KrakenUniq: confident and fast metagenomics classification using unique k-mer counts

Dreycey Albin (da39)

April 16, 2019

Scientific question

The scientific question and goal for this paper focuses on *delivering a metagenomic classifier that minimizes false positives, which commonly occurs when reads are mapped to low complexity regions of a genome of an organism not actually present in the sample.*

Hypothesis tested

The hypothesis tested in this research is that *the probabilistic HyperLogLog algorithm may be used to quickly calculate cardinalities for each taxon, thereby speeding the ability to rid of false positives.* In addition, the very idea that this may reduce false positives is a hypothesis. The authors test this hypothesis by comparing their outcomes to 21 simulated and 10 biological datasets. They subsequently compate KrakenUniqs ability to classify the bacteria comapared to 11 other metagenomic classifiers.

Methods/Results

The methods for KrakenUniq seem to fall into 3 major categories: (1) The use of the HyperLogLog algorithm to quickly estimate taxon cardinalities; (2) Testing KrakenUniq on 21 simulated datasets; and lastly, (3) testing KrakenUniq on 11 biological datasets.

0.1 The HyperLogLog Algorithm

The first major section in the paper was focused on the HyperLogLog algorithm, which uses a probabilistic method to calculate the cardinality of a set. In the case of KrakenUniq, this is used to calculate the number of unique kmers for each taxon mapped. In essence, it makes sense that reads may be mapped to genomes of organisms not actually present in the sample. By extension, when this happens, only a small portion of the genome will actually have reads mapped, thereby meaning that the reads that are mapped have less sequence diversity than would be seen with reads mapped to an organism actually in the sample. The algorithm in KrakenUniq works by first mapping the reads to each taxon/ Thereafter the cardinality is calculated by the HyperLogLog algorithm, which works like this: Kmers for a given taxon are mapped to bitstrings using a locaity sensitive hash function; note the proabability of finding a certian number of leading zeros decreases as the number of leading zeros increases. However, the larger the set, the increased probability you will find a kmer mapped to a bitstring with a higher number of leading zeros, and this can therefore be used as a rough estimate for cardinality:

$$2^{k+1};\ where\ k\ =\ number\ of\ leading\ zeros$$

0.2 Simulated Datasets

Two metrics were used to compare the different metagenomic classifiers. These are the harmonic mean of precision, F1, and the recall at a dalse discovery rate of 5%. The authors found that, on average, the recall increases by 4-9% and the F1 score increases by 2-3%. It is noted that the number of unique reads is decreased within the simulated data sets, as compared to real biological datasets. Testing KrakenUniq against the 11 other classifiers showed that KrakenUniq ussually outperforms all competitors, depending on the database used (the nt database works better). BLAST outperforms KrakenUniq when KrakenUniq uses the 'std' database, but it is shown that KrakenUniq with the 'nt' database outperforms all other programs (although CLARK also ranks higher). To test on actual sample output, the authors generated a third dataset by sampling reads from the Sequence Read Archive (SRA). For this 280 datasets from 280 different species were chosen, and then obtained 34 million read pairs in total. Using these datasets, they saw that the recall increased with higher coverage, yet so

did the potential of the false positive rate. Because of the latter finding, they are pushed to use higher thresholds when the coverage is higher.

0.3 Biological Datasets

To test real biological data, the authors reanalyzed patient samples from previously described series of neurological infections. In the original study the mass of the spinal cords and the brains were taken into account, and 4 of the samples were be analyzed using metagnomics. The authors of wanted test KrakUniqs ability to rerun the analysis that was shown in the original study. In two cases, the reads are very low, but the number of unique kmers is still very high, showing that the reads are most likely from different regions of the taxons genome. Overall, the authors were able to classify the samples accurately, even when the samples had a lower read count.

Key implications of the results

KrakenUniq provides a fast method for being able to classify metagenomic samples while decreasing the false postive rate of the classified samples. This method allows for a competitive advanage over other methods by ridding of taxon calls that don't have reads that span the genome of the taxon, yet while not having to first map the reads to the genomes of each species.