In Silico Identification of 16S rRNA Outliers from Sequence-Based Bacterial Identification Studies

Tara Minucci†, Alexander Kirst‡, and Dr. Jonathan Blake‡
†Department of Biological Sciences, ‡Department of Mathematics and Computer Science
275 Mount Carmel Avenue, Hamden, CT 06518-1908

Abstract

Recent advances in gene sequencing have provided researchers with a relatively inexpensive and fast technique for classifying bacteria. Researchers often use the results obtained from BLAST searches on these sequences to suggest similarities, and build phylogenetic trees. BLAST searches, however, are optimized for speed, and not accuracy. The gold standard for establishing sequence similarity is pairwise alignment. Using sequences obtained from Quiniipiac University researchers, this project will build alignment tools to identify outliers, automating alignment-based sequence identification and phylogenetic tree building. Five candidate outliers were identified as a result of this study.

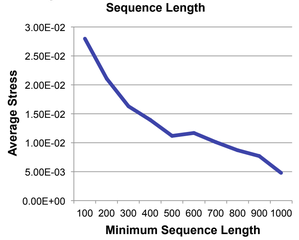
Introduction

Due to rapid advances in the sequencing process, many bacterial studies are characterized by sequencing (and subsequent processing) of unknown organisms. Researchers typically select the 16S Ribosomal RNA sequence as the target of their studies because the sequence is ubiquitous (across bacterial and archaeal genomes), highly functionally conserved, and of sufficient length for meaningful bioinformatics analysis[1]. Researchers often use BLAST searches through databases at the National Center for Biotechnology Information (NCBI) to classify unknown sequences. BLAST is used for NCBI database searches because it is optimized for speed, and the NCBI databases are huge. The gold standard for sequence similarity measures, however, is pairwise sequence alignment. Pairwise alignment determines optimal alignment and similarity scores, and thus is much slower. With the introduction of 16S-specific databases[2], however, reasonably fast searches using optimal techniques become feasible. As the first step in building tools for researchers (faculty and students) at Quinnipiac to identify 16S sequences using this gold standard, we are investigating previously sequenced 16S data to identify outliers in the dataset.

Materials and Methods

Data collection consisted of obtaining 16S rRNA sequences from Quinnipiac University faculty research. A total of 404 sequences were collected from Dr. Lisa Kaplan, Dr. Christian Eggers, and Dr. Donnasue Graesser. Each sequence was converted into FASTA format and a common header naming scheme was used. The sequences ranged widely in quality. Shorter sequences, and sequences with many unknown positions, skewed similarity scores and thus masked true outliers. We iteratively clustered the data with increasing minimum sequence lengths, and noticed that the removal of shorter sequences resulted in a large decrease in measured error (stress). Figure 1 shows this decrease. We chose 1000 as a minimum sequence length, giving us 161 sequences.

Figure 1: Stress as a function of Minimum Sequence Length



Sequences were aligned using the Smith/Waterman algorithm with Gotoh's improvements[3]. Similarity scores were converted to distances, and a clustering library[4] was used to convert these distance scores into a phylogenetic tree. Multidimensional scaling (MDS) analysis[5] was used as an additional visualization tool to observe outliers in the fixed data set, and was used to determine stress in the previous analysis.

Potential outliers were then compared to the Living Tree Project (LTP) database[2] of non-redundant bacterial and archaeal 16S sequences. Data was visualized using the clustering and MDS processes described above.

Results

The Phylogenetic tree created by the clustering algorithm on the 161 sequences in the working set identified 5 sequences as potential outliers. Figure 2 shows the top half of the full tree with the 5 outlier sequences enlarged. It is clear from the tree that these 5 sequences are separate from the rest of the clustered data.

Figure 2. A segment of the phylogenetic tree created using the 161 sequence data set. The top portion of the tree is enlarged to highlight the five potential outliers.

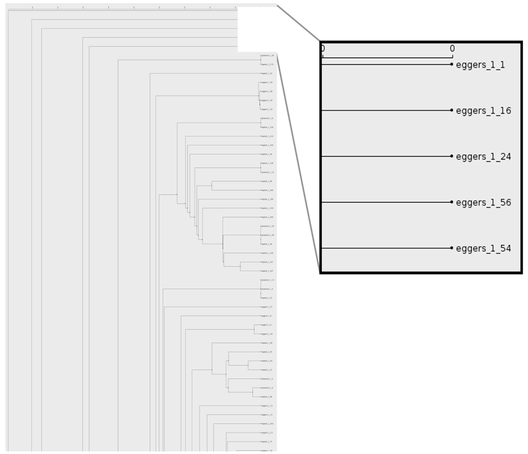
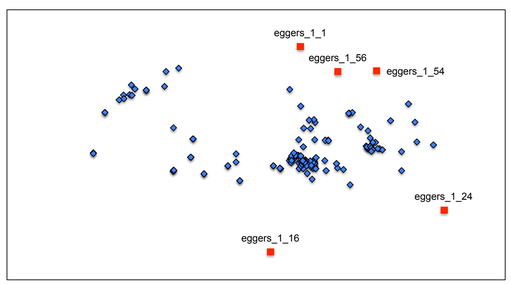


Figure 3. A Multidimensional scaling analysis of the curated dataset of 161 16S sequences with candidate outliers highlighted.



In addition to the tree visualization, we also used MDS on the inter-point distances to reduce the final dimension to 2D and plotted the results. Figure 3 shows this plot with the 5 candidate outliers clearly identified as such.

Having identified the outliers, we then considered why they were outliers. We recognize that sequences could be outliers because of contamination or error, so we aligned these sequences with those in the LTP database to see if they were in fact 16S sequences. After aligning the 5 outliers to the LTP database, and considering the top 5 hits for each outlier, we found three possible results. The top 5 hits could:

1. agree at the family and genus level
2. agree at the family level
3. show no agreement

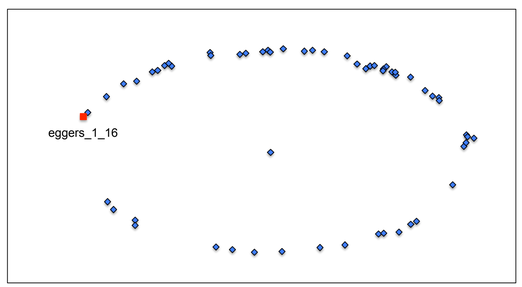
An example of an outlier from each of these categories is shown in Table 1.

Sequence Header	Target Family	Target Species
eggers_1_16	Spirochaetaceae	Borreliella spielmanii
eggers_1_16	Spirochaetaceae	Treponema amylovorum
eggers_1_16	Spirochaetaceae	Borrelia afzelii
eggers_1_16	Spirochaetaceae	Borrelia lusitaniae
eggers_1_1	Lactobacillaceae	Lactobacillus brantae
eggers_1_1	Lactobacillaceae	Lactobacillus hominis
eggers_1_1	Syntrophomonadaceae	Syntrophomonas wolfei subsp. saponavida
eggers_1_1	Helicobacteraceae	Helicobacter fennelliae
eggers_1_54	Staphylococcaceae	Staphylococcus xylosus
eggers_1_54	Staphylococcaceae	Staphylococcus saprophyticus subsp. bovis
eggers_1_54	Staphylococcaceae	Staphylococcus saprophyticus subsp. saprophyticus
eggers_1_54	Staphylococcaceae	Staphylococcus cohnii subsp. cohnii
eggers_1_54	Staphylococcaceae	Staphylococcus arlettae

Table 1. Exemplar sequences from each of the three outlier categories showing the top 5 LTP database hits. Scaled identity represents alignment length and sequence identity scaled by the number of gaps in the alignment.

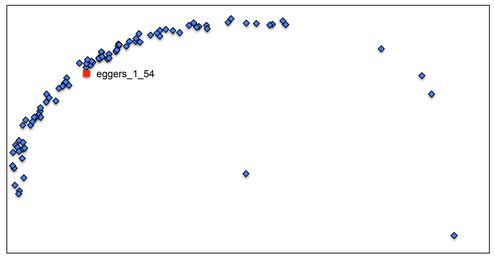
To further investigate outlier sequences that agree at the family or family/genus level with sequences in the LTP, we generated phylogenetic trees and MDS plots of the outliers and their respective families or genera. The outlier sequences were not distinguished in either case by the trees, but the MDS plots yielded potentially more interesting results.

Figure 4. A Multidimensional scaling analysis of the sequence labeled eggers_1_16 with the sequences in the LTP database from the family Spirochaetaceae.



The MDS plot of the sequence labeled eggers_1_16 with sequences from the LTP from the family Spirochaetaceae is shown in Figure 4. The MDS plot of the sequence labeled eggers_1_54 with sequences from the LTP from the genus Staphylococcus is shown in Figure 5. In both cases the target sequence is highlighted in the plot, and the highlighted sequence borders an open space in the plot.

Figure 5. A Multidimensional scaling analysis of the sequence labeled eggers_1_54 with the sequences in the LTP database from the genus Staphylococcus.



Conclusions

Our research has identified 5 outlier sequences from a cleaned dataset of 161 16S sequences collected at Quinnipiac University. These 5 sequences present as outliers in both phylogenetic tree analysis and MDS. Cleaning the dataset removed sequences that were outliers due to their length or lack of fidelity. Further investigation should focus on these 5 identified sequences.

Outlier sequences were aligned to the LTP database, and we have identified three categories by which outlier 16S sequence alignment results from the LTP database can be classified. Table 2 shows the alignment results of the 5 outlier sequences to the LTP database.

Outlier Sequence	Family Agreement	Genus Agreement
eggers_1_1	None	None
eggers_1_16	Spirochaetaceae	None
eggers_1_24	None	None
eggers_1_54	Staphylococcaceae	Staphylococcus
eggers_1_56	Staphylococcaceae	Staphylococcus

Table 2. Family/genus predictions for outliers when compared to the LTP database.

Acknowledgments

We would like to thank Quinnipiac University professors Dr. Lisa Kaplan, Dr. Christian Eggers, and Dr. Donnasue Graesser for providing 16S rRNA sequences from their research. We would also like to thank Dr. Joel Vaughan for his statistical consultations.

References

[1] Janda J. and Abbot S., (2007) 16S rRNA Gene Sequencing for Bacterial Identification in the Diagnostic Laboratory: Pluses, Perils, and Pitfalls, J. Clinical Microbiology. 2761–2764.
[2] Quast C, et. al. (2013) The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. Nucl. Acids Res. 41 (D1): D590-D596.
[3] Gotoh, O (1982). An improved algorithm for matching biological sequences. J. Mol. Bio.162: 705.
[4] hierachchical-clustering-java obtained from https://github.com/lbehnke/hierarchical-clustering-java downloaded April 2016.
[5] Pich, Christian. (2009). Applications of Multidimensional Scaling to Graph Drawing. Konstanz, University of Konstanz.