Finding MEMs with the RLBWT

Christina Boucher[1], Travis Gagie[2], Marco Oliva[1], and Massimiliano Rossi[1]

[1] CISE Department, University of Florida, US
[2] Faculty of Computer Science, Dalhousie University, Canada

In the past decades, sequencing data has been produced at an unprecedented pace, thanks to various sequencing projects [9–14]. FM-index based read-aligners [4, 6, 7] played a fundamental role in the analysis of those data. The alignment process, first build an index for one or a few genomes, then align the reads, typically, using the seed-and-extend paradigm. We first use the index to find exact matches of the reads in the reference, called seeds, then we extend the seeds using dynamic programming to find approximate matches. Maximal exact matches (MEMs) are exact matches — between the reference and the reads — that cannot be extended neither on the left nor on the right. They are proved to be effective candidates as seeds [5, 8].

Using only few genomes as reference can bias the results of the alignment, leading, for example, to medical misdiagnosis. Hence, increasing the number of reference genomes could alleviate the bias problem. The FM-index allows to index the references by storing them in a compressed space using their run-length compressed Burrows-Wheeler Transforms (RLBWTs). However, the functionalities of the FM-index to support MEM-finding are more challenging to compress.

A recent major breakthrough in text indexing is the r-index [2], which supports exact pattern matching in $\mathcal{O}(r)$ space, where r is the size of the RLBWT and n is the length of the original text. In [3] it was shown how to build the r-index efficiently for very large high-repetitive datasets, by using the prefix-free parsing (PFP). The r-index, theoretically, can be enhanced to include functionalities to support MEM-finding in $\mathcal{O}(r \log(n/r))$ space, but its design is complex and has not been implemented. In [1] the RLBWT is augmented with r thresholds such that, if we have fast random access to the text, then we can quickly compute the matching statistics of a pattern, from which we can easily find MEMs. The proposed solution is simple and promises to be practical — but they did not say how to compute their thresholds efficiently!

We show how to compute the thresholds with PFP in $\mathcal{O}(n)$ time and space bounded by the size of the PFP, which is often significantly smaller than the text, while simultaneously building the r-index. As a byproduct, we also show how to compute the longest common prefix (LCP) array, another important data structure in text indexing, in the same time and space bounds. We have implemented our algorithm and proven its practicality experimentally by building the thresholds and r-indexes on several datasets, including 1000 human chromosome 19s and 10,000 *Salmonella* genomes. Finally, we have demonstrated the utility indexing a large number of genomes comparing MEM-finding with a single-genome reference, observing that on 500 genomes we find 3.7% more sequence reads with a MEM of length greater than 50.

References

1. Bannai, H., Gagie, T., I, T.: Refining the r-index. Theor. Comput. Sci. **812**, 96–108 (2020)
2. Gagie, T., Navarro, G., Prezza, N.: Fully functional suffix trees and optimal text searching in bwt-runs bounded space. J. ACM **67**(1), 2:1–2:54 (2020)
3. Kuhnle, A., Mun, T., Boucher, C., Gagie, T., Langmead, B., Manzini, G.: Efficient construction of a complete index for pan-genomics read alignment. Journal of Computational Biology **27**(4), 500–513 (2020)
4. Langmead, B., Salzberg, S.L.: Fast gapped-read alignment with Bowtie 2. Nature Methods **9**(4), 357–359 (2012)
5. Li, H.: Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM (2013)
6. Li, H., Durbin, R.: Fast and accurate short read alignment with Burrows–Wheeler Transform. Bioinformatics **25**(14), 1754–1760 (2009)
7. Li, R., Zhu, H., Ruan, J., Qian, W., Fang, X., Shi, Z., Li, Y., Li, S., Shan, G., Kristiansen, K., Li, S., Yang, H., Wang, J., Wang, J.: De novo assembly of human genomes with massively parallel short read sequencing. Genome Research **20**(2), 265–272 (2010)
8. Miclotte, G., Heydari, M., Demeester, P., Rombauts, S., de Peer, Y.V., Audenaert, P., Fostier, J.: Jabba: hybrid error correction for long sequencing reads. Algorithms Mol. Biol. **11**, 10 (2016)
9. Obrien, S.J., Haussler, D., Ryder, O.: The birds of Genome10K. Gigascience **3**, 32 (2014)
10. The 1000 Genomes Project Consortium: A global reference for human genetic variation. Nature **526**, 68–74 (2015)
11. The 1001 Genomes Consortium: Epigenomic Diversity in a Global Collection of Arabidopsis thaliana Accessions. Cell **166**(2), 492–505 (2016)
12. The National Center for Biotechnology Information: SARS-CoV-2 (Severe acute respiratory syndrome coronavirus 2) Sequences
13. Turnbull, C., et al.: The 100,000 genomes project: bringing whole genome sequencing to the nhs. British Medical Journal **361** (2018)
14. Zhou, S., Bechner, M., Place, M., Churas, C., Pape, L., , Leong, S., Runnheim, R., Forrest, D., Goldstein, S., Livny, M.: Validation of rice genome sequence by optical mapping. BMC Genomics **8**(1), 278 (2007)