UniFrac: a New Phylogenetic Method for Comparing Microbial Communities
Outline

- Background
- What is UniFrac?
- Materials and Methods
- Results
- Discussion
- Questions
The vast majority of microbes cannot be cultured with current methods

- Only half (26) out of the 52 major bacterial lineages have cultured representatives
- Most of those culturable strains are only distantly related to the dominant phylotypes

Authors’ Claim: "[O]ur sole source of information about the biology of much of the diversity of life is the environmental distribution of sequences."
Current Methods - Statistical

- Sørenson and Jaccard indices of group overlap
- LibShuff and \( \int \)-LibShuff
- Hierarchical clustering and ordination based on the distribution of sequences
Limitations

- Don't account for different degrees of similarity between sequences
  - If the 16S is within 95-98% identity, the sequences are treated equally, even if looking at the entire sequence shows a 3-40% sequence divergence.
- Results in a substantial loss of information
Current Methods - Phylogenetic

- $P$ test
- $F_{ST}$ test
Current Methods - Phylogenetic

- Limitations
  - Only have been applied to detect significant differences
  - Cannot compare multiple samples simultaneously
  - Do not account for branch lengths
What is UniFrac?

- “Unique fraction metric”
- A new, lineage-based phylogenetic distance method
  - Measures the phylogenetic distance between sets of taxa in a phylogenetic tree as the fraction of the branch length of the tree that leads to descendants from either one environment or the other, but not both
What is UniFrac?

- A true distance measurement
  - Can compare multiple samples simultaneously
  - Can be used with standard multivariate statistics (e.g., UPGMA or PCA)
- Web service @
  [http://bmf2.colorado.edu/unifrac/index.psp](http://bmf2.colorado.edu/unifrac/index.psp)
- Downloadable software module for Python built on an old version of PyCogent
Metric Characteristics

- Captures the amount of evolution unique to each state
- Reflects changes in one environment that would be harmful in the other(s)
- It XORs the branches in the set of environments and compares the remaining branch lengths
Basic Idea

\[ D = 1 \]

\[ D \approx 0.5 \]

Image from http://bmf2.colorado.edu/unifrac/help.psp
Basic Concept

- Similar environments can be translated into each other with few changes in species
  - E.g., seawater from two bays in California, as we talked about last week in the PHACCS paper
- Divergent environments should each have a few species that can't survive (or can't survive well) in the other environment(s)
  - E.g., water from hot springs versus water from the Arctic
How does culturing affect similarities between samples?

- Do cultured samples really reflect the environments they come from?
- Authors’ Test: Compare various cultured samples versus the environments they came from to see how similar they are.
How cosmopolitan are bacterial lineages?

- Most lineages seem to be pretty worldwide
- Is the same true for extremeophiles such as Arctic/Antarctic bacteria?
  - Aside: we saw an example of this in the Salinibacter ruber paper last week--two very similar extremeophiles were found in two widely dispersed environments: one in California, one in Spain
  - Authors’ Test: Tested whether psychrophilic (cold-loving) bacteria could cross the equatorial sea
Questions To Be Answered

- Are marine ice, sediment, and seawater three distinct, homogeneous habitats?
  - Generally treated separately in the literature--should they be?
  - Authors’ Test: Compared marine ice, sediment, and seawater to see how the various samples clustered together.
Uses of UniFrac

- Tell whether two communities differ significantly through Monte Carlo simulations
  - Authors did this by keeping the tree intact and randomly assigning environmental labels
- Produce a distance matrix to describe the phylogenetic differences between microbial communities
Materials and Methods

- Samples examined
  - Started with 23 small-subunit rRNA sequence libraries from 12 different marine studies
  - After data cleaning, removed 3 of those samples due to lack of sequence information
  - Included cultured and uncultured samples from seawater, sea ice, and marine sediment
# Sample Descriptions

<table>
<thead>
<tr>
<th>Sample</th>
<th>Reference</th>
<th>No. of sequences</th>
<th>Water column depth (m)</th>
<th>Sediment depth (cm)</th>
<th>Latitude, longitude</th>
<th>Temp (°C)</th>
</tr>
</thead>
<tbody>
<tr>
<td>SRUI</td>
<td>38</td>
<td>79</td>
<td>155</td>
<td>0-1.1</td>
<td>76°58’N, 15°34’E</td>
<td>2.6</td>
</tr>
<tr>
<td>STU2</td>
<td>25</td>
<td>33</td>
<td>6,400</td>
<td></td>
<td>40°06’N, 144°11’E</td>
<td></td>
</tr>
<tr>
<td>SNU3</td>
<td>4</td>
<td>36</td>
<td>709-940</td>
<td>1-2</td>
<td>66°S, 143°E</td>
<td></td>
</tr>
<tr>
<td>SNU4</td>
<td>4</td>
<td>31</td>
<td>709-940</td>
<td></td>
<td>66°S, 143°E</td>
<td></td>
</tr>
<tr>
<td>SNU5</td>
<td>5</td>
<td>101</td>
<td>761</td>
<td>0-0.4</td>
<td>66°32’S 143°38’E</td>
<td></td>
</tr>
<tr>
<td>SNU6</td>
<td>5</td>
<td>146</td>
<td>761</td>
<td>1.5-2.5</td>
<td>66°32’S 143°38’E</td>
<td></td>
</tr>
<tr>
<td>SNU7</td>
<td>5</td>
<td>231</td>
<td>761</td>
<td>20-21</td>
<td>66°32’S 143°38’E</td>
<td></td>
</tr>
<tr>
<td>WRU8</td>
<td>2</td>
<td>87</td>
<td>55, 131</td>
<td></td>
<td>72-88°N, 51-350°E</td>
<td></td>
</tr>
<tr>
<td>WTC9</td>
<td>8</td>
<td>36</td>
<td>1</td>
<td></td>
<td>54°09’N, 7°52’E</td>
<td></td>
</tr>
<tr>
<td>WTU10</td>
<td>22</td>
<td>75</td>
<td>1-2</td>
<td></td>
<td>40°N, 73°E</td>
<td>23.8-29.2</td>
</tr>
<tr>
<td>WTC11</td>
<td>22</td>
<td>21</td>
<td>1-2</td>
<td></td>
<td>40°N, 72°E</td>
<td>23.8-29.2</td>
</tr>
<tr>
<td>WTU12</td>
<td>1</td>
<td>544</td>
<td></td>
<td></td>
<td>42°N, 71°E</td>
<td>16</td>
</tr>
<tr>
<td>WPU13</td>
<td>11</td>
<td>17</td>
<td>10</td>
<td></td>
<td>32°37’N 64°57’W</td>
<td></td>
</tr>
<tr>
<td>WPU14</td>
<td>11</td>
<td>40</td>
<td>100, 500</td>
<td></td>
<td>31°34’N 64°57’W</td>
<td></td>
</tr>
<tr>
<td>INC15</td>
<td>3</td>
<td>58</td>
<td></td>
<td></td>
<td>68°S, 78°E</td>
<td></td>
</tr>
<tr>
<td>IRU16</td>
<td>6</td>
<td>62</td>
<td></td>
<td></td>
<td>80°N, 0°E</td>
<td></td>
</tr>
<tr>
<td>IRC17</td>
<td>6</td>
<td>109</td>
<td></td>
<td></td>
<td>80°N, 0°E</td>
<td></td>
</tr>
<tr>
<td>INU18</td>
<td>6</td>
<td>20</td>
<td></td>
<td></td>
<td>70°S, 15°E</td>
<td></td>
</tr>
<tr>
<td>INC19</td>
<td>6</td>
<td>87</td>
<td></td>
<td></td>
<td>70°S, 15°E</td>
<td></td>
</tr>
<tr>
<td>INU20</td>
<td>7</td>
<td>75</td>
<td></td>
<td></td>
<td>62-72°S, 74-165°E</td>
<td></td>
</tr>
</tbody>
</table>

*The first character in the sample name designates the environment type (S, marine sediment; W, water; and I, ice). The second character indicates the geographic location (R, Arctic; N, Antarctic; T, temperate; and P, tropical). The third character indicates whether the sequences were derived from cultured isolates (C) or environmental clones (U).*
Data Analysis (Development Details)

- Written in Python on a Mac running OS X, including a UPGMA and a PCA implementation
- Pulled down sequences from Genbank
- Aligned sequences using the Arb⁠¹ tools and manual curation
  - A GUI tool for sequence database handling and data analysis
- Continued to use Arb to fill in a tree via parsimony
- Used the UniFrac tools to perform significance tests, UGPMA clustering, and PCA

Examine how the number and evenness of sequences affect UPGMA clustering

Ran several simulations using 100 reduced trees
- Each environment was randomly assigned a specific number of sequences defined at the start of each experiment
- Sequences (and environments) were added or removed as appropriate for each iteration
List of samples that are not significantly different at a cutoff of $P$-value $\geq 0.05$

Kind of hard to see the clusters here
P-Value Clusters

Image generated by Graphviz using the FDP algorithm. Edge lengths NOT proportional to the real distance between the samples.
UPGMA Tree

**FIG. 2.** UPGMA cluster of marine samples. The number of sequences that represent each environment is indicated next to the sample name, as well as the symbol with which the sample is represented in Fig. 3.

**TABLE 3.** UPGMA jackknifing results

<table>
<thead>
<tr>
<th>Node</th>
<th>17</th>
<th>20</th>
<th>31</th>
<th>36</th>
<th>40</th>
<th>58</th>
</tr>
</thead>
<tbody>
<tr>
<td>N1</td>
<td>3</td>
<td>14</td>
<td>31</td>
<td>27</td>
<td>12</td>
<td>NA</td>
</tr>
<tr>
<td>N2</td>
<td>8</td>
<td>1</td>
<td>29</td>
<td>33</td>
<td>48</td>
<td>63</td>
</tr>
<tr>
<td>N3</td>
<td>1</td>
<td>8</td>
<td>7</td>
<td>11</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>N4</td>
<td>14</td>
<td>16</td>
<td>11</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>N5</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>27</td>
<td>37</td>
</tr>
<tr>
<td>N6</td>
<td>27</td>
<td>36</td>
<td>57</td>
<td>67</td>
<td>53</td>
<td>63</td>
</tr>
<tr>
<td>N7</td>
<td>23</td>
<td>23</td>
<td>36</td>
<td>44</td>
<td>52</td>
<td>66</td>
</tr>
<tr>
<td>N8</td>
<td>22</td>
<td>17</td>
<td>17</td>
<td>39</td>
<td>31</td>
<td>37</td>
</tr>
<tr>
<td>N9</td>
<td>52</td>
<td>58</td>
<td>64</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>N10</td>
<td>8</td>
<td>16</td>
<td>79</td>
<td>96</td>
<td>94</td>
<td>100</td>
</tr>
<tr>
<td>N11</td>
<td>6</td>
<td>12</td>
<td>40</td>
<td>46</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>N12</td>
<td>13</td>
<td>31</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>N13</td>
<td>16</td>
<td>38</td>
<td>41</td>
<td>38</td>
<td>64</td>
<td>79</td>
</tr>
<tr>
<td>N14</td>
<td>34</td>
<td>50</td>
<td>29</td>
<td>23</td>
<td>12</td>
<td>6</td>
</tr>
<tr>
<td>N15</td>
<td>69</td>
<td>77</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>N16</td>
<td>18</td>
<td>40</td>
<td>27</td>
<td>28</td>
<td>28</td>
<td>21</td>
</tr>
<tr>
<td>N17</td>
<td>24</td>
<td>35</td>
<td>43</td>
<td>46</td>
<td>37</td>
<td>50</td>
</tr>
<tr>
<td>N18</td>
<td>97</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
</tr>
</tbody>
</table>

*For each node in the UPGMA tree (Fig. 2) (rows), the numbers show the percentages of trials (⁴ = 100) that the node occurred in when each environment was represented by only 17, 20, 31, 36, 40, or 58 sequences (columns). The node names correspond to the node labels in Figure 2. NA, not available.*
PC1 divides ice/cultured from other uncultured PC2 divides open sea from sediment/coastal

PC3 divides seawater from sediment and ice PC4 -- I’m not sure. Sediment depth, maybe?
What does this mean?

- See biologically meaningful patterns
  - Support previous observations
  - Reveal characteristics of marine microbial communities
- Patterns mentioned
  - Cultured vs. Uncultured
  - Geography Doesn’t Matter
  - Seawater Varies
Pattern 1: Cultured vs. Uncultured

- Ice samples cluster together
  - Cultured or uncultured
  - Supports the hypothesis that most of the bacteria in ice are culturable
  - Well supported by jackknife values (100% in some cases)
Cultured sediment and seawater tend to cluster with the ice samples

- NOT with the original environment.
- Separated out by the first principal component along with ice
- Suggests that bacteria grown in sea ice or pure culture share some property the rest do not
Pattern 2: Geography Doesn’t Matter

- A minor player in structuring communities
- Habitat means more
  - Arctic, Antarctic, and Japanese marine sediment samples all cluster together
  - Supports rejecting the hypothesis that psychophilic bacteria cannot cross the equator
- Caveat: Jackknife values are poor
  - We need more data to really tell how strong this pattern is
Pattern 3: Seawater Varies

- Uncultured ice and sediment tend to cluster together in distinct clusters
- Seawater does not cluster as strongly, but rather forms biological subclusters
  - Arctic seawater sample with Arctic ice cluster
  - Coastal seawater communities
  - Open ocean communities
  - “Terrestrially impacted” seawater
- Seawater should not be treated as a homogenous environment
UniFrac is useful:
- Detects biologically meaningful patterns.
- Accepts diverse kinds and sources of data.
- Suitable for large-scale comparisons between environments.
- Can weight sequences in samples by abundance.
- Small sample sizes can be enough to find associations.
- As more data is accumulated, UniFrac will grow more useful.
UniFrac can:

- Provide a unified framework for explaining previous observations.
- Allow broad conclusions to be drawn about sample similarity.
- Shed light on the biological factors that structure microbial communities.
Weighted UniFrac

- Equation

\[ u = \sum_{i}^{n} b_i \times \left| \frac{A_i}{A_T} - \frac{B_i}{B_T} \right| \]

- For all branches, multiply the current branch length by the difference in proportion of descendant branches from each community
Fast UniFrac

- New implementation of the same concepts
- Much faster (up to two orders of magnitude)
  - BLAST-based phylogeny generation
- Many UI changes
  - 3D Visualization of PCA analysis
  - Better visualizations of the first 10 PCs
  - Better import capabilities
- Parallelization
- Part of PyCogent 1.3 and later versions
Fast Unifrac

Unanswered Questions

1. Do the *jackknifing results* make sense?
   - The numbers for N10 really bother me.
2. Yours?
Thank-You!