Study finds growth of genomic databases affects species accuracy

October 30 2018, by Mike Williams

Credit: CC0 Public Domain

There are many ways to slice and dice genomic data to identify a species of bacteria, or at least find its close relatives. But fast techniques to sequence genomes have flooded the public databases and in a biased

fashion, containing lots of genomic data about some species and not enough about others, according to a Rice University computer scientist.

Todd Treangen and his colleagues tested taxonomic classification methods that match genomic sequences from bacteria of interest with those recorded in large databases to identify species. In the process, they charted a path toward improved accuracy and sensitivity.

Treangen is senior author of a study published this month in *Genome Biology* that demonstrates how changes over time in a widely used federal database, the National Center for Biotechnology Information's RefSeq, have influenced the accuracy of metagenomic classification methods.

A primary concern for Treangen, an expert in metagenomics—the study of genetic material from environmental samples—is maintaining the ability to quickly identify bacteria that pose a threat to public health.

Big data is uniquely positioned to do this—but there's so much of it. At present, he said, low-cost and high-throughput DNA shotgun sequencing machines, which read short DNA sequences from collections of microorganisms, have resulted in the doubling of genomic data in RefSeq every two to three years.

"I initially thought more data is always better for these methods," said Treangen, who joined Rice this year from the University of Maryland Institute for Advanced Computer Studies. "You would expect that there would be no penalty, because database growth is good." However, the researchers found that bacterial data in RefSeq has an outsized effect at the species level of the taxonomic hierarchy, which is growing at a breakneck pace.

That's a problem for researchers who combine two common techniques

to identify what they find. One is called k-mer-based classification, which identifies short DNA sequences from all the organisms in a bacterial sample via exact matches.

"Most of the methods that have made the problem computationally feasible rely on k-mers, which are exact matches of length 'k,' or a key in to the microbes contained in the database," he said. "If a sequenced read perfectly matches something in the database, the intuition is that you can say what that is with great precision and shortcut more expensive computational approaches."

A commonly used technique with k-mer-based classification is lowest common ancestor (LCA) assignment, he said. LCA compares samples to sequences that share a match, assigning them if necessary to a higher level in the taxonomy, such as a genus rather than a species. But this may not be specific enough for researchers trying to pin down a pathogen, he said.

In fact, the study found a k-mer-based classification tool called Bracken, which uses Bayesian statistics to infer the best match for a sequence, helped mitigate the imbalance. Even so, it struggled to identify genomes with close relatives, but not perfect matches, in the database.

Treangen said well-funded research into particular pathogens is a necessity and has greatly aided rapid-outbreak detection and tracking, but it ultimately biases public databases like RefSeq.

"For instance, there's an immense bias toward foodborne pathogens," he said. "Society wants to know a lot about Salmonella, and rightfully so. The FDA, and specifically GenomeTrakr, have aided in the sequencing of thousands of relevant pathogens and have added them directly to the reference database."

However, he said that skews the reference database toward particular genera and families of microbes in a way that affects the accuracy and sensitivity of fast taxonomic-classification tools like Kraken that use k-mer and LCA-based approaches.

Treangen said the best recent example of a false positive identification is a study that initially reported evidence of anthrax bacteria in New York City's subways. The study, based on sequenced genomes from samples, was later revised to reflect mismatches that falsely identified the sequences as Bacillus anthracis.

While a focus on public health is a key priority, Treangen said novel techniques able to cope with database growth and noise, coupled with an increased breadth of sequenced genomes, is needed for continued improvements in the field. "For example, microorganisms from the soil and ocean are severely under-sampled," he said. "There remain a lot of microbes that we need to continue to sequence to better fill out public databases, and that will ultimately help our ability to accurately classify microbes from complex samples."

More information: Daniel J. Nasko et al. RefSeq database growth influences the accuracy of k-mer-based lowest common ancestor species identification, *Genome Biology* (2018). DOI: 10.1186/s13059-018-1554-6

Provided by Rice University