Dreycey Albin

1/16/2019

Computational Microbial Forensics; Paper review #2

RefSeq database growth influences the accuracy of k-mer-based lowest common ancestor species identification

Introduction. *The focus of the research in this article was focused on how different versions of the RefSeq database affect the output of k-mer based taxonomic identification tools*. In particular, the researchers used Kraken as the representative k-mer tool because of the favorable F_1 Score, which is measure that establishes both precision and recall. In addition, Bracken, a Bayesian method that further refines Kraken results, was used to further test the ability of Kraken to make species level assessments. Because of the increasing number of species being added to the RefSeq database, and a decreasing diversity of genera being added, it was hypothesized the updated databases disrupted the LCA (Least Common Ancestor) approach used by K-mer methods. This is important as it may be better to use older database versions for correctly verifying a sample using K-mer methods, and it may indicate that the influence on accuracy of these methods may continue to get worse as RefSeq growth continues.

Methods. To obtain different versions of the RefSeq databases, which are not publicly available, a mixture of Perl and Python scripts were used on version 84 of RefSeq. These scripts utilized the catalog file to retrieve sequences present in the previous versions of RefSeq. Using the Biowulf cluster through NIH, every 10 versions of RefSeq, up to 80, were used to test Kraken and Bracken. There were several benchmarks used to test the k-mer programs: (1) Correctness of species and genus level calls; (2) Counting the number of each type of call (species, genus, domain-family, and unclassified); (3) Ability to correctly distinguish between similar species; and (4) The ability to perform an analysis on metagenomic samples. Using the mentioned benchmark tests allowed for the researchers to obtain a quantitative assessment for how RefSeq growth influences the accuracy of K-mer based methods for taxonomic classifications.

Results. Indicated throughout the article, there are three essential results reported. First, it was found that the number of species added to the RefSeq database has been outpacing the number of genus' added to the database. This causes the Simpson's diversity index to be higher for that for species, and also shows a stagnation in the diversity for the number of genera in the RefSeq database. Second, in all of the benchmark cases, it was shown that the number of unclassified hits decreases as the RefSeq database grows, yet the number of hits at the species level decreases. This is because of similar species being in the database, causing the LCA algorithm to arrive at the genus level. Lastly, using the metagenomic samples and *Bacillus* species as similar species, it was shown that Bracken may help identify the species hits, yet it struggles with correctly identifying the species.

Key Results and Hypothesis Testing. The results in the study convincingly show that the original hypothesis was correct: The growth of the RefSeq database has an effect on the accuracy of K-mer Based methods (in specific, Kraken). Furthermore, testing different versions of the database showed an optimal versions of the database, which may point to an underlying flaw in the LCA approach that Kraken used for searching a database for taxonomically related species.