Mapping-friendly Sequence Reductions to process compressed genomic data

Roland Faure, Baptiste Hilaire, Dominique Lavenier

Abstract

Mapping-friendly Sequence Reductions to process compressed genomic data

Roland Faure[1,2]*, Baptiste Hilaire[1,] Dominique Lavenier[1]

[1]*Univ. Rennes, Inria RBA, CNRS UMR 6074, Rennes, France*
[2]*Evolution Biologique et Ecologie, Université libre de Bruxelles (ULB), Brussels, Belgium*
***Corresponding author**: roland.faure@irisa.fr*

Abstract

Efficiently managing vast DNA datasets necessitates the development of highly effective sequence compression techniques to reduce storage and computational requirements. We propose here to explore the potential of a lossy compression technique, Mapping-friendly Sequence Reductions (MSRs).

MSRs were introduced in [1] as a generalization of homopolymer compression to improve the accuracy of alignment tools. Essentially, MSRs deterministically transform sequences into shorter counterparts, in such a way that if an original query and a target sequence align, their reduced forms will align as well. While homopolymer compression is one example of an MSR, numerous others exist, potentially offering substantial sequence length reduction—such as retaining only bases between 'A' and 'T' (on average, 1 base out of 16): AACAGTGACACTAAACT → GCC. These rapid computations yield lossy representations of the originals. Notably, the reduced sequences can be stored, aligned, assembled, and indexed much like regular sequences.

MSRs could be used to improve the efficiency of taxonomic classification tools, by indexing and querying reduced sequences. Our experimentation with a toy example, a mixture of 10 *E. coli* strains, demonstrates that this approach can yield greater precision than indexing and querying a reduced portion of k-mers (typically minimizers). Specifically, using the reduction described above, 76% of reduced 31-mers were unique, whereas only 47% of not-reduced 31-mers were.

In our presentation, we will also explore other tasks that could benefit from sequence reduction, such as mapping, genome assembly, and structural variant detection.

References

[1] Blassel L, Medvedev P, Chikhi R. Mapping-friendly sequence reductions: Going beyond homopolymer compression. iScience. 2022 Oct 13;25(11):105305.