INNOVATING CROSS-DOMAIN SOLUTIONS TO DETECT EMERGING BIOLOGICAL THREATS

Enabling The Rapid Search Of Petabyte-scale Dna Sequence Databases With Bloom Filters

Jason Gans Los Alamos National Laboratory

At approximately sixty petabases and doubling every 24 months, the Sequence Read Archive (SRA) is the largest repository of biological sequence data in the world. Reflecting the diversity of modern genomic science, the SRA is an invaluable data trove of nucleotide sequences from diverse sample types, including infectious diseases, clinical and environmental microbial communities, and gene expression studies. In addition to sequence data supporting published studies in the scientific literature, the SRA also contains information that can reveal unpublished activities occurring in laboratories around the world (due to laboratory contamination that can occur before, and during, the sequencing of biological samples). However, the large, multi-petabyte size of the SRA has limited the development of software tools and services to search the SRA with user-defined query sequences. This inability to search the SRA prevents the rapid identification of sequencing datasets that may contain sequences of interest and diminishes the utility of the SRA.

We have addressed this challenge and enabled rapid sequence searching by creating a compressed, searchable version of the SRA using Bloom filters. A Bloom filter is a computer science data structure that enables approximate tests of set membership (e.g., is a given DNA k-mer present in the set of k-mers representing an SRA record?). Bloom filters provide a tradeoff between reduced storage requirements and the probability of a false positive result for a set membership test. Drawing from the well-established literature on methods for storing and searching biological sequences in Bloom filters, this effort used cloud computing and a highly optimized Bloom filter creation pipeline to convert over 99% of SRA records into Bloom filters. The resulting Bloom filter database requires only 375 TB of disk space, much less than the tens of petabytes required to store the complete SRA. Applications of this new search capability include identifying SRA records that contain (a) poorly characterized or newly discovered taxa that have not been included in the taxonomic inventory of SRA records maintained by the NIH, and (b) evidence of genetic engineering (by searching for codon-optimized variants of naturally occurring genes). In addition to bioinformatic applications, we will also present recent progress in lossless compression algorithms for groups of Bloom filters, and “lessons learned” running large scale (> 1000 compute instance) cloud-based calculations over multiple months.

Los Alamos National Laboratory LDRD program