The interplay between Burrows-Wheeler Transform and Phylogenetic Compression

Contexte

The publication of the Burrows-Wheeler Transform (BWT) [1] belongs to the seminal papers in computer science. BWT is a reversible transform that permutes characters in a given string in such a way that the string becomes better compressible. BWT forms the main building block of widely used compressors such as BZIP2, and the transform has further enabled the emergence of modern fulltext indexes, such as the FM-index [2], r-index [3], or MOVI [4]. The FM index was in the center of the read mapping revolution of 2019, where programs such as BWA [5], Bowtie [6], SOAP2 [7] finally enabled DNA sequencing read alignment to human genomes on ordinary computers, making BWT one of the very impactful discoveries in bioinformatics. Over the years, BWT has been subject of extensive research, both from theoretical and practical perspective, and has been shown to also provide an important measure of sequence complexity (the r-measure) [8].

However, despite its clear utility, BWT is difficult to scale to modern bioinformatics data ranging from terabytes to petabytes. Addressing these scaling issues, a recent framework called RopeBWT3 [9] has been shown, under special circumstances, to perform a BWT at the Tbp scale; however, the key demonstration was done using a particularly repetitive bacterial genome collection previously subjected to phylogenetic compression [10], another reordering-based preprocessing technique for boosting compressibility. This resulted in a compression of 7.3 Tbp of commonly studied bacteria to only 27.6 GB. The observed performance gains may be largely explained by the fact that phylogenetic compression substantially improves BWT, possibly suggesting the following two-step protocol for genome collections: 1) genome-level reordering using phylogenetic compression, followed by 2) character-level reordering using the Burrows-Wheeler Transform. However, while reordering of sequences prior to BWT has been shown to be highly beneficial for 100bp sequencing reads [11], [12], the impact for bacterial genomes (each several Mpb long) remains unclear and proving or disproving this will require novel mathematical and computational techniques.

Objectifs

The goal of this project is to study the interplay between the Burrows-Wheeler Transform and Phylogenetic Compression. As a first objective, the student will study the impact of phylogenetically informed reordering on the compression capabilities of the BWT. The second objective of the project is to formalize phylogenetic compression as a “phylogenetic transform”, a stringological transform conceptually similar to the BWT. Overall, the project is expected to provide fundamental insights about the applicability of the BWT to modern biological data, and its results may be instrumental for software tools based on phylogenetic compression-based frameworks, and may have impacts in areas such as search engines for sequencing data and infectious disease diagnostics.

REFERENCES

[1] M. Burrows and D. J. Wheeler, “A Block-sorting Lossless Data Compression Algorithm,” 1994.
[2] P. Ferragina and G. Manzini, “Opportunistic data structures with applications,” in Proceedings 41st Annual Symposium on Foundations of Computer Science, Nov. 2000, pp. 390–398.
[3] T. Gagie, G. Navarro, and N. Prezza, “Optimal-Time Text Indexing in BWT-runs Bounded Space,” in Proceedings of the 2018 Annual ACM-SIAM Symposium on Discrete Algorithms (SODA), in Proceedings. , Society for Industrial and Applied Mathematics, 2018, pp. 1459–1477.
[4] M. Zakeri, N. K. Brown, O. Y. Ahmed, T. Gagie, and B. Langmead, “Movi: a fast and cache-efficient full-text pangenome index,” bioRxivorg, Nov. 2023, doi: 10.1101/2023.11.04.565615.
[5] H. Li and R. Durbin, “Fast and accurate short read alignment with Burrows-Wheeler transform,” Bioinformatics, p. btp324, May 2009.
[6] B. Langmead, C. Trapnell, M. Pop, and S. L. Salzberg, “Ultrafast and memory-efficient alignment of short DNA sequences to the human genome,” Genome Biol., vol. 10, no. 3, p. R25, Mar. 2009.
[7] R. Li et al., “SOAP2: an improved ultrafast tool for short read alignment,” Bioinformatics, vol. 25, no. 15, pp. 1966–1967, Aug. 2009.
[8] G. Navarro, “Indexing Highly Repetitive String Collections, Part I: Repetitiveness Measures,” ACM Comput. Surv., vol. 54, no. 2, pp. 1–31, Mar. 2021.
[9] H. Li, “BWT construction and search at the terabase scale,” arXiv [q-bio.GN], Sep. 01, 2024. [Online]. Available: http://arxiv.org/abs/2409.00613
[10] Karel Břinda et al., “Efficient and Robust Search of Microbial Genomes via Phylogenetic Compression,” bioRxiv, Apr. 2023, doi: 10.1101/2023.04.15.536996.
[11] D. Cenzato, V. Guerrini, Z. Lipták, and G. Rosone, “Computing the optimal BWT of very large string collections,” in 2023 Data Compression Conference (DCC), IEEE, Mar. 2023, pp. 71–80.
[12] D. Cenzato and Z. Lipták, “A survey of BWT variants for string collections,” Bioinformatics, vol. 40, no. 7, May 2024, doi: 10.1093/bioinformatics/btae333.