Proof that the unweighted UniFrac measurement is conformant to the triangle inequality

Ruth Grace Wong[1], Gregory B, Gloor[1]
1 Department of Biochemistry, Univesrity of Western Ontario, London, Ontario, Canada
¤a Western University, 1151 Richmond Street, London, Ontario, Canada, N6A 3K7
*** E-mail: Corresponding ggloor@uwo.ca**

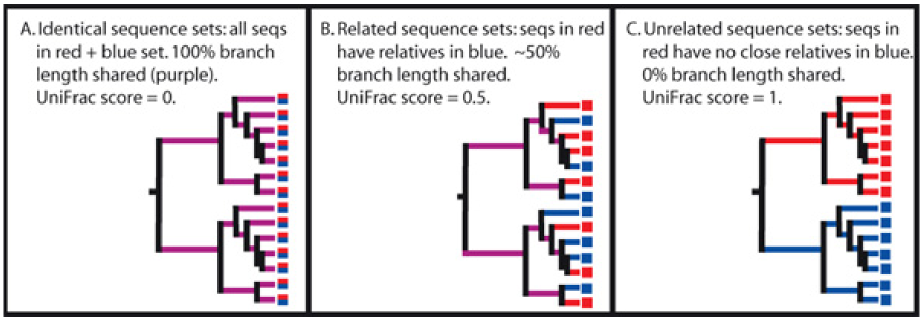
Abstract

Here we provide a written proof that the unweighted UniFrac method is conformant with the triangle inequality: one of the requirements for a proper distance metric.

Introduction

The UniFrac measurement is widely used in microbiome research to measure the difference between two microbiome samples. Calculating the measurement requires a table with counts of the number of reads for each Operational Taxonomic Unit (OTU) found in each sample, along with a phylogenetic tree with all the OTUs at the tree tips. The measurement itself (performed on 2 samples at a time) is the branch lengths in the phylogenetic tree that are in one sample but not the other, divided by the total branch lengths present in both samples.

Figure 1. Examples of the UniFrac calculation


A. Identical sequence sets: all seqs in red + blue set. 100% branch length shared (purple). UniFrac score = 0.
B. Related sequence sets: seqs in red have relatives in blue. ~50% branch length shared. UniFrac score = 0.5.
C. Unrelated sequence sets: seqs in red have no close relatives in blue. 0% branch length shared. UniFrac score = 1.

For example, in the depiction, red branch lengths represent parts of the tree that extend from OTUs that are only present in one sample, blue branch lengths represent parts of the tree that extend from OTUs that are only present in the other sample, and purple branch lengths represent parts of the tree that extend from OTUs present in both samples. In the first image,

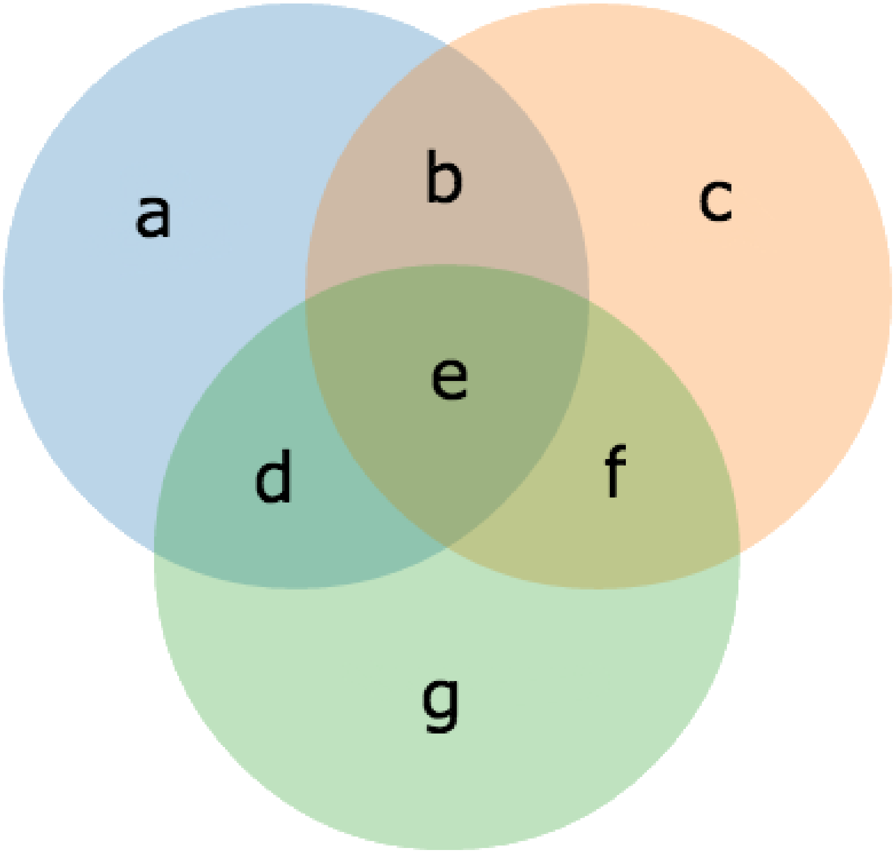
Here we present a proof that the UniFrac measurement is conformant with the triangle inequality, showing that it is a proper distance metric.

Proof

We will examine perform the proof using three samples, 1, 2, and 3, denoting the distances between them as d_12, d_23, and d_13. The triangle inequality is satisfied if we can show that $d_12 + d_23 \geq d_13$.

We can visualize the UniFrac method by putting all the branch lengths of the phylogenetic tree into a Venn Diagram.

Figure 2. Venn diagram of which samples the branch lengths in the phylogenetic tree belong to. The blue circle represents sample 1, the orange circle represents sample 2, and the green circle represents sample 3.



Each circle of the Venn Diagram represents one sample. The region denoted by a represents all the branch lengths of the tree that stem from OTUs present in sample 1 but not any of the other samples. The region denoted by c represents all the branch lengths of the tree that stem from OTUs present in sample 2 but not any of the other samples. The region denoted by g represents all the branch lengths of the tree that stem from OTUs present in sample 3 but not any of the other samples. The regions b and e represent the branch lengths shared by samples 1 and 2, the regions e and f represent the branch lengths shared by samples 2 and 3, the regions d and e represent the branch lengths shared by samples 1 and 3, and the region e represents the branch lengths shared by all three samples.

The UniFrac distances between the samples are as follows:

$$d_{12} = \frac{sample\_1\_and\_2\_unshared\_branchlengths}{sample\_1\_and\_2\_total\_branch\_lengths}$$

$$d_{23} = \frac{sample\_2\_and\_3\_unshared\_branchlengths}{sample\_2\_and\_3\_total\_branch\_lengths}$$

$$d_{13} = \frac{sample\_1\_and\_3\_unshared\_branchlengths}{sample\_1\_and\_3\_total\_branch\_lengths}$$

In terms of the regions of the Venn Diagram, this can be rewritten as:

$$d_{12} = \frac{a+c+d+f}{a+b+c+d+e+f)}$$

$$d_{23} = \frac{b+c+d+g}{b+c+d+e+f+g}$$

$$d_{13} = \frac{a+b+f+g}{a+b+d+e+f+g}$$

The Triangle Inequality can be formulated as follows: $d_{12} + d_{23} \geq d_{13}$

Substituting in the regions of the Venn Diagram:

$$\frac{a+c+d+f}{a+b+c+d+e+f)} + \frac{b+c+d+g}{b+c+d+e+f+g} \geq \frac{a+b+f+g}{a+b+d+e+f+g}$$

$$\frac{a+c+d+f}{a+b+c+d+e+f)} + \frac{b+c+d+g}{b+c+d+e+f+g} - \frac{a+b+f+g}{a+b+d+e+f+g} \geq 0$$

Multiplying each term by $(a+b+c+d+e+f)(b+c+d+e+f+g)(a+b+d+e+f+g)$ to cancel out the denominators: $(a^2b+a^2c+a^2d+a^2e+a^2f+a^2g+ab^2+2abc+3abd+2abe+3abf+2abg+ac^2+3acd+2ace+3acf+2acg+2ad^2+3ade+4adf+3adg+ae^2+3aef+2aeg+2af^2+3afg+ag^2+b^2c+b^2d+b^2f+bc^2+3bcd+2bce+3bcf+2bcg+2bd^2+2bde+4bdf+2bdg+2bef+2bf^2+2bfg+c^2d+c^2e+c^2f+c^2g+2cd^2+3cde+4cdf+3cdg+ce^2+3cef+2ceg+2cf^2+3cfg+cg^2+d^3+2d^2e+3d^2f+2d^2g+de^2+4def+2deg+3df^2+4dfg+dg^2+e^2f+2ef^2+2efg+f^3+2f^2g+fg^2)+$

$(a^2b+a^2c+a^2d+a^2g+2ab^2+3abc+4abd+2abe+2abf+3abg+ac^2+3acd+2ace+2acf+2acg+2ad^2+2ade+2adf+3adg+2aeg+2afg+ag^2+b^3+2b^2c+3b^2d+2b^2e+2b^2f+2b^2g+bc^2+4bcd+3bce+3bcf+3bcg+3bd^2+4bde+4bdf+4bdg+be^2+2bef+3beg+bf^2+3bfg+bg^2+c^2d+c^2e+c^2f+c^2g+2cd^2+3cde+3cdf+3cdg+ce^2+2cef+2ceg+cf^2+2cfg+cg^2+d^3+2d^2e+2d^2f+2d^2g+de^2+2def+3deg+df^2+3dfg+dg^2+e^2g+2efg+eg^2+f^2g+fg^2)-$

$(a^2b - a^2c - a^2d - a^2e - a^2f - a^2g - 2ab^2 - 3abc - 3abd - 3abe - 4abf - 3abg - ac^2 - 2acd - 2ace - 3acf - 2acg - ad^2 - 2ade - 3adf - 2adg - ae^2 - 3aef - 2aeg - 2af^2 - 3afg - ag^2 - b^3 - 2b^2c - 2b^2d - 2b^2e - 3b^2f - 2b^2g - bc^2 - 2bcd - 2bce - 4bcf - 3bcg - bd^2 - 2bde - 4bdf - 3bdg - be^2 - 4bef - 3beg - 3bf^2 - 4bfg - bg^2 - c^2f - c^2g - 2cdf - 2cdg - 2cef - 2ceg - 2cf^2 - 3cfg - cg^2 - d^2f - d^2g - 2def - 2deg - 2df^2 - 3dfg - dg^2 - e^2f - e^2g - 2ef^2 - 3efg - eg^2 - f^3 - 2f^2g - fg^2) \geq 0$

This simplifies to: $a^2b + a^2c + a^2d + a^2g + ab^2 + 2abc + 4abd + abe + abf + 2abg + ac^2 + 4acd + 2ace + 2acf + 2acg + 3ad^2 + 3ade + 3adf + 4adg + 2aeg + 2afg + ag^2 + +b^2c + 2b^2d + bc^2 + 5bcd + 3bce + 2bcf + 2bcg + 4bd^2 + 4bde + 4bdf + 3bdg + bfg + 2c^2d + 2c^2e + c^2f + c^2g + 4cd^2 + 6cde + 5cdf + 4cdg + 2ce^2 + 3cef + 2ceg + cf^2 + 2cfg + cg^2 + 2d^3 + 4d^2e + 4d^2f + 3d^2g + 2de^2 + 4def + 3deg + 2df^2 + 4dfg + dg^2 + +efg + 2f^2g \geq 0$

All the terms are positive, and all the regions contain positive branch lengths, therefore the inequality is true.

Acknowledgments

Thanks to Mingchun Liu for helping me with the proof.