Optimized K-mer Matching For Million-Genome Collections On Laptops

Francesca Brunetti, Karel Břinda

SeqBIM

Abstract

Optimized K-mer Matching For Million-Genome Collections On Laptops

Francesca Brunetti[1,2]*, Karel Břinda[1]

[1] *Inria, Irisa, University of Rennes, Rennes, France*

[2] *Department of Public Health and Infectious Diseases, Sapienza University of Rome, Rome, Italy*

***Corresponding author**: francesca.brunetti@uniroma1.it

Abstract

Bacteria play a crucial role in human health, and their rapid identification is essential for timely interventions in cases of infections or outbreaks. The rapid progress in genome sequencing technologies has led to the creation of large bacterial genome collections, such as 661k [1] (n = 661,405) and AllTheBacteria [2] (n = 2,440,377). These databases have significant potential for applications such as rapid diagnostics of antibiotic resistance [3] and epidemiological surveillance [4], provided they can be efficiently searched in real time at points of care on portable devices.

However, real-time alignment across million-genome collections on portable devices is not possible with the available methods. While traditional tools such as BLAST [5] offer near-real-time search, they depend on a cluster and search only a limited subset of available genomes [6]. State-of-the-art k-mer indexes [7–9] struggle to scale to millions of genomes [10,11] or require clusters [8,9]. LexicMap [12] provides a moderately efficient BLAST-like search (15 mins per plasmid), but with resource requirements unsuitable to portable devices (AllTheBact index >4TB disk). The most promising has been Phylign [6], which manages alignments to the entire 661k collection on a standard laptop (index 100 GB disk, 14 GB RAM, 0.25 min per plasmid with batching). However, Phylign's suitability to time-critical decision-making at points of care [13] is limited by its dependency on COBS [13], which is fast only for short queries or near-exact matches. Although phylogenetic compression could be combined with more modern low-level indexes [10,11], no established methodology yet exists to guide their selection, mode of use, and parameter choice.

In this presentation, we will introduce a methodology for k-mer matching across million genome collections on portable devices using phylogenetic compression. We will begin by translating selected biological and epidemiological questions into k-mer-based algorithms and explore four types of atomic k-mer queries required across different types of questions. Next, we will summarize techniques for integrating k-mer matching within large-scale workflows using phylogenetic compression such as Phylign. We will then present results from our performance evaluation of three state-of-the-art k-mer indexes: Fulgor [11], Themisto [10], and Metagraph [9]. Finally, we will show that replacing COBS with phylogenetically compressed Fulgor reduces k-mer matching time by a factor of >6x for long sequences, while maintaining the comparable space requirements. Overall, our work provides practical guidelines for biologists and epidemiologists for applying k-mer-based search to their specific biological questions, on standard laptop computers.

References

1. Blackwell GA, Hunt M, Malone KM, Lima L, Horesh G, Alako BTF, et al. Exploring bacterial diversity via a curated and searchable snapshot of archived DNA sequences. PLoS Biol. 2021;19: e3001421.
2. Hunt M, Lima L, Shen W, Lees J, Iqbal Z. AllTheBacteria - all bacterial genomes assembled, available and searchable. bioRxiv. 2024. p. 2024.03.08.584059. doi:10.1101/2024.03.08.584059
3. Břinda K, Callendrello A, Ma KC, MacFadden DR, Charalampous T, Lee RS, et al. Rapid inference of antibiotic resistance and susceptibility by genomic neighbour typing. Nature Microbiology. 2020;5: 455–464.
4. Gardy JL, Loman NJ. Towards a genomics-informed, real-time, global pathogen surveillance system. Nat Rev Genet. 2018;19. doi:10.1038/nrg.2017.88
5. Basic local alignment search tool. J Mol Biol. 1990;215: 403–410.
6. Břinda K, Lima L, Pignotti S, Quinones-Olvera N, Salikhov K, Chikhi R, et al. Efficient and Robust Search of Microbial Genomes via Phylogenetic Compression. bioRxiv. 2024. doi:10.1101/2023.04.15.536996
7. Marchet C, Boucher C, Puglisi SJ, Medvedev P, Salson M, Chikhi R. Data structures based on -mers for querying large collections of sequencing data sets. Genome Res. 2021;31: 1–12.
8. Bradley P, den Bakker HC, Rocha EPC, McVean G, Iqbal Z. Ultrafast search of all deposited bacterial and viral genomic data. Nat Biotechnol. 2019;37: 152–159.
9. Karasikov M, Mustafa H, Danciu D, Zimmermann M, Barber C, Rätsch G, et al. Indexing all life's known biological sequences. bioRxiv. bioRxiv; 2020. doi:10.1101/2020.10.01.322164
10. Alanko JN, Vuohtoniemi J, Mäklin T, Puglisi SJ. Themisto: a scalable colored k-mer index for sensitive pseudoalignment against hundreds of thousands of bacterial genomes. Bioinformatics. 2023;39: i260–i269.
11. Fan J, Singh NP, Khan J, Pibiri GE, Patro R. Fulgor: A fast and compact -mer index for large-scale matching and color queries. bioRxiv. 2023. doi:10.1101/2023.05.09.539895
12. Shen W, Iqbal Z. LexicMap: efficient sequence alignment against millions of prokaryotic genomes. bioRxiv. 2024. p. 2024.08.30.610459. doi:10.1101/2024.08.30.610459
13. Bingmann T, Bradley P, Gauger F, Iqbal Z. COBS: A Compact Bit-Sliced Signature Index. String Processing and Information Retrieval. Cham: Springer International Publishing; 2019. pp. 285–303.