SARS-CoV-2 Coronavirus Data Compression Benchmark

Innar Liiv

Tallinn University of Technology, 15A Akadeemia Rd, 12618 Tallinn, Estonia
innar.liiv@taltech.ee

Abstract. This paper introduces a lossless data compression competition that benchmarks solutions (computer programs) by the compressed size of the 44,981 concatenated SARS-CoV-2 sequences, with a total uncompressed size of 1,339,868,341 bytes. The data, downloaded on 13 December 2020, from the severe acute respiratory syndrome coronavirus 2 data hub of ncbi.nlm.nih.gov is presented in FASTA and 2Bit format. The aim of this competition is to encourage multidisciplinary research to find the shortest lossless description for the sequences and to demonstrate that data compression can serve as an objective and repeatable measure to align scientific breakthroughs across disciplines. The shortest description of the data is the best model; therefore, further reducing the size of this description requires a fundamental understanding of the underlying context and data. This paper presents preliminary results with multiple well-known compression algorithms for baseline measurements, and insights regarding promising research avenues. The competition’s progress will be reported at https://coronavirus.innar.com and the benchmark is open for all to participate and contribute.

Keywords: lossless data compression · benchmark · SARS-CoV-2

1 Motivation

Marvin Minsky considered Kolmogorov, Chaitin, and Solomonoff’s algorithmic information theory “the most important discovery since Gödel” and conjectured that “practical approximations to [their theory]... would make better predictions than anything we have today” [21].

This competition is intended to encourage multidisciplinary research, in the spirit of Kolmogorov, Chaitin, and Solomonoff’s theory, to develop the shortest lossless description for the sequences of SARS-CoV-2. A successful result will serve as a demonstration that data compression can offer an objective and repeatable measure to align scientific breakthroughs across disciplines. The shortest description of a dataset is the best model. Further compression of the sequences of SARS-CoV-2 will require a fundamental understanding of the data and its context.
2 Related Work

The main theoretical underpinnings of this benchmark are the practical approximations to Kolmogorov–Chaitin–Solomonoff complexity by Li and Vitanyi [16] and the minimum description length principle [24,9].

Kolmogorov complexity is the length of the shortest effective description of an object [14]. Therefore, the idealistic goal of the SARS-CoV-2 coronavirus data compression benchmark is to find the Kolmogorov complexity of SARS-CoV-2 sequences. Since doing so requires an infinite amount of work, as a practical approximation, the smallest archive plus the decompressor is considered a computable proxy. Matt Mahoney has written an inspiring and excellent rationale for a large text compression benchmark [19] with an extended discussion about the connections between intelligence and compression.

Several lossless compression benchmarks have been proposed over the years [3,11], the most well-known by Marcus Hutter, who offered €500,000 as a challenge prize [14]. The compression of genetic sequences, as a specific niche of data compression research, has been a popular topic for more than 25 years [7,8]. An interested reader is referred to two comprehensive surveys about data compression methods for biological sequences [2,10]. Kryukov et al. have recently presented a comprehensive evaluation of reference-free compressors for FASTA-formatted sequences [15] and developed a sequence compression benchmark database.

De Maio et al. have identified several oddities specific to SARS-CoV-2 sequencing data, which may “arise from specific combinations of sample preparation, sequencing technology, and consensus calling approaches” [6]. Such aspects, and other systematic errors typical to sequence data [20], can support the design of a specific compression strategy.

3 The Task

Losslessly compress the 1.25GB file coronavirus.fasta [17] or its 2bit representation equivalent coronavirus.2bit (0.31 GB) [17] to less than 1,238,330 bytes (the current smallest compressed size of the dataset, including the decompressor).

4 The Data

The data is presented in FASTA and 2Bit (UCSC-twobit [12]) format, consisting of 44,981 concatenated SARS-CoV-2 sequences with a total uncompressed size of 1,339,868,341 bytes [17]. It was downloaded on 13 December 2020 from the severe acute respiratory syndrome coronavirus 2 data hub of ncbi.nlm.nih.gov [23]. Each participant can choose which file to use—that is, the compressor does not have to work on both datasets.
5 Setting the Scene

The challenge is to compress 44,981 concatenated SARS-CoV-2 sequences. To provide a slightly simpler example, more susceptible to manual observation, the compression results for one sequence (reference sequence NC\textunderscore 045512 \[29\]) are presented in Table 1.

| Bytes | File     | Format | Compressor | Parameters |
|-------|----------|--------|------------|------------|
| 7233  | NC\_045512 | FASTA  | cmix \[13\] |            |
| 7277  | NC\_045512 | FASTA  | paq8l \[18\] | -8         |
| 7277  | NC\_045512 | FASTA  | GeCo3 \[28\] | -l 1 \-lr 0.06 \-hs 8 |
| 7308  | NC\_045512 | 2Bit   | cmix \[13\] |            |
| 7337  | NC\_045512 | 2Bit   | brotli \[1\] | -q 10      |
| 7346  | NC\_045512 | 2Bit   | paq8l \[18\] | -8         |
| 7355  | NC\_045512 | 2Bit   | zstd \[4\] | -19        |
| 7369  | NC\_045512 | 2Bit   | bcm \[22\] | -9         |
| 7376  | NC\_045512 | 2Bit   | gzip \[26\] | -9         |
| 7508  | NC\_045512 | 2Bit   | xz \[5\] | -9         |
| 7517  | NC\_045512 | 2Bit   | zip \[26\] | -9         |
| 7524  | NC\_045512 | 2Bit   | Uncompressed |          |
| 7545  | NC\_045512 | 2Bit   | rar \[25\] | m5         |
| 7802  | NC\_045512 | FASTA  | bcm \[22\] | -9         |
| 7868  | NC\_045512 | 2Bit   | bzip2 \[27\] | -9         |
| 8399  | NC\_045512 | FASTA  | brotli \[1\] | -q 11      |
| 8519  | NC\_045512 | FASTA  | zstd \[4\] | -19        |
| 8801  | NC\_045512 | FASTA  | bzip2 \[27\] | -9         |
| 9000  | NC\_045512 | FASTA  | xz \[5\] | -9         |
| 9598  | NC\_045512 | FASTA  | gzip \[26\] | -9         |
| 9623  | NC\_045512 | FASTA  | rar \[25\] | m5         |
| 9738  | NC\_045512 | FASTA  | zip \[26\] | -9         |
| 30416 | NC\_045512 | FASTA  | Uncompressed |          |
6 Preliminary Results

Table 2 presents the current compression results for 44,981 SARS-CoV-2 sequences (with a total uncompressed size of 1,339,868,341 bytes) sorted by the number of bytes (with fewer bytes meaning better compressibility), acting as the baseline measurement for the challenge. The bytes column in Table 2 did not include the size of the decompressor, which will be considered in the final benchmark. Considering the total size of the compressed archive and the decompressor (instead of just considering the compressed archive), the PAQ8L compressor [18] by Matt Mahoney performed the best, with the best results achieved using the 2Bit format of the dataset. The resulting compressed archive for PAQ8L, including the compressed decompression executable, has a total size of 1,238,330 (1,207,839+30,491) bytes. The CMIX compressor [13] by Byron Knoll resulted in a smaller compressed archive (988,958), but the total size, including the compressed decompressor, is 1,282,852 (988,958+293,894).

Table 2. Compression results for 44,981 SARS-CoV-2 sequences (fewer bytes is better)

| Bytes        | File       | Format | Compressor | Parameters |
|--------------|------------|--------|------------|------------|
| 988,958      | Coronavirus| 2Bit   | cmix [13]  |            |
| 1,207,839    | Coronavirus| 2Bit   | paq8l [18] | -8         |
| 1,425,590    | Coronavirus| FASTA  | paq8l [18] | -8         |
| 1,985,384    | Coronavirus| 2Bit   | xz         | -9         |
| 2,022,796    | Coronavirus| FASTA  | xz         | -9         |
| 2,043,140    | Coronavirus| 2Bit   | bcm [22]   | -9         |
| 2,044,664    | Coronavirus| 2Bit   | rar [25]   | m5         |
| 2,050,285    | Coronavirus| FASTA  | GeCo3 [28]| -l 1 -lr 0.06 -hs 8 |
| 2,367,487    | Coronavirus| 2Bit   | zstd [4]   | -19        |
| 2,728,490    | Coronavirus| FASTA  | bcm [22]   | -9         |
| 2,871,864    | Coronavirus| 2Bit   | brotli [1]| -q 10      |
| 2,871,864    | Coronavirus| FASTA  | brotli [1]| -q 10      |
| 4,217,341    | Coronavirus| FASTA  | zstd [4]   | -19        |
| 5,924,805    | Coronavirus| FASTA  | rar [25]   | m5         |
| 67,575,178   | Coronavirus| 2Bit   | gzip [24]  | -9         |
| 67,575,325   | Coronavirus| 2Bit   | zip [26]   | -9         |
| 75,530,790   | Coronavirus| 2Bit   | bzip2 [27]| -9         |
| 75,530,790   | Coronavirus| FASTA  | bzip2 [27]| -9         |
| 77,356,405   | Coronavirus| FASTA  | gzip [26]  | -9         |
| 77,356,550   | Coronavirus| FASTA  | zip [26]   | -9         |
| 332,133,731  | Coronavirus| 2Bit   | Uncompressed|          |
| 1,339,868,341| Coronavirus| FASTA  | Uncompressed|          |
7 Conclusions

The sequences of the SARS-CoV-2 coronavirus are compressible. Further compression will require a mix of novel and creative approaches: moving beyond the state of the art of data compression or understanding the patterns and relationships within parts of sequences and between sequences.

The SARS-CoV-2 coronavirus data compression benchmark has a vital multidisciplinary aspect: the objective and repeatable measure in this challenge can help to align scientific breakthroughs across disciplines. At the end of the day, different theories and models to understand the coronavirus are measurable through the shortest description of the dataset.

In addition, the scientific momentum around and attention paid to SARS-CoV-2 can be applied to support breakthroughs by the data compression community and advance the state of the art of compression. The techniques used for improving the compression of SARS-CoV-2 datasets can feed back to better understanding the underlying mechanisms of the coronavirus.

References

1. Alakuijala, J., Farruggia, A., Ferragina, P., Kliuchnikov, E., Obryk, R., Szabadka, Z., Vandevenne, L.: Brotli: A general-purpose data compressor. ACM Transactions on Information Systems (TOIS) 37(1), 1–30 (2018)
2. Bonfield, J.K., Mahoney, M.V.: Compression of fastq and sam format sequencing data. PloS one 8(3), e59190 (2013)
3. Broukhis, L.: The human knowledge compression prize. URL http://mailcom.com/challenge/ (1996)
4. Collet, Y., Kucherawy, M.: Zstandard compression and the application/zstd media type. RFC 8478 (2018)
5. Collin, L.: Xz utils. URL http://tukaani.org/xz/ (2020)
6. De Maio, N., Walker, C., Borges, R., Weilguny, L., Sledkowicz, G., Goldman, N.: Issues with sars-cov-2 sequencing data. virological (2020)
7. Grumbach, S., Tahi, F.: Compression of dna sequences. In: [Proceedings] DCC93: Data Compression Conference. pp. 340–350. IEEE (1993)
8. Grumbach, S., Tahi, F.: A new challenge for compression algorithms: genetic sequences. Information Processing & Management 30(6), 875–886 (1994)
9. Grünewald, P.D.: The minimum description length principle. MIT press (2007)
10. Hosseini, M., Pratas, D., Pinho, A.J.: A survey on data compression methods for biological sequences. Information 7(4), 56 (2016)
11. Hutter, M.: The human knowledge compression prize. URL http://prize.hutter1.net (2006)
12. Karolchik, D., Baertsch, R., Diekhans, M., Furey, T.S., Hinrichs, A., Lu, Y., Roskin, K.M., Schwartz, M., Sugnet, C.W., Thomas, D.J., et al.: The ucsc genome browser database. Nucleic acids research 31(1), 51–54 (2003)
13. Knoll, B.: Cmix data compression program. URL https://www.byronknoll.com/cmix.html (2019)
14. Kolmogorov, A.N.: Three approaches to the quantitative definition of information. Problems of information transmission 1(1), 1–7 (1965)
15. Kryukov, K., Ueda, M.T., Nakagawa, S., Imanishi, T.: Sequence compression benchmark (scb) database—a comprehensive evaluation of reference-free compressors for fasta-formatted sequences. GigaScience 9(7), gia072 (2020)
16. Li, M., Vitányi, P.: An introduction to Kolmogorov complexity and its applications. Springer (1997)
17. Liiv, I.: Sars-cov-2 coronavirus data compression benchmark. URL https://coronavirus.innar.com (2020)
18. Mahoney, M.: Paq8 data compression program. URL http://mattmahoney.net/dc/ (2007)
19. Mahoney, M.: Rationale for a large text compression benchmark. URL http://mattmahoney.net/dc/rationale.html (2009)
20. Meacham, F., Boffelli, D., Dhabhi, J., Martin, D.L., Singer, M., Pachter, L.: Identification and correction of systematic error in high-throughput sequence data. BMC bioinformatics 12(1), 451 (2011)
21. Minsky, M.: Panel: The limits of understanding. URL https://www.youtube.com/watch?v=DfY-DRsE86s&t=1h30m02s (2014)
22. Muravyov, I.: Bcm data compression program. URL https://github.com/encode84/bcm (2020)
23. NCBI: Severe acute respiratory syndrome coronavirus 2 data hub. URL https://www.ncbi.nlm.nih.gov/labs/virus/vssi/#/virus?Seq-Type[]=Nucleotide&VirusLineage[]=SARS-CoV-2,%20taxid:2697049 (2020)
24. Rissanen, J.: Modeling by shortest data description. Automatica 14(5), 465–471 (1978)
25. Roshal, A.: Rar data compression program. URL https://www.rarlab.com/ (2020)
26. Salomon, D.: Data compression: the complete reference. Springer Science & Business Media (2004)
27. Seward, J.: Bzip2 data compression program. URL https://www.sourceforge.org/bzip2/ (2020)
28. Silva, M., Pratas, D., Pinho, A.J.: Efficient dna sequence compression with neural networks. GigaScience 9(11), gia119 (2020)
29. Wu, F., Zhao, S., Yu, B., Chen, Y.M., Wang, W., Song, Z.G., Hu, Y., Tao, Z.W., Tian, J.H., Pei, Y.Y., et al.: A new coronavirus associated with human respiratory disease in china. Nature 579(7798), 265–269 (2020)