Lab Report 4

Zachary Foster

May 13, 2014

We acquired 454 reads from 16S samples form our instructors. We used `mothur` to analyze the reads using default settings, except where stated otherwise. First we demultiplexed and quality filtered the reads using the `trim.seqs` command. Sequences were filtered out if they had an average phred score of less than 25, contained any ambiguous bases, had an homo-polymer run greater than 10bp, or they were shorter than 200bp. Reads were also trimmed up to any bases with a phred score less 20. Due to the short sequences produced by high throughput sequencers and the limited amount of phylogenetic signal resulting, it is important that stringent quality filtering is applied. Read errors can artificially inflate diversity estimates and compromise alignments.

I then extracted all of the reads from the sample ALP12GB; from this point all further analyses used this set of reads. The `unique.seqs` command was used to remove duplicate sequences in order to make further steps more computationally efficient. There were 509 high quality reads, 320 of which represent unique sequences. We then used the `chimera.uchime` to remove any chimeras, but none were found. `filter.seqs` was used to filter the badly aligned positions in the alignments using the the `vertical=T` option. This process removed 49600 of 50000 columns, resulting in an alignment of 400bp. We used the Ribosomal Database Project Naive Bayes Classifier implemented by the `classify.seqs` with a `cutoff` of 80 and assigned the resulting phylum level assignments to their respective reads. We removed non-Bacterial sequences using the `remove.lineage` command.

We made phylogenetic trees using three methods: `clustalw`, `FastTree`, and `PhyML`, which implement neighbor-joining 1, mixed distance and maximum likelihood 2, and maximum likelihood respectively 3. We used a R script `annotate_tree.R` to apply the phylum level identifications to the trees to aid their display. The three methods of phylogentic inference produced substantially different trees. The methods that incorporated maximum likelihood (PhyML and FastTree to a lesser extent) produced more believable trees with more hierarchical structure that correspond to the large span of diversity that is probably present, but took longer to run.

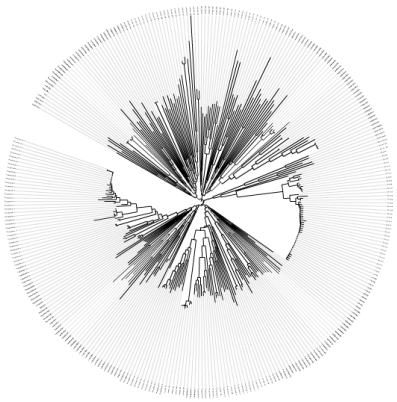

Figure 1: Clustal

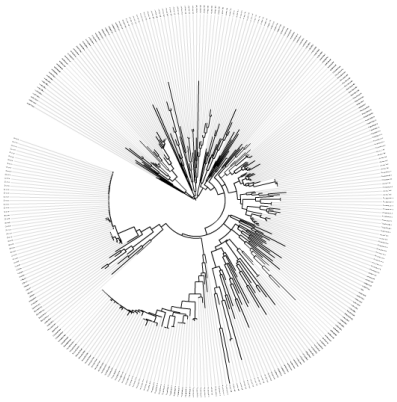

Figure 2: FastTree

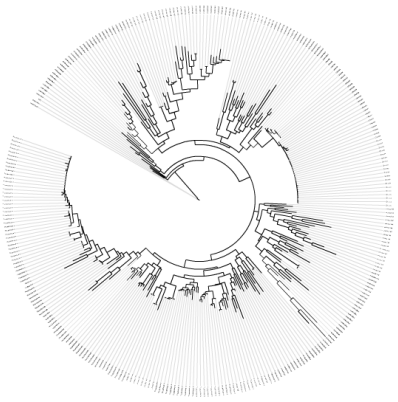

Figure 3: PhyML