy) + \int_{T \setminus T(u,\delta)} d(u, y)\, P(dy),$$

and note that

$$\lim_v \frac{1}{d(u, v)} \left[\int_T d(v, y)^2 P(dy) - \int_T d(u, y)^2 P(dy) \right] = 2D(u, \delta),$$

where the limit is taken over $v \to u$, $v \in T(u, \delta)$. Note that if u is in the interior of an edge $[a, b]$ and b is in the direction δ from u, u is in the direction α from a, and

u is in the direction β from b, then

$$
\begin{aligned}
D(u, \delta) &= -\int_{T \backslash T(b, \beta)} d(u, y)\, P(dy) - \int_{(u,b)} d(u, y)\, P(dy) \\
&\quad + \int_{T \backslash T(a, \alpha)} d(u, y)\, P(dy) + \int_{(a,u)} d(u, y)\, P(dy) \\
&= -\int_{T \backslash T(b, \beta)} d(a, y)\, P(dy) + d(a, u) P(T \backslash T(b, \beta)) \\
&\quad - \int_{(u,b)} d(a, y)\, P(dy) + d(a, u) P((u, b)) \\
&\quad + \int_{T \backslash T(a, \alpha)} d(a, y)\, P(dy) + d(a, u) P(T \backslash T(a, \alpha)) \\
&\quad + d(a, u) P((a, u)) - \int_{(a,u)} d(a, y)\, P(dy) \\
&= D(a, \alpha) + d(a, u).
\end{aligned}
$$

If for some vertex u of the reference tree $D(u, \delta) \geq 0$ for all directions δ associated with u, then u is the barycenter (this case includes the trivial one in which u is a leaf and all the mass of P is concentrated on u). If there is no such vertex, then there must be a unique pair of neighboring vertices a and b such that $D(a, \alpha) < 0$ and $D(b, \beta) < 0$, where α is the direction from a pointing towards b and β is the direction from b pointing towards a. In that case, the barycenter must lie on the edge between a and b, and it follows from the calculations above that the barycenter is the point $u \in (a, b)$ such that $d(a, u) = -D(a, \alpha)$.

6. Assessing significance

To assess the significance of the observed distance between the probability distributions associated with two samples of placed reads of size m and n, we use an idea familiar from classical nonparametric statistics: we imagine creating all $\binom{m+n}{m}$ pairs of "samples" that arise from placing m of the reads into one sample and the remaining n into the other, computing the distances between the two probability distributions on the reference tree that result from the placed reads, and determining what proportion of these distances exceed the distance observed in the data. This proportion may be thought of as a p-value for a test of the null hypothesis of no clustering against an alternative of some degree of clustering.

Of course, for most values of m and n it is infeasible to actually perform this exhaustive listing of distances. We observe that if, as above, the pooled collection of placed reads from both samples is $\pi_1, \ldots, \pi_{m+n}$, $I \subseteq \{1, \ldots, m+n\}$ is a uniformly distributed random subset with cardinality m (that is, all $\binom{m+n}{m}$ values are equally likely), $J := I^c$ is the complement of I, $\tilde{P}$ is the random probability distribution $\frac{1}{m} \sum_{i \in I} \pi_i$, and $\tilde{Q}$ is the random probability distribution $\frac{1}{n} \sum_{j \in J} \pi_j$, then the proportion of interest is simply the probability that the distance between $\tilde{P}$ and $\tilde{Q}$ exceeds the distance between P and Q. We can approximate this probability in the obvious way by taking independent replicates of (I, J) and hence of $(\tilde{P}, \tilde{Q})$ and looking at the proportion of them that result in distances greater than the observed one. We illustrate this Monte Carlo approximation procedure in Section 8.

7. GAUSSIAN APPROXIMATION

Although the Monte Carlo approach to approximating a p-value in Section 6 is conceptually straightforward, it is tempting to explore whether there are further approximations to the outcome of this procedure that give satisfactory results but require less computation.

Recall that $\pi_1, \ldots, \pi_{m+n}$ is the pooled collection of placed reads and that $\tilde{P} = \frac{1}{m} \sum_{i \in I} \pi_i$ and $\tilde{Q} = \frac{1}{n} \sum_{j \in J} \pi_j$, where I is a uniformly distributed random subset of $\{1, \ldots, m+n\}$ and J is its complement. Write

$$G_k(u) := \pi_k(\tau(u)) \text{ for any } u \in T, \ 1 \le k \le m+n,$$

where we recall that $\tau(u)$ is the tree below u relative to the root ρ. Define a T-indexed stochastic process $X = (X(u))_{u \in T}$ by

$$
\begin{aligned}
X(u) &:= \tilde{P}(\tau(u)) - \tilde{Q}(\tau(u)) \\
&= \frac{1}{m} \sum_{i \in I} G_i(u) - \frac{1}{n} \sum_{j \in J} G_j(u).
\end{aligned}
$$

Then,

$$Z_p(\tilde{P}, \tilde{Q}) = \left[\int_T |X(u)|^p \ \lambda(du) \right]^{\frac{1}{p} \wedge 1}.$$

If H_k, $1 \le k \le m+n$, is the indicator random variable for the event $\{k \in I\}$, then

$$X(u) = \sum_{k=1}^{m+n} \left[\left(\frac{1}{m} + \frac{1}{n} \right) H_k - \frac{1}{n} \right] G_k(u).$$

Writing $\mathbb{E}$, $\mathbb{V}$, and $\mathbb{C}$ for expectation, variance, and covariance, we have

$$\mathbb{E}[H_i] = \frac{m}{m+n},$$

$$\mathbb{V}[H_i] = \frac{m}{m+n} \frac{n}{m+n},$$

and

$$\mathbb{C}[H_i, H_j] = -\frac{1}{m+n-1} \frac{m}{m+n} \frac{n}{m+n}, \quad i \ne j.$$

It follows that

$$\mathbb{E}[X(u)] = 0$$

and
$$\mathbb{C}[X(u), X(v)]$$

$$= \frac{1}{mn} \left(\sum_i G_i(u)G_i(v) - \frac{1}{m+n-1} \sum_{i \neq j} G_i(u)G_j(v) \right)$$

$$\approx \frac{1}{mn} \left[\sum_i G_i(u)G_i(v) - \frac{1}{m+n} \sum_{i,j} G_i(u)G_j(v) \right]$$

$$= \frac{1}{mn} \left[\sum_i G_i(u)G_i(v) - (m+n) \left(\frac{1}{m+n} \sum_i G_i(u) \right) \left(\frac{1}{m+n} \sum_j G_j(v) \right) \right]$$

$$= \frac{1}{mn} \left[\sum_i \left(G_i(u) - \bar{G}(u) \right) \left(G_i(v) - \bar{G}(v) \right) \right]$$

$$=: \Gamma(u,v)$$

when $m + n$ is large, where $\bar{G}(u) := \frac{1}{m+n} \sum_k G_k(u)$.

Remark 7.1. *In the case of point placements, with the probability distribution π_k being the point mass at $w_k \in T$ for $1 \leq k \leq m+n$, then*

$$\Gamma(u,v) = \frac{1}{mn} \left[\sum_k \#\{k : u \in [\rho, w_k], \, v \in [\rho, w_k]\} \right.$$

$$\left. - \frac{1}{m+n} \#\{k : u \in [\rho, w_k]\} \#\{k : v \in [\rho, w_k]\} \right].$$

By a standard central limit theorem for exchangeable random variables, the process X is approximately Gaussian with covariance kernel Γ when $m + n$ is large. A straightforward calculation shows that we may construct a Gaussian process ξ with covariance kernel Γ by taking independent standard Gaussian random variables $\eta_1, \ldots, \eta_{m+n}$ and setting

$$\xi(u) = \frac{1}{\sqrt{mn}} \left[\sum_i G_i(u)\eta_i - \frac{1}{m+n} \left(\sum_i G_i(u) \right) \left(\sum_i \eta_i \right) \right].$$

It follows that the distribution of $Z_p(\tilde{P}, \tilde{Q})$ is approximately that of the random variable

$$(3) \qquad \left[\int_T |\xi(u)|^p \, \lambda(du) \right]^{\frac{1}{p} \wedge 1},$$

One can repeatedly sample the normal random variates η_i and numerically integrate (3) to approximate the distribution of this integral.

There is an even simpler approach for the case $p = 2$. Let μ_k^2, $k = 1, 2, \ldots$, and ψ_k, $k = 1, 2, \ldots$, be the positive eigenvalues and corresponding normalized eigenfunctions of the non-negative definite, self-adjoint, compact operator on $L^2(\lambda)$ that maps the function f to the function $\int_T \Gamma(\cdot, v)f(v) \, \lambda(dv)$. The functions $\mu_k \psi_k$, $k = 1, 2, \ldots$, form an orthonormal basis for the reproducing kernel Hilbert space associated with Γ and the Gaussian process ξ has the Karhunen-Loève expansion

$$\xi(u) = \sum_k \mu_k \psi_k(u) \eta_k,$$

where η_k, $k = 1, 2, \ldots$, are independent standard Gaussian random variables – see (Jain and Marcus, 1978) for a review of the theory of reproducing kernel Hilbert spaces and the Karhunen-Loève expansion.

Therefore,

$$\int_T |\xi(u)|^2 \, \lambda(du) = \sum_k \mu_k^2 \eta_k^2,$$

and the distribution of $Z_2^2(\tilde{P}, \tilde{Q})$ is approximately that of a certain positive linear combination of independent χ_1^2 random variables.

The eigenvalues of the operator associated with Γ can be found by calculating the eigenvalues of a related matrix as follows. Define an $(m+n) \times (m+n)$ non-negative definite, self-adjoint matrix M given by

$$M_{ij} := \frac{1}{mn} \int_T \left(G_i(u) - \bar{G}(u)\right)\left(G_j(u) - \bar{G}(u)\right) \, \lambda(du).$$

Note that if we have point placements at locations $w_k \in T$ for $1 \leq k \leq m+n$ as in Remark 7.1, then

$$M = \frac{1}{mn}\left(I - \frac{1}{m+n}\mathbf{1}\mathbf{1}^\top\right)N\left(I - \frac{1}{m+n}\mathbf{1}\mathbf{1}^\top\right),$$

where I is the identity matrix, $\mathbf{1}$ is the vector which has 1 for every entry, and the matrix N has (i, j) entry given by the distance from the root to the "most recent common ancestor" of w_i and w_j.

Suppose that x is an eigenvector of M for the positive eigenvalue ν^2. Set

(4)
$$\psi(u) := \sum_j \left(G_j(u) - \bar{G}(u)\right) x_j.$$

Observe that

$$\int_T \Gamma(u, v)\psi(v) \, \lambda(dv)$$

$$= \frac{1}{mn} \int_T \left[\sum_i \left(G_i(u) - \bar{G}(u)\right)\left(G_i(v) - \bar{G}(v)\right)\right] \sum_j \left(G_j(v) - \bar{G}(v)\right) x_j \, \lambda(dv)$$

$$= \sum_i \left(G_i(u) - \bar{G}(u)\right) \sum_j M_{ij} x_j$$

$$= \sum_i \left(G_i(u) - \bar{G}(u)\right) \nu^2 x_i$$

$$= \nu^2 \psi(u),$$

and so ψ is an (unnormalized) eigenfunction of the operator on $L^2(\lambda)$ defined by the covariance kernel Γ with eigenvalue ν^2.

Conversely, suppose that μ^2 is an eigenvalue of the operator with eigenfunction ϕ. Set

$$x_j := \int_T \left(G_j(v) - \bar{G}(v)\right) \phi(v) \, \lambda(dv).$$

Then,

$$\sum_j M_{ij} x_j$$

$$= \sum_j \frac{1}{mn} \int_T \left(G_i(u) - \bar{G}(u)\right) \left(G_j(u) - \bar{G}(u)\right) \lambda(du)$$

$$\times \int_T \left(G_j(v) - \bar{G}(v)\right) \phi(v) \lambda(dv)$$

$$= \int_T \left(G_i(u) - \bar{G}(u)\right)$$

$$\times \left[\int_T \frac{1}{mn} \sum_j \left(G_j(u) - \bar{G}(u)\right) \left(G_j(v) - \bar{G}(v)\right) \phi(v) \lambda(dv) \right] \lambda(du)$$

$$= \int_T \left(G_i(u) - \bar{G}(u)\right) \left[\int_T \Gamma(u, v) \phi(v) \lambda(dv) \right] \lambda(du)$$

$$= \int_T \left(G_i(u) - \bar{G}(u)\right) \mu^2 \phi(u) \lambda(du)$$

$$= \mu^2 x_i,$$

so that μ^2 is an eigenvalue of M with (unnormalized) eigenvector of x.

It follows that the positive eigenvalues of the operator associated with Γ coincide with those of the matrix M and have the same multiplicities.

However, we don't actually need to compute the eigenvalues of M to implement this approximation. Because M is orthogonally equivalent to a diagonal matrix with the eigenvalues of M on the diagonal, we have from the invariance under orthogonal transformations of the distribution of the random vector $\eta := (\eta_1, \ldots, \eta_{m+n})^\top$ that $\sum_k \mu_k^2 \eta_k^2$ has the same distribution as $\eta^\top M \eta$. Thus, the distribution of the random variable $Z_2^2(\tilde{P}, \tilde{Q})$ is approximately that of $\sum_{ij} M_{ij} \eta_i \eta_j$. In the example application of Section 8, this provides a reasonable though not perfect approximation (Figure 6).

One might hope to go even further in the $p = 2$ case and obtain an analytic approximation for the distribution $\sum_k \mu_k^2 \eta_k^2$ or a useful upper bound for its right tail.

It is shown in (Hwang, 1980) that if we order the positive eigenvalues so that $\mu_1^2 \geq \mu_2^2 \geq \ldots$ and assume that $\mu_1^2 > \mu_2^2$, then

$$\mathbb{P}\left\{ \sum_k \mu_k^2 \eta_k^2 \geq r \right\} \sim \sqrt{\frac{2}{\pi}} \mu_1 \prod_{k>1} \left(1 - \frac{\mu_k^2}{\mu_1^2}\right)^{-\frac{1}{2}} r^{-\frac{1}{2}} \exp\left(-\frac{r}{2\mu_1^2}\right),$$

in the sense that the ratio of the two sides converges to one as $r \to \infty$. It is not clear what the rate of convergence is in this result and it appears to require a detailed knowledge of the spectrum of the matrix M to apply it.

Gaussian concentration inequalities such as Borell's inequality (see, for example, Section 4.3 of (Bogachev, 1998)) give bounds on the right tail that only require a knowledge of $\mathbb{E}[(\sum_k \mu_k^2 \eta_k^2)^{\frac{1}{2}}]$ and μ_1^2, but these bounds are far too conservative for the example in Section 8.

There is a substantial literature on various series expansions of densities of positive linear combinations of independent χ_1^2 random variables. Some representative papers are (Robbins and Pitman, 1949; Gurland, 1955; Pachares, 1955; Ruben,

1962; Kotz et al., 1967; Gideon and Gurland, 1976). However, it seems that applying such results would also require a detailed knowledge of the spectrum of the matrix M as well as a certain amount of additional computation to obtain the coefficients in the expansion and then to integrate the resulting densities, and this may not be warranted given the relative ease with which it is possible to repeatedly simulate the random variable $\eta^\top M \eta$.

8. Example application

To demonstrate the use of the Z_p metric in an example application, we investigated variation in expression levels for the *psbA* gene for an experiment in the Sargasso Sea (Vila-Costa et al., 2010). Metatranscriptomic data was downloaded from the CAMERA website (http://camera.calit2.net/), and a *psbA* alignment was supplied by Robin Kodner. Searching and alignment was performed using HMMER (Eddy, 1998), a reference tree was inferred using RAxML (Stamatakis, 2006), and phylogenetic placement was performed using pplacer (Matsen et al., 2010). Although not yet formally released software, the code used to calculate the KR metrics is available at http://github.com/matsen/mokaphy.

Visual inspection of the trees fattened by number of placements showed the same overall pattern with some minor differences (Figure 1 and 2). Application of the KR metric revealed a significant difference between the two samples. The value of the Z_1 for this example (using spread placements) was 0.006601; this is far out on the tail of the distribution (Figure 6), and is in fact larger than any of the 1000 replicates generated via shuffling or the Gaussian-based approximation.

Such a low p-value prompts the question of whether the two distributions could be distinguished using a "weaker" methodology. We observe that the density plot of the distributions of pairwise distances (Figure 3) does not reveal an obvious difference in the pairwise distances and that the two barycenters are close together (Figure 4).

It was not intuitively obvious to us how varying p would affect the distribution of the Z_p distance under the null hypothesis of no clustering. To investigate this question, we plotted the observed distance along with boxplots of the null distribution for a collection of different p (Figure 5). It is apparent that there is a consistent conclusion over a wide range of values of p.

One can also visualize the difference between the two samples by drawing the tree such that the branch thicknesses represent the minimal amount of "mass" which flows through that edge in the optimal transport of mass implicit in the computation of $Z_1(P, Q)$ (Figure 7). The color of the edge represents the sign of the movement.

9. Conclusion

As sequencing becomes faster and less expensive, it will become increasingly common to have a collection of large data sets for a given gene. Phylogenetic placement can furnish an evolutionary context for query sequences, resulting in each data set being represented as a probability distribution on a phylogenetic tree. The Kantorovich-Rubinstein metric is a natural means to compare those probability distributions. In this paper we showed that the UniFrac metric is the phylogenetic Kantorovich-Rubinstein metric for point placements. We explored Zolotarev-type generalizations of the KR metric, showed how to approximate the

limiting distribution and made connections with the analysis of variance in the $p = 2$ case.

The phylogenetic KR metric and its generalizations can be used any time one wants to compare two probability distributions on a tree. However, our software implementation is designed with metagenomic and metatranscriptomic investigations in mind; for this reason it is tightly integrated with the phylogenetic placement software pplacer (Matsen et al., 2010). With more than two genes, techniques such as principal components analysis could be applied to the pairwise distances to cluster environments based on the KR distances as has been done with UniFrac (Lozupone and Knight, 2005; Lozupone et al., 2008; Hamady et al., 2009). In a large scale metagenomic analysis with several genes, the KR and related metric could be used to scan for "discriminating" genes, i.e. genes whose frequency differs between environments.

One potential future extension not explored here is to partition the tree into subsets in a principal components fashion for a single data set. Recall that (4) gives a formula for the eigenfunctions of the covariance kernel Γ given the eigenvectors of M. For any k, one could partition the tree into subsets based on the sign of the product of the first k eigenfunctions, which would be analogous to partitioning Euclidean space by the hyperplanes associated with the first k eigenvectors in traditional principal components analysis.

Future methods will also need to take details of the DNA extraction procedure into account. Recent work shows that current lab methodology is unable to recover absolute mixture proportions due to differential ease of DNA extraction between organisms (Morgan et al., 2010). However, relative abundance between samples for a given organism with a fixed laboratory protocol potentially can be measured, assuming consistent DNA extraction protocols are used. An important next step is to incorporate such organism-specific biases into the sort of analysis described here.

10. Tables and figures

Acknowledgements

References

Åkerborg, Ö., B. Sennblad, L. Arvestad, and J. Lagergren (2009). Simultaneous Bayesian gene tree reconstruction and reconciliation analysis. *PNAS 106*(14), 5714.

Ambrosio, L., N. Gigli, and G. Savaré (2008). *Gradient flows in metric spaces and in the space of probability measures* (Second ed.). Lectures in Mathematics ETH Zürich. Basel: Birkhäuser Verlag.

Baker, B. and J. Banfield (2003). Microbial communities in acid mine drainage. *FEMS Microbiol. Ecol. 44*(2), 139–152.

Berger, S. and A. Stamatakis (2009). Evolutionary Placement of Short Sequence Reads. *Submitted to Sys. Biol.; http://arxiv.org/abs/0911.2852*.
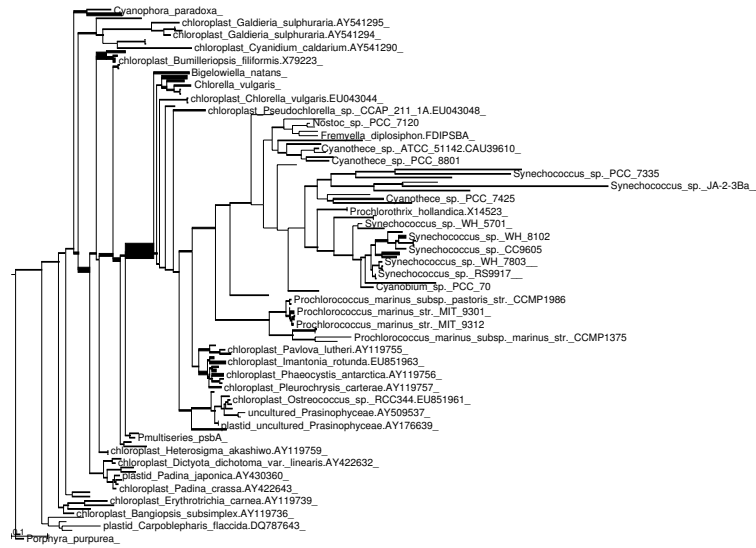
FIGURE 1. Tree with branches thickened as a linear function of the number of placements in the control sample placed on that edge.
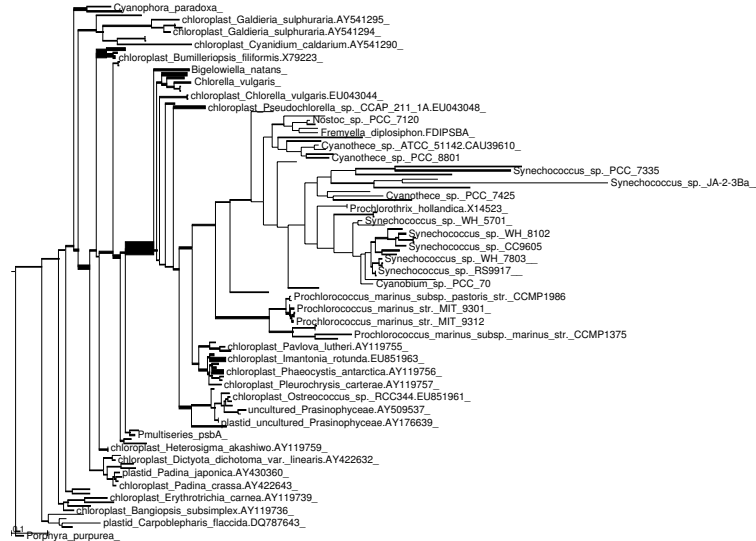

FIGURE 2. Tree as in Figure 1 but for the DMSP-treated sample.

Berglund-Sonnhammer, A., P. Steffansson, M. Betts, and D. Liberles (2006). Optimal gene trees from sequences and species trees using a soft interpretation of parsimony. *J. Mol. Evol. 63*(2), 240–250.

Bik, E., P. Eckburg, S. Gill, K. Nelson, E. Purdom, F. Francois, G. Perez-Perez, M. Blaser, and D. Relman (2006). Molecular analysis of the bacterial microbiota in the human stomach. *Proceedings of the National Academy of Sciences of the*
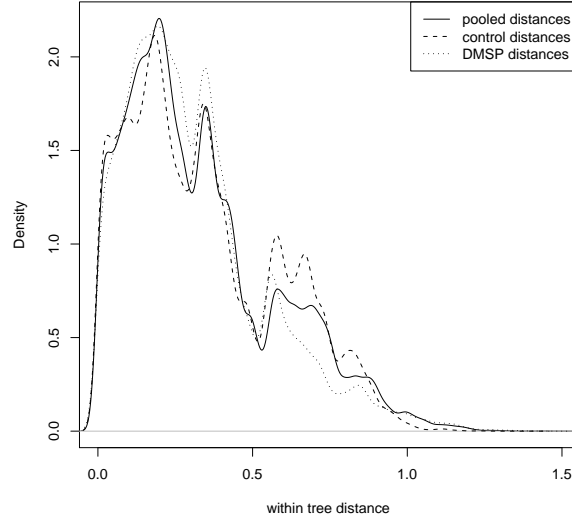
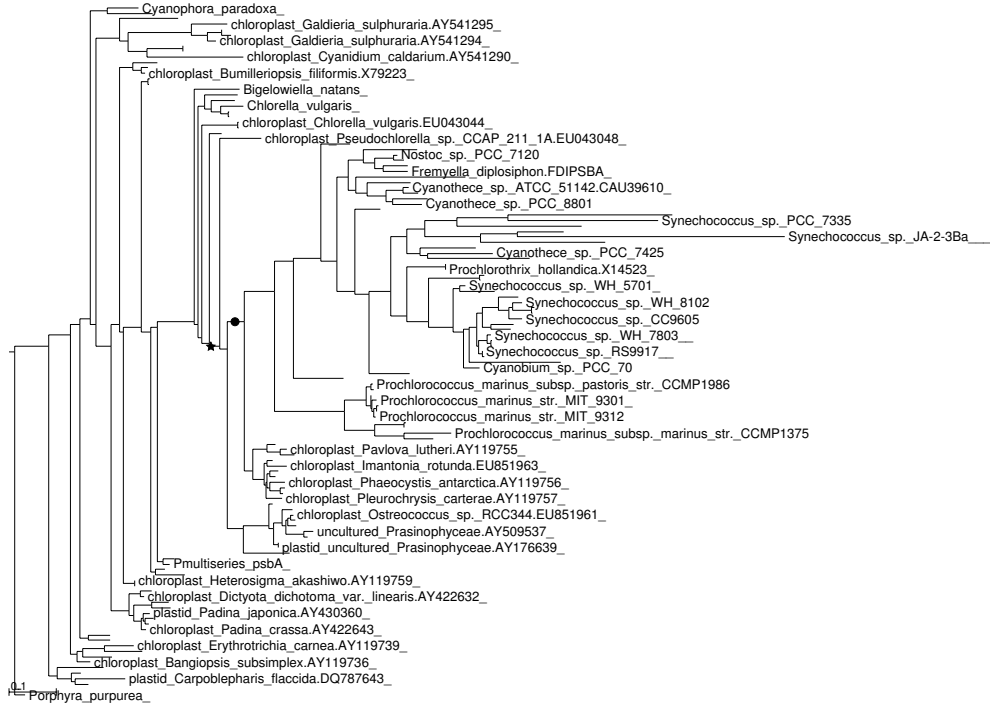FIGURE 3. Density plot of the distribution of pairwise distances for the sample data sets.

FIGURE 4. Dendrogram with barycenters marked. Circle is the control sample, and star is the sample treated with DMSP.
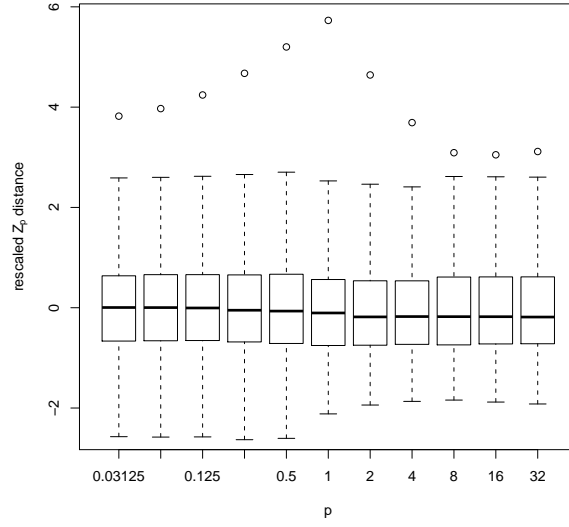
FIGURE 5. Plot showing sample (point) and randomized ranges (box-and-whisker). Outliers eliminated for clarity. For each p, the distribution was rescaled by subtracting the mean and dividing by the standard deviation.
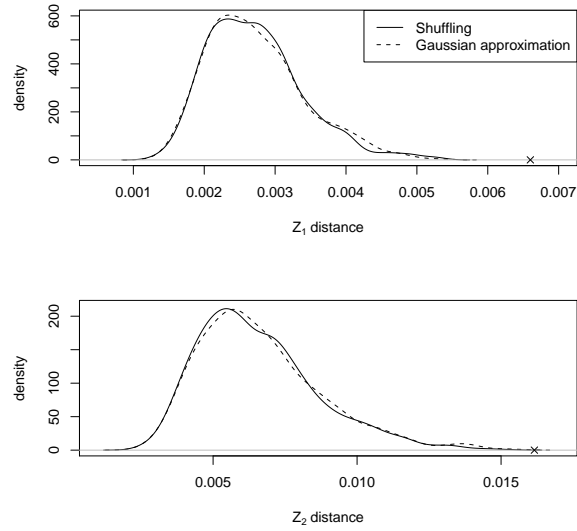

FIGURE 6. Comparison of the distribution of Z_1 and Z_2 distances obtained by shuffling, Gaussian approximation, and the observed value (marked with x) in the example case.
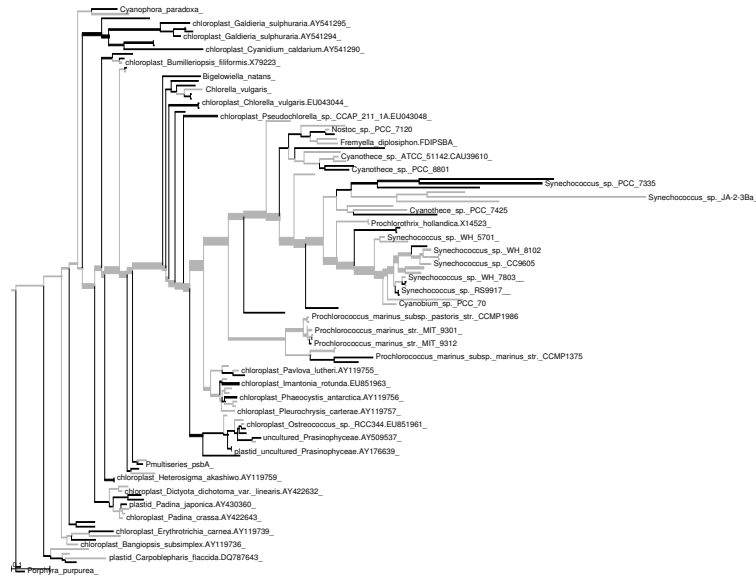
FIGURE 7. A tree displaying the optimal movement of mass for the KR metric. When moving from the first probability distribution to the second, edges marked in gray have mass moving towards the root, while those marked in black have mass moving towards the leaves. Thickness shows the quantity of mass moving through that edge.

United States of America 103(3), 732.

Billera, L., S. Holmes, and K. Vogtmann (2001). Geometry of the space of phylogenetic trees. *Adv. Appl. Math. 27*(4), 733–767.

Bogachev, V. I. (1998). *Gaussian measures*, Volume 62 of *Mathematical Surveys and Monographs*. Providence, RI: American Mathematical Society.

Bridson, M. R. and A. Haefliger (1999). *Metric spaces of non-positive curvature*, Volume 319 of *Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]*. Berlin: Springer-Verlag.

Burago, D., Y. Burago, and S. Ivanov (2001). *A course in metric geometry*, Volume 33 of *Graduate Studies in Mathematics*. Providence, RI: American Mathematical Society.

Chen, K., D. Durand, and M. Farach-Colton (2000). NOTUNG: a program for dating gene duplications and optimizing gene family trees. *J. Comp. Biol. 7*(3-4), 429–447.

Eddy, S. (1998). Profile hidden Markov models. *Bioinformatics 14*(9), 755–763.

Felsenstein, J. (2004). *Inferring Phylogenies*. Sunderland, MA: Sinauer Press.

Gideon, R. A. and J. Gurland (1976). Series expansions for quadratic forms in normal variables. *J. Amer. Statist. Assoc. 71*(353), 227–232.

Gill, S., M. Pop, R. DeBoy, P. Eckburg, P. Turnbaugh, B. Samuel, J. Gordon, D. Relman, C. Fraser-Liggett, and K. Nelson (2006). Metagenomic analysis of the human distal gut microbiome. *Science 312*(5778), 1355–1359.

Gurland, J. (1955). Distribution of definite and of indefinite quadratic forms. *Ann. Math. Statist. 26*, 122–127.

Hamady, M., C. Lozupone, and R. Knight (2009). Fast UniFrac: facilitating high-throughput phylogenetic analyses of microbial communities including analysis of pyrosequencing and PhyloChip data. *ISME-J 4*(1), 17–27.

Hwang, C.-R. (1980). Gaussian measure of large balls in a Hilbert space. *Proc. Amer. Math. Soc. 78*(1), 107–110.

Jain, N. C. and M. B. Marcus (1978). Continuity of sub-Gaussian processes. In *Probability on Banach spaces*, Volume 4 of *Adv. Probab. Related Topics*, pp. 81–196. New York: Dekker.

Kotz, S., N. L. Johnson, and D. W. Boyd (1967). Series representations of distributions of quadratic forms in normal variables. I. Central case. *Ann. Math. Statist. 38*, 823–837.

Lozupone, C., M. Hamady, B. Cantarel, P. Coutinho, B. Henrissat, J. Gordon, and R. Knight (2008). The convergence of carbohydrate active gene repertoires in human gut microbes. *PNAS 105*(39), 15076.

Lozupone, C., M. Hamady, S. Kelley, and R. Knight (2007). Quantitative and qualitative {beta} diversity measures lead to different insights into factors that structure microbial communities. *Appl. Environ. Microbiol. 73*(5), 1576.

Lozupone, C. and R. Knight (2005). UniFrac: a new phylogenetic method for comparing microbial communities. *Appl. environ. microbio. 71*(12), 8228.

Matsen, F. A., R. B. Kodner, and E. V. Armbrust (2010). pplacer: linear time maximum-likelihood and bayesian phylogenetic placement of sequences onto a fixed reference tree. *Submitted to BMC Bioinfo.*.

Monier, A., J. Claverie, and H. Ogata (2008). Taxonomic distribution of large DNA viruses in the sea. *Genome Biol. 9*(7), R106.

Morgan, J., A. Darling, and J. Eisen (2010). Metagenomic sequencing of an in vitro-simulated microbial community. *PLoS ONE 5*(4), e10209. doi:10.1371/journal.pone.0010209.

Pachares, J. (1955). Note on the distribution of a definite quadratic form. *Ann. Math. Statist. 26*, 128–131.

Purdom, E. (2008). Analyzing data with graphs: Metagenomic data and the phylogenetic tree.

Rachev, S. T. (1991). *Probability metrics and the stability of stochastic models.* Wiley Series in Probability and Mathematical Statistics: Applied Probability and Statistics. Chichester: John Wiley & Sons Ltd.

Rachev, S. T. and L. Rüschendorf (1998). *Mass transportation problems. Vol. I.* Probability and its Applications (New York). New York: Springer-Verlag. Theory.

Robbins, H. and E. J. G. Pitman (1949). Application of the method of mixtures to quadratic forms in normal variates. *Ann. Math. Statistics 20*, 552–560.

Ruben, H. (1962). Probability content of regions under spherical normal distributions. IV. The distribution of homogeneous and non-homogeneous quadratic functions of normal variables. *Ann. Math. Statist. 33*, 542–570.

Schloss, P. and J. Handelsman (2006). Introducing TreeClimber, a test to compare microbial community structures. *Applied and environmental microbiology 72*(4), 2379.

Slatkin, M. and W. P. Maddison (1989). A cladistic measure of gene flow inferred from the phylogenies of alleles. *Genetics 123*(3), 603–613.

Stamatakis, A. (2006). RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics 22*(21), 2688.

Vila-Costa, M., J. M. Rinta-Kanto, S. Sun, S. Sharma, R. Porestky, and M. A. Moran (2010). Transcriptomic analysis of a marine bacterial community enriched with dimethylsulfoniopropionate (dmsp). *ISME-J. In press.*.

Villani, C. (2003). *Topics in optimal transportation*, Volume 58 of *Graduate Studies in Mathematics*. Providence, RI: American Mathematical Society.

Villani, C. (2009). *Optimal transport*, Volume 338 of *Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]*. Berlin: Springer-Verlag. Old and new.

Von Mering, C., P. Hugenholtz, J. Raes, S. Tringe, T. Doerks, L. Jensen, N. Ward, and P. Bork (2007). Quantitative phylogenetic assessment of microbial communities in diverse environments. *Science 315*(5815), 1126.

E-mail address: evans@stat.berkeley.edu

DEPARTMENT OF STATISTICS UNIVERSITY OF CALIFORNIA 367 EVANS HALL #3860 BERKELEY, CA 94720-3860

E-mail address: matsen@fhcrc.org

PROGRAM IN COMPUTATIONAL BIOLOGY, FRED HUTCHINSON CANCER RESEARCH CENTER, 1100 FAIRVIEW AVE. N. M1-B514, P.O. BOX 19024, SEATTLE, WA 98109-1024