TreeLign: Automatic Update of 16s rRNA Phylogenetic Tree and Multiple Alignment in Metagenomics

Shaojie Zhang\textsuperscript{1,*}, Yuan Li\textsuperscript{1}, Aaron Halpern\textsuperscript{2,3}

1, School of Electric Engineering and Computer Science, University of Central Florida, Orlando, FL 32816; \{shzhang, liy\}@cs.ucf.edu.
2. J. Craig Venter Institute, Rockville, MD 20850;
3. Present Address: Complete Genomics, 2071 Stierlin Court, Mountain View, CA 94043; ahalpern@completegenomics.com.
* To whom correspondence should be addressed.

\textbf{Keywords:} 16s rRNA, phylogenetic classification, multiple alignment, maximum parsimony

\textbf{Abstract}

Recently, a variety of metagenomics projects have produced a large amount of 16S ribosomal RNA (16S rRNA) sequences. Fast and accurate taxonomic classification and phylogenetic assessment of these 16S rRNA sequences have become a challenge in metagenomics. One classification method is based on the “best hit” of BLAST. But, this method may lead to wrong assignment because it only compares with leaves of the phylogenetic tree. Another attempt is to use maximum likelihood (ML) evaluating each possible phylogenetic positions in the tree with a predefined multiple alignment, which aligns the target 16S rRNA sequence with reference alignment. The quality of the fixed alignment will affect the ML evaluations. Here we propose a new computational method, TreeLign, to classify 16S rRNA sequences with the reference phylogenetic tree and multiple alignment. By using Fitch algorithm, TreeLign first generates the profiles of each interior nodes based on reference phylogeny and multiple alignment. Then, TreeLign tries every possible branch on the phylogeny that the query sequence can be inserted with Maximum Parsimony (MP) score. Finally, by iteratively reassigning phylogenetic positions for each sequence in the reference alignment, TreeLign is able to improve the accuracy of both reference alignment and the corresponding phylogenetic tree with the minimum number of mutations.

We tested the assignment accuracy of TreeLign on the core set of Greengene's 16S rRNA data. We first computed the distance between the original phylogenetic position and that reassigned by TreeLign after leaving the sequence out of the phylogeny. The results show that 45.20% sequences are exactly re-assigned back to the original locations and the average distance of assignment is equivalent to 2.46% of the tree's diameter. This implies that in general the assignment of TreeLign is consistent with the phylogenetic tree of Greengene's core set, though with a few disagreements. TreeLign can further improve total number of mutations on the core set of Greengene’s 16S rRNA data. TreeLign decreases MP score on the phylogeny by 12.33%, through iteratively removing every sequences once and then reassigning them back.