kmerprof : A Comparative *k*-mer Tool Tailored for Distributed-Memory Parallel Computers

Brett Youtsey[1], Steven Hofmeyr[2], Patrick Chain[1], Migun Shakya[1]

1. Bioscience Division, Los Alamos National Laboratory, Los Alamos, NM 87544
2. Computing Sciences, Lawrence Berkeley National Laboratory, Berkeley, CA 94720

Contact: byoutsey@lanl.gov, migun@lanl.gov

Introduction

Publicly available shotgun metagenomic samples have increased in both size and volume in recent years (Figure 1). Traditional ways of analyzing and comparing metagenomes require reference databases or assembling them into contigs, and as a result a large fraction of the sequences remain uncharacterized. To be able to use most of the sequence data for comparisons, reference free methods such as the ones that use k-mers can be effective. However, available k-mer based tools are usually not scalable and cannot handle large numbers and sizes of metagenomes. Here, we introduce *kmerprof,* a comparative *k*-mer tool tailored for distributed-memory parallel computers.

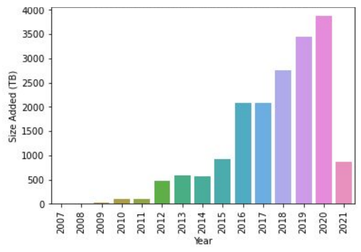

Figure 1: Growth in the volume of SRA records over the past decade as reported by NCBI[1] as of 03/21.

Workflow

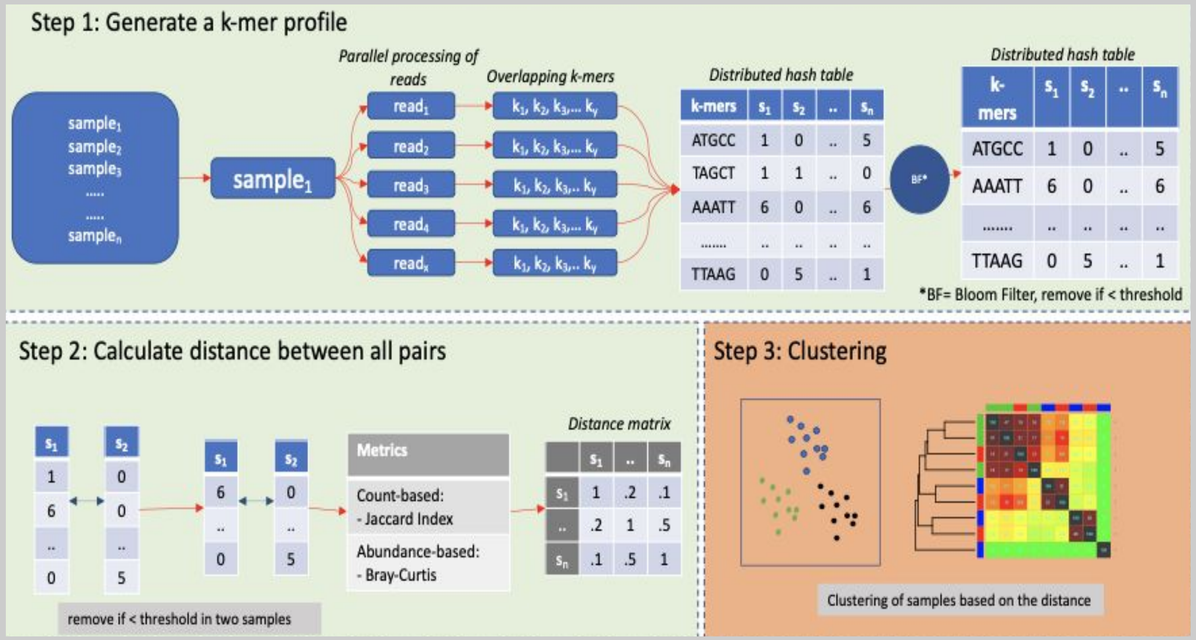

Figure 2: Diagram showing the workflow of kmerprof. *Step 1:* reads in each sample are split for parallel processing and the abundances of each k-mer are stored in a distributed hash table. The hash table is then passed through an abundance filter to remove k-mers seen less than a specified abundance threshold. *Step 2*: kmerprof performs pairwise calculations of the Jaccard Index & Bray-Curtis Dissimilarity for each sample's filtered hash table. *Step 3*: samples are clustered with the resulting distance matrix.

Results

We ran kmerprof on the Cray XC40 supercomputer, Cori[2] on KNL nodes. Each KNL compute node has a single-socket Intel Xeon Phi Processor 7250 processor with 68 cores per node at 1.4 GHz and 96 GB of DDR4 2400 MHz memory.

Performance

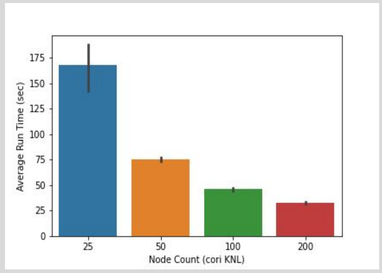

Figure 3: Kmeprof runtime benchmark for 99 sample HMP metagenomic dataset (309 GB). Each node count was run in triplicate with error bars as the standard deviation.

Clustering

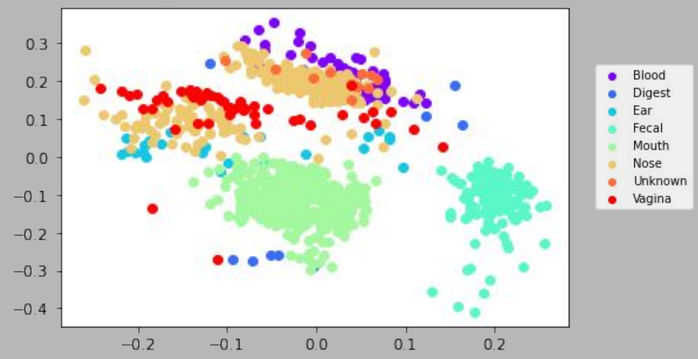

Figure 4: NMDS plot generated from the jaccard *k*-mer distance matrix (*k=21* bp) of a metagenomic dataset from the Human Microbiome Project[3] (n=882, 8.7 TB). The dataset contains samples from 19 different body sites collapsed into 8 general regions. Digest: digestive tract.

Benchmarking Accuracy with Genomic Distances

Taxa Rank	Samples	Subgroups	Distance
Genus: *escherichia*	74	*e. coli* & *e. fegusonii*	109
Family: *enterobacteriaceae*	97	*citrobacter , klebsiella, escherichia, salmonella,* & *enterobacter*	143
Order: *enterobacteriales*	99	*enterobacteriaceae, erwiniaceae, pectobacteriaceae,* & *yersiniaceae*	127
Class: *gammaproteobacteria*	98	*enterobacteriales, pseudomonadales, pasteurellales,* & *xanthomonadales*	135
Phylum: *proteobacteria*	100	*gammaproteobacteria, epsilonproteobacteria, alphaproteobacteria,* & *betaproteobacteria*	139

Table 1: Comparison between *k*-mer and alignment-based clustering of 5 separate genomic datasets of ascending taxonomic rank. kmerprof was run (*k* = 32 bp) on SRA records. GTDBTk[4] generated trees from each SRA's corresponding RefSeq[5] assembly. To compare the two resulting trees the Robinson- Foulds Distance was calculated.

Conclusions/Future Directions

- Kmerprof can store *k*-mer distributions of large and diverse datasets with distributed memory and generate novel hierarchies
- Co-assemble metagenomic samples clustered by kmerprof and assess assembly quality
- Use supervised machine learning to extract signature *k*-mers that best identify sample clusters

LA-UR-21-27520

References

1. https://trace.ncbi.nlm.nih.gov/Traces/sra/sra.cgi?view=announcement
2. https://docs.nersc.gov/systems/cori/
3. Turnbaugh, P. J., Ley, R. E., Hamady, M., Fraser-Liggett, C. M., Knight, R., & Gordon, J. I. (2007). The human microbiome project. *Nature*, *449*(7164), 804-810.
4. Chaumeil, P. A., Mussig, A. J., Hugenholtz, P., & Parks, D. H. (2020). GTDB-Tk: a toolkit to classify genomes with the Genome Taxonomy Database.
5. O'Leary, N. A., Wright, M. W., Brister, J. R., Ciufo, S., Haddad, D., McVeigh, R., ... & Pruitt, K. D. (2016). Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic acids research*, *44*(D1), D733-D745.