Abstract

SeqBIM

Towards space-efficient data structures for large genome-distance matrices with quick retrieval

Léo Ackermann[1]*, Pierre Peterlongo[1], Karel Břinda[1]

[1] *Univ. Rennes, Inria, CNRS, IRISA - UMR 6074, Rennes, F-35000 France*
***Corresponding author**: leo.ackermann@inria.fr

Abstract

The explosion of genomic data presents a significant computational challenge for all downstream analyses. Particularly challenging are microbial collections, which already encompass several millions of genomes (e.g., AllTheBacteria [1]: 2.4M bacterial isolates; GISAID [2]: 15M SARS-CoV-2 sequences) and are further growing exponentially [3]. Many standard bioinformatics analyses lean on pairwise distances between genomes, including phylogeny inference [4, 5, 6] and genome clustering [7]. As a result, the scalability of multiple downstream analyses relies on the efficient computation and storage of pairwise distance matrices.

However, while the efficient computation of distances have been addressed by modern sketching-based methods such as Mash [8] and successors (see [9] for a survey), the storage and indexing of the resulting matrices remain a significant challenge. In fact, due to their quadratic size in the number of genomes, these matrices already surpass most storage capacities (e.g., 24 TB required for AllTheBacteria) and are thus heavily truncated when stored (e.g., [10]). This calls for a dedicated data structure that would be space-efficient and support near-constant-time distance retrieval queries. However, despite all the recent advances in compression of restricted families of matrices, such as sparse or low-rank ones [11, 12, 13], to the best of our knowledge, no scalable method is currently available for large genome-distance matrices.

In this presentation, we will discuss our ongoing work on subquadratic compression of distance matrices of large bacterial genome collections. Building upon insights from [14, 3], our approach takes advantage of the peculiar structure of those collections, that can extensively be explained by their underlying phylogeny. As a first step, we focus on phylogenetic trees as a candidate backbone for a compact data structure for pairwise distances. During this talk, we will show that collections of genomes following the infinite-site model can be stored in linear space supporting constant-time queries. We will then demonstrate our preliminary results on the tradeoffs that exist between metric distortion and storing cost in practical use cases. Overall, our work draws a path towards practical data structures that would be applicable to collections of millions of genomes with only negligible distance data distortion.

Keywords. Distance matrices, Compression, Phylogeny, Compact data structures

References

[1] Martin Hunt, Leandro Lima, Wei Shen, John Lees, and Zamin Iqbal. Allthebacteria - all bacterial genomes assembled, available and searchable. March 2024.

[2] Stefan Elbe and Gemma Buckland-Merrett. Data, disease and diplomacy: Gisaid's innovative contribution to global health. *Global Challenges*, 1(1):33–46, January 2017.

[3] Karel Břinda, Leandro Lima, Simone Pignotti, Natalia Quinones-Olvera, Kamil Salikhov, Rayan Chikhi, Gregory Kucherov, Zamin Iqbal, and Michael Baym. Efficient and robust search of microbial genomes via phylogenetic compression. April 2023.

[4] R.R. Sokal, C.D. Michener, and University of Kansas. *A Statistical Method for Evaluating Systematic Relationships*. University of Kansas science bulletin. University of Kansas, 1958.

[5] N Saitou and M Nei. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution*, July 1987.

[6] D. Bryant. Neighbor-net: An agglomerative method for the construction of phylogenetic networks. *Molecular Biology and Evolution*, 21(2):255–265, August 2003.

[7] Roberto Tagliaferri, Alberto Bertoni, Francesco Iorio, Gennaro Miele, Francesco Napolitano, Giancarlo Raiconi, and Giorgio Valentini. A review on clustering and visualization methodologies for genomic data analysis. 2007.

[8] Brian D. Ondov, Todd J. Treangen, Páll Melsted, Adam B. Mallonee, Nicholas H. Bergman, Sergey Koren, and Adam M. Phillippy. Mash: fast genome and metagenome distance estimation using minhash. *Genome Biology*, 17(1), June 2016.

[9] Guillaume Marçais, Brad Solomon, Rob Patro, and Carl Kingsford. Sketching and sublinear data structures in genomics. *Annual Review of Biomedical Data Science*, 2(1):93–118, July 2019.

[10] John A. Lees, Gerry Tonkin-Hill, Zhirong Yang, and Jukka Corander. Mandrake: visualizing microbial population structure by embedding millions of genomes into a low-dimensional representation. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 377(1861), August 2022.

[11] H. Cheng, Z. Gimbutas, P. G. Martinsson, and V. Rokhlin. On the compression of low rank matrices. *SIAM Journal on Scientific Computing*, 26(4):1389–1404, January 2005.

[12] Urban Borštnik, Joost VandeVondele, Valéry Weber, and Jürg Hutter. Sparse matrix multiplication: The distributed block-compressed sparse row library. *Parallel Computing*, 40(5):47–58, 2014.

[13] Jeremiah Willcock and Andrew Lumsdaine. Accelerating sparse matrix computations via data compression. In *Proceedings of the 20th Annual International Conference on Supercomputing*, ICS '06, page 307–316, New York, NY, USA, 2006. Association for Computing Machinery.

[14] Po-Ru Loh, Michael Baym, and Bonnie Berger. Compressive genomics. *Nature Biotechnology*, 30(7):627–630, July 2012.