Compressed indexing of (ultra) large viral alignments

Mikael Salson, Thomas Baudeau, Arthur Boddaert, Awa Bousso Gueye, Laurent Bulteau, Yohan Hernandez-Courbevoie, Camille Marchet, Nan Pan, Sebastian Will, Yann Ponty

Abstract

SeqBIM

Compressed indexing of (ultra) large viral alignments

Mikaël Salson[1*], Thomas Baudeau[1], Arthur Boddaert[2], Awa Bousso Gueye[2], Laurent Bulteau[3], Yohan Hernandez--Courbevoie[2], Camille Marchet[1], Nan Pan[4], Sebastian Will[4], Yann Ponty[4*]

[1] *CNRS, UMR 9189 CRIStAL, Lille University*
[2] *Lille University*
[3] *CNRS, UMR 8049 LIGM, Gustave Eiffel University*
[4] *CNRS, UMR 7161 LIX, Ecole Polytechnique, Institut Polytechnique de Paris*
***Corresponding authors**: mikael.salson@univ-lille.fr; yann.ponty@lix.polytechnique.fr

Abstract

Recent outbreaks motivate a systematic collection of pathogenic genomes, including a strong focus on viruses, in order to accelerate their study and monitor the apparition/spread of variants. Owing to their limited length and temporal locality, viral genomes are usually organized within large multiple sequence alignments (MSA). Those alignments are usually horizontally compact and largely homogeneous at the level of individual columns, yet less so at a linear level due to local variations, hindering the performances of sequential compression algorithms.

The analysis of MSAs allows for the evaluation of evolutionary distances and position-specific and other higher-order statistics. In particular, in the context of viruses whose genetic material consists of single-stranded nucleic acids (ssRNA viruses), evolutionary constraints at the structural level can be revealed by covariation analysis. Such analyses motivate the analysis of the joint content of columns pairs, to assess the propensity of genomic positions to form a base pair mediated by hydrogen bonds. Ultimately, they enable a reconstruction of RNA architecture(s), potentially revealing targets for future drugs [1].

In this work, we introduce a new compressed index, called `CREMSA` (`C`olumn-wise `R`unlength `E`ncoding for `M`ultiple `S`equence `A`lignments) which both greatly reduces the storage required to store column-wise redundant MSAs. Contrasting with earlier efforts, solely focusing on the file-level compression of an MSA [2], our index enabling direct and efficient access to column-wise statistics (no decompression needed). Our index processes columns independently, and replaces each column content with a compressed bit vector [3] storing the offsets where a new nucleotide occurs (storing the new nucleotide in a separate array). Doing so, it exploits the presence of long runs of nucleotides (or gap) that are frequent in viral alignments, saving space while speeding-up queries. The data structure enables access to individual sequences of length n in $O(n)$ time, and (multiple) column-wise statistics in $O(r)$ time, where r is the total number of runs in the column(s).

We performed a preliminary analysis of a dataset of 10^6 SARS-CoV 2 genomes

from the NCBI to demonstrate the usability of the approach. The method directly reduces the disk space from 30 GB to 53 MB (99.8% deflation), an impressive feat of compression, even if compared to a baseline gzip compression (2.7GB/98% deflation; 51× larger than `CREMSA`). In principle, the compression ratio could further benefit from a custom sort procedure, minimizing the sum of Hamming distances over consecutive genomes/rows, but we unfortunately proved the associated problem to be NP-hard. As a practical tradeoff, we sorted rows according to an accepted phylogenetic tree (ordering genomes by their Pango lineage, to contiguously present genomes of a given clade), leading to a substantial improvement (30MB, a further 44% reduction). From the `CREMSA` representation, we compute the popular RNAalifold [4] conservation score between every pair of MSA columns (66M pairs of 1Mbp columns) within 7h and 100MB RAM on a desktop computer (Intel i7-12700, 64GB RAM), and found 251 pairs of columns associated with a conservation score greater than 0.5, indicating potential evolutionarily-pressured base pairs.

References

[1] Sandra Triebel, Kevin Lamkiewicz, Nancy Ontiveros, Blake Sweeney, Peter F. Stadler, Anton I. Petrov, Michael Niepmann, and Manja Marz. Comprehensive survey of conserved rna secondary structures in full-genome alignment of hepatitis c virus. *Scientific Reports*, 14(1), July 2024.

[2] Sebastian Deorowicz, Joanna Walczyszyn, and Agnieszka Debudaj-Grabysz. CoMSA: compression of protein multiple sequence alignment files. *Bioinformatics*, 35(2):227–234, 07 2018.

[3] Simon Gog, Timo Beller, Alistair Moffat, and Matthias Petri. From theory to practice: Plug and play with succinct data structures. In *13th International Symposium on Experimental Algorithms, (SEA 2014)*, pages 326–337, 2014.

[4] Stephan H Bernhart, Ivo L Hofacker, Sebastian Will, Andreas R Gruber, and Peter F Stadler. Rnaalifold: improved consensus structure prediction for rna alignments. *BMC Bioinformatics*, 9(1), November 2008.