Searching DNA Databases on Mobile Devices

Andrew J. Mikalsen

Computer Science and Engineering
University at Buffalo
ajmikals@buffalo.edu

1 Motivation

DNA sequencing is the process of digitizing a DNA molecule by determining the corresponding sequence of A, C, G, and T symbols, called a DNA read. While once confined to dedicated laboratories, with the introduction of *portable DNA sequencers* [9], DNA sequencing can be conducted in the field and in real-time. Since their inception, these devices have lead researchers and practitioners to envision radical new applications, including rapid medical diagnosis [10], tracking infectious diseases [8], biological threat detection [2], and bioprospecting [5].

Yet, these applications rely on a form of DNA analysis called metagenomic classification, where the goal is to identify the organism that a DNA read belongs to. This is an extremely resource intensive task that performs expensive statistical and combinatorial processing while searching for substrings in a massive reference database of known DNA sequences. This database can easily reach terabytes to petabytes in size, and because hundreds to thousands of queries are needed to classify a single DNA read, searches must be very efficient. Moreover, realistic applications demand real-time processing in unreliable network conditions, which often means that the analysis must be conducted locally on a low-energy and small-form-factor mobile device [8, 4, 7, 6].

2 Substring Index

The main challenge in mobile DNA analysis is managing the reference database, which contains the genomes of all known organisms. The key operation is *exact pattern matching*, which given a string, finds all places where that string occurs as a substring in the database. Given a DNA read to classify, metagenomic classifiers take various fragments of the read and identify which genomes contain those fragments as substrings [11, 3, 1]. Then, these matches are used to classify the DNA read via statistical and combinatorial analysis.

Even high-end mobile devices are extremely memory limited. For instance, the NVIDIA Jetson Orin Nano has only 8 GB of RAM. At the same time, the reference database can easily reach terabytes to petabytes in size. Consequently, even with sophisticated compression techniques commonly found in the literature [3, 1], it is infeasible to store the entire database in main memory. This talk will present the *compacted string B-tree* (CSBT), our external memory substring index. The CSBT performs exact pattern matching queries with the optimum I/O cost in theory while being very efficient in practice. The CSBT maintains a static B^+-tree of suffix offsets into the indexed strings along with auxiliary information, such as the lengths of matches between neighboring suffixes, using a space-efficient encoding. With this information, queries only need to match against a single substring of the database per level of the tree. Our talk will also show how we use the CSBT to enable real-time metagenomic classification on memory-limited systems.

3 Memory and I/O Management

Still, real-time DNA analysis cannot be achieved by optimized substring indexes on their own. Not only should the database and index be cached when possible, but queries should be performed concurrently to utilize the entire I/O bandwidth. Even though these are common optimizations, the nature of metagenomic classification software complicates the typical approaches. The locality found in the index is notably distinct from that in the database. Moreover, as metagenomic classifiers must be highly parallel, the I/O and parallel schedulers should be co-designed. Our talk will include our approaches.

References

[1] L. Depuydt et al. "Run-Length Compressed Metagenomic Read Classification with SMEM-finding and Tagging". In: *bioRxiv* (2025).
[2] J.L. Gardy and N.J. Loman. "Towards a Genomics-Informed, Real-Time, Global Pathogen Surveillance System". In: *Nature Reviews Genetics* 19 (2018), pp. 9–20.
[3] D. Kim et al. "Centrifuge: Rapid and Sensitive Classification of Metagenomic Sequences". In: *Genome Research* 26.12 (2016), pp. 1721–1729.
[4] S.Y. Ko, L. Sassoubre, and J. Zola. "Applications and Challenges of Real-Time Mobile DNA Analysis". In: *International Workshop on Mobile Computing Systems & Applications (HotMobile)*. 2018, pp. 1–6.
[5] A. Latorre-Pérez et al. "A Round Trip to the Desert: In Situ Nanopore Sequencing Informs Targeted Bioprospecting". In: *Frontiers in Microbiology* 12 (2021).
[6] A.J. Mikalsen and J. Zola. "Coriolis: Enabling Metagenomic Classification on Lightweight Mobile Devices". In: *Intelligent Systems for Molecular Biology (ISMB)*. 2023, pp. i66–i75.
[7] M. Oliva et al. "Portable Nanopore Analytics: Are We There Yet?" In: *Bioinformatics* 36.16 (2020), pp. 4399–4405.
[8] J. Quick et al. "Real-Time, Portable Genome Sequencing for Ebola Surveillance". In: *Nature* 530.7589 (2016), pp. 228–232.
[9] Oxford Nanopore Technologies. *MinION*. https://nanoporetech.com/products/minion. 2025.
[10] M.C. Walter et al. "MinION as Part of a Biomedical Rapidly Deployable Laboratory". In: *Journal of Biotechnology* 250 (2017), pp. 16–22.
[11] D.E. Wood and S.L. Salzberg. "Kraken: Ultrafast Metagenomic Sequence Classification Using Exact Alignments". In: *Genome Biology* 15.3 (2014), R46.