Supplemental Methods.

Here we prove that the unweighted and weighted Unifrac metrics are in general metrics. The proofs are given in a general context that includes both of these metrics as special cases.

Case 1: Unweighted case.

Consider a fixed and given set of indices $\mathcal{L}$ and a finite rooted tree, where leaves are labeled using subsets of $\mathcal{L}$ (Fig. S1). Furthermore, assume there is a certain weight function W that associates to each edge e a finite quantity $W(e) \geq 0$. In the context of the unweighted Unifrac metric, $\mathcal{L}$ is the set of environmental samples that each leaf (i.e. sequence) is found in, and $W(e)$ is just the branch length of e.

In what follows, we use the letters a, b, c and l to denote generic indices in $\mathcal{L}$. Furthermore, we use e to denote a generic edge in the tree. To each e, let $\mathcal{L}(e)$ be the union of the labels of all the leaves that descend from e (Fig. S1). We write $l \in \mathcal{L}(e)$ to denote that l is in $\mathcal{L}(e)$. Otherwise, we write $l \notin \mathcal{L}(e)$.

We show that the tree and weight function always induce a pseudo-metric over $\mathcal{L}$ and give a necessary and sufficient condition for it to be a metric. In particular, the unweighted UniFrac metric defines a pseudo-metric on the space of environments, but a metric on the "location" of those environments relative to a tree (Fig. S1).

Before stating the main definition and result, we introduce a notation clarified via examples. Define S_{ab} as the sum of the weights of all the e's such that $a \in \mathcal{L}(e)$ and $b \in \mathcal{L}(e)$. On the other hand, using the symbol $\neg$ to denote negation, define $S_{a(\neg b)(\neg c)}$ as the sum of the weights of all the edges e such that $a \in \mathcal{L}(e)$, $b \notin \mathcal{L}(e)$ and $c \notin \mathcal{L}(e)$.

Define the d_1-distance between a and b as the quantity:

$$d_1(a,b) = \begin{cases} 0 & \text{, when } S_{a(\neg b)} + S_{(\neg a)b} = 0; \\ \frac{S_{a(\neg b)} + S_{(\neg a)b}}{S_{ab} + S_{a(\neg b)} + S_{(\neg a)b}} & \text{, when } S_{a(\neg b)} + S_{(\neg a)b} > 0. \end{cases}$$

Theorem 1. *For all a, b and c the following properties apply:*

(i) $d_1(a,b) = d_1(b,a) \geq 0$;

(ii) *If $a = b$ then* $d_1(a,b) = 0$;

(iii) $d_1(a,b) \leq d_1(a,c) + d_1(c,b)$.

In particular, d_1 defines a pseudo-metric over $\mathcal{L}$. Furthermore, d_1 is a metric if and only if $S_{a(\neg b)} + S_{(\neg a)b} > 0$, for all $a \neq b$.

Proof. The first and second properties are direct from the definition. To show that d_1 is a pseudo-metric, it only remains to show property (iii) i.e. that d_1 satisfies the triangular inequality. For this, it suffices to consider the following three cases:

(1) $d_1(a,b) = 0$,

(2) $d_1(a,c) = 0$ or $d_1(c,b) = 0$, and

(3) $d_1(a,b) > 0$, $d_1(a,c) > 0$ and $d_1(c,b) > 0$.

The first case is obvious because d_1 is nonnegative.

For the second case, assume without loss of generality that $d_1(a,c) = 0$ i.e. that $S_{a(\neg c)} + S_{(\neg a)c} = 0$. This means that $a \in \mathcal{L}(e)$ if and only if $c \in \mathcal{L}(e)$, for each edge e such that $W(e) > 0$. Therefore:

$$\begin{aligned}
S_{a(\neg b)} &= \sum_{e:\, a\in\mathcal{L}(e),\, b\notin\mathcal{L}(e)} W(e) = \sum_{e:\, c\in\mathcal{L}(e),\, b\notin\mathcal{L}(e)} W(e) = S_{c(\neg b)};\\
S_{(\neg a)b} &= \sum_{e:\, a\notin\mathcal{L}(e),\, b\in\mathcal{L}(e)} W(e) = \sum_{e:\, c\notin\mathcal{L}(e),\, b\in\mathcal{L}(e)} W(e) = S_{(\neg c)b};\\
S_{ab} &= \sum_{e:\, a\in\mathcal{L}(e),\, b\in\mathcal{L}(e)} W(e) = \sum_{e:\, c\in\mathcal{L}(e),\, b\in\mathcal{L}(e)} W(e) = S_{cb};
\end{aligned}$$

from which we deduce that $d_1(a,b) = d_1(c,b)$; in particular, since $d_1(a,c) = 0$, $d_1(a,b) \le d_1(c,b) + d_1(a,c)$. This shows the triangular inequality for the second case.

For the third and last case, consider the partition of the edges in the tree according to the Venn diagram in Figure S2. Following the diagram, define $S_1 = S_{a(\neg b)(\neg c)}$, $S_4 = S_{ab(\neg c)}$, $S_7 = S_{abc}$, etc and notice that:

$$\begin{array}{lllll}
S_{a(\neg b)} = (S_1+S_6) & ; & S_{(\neg a)b} = (S_2+S_5) & ; & S_{ab} = (S_4+S_7);\\
S_{a(\neg c)} = (S_1+S_4) & ; & S_{(\neg a)c} = (S_3+S_5) & ; & S_{ac} = (S_6+S_7);\\
S_{c(\neg b)} = (S_3+S_6) & ; & S_{(\neg c)b} = (S_2+S_4) & ; & S_{cb} = (S_5+S_7).
\end{array}$$

Therefore:

$$\begin{aligned}
d_1(a,b) &= \frac{S_1+S_2+S_5+S_6}{S_1+S_2+S_4+S_5+S_6+S_7};\\
d_1(a,c) &= \frac{S_1+S_3+S_4+S_5}{S_1+S_3+S_4+S_5+S_6+S_7};\\
d_1(c,b) &= \frac{S_2+S_3+S_4+S_6}{S_2+S_3+S_4+S_5+S_6+S_7}.
\end{aligned}$$

Based on the previous identities, if one defines

$$\begin{aligned}
p &= (S_1+S_2+S_4+S_5+S_6+S_7)\cdot(S_1+S_3+S_4+S_5+S_6+S_7)\cdot(S_2+S_3+S_4+S_5+S_6+S_7);\\
q &= S_1^2S_2 + S_1^2S_3 + S_1^2S_4 + S_1^2S_6 + S_1S_2^2 + S_1S_3^2 + 3S_1S_4^2 + S_1S_6^2\\
&\quad + S_2^2S_3 + S_2^2S_4 + S_2^2S_5 + S_2S_3^2 + 3S_2S_4^2 + S_2S_5^2 + 2S_3^2S_4 + S_3^2S_5\\
&\quad + S_3^2S_6 + 2S_3^2S_7 + 4S_3S_4^2 + S_3S_5^2 + S_3S_6^2 + 2S_3S_7^2 + 2S_4^2S_4 + 4S_4^2S_5\\
&\quad + 4S_4^2S_6 + 4S_4^2S_7 + 2S_4S_5^2 + 2S_4S_6^2 + 2S_4S_7^2 + 2S_1S_2S_3 + 4S_1S_2S_4\\
&\quad + 2S_1S_2S_5 + 2S_1S_2S_6 + 2S_1S_2S_7 + 4S_1S_3S_4 + 2S_1S_3S_5 + 2S_1S_3S_6\\
&\quad + 2S_1S_3S_7 + 3S_1S_4S_5 + 4S_1S_4S_6 + 3S_1S_4S_7 + S_1S_5S_6 + S_1S_6S_7\\
&\quad + 4S_2S_3S_4 + 2S_2S_3S_5 + 2S_2S_3S_6 + 2S_2S_3S_7 + 4S_2S_4S_5 + 3S_2S_4S_6\\
&\quad + 3S_2S_4S_7 + S_2S_5S_6 + S_2S_5S_7 + 5S_3S_4S_5 + 5S_3S_4S_6 + 6S_3S_4S_7\\
&\quad + 2S_3S_5S_6 + 3S_3S_5S_7 + 3S_3S_6S_7 + 4S_4S_5S_6 + 4S_4S_5S_7 + 4S_4S_6S_7;
\end{aligned}$$

then $p \cdot \{d_1(a,c) + d_1(c,b) - d_1(a,b)\} = q$. Since $p > 0$ and $q \ge 0$, d_1 satisfies the triangular inequality, proving that d_1 is a pseudo-metric.

Due to properties (I)-(III), d_1 is a metric only when $d_1(a,b) > 0$, for all $a \ne b$. From the given definition, it is clear that this last property applies only when $S_{a(\neg b)} + S_{(\neg a)b} > 0$, for all $a \ne b$, which completes the proof of the theorem. □

Case 2: Weighted case.

Consider now a finite rooted tree where leaves are labeled using multi-subsets of $\mathcal{L}$ (Fig. S3). As before, assume there is a certain weight function W that associates to each edge e a finite quantity $W(e) \ge 0$. In addition, for each $l \in \mathcal{L}$, assume there is a weight function W_l that associates to each e a nonnegative and finite quantity $W_l(e)$. For the weighted Unifrac metric, $W(e)$ is the branch length of e, and $W_l(e)$ is the fraction of times that l occurs as an element in the multi-sets associated with the leaves that descend from e, relative to the number of times that l occurs in all of the multi-sets (Fig. S3).

The d_2-distance between a and b is now defined as the quantity:

$$d_2(a,b) = \sum_e W(e) \cdot |W_a(e) - W_b(e)|.$$

THEOREM 2. *d_2 defines a pseudo-metric over $\mathcal{L}$. Furthermore, d_2 is a metric if and only if, for all $a \neq b$, there exists e such that $W(e) \cdot |W_a(e) - W_b(e)| > 0$.*

PROOF. It is immediate from the definition that $d_2(a,b) \geq 0$ and if $a = b$ then $d_2(a,b) = 0$. On the other hand, using that $|x+y| \leq |x| + |y|$, for any pair of real numbers x and y, we obtain that

$$\begin{aligned} d_2(a,b) &= \sum_e W(e) \cdot |W_a(e) - W_b(e)|; \\ &= \sum_e W(e) \cdot \big|\big(W_a(e) - W_c(e)\big) + \big(W_c(e) - W_b(e)\big)\big|; \\ &\leq \sum_e W(e) \cdot |W_a(e) - W_c(e)| + \sum_e W(e) \cdot |W_c(e) - W_b(e)|; \\ &= d_2(a,c) + d_2(c,b). \end{aligned}$$

This shows that d_2 is a pseudo-metric. In particular, d_2 is a metric only when $d_2(a,b) > 0$, for all $a \neq b$. Since this last property holds only when, for all $a \neq b$, there is e such that $W(e) \cdot |W_a(e) - W_b(e)| > 0$, the theorem follows. □

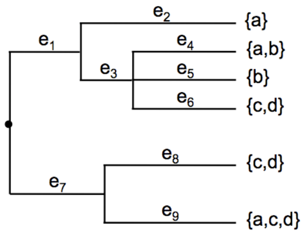

Figure S1. Rooted tree in which leaves are labeled using subsets of $\mathcal{L} = \{a,b,c,d\}$. Notice that $\mathcal{L}(e_1) = \{a,b,c,d\}$ and $\mathcal{L}(e_2) = \{a\}$. On the other hand, $S_{(\neg a)b} = W(e_5)$, $S_{a(\neg b)} = W(e_2) + W(e_7) + W(e_9)$, and $S_{ab} = W(e_1) + W(e_3) + W(e_4)$. Relative to this tree and weight function W, if $S_{(\neg a)b} + S_{a(\neg b)} > 0$ then:

$$d_1(a,b) = \frac{W(e_2) + W(e_5) + W(e_7) + W(e_9)}{W(e_1) + W(e_2) + W(e_3) + W(e_4) + W(e_5) + W(e_7) + W(e_9)}.$$

Due to Theorem 1, d_1 is a pseudo-metric. In this particular case, however, d_1 is not a metric because, for instance, $c \neq d$ and yet $d_1(c,d) = 0$. This is consistent with the fact that c and d cannot be told apart in any leave. Hence, in a sense, c and d are located at the same "place" on the tree.

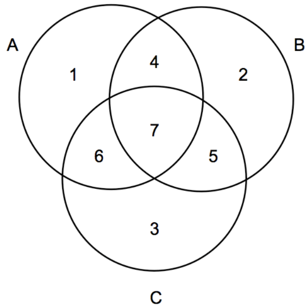

Figure S2. Venn diagram used to prove the triangular inequality in THEOREM 1. A is the set of all those edges e such that $a \in \mathcal{L}(e)$. Similarly, $B = \{e : b \in \mathcal{L}(e)\}$ and $C = \{e : c \in \mathcal{L}(e)\}$. The various possible intersections between A, B and C or their complements are denoted numerically.

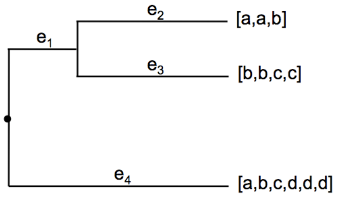

Figure S3. Tree in which leaves are labeled using multi-subsets of $\mathcal{L} = \{a, b, c, d\}$. (A multi-set is an unordered list of possibly repeated objects. We use square-brackets to denote multi-sets. For instance, $[a, a, b]$ is a multi-set that contains twice a and once b. Since multi-sets are unordered, $[a, a, b] = [b, a, a] = [a, b, a]$.) Relative to the above tree and weight functions W, W_a, W_b, W_c and W_d, the d_2-distance between a and b is:

$$d_2(a, b) = \sum_{i=1}^{4} W(e_i) \cdot |W_a(e_i) - W_b(e_i)|.$$

For the weighted Unifrac metric, we have that $W_a(e_1) = 2/3$, $W_b(e_1) = 3/4$, $W_a(e_2) = 2/3$, $W_b(e_2) = 1/4$, $W_a(e_3) = 0$, $W_b(e_3) = 1/2$, $W_a(e_4) = 1/3$, and $W_b(e_4) = 1/4$. As a result, if $W(e)$ denotes the branch-length of e then the weighted Unifrac distance between a and b is:

$$d_2(a, b) = \frac{W(e_1)}{12} + \frac{5 \cdot W(e_2)}{12} + \frac{W(e_3)}{2} + \frac{W(e_4)}{12}.$$