DNA databases: New method cuts indexing from weeks to hours, searches to minutes

June 28 2021

Credit: CC0 Public Domain

Rice University computer scientists are sending RAMBO to rescue genomic researchers who sometimes wait days or weeks for search results from enormous DNA databases.

DNA sequencing is so popular, genomic datasets are doubling in size every two years, and the tools to search the data haven't kept pace. Researchers who compare DNA across genomes or study the evolution of organisms like the virus that causes COVID-19 often wait weeks for software to index large, "metagenomic" databases, which get bigger every month and are now measured in petabytes.

RAMBO, which is short for "repeated and merged bloom filter," is a new method that can cut indexing times for such databases from weeks to hours and search times from hours to seconds. Rice University computer scientists presented RAMBO last week at the Association for Computing Machinery data science conference SIGMOD 2021.

"Querying millions of DNA sequences against a large database with traditional approaches can take several hours on a large compute cluster and can take several weeks on a single server," said RAMBO co-creator Todd Treangen, a Rice computer scientist whose lab specializes in metagenomics. "Reducing database indexing times, in addition to query times, is crucially important as the size of genomic databases are continuing to grow at an incredible pace."

To solve the problem, Treangen teamed with Rice computer scientist Anshumali Shrivastava, who specializes in creating algorithms that make big data and machine learning faster and more scalable, and graduate students Gaurav Gupta and Minghao Yan, co-lead authors of the peer-reviewed conference paper on RAMBO.

RAMBO uses a data structure that has a significantly faster query time than state-of-the-art genome indexing methods as well as other advantages like ease of parallelization, a zero false-negative rate and a low false-positive rate.

"The search time of RAMBO is up to 35 times faster than existing

methods," said Gupta, a doctoral student in electrical and computer engineering. In experiments using a 170-terabyte dataset of microbial genomes, Gupta said RAMBO reduced indexing times from "six weeks on a sophisticated, dedicated cluster to nine hours on a shared commodity cluster."

Yan, a Ph.D student in computer science, said, "On this huge archive, RAMBO can search for a gene sequence in a couple of milliseconds, even sub-milliseconds using a standard server of 100 machines."

RAMBO improves on the performance of Bloom filters, a half-century-old search technique that has been applied to genomic sequence search in a number of previous studies. RAMBO improves on earlier Bloom filter methods for genomic search by employing a probabilistic data structure known as a count-min sketch that "leads to a better query time and memory trade-off" than earlier methods, and "beats the current baselines by achieving a very robust, low-memory and ultrafast indexing data structure," the authors wrote in the study.

Gupta and Yan said RAMBO has the potential to democratize genomic search by making it possible for almost any lab to quickly and inexpensively search huge genomic archives with off-the-shelf computers.

"RAMBO could decrease the wait time for tons of investigations in bioinformatics, such as searching for the presence of SARS-CoV-2 in wastewater metagenomes across the globe," Yan said. "RAMBO could become instrumental in the study of cancer genomics and bacterial genome evolution, for example."

More information: Gaurav Gupta et al, Fast Processing and Querying of 170TB of Genomics Data via a Repeated And Merged BloOm Filter (RAMBO), *Proceedings of the 2021 International Conference on*

Management of Data (2021). DOI: 10.1145/3448016.3457333