Centrifuge: rapid and accurate classification of metagenomic sequences

Daehwan Kim[1], Li Song[1,3], Florian P. Breitwieser[1] and Steven L. Salzberg[1,2,3]

[1]Center for Computational Biology, McKusick-Nathans Institute of Genetic Medicine
[2]Dept. of Biomedical Engineering, Johns Hopkins Schools of Medicine and Engineering
[3]Dept. Of Computer Science, Johns Hopkins University, Baltimore, MD

Abstract

Centrifuge is a very rapid and memory-efficient system for the classification of DNA sequences from microbial samples, with better sensitivity than and comparable accuracy to other leading systems. The system uses a novel indexing scheme based on the Burrows-Wheeler transform (BWT) and the Ferragina-Manzini (FM) index, optimized specifically for the metagenomic classification problem. Centrifuge requires a relatively small index (e.g., 2.9 GB for ~2,800 bacterial genomes) yet provides very fast classification speed, allowing it to process a typical DNA sequencing run within an hour. Together these advances enable timely and accurate analysis of large metagenomics data sets on conventional desktop computers.

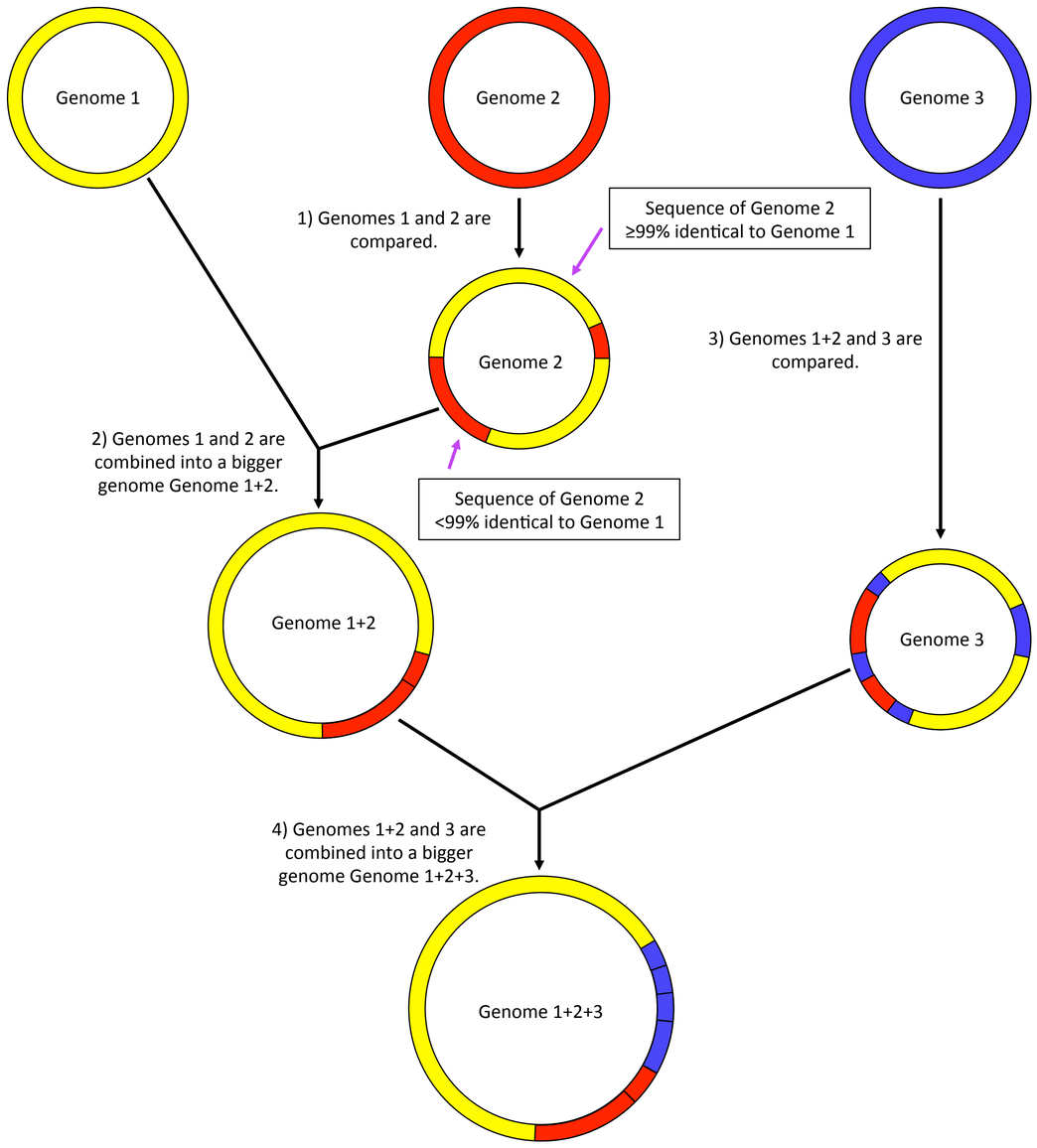

Figure 1. Compression of genomes belonging to the same species before indexing. Sequences ≥99% identical are merged.

Species name	Number of genomes	Total size (Mbp)	Total size after compression (Mbp)	Compression ratio (%)
Chlamydia trachomatis	73	85.9	1.9	98
Escherichia coli	62	315.8	58.4	82
Helicobacter pylori	53	86.2	68.8	20
Staphylococcus aureus	49	138.9	22.1	84
Salmonella enterica	43	209.4	34.8	83
Listeria monocytogenes	31	115.5	17.4	85
Streptococcus pneumoniae	25	52.8	10.1	81
Mycobacterium tuberculosis	21	92.5	8.7	91
Streptococcus pyogenes	19	35.2	7.3	79
Bacillus anthracis	9	32.7	5.6	83

Table 1. Compression ratio for genomes at the species level for 10 common bacterial genomes in GenBank.

Genome compression

Most metagenomics classification programs either require a very large index or suffer from slow classification rates. In order to keep the index size small, we first compress the genomes from the same species before indexing (Figure 1 and Table 1). We then use memory-efficient indexing schemes based on the BWT/FM index for these compressed (or reduced) sequences, which we further customized for lower memory usage. The BWT/FM index allows very fast search [1], which we also optimized specifically for classification.

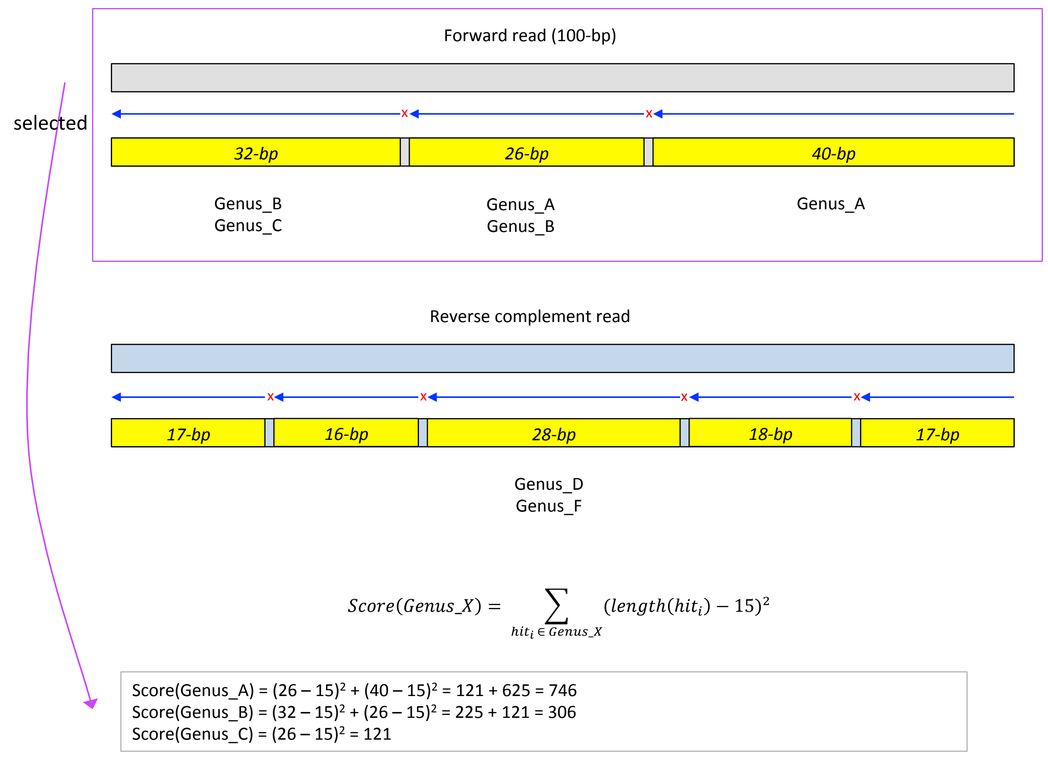

Figure 2. Classification of reads.

Classification

Centrifuge classifies reads using a fast alignment method similar to [1] but using our novel index (Figure 2). We compared Centrifuge with the leading classification programs using reads simulated from 2,800 bacterial genomes (100-bp long with a per-base error rate of 3%). Centrifuge provides the highest sensitivity (97.4%) and high precision (98%), as shown in Table 2.

Classifier	Genus sensitivity	Genus precision	Speed (reads/min)	Memory usage (GB of RAM)
PhymmBL [2]	90.3	98.4	48	3.0
Megablast [3]	94.8	98.2	363	3.0
Kraken [4]	92.0	99.9	1,217,039	73.6
Centrifuge	97.4	98.0	546,946	2.9

Table 2. Classification sensitivity, precision, and speed, memory usage based on simulated reads.

Acknowledgements

This work is supported in part by NIH grants R01-HG006102 and R01-HG006677 and by ARO grant W911NF-1410490.

References

1. Langmead, B. & Salzberg, S.L. **Fast gapped-read alignment with Bowtie 2**. Nat Methods 9, 357-359 (2012).
2. Brady A, Salzberg SL: **PhymmBL expanded: confidence scores, custom databases, parallelization and more.** Nat Methods 2011, 8:367.
3. Zhang Z, Schwartz S, Wagner L, Miller W: **A greedy algorithm for aligning DNA sequences.** J Comput Biol 2000, 7:203-214.
4. Wood DE, Salzberg SL: Kraken: **ultrafast metagenomic sequence classification using exact alignments.** Genome Biol 2014, 15:R46..