Proof of concept Zstandard user-provided dictionary compression for FASTA files

Michael Persico[1]

1 Department of Biology, Concordia University, Montreal, Quebec, Canada,

☯These authors contributed equally to this work.

¤Current Address: Dept/Program/Center, Institution Name, City, State, Country

†Deceased

¶Membership list can be found in the Acknowledgments sections

Abstract

Background

Zstandard (Zstd) represents a lossless data compression mechanism that is highly configurable and is aimed at coupling high compression ratios with fast compression/decompression performance. Previous studies have paired specific Zstd configurations with various file formats in bioinformatics to reduce total data volume. This paper presents a "training mode" pipeline, written in the Julia programming language, wherein a custom compression dictionary is generated from a sample FASTA set in order to explore further compression improvements and compare them to the compression performance of Xz, Zlib, Bzip2, and Lz4 compressors.

Results

Conclusions

Introduction

The storage of biological data has represented a significant topic of research, with a number of challenges presented over subsequent generations of technological development, with current challenges discovered as the volume and complexity of data continues to increase[1, 2].

Materials and methods

A list of direct and indirect dependencies can be found in the repository's Manifest.toml file.

Results

Discussion

Conclusion

Supporting information

S1 Fig. Bold the title sentence. Add descriptive text after the title of the item (optional).

S2 Fig. Lorem ipsum. Maecenas convallis mauris sit amet sem ultrices gravida. Etiam eget sapien nibh. Sed ac ipsum eget enim egestas ullamcorper nec euismod ligula. Curabitur fringilla pulvinar lectus consectetur pellentesque.

S1 File. Lorem ipsum.

S1 Video. Lorem ipsum. Maecenas convallis mauris sit amet sem ultrices gravida. Etiam eget sapien nibh. Sed ac ipsum eget enim egestas ullamcorper nec euismod ligula. Curabitur fringilla pulvinar lectus consectetur pellentesque.

S1 Appendix. Lorem ipsum. Maecenas convallis mauris sit amet sem ultrices gravida. Etiam eget sapien nibh. Sed ac ipsum eget enim egestas ullamcorper nec euismod ligula. Curabitur fringilla pulvinar lectus consectetur pellentesque.

S1 Table. Lorem ipsum. Maecenas convallis mauris sit amet sem ultrices gravida. Etiam eget sapien nibh. Sed ac ipsum eget enim egestas ullamcorper nec euismod ligula. Curabitur fringilla pulvinar lectus consectetur pellentesque.

Acknowledgments

I would like to express my sincere gratitude to Professor David Walsh for his advice that helped shape the research as it progressed; My friends and family that supported me throughout my studies; The Julia community for their support before and throughout the writing of this paper; The Quarto developers for producing the Quarto publishing system[3] along with the PLOS template used for this paper; Luca Di Maio and contributors to the Distrobox tool for the ease of setting up the containerized development environments[4]; Maximiliano Sandoval and contributors to the Citations app for managing the paper's bibliography[5], and all other persons that had indirectly assisted through their programs and research.

References

1. D'Argenio V. The high-throughput analyses era: are we ready for the data struggle? High-throughput. 2018;7(1):8.
2. Li Y, Chen L. Big biological data: challenges and opportunities. Genomics, proteomics & bioinformatics. 2014;12(5):187.
3. Allaire JJ, Teague C, Scheidegger C, Xie Y, Dervieux C. Quarto; 2022. Available from: https://github.com/quarto-dev/quarto-cli.

4. Di Maio L. Distrobox;. Available from: https://github.com/89luca89/distrobox.
5. Sandoval M. Citations – apps for Gnome;. Available from: https://apps.gnome.org/app/org.gnome.World.Citations/.