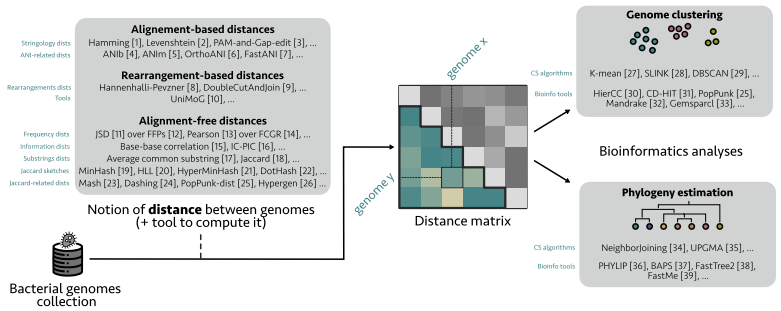
Subquadradic Storage of Pairwise Distance Matrices of Large Microbial Genome Collections

Léo Ackermann[※], Pierre Peterlongo[※], Karel Břinda[※]

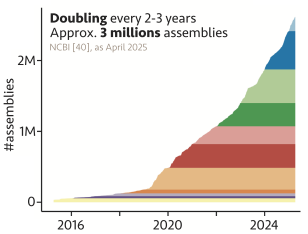
[※]Univ. Rennes, Inria, CNRS, IRISA - UMR 6074, Rennes, F-350

C.1 Distance matrices underlie major bioinformatics analyses



C.2 Their storage has become infeasible

Exponential growth of microbial data



Distance matrices grow quadratically as fast

3M assemblies ⇒ 4.5×10^{12} distances ⇒ **18 TB**

Limits of the state-of-the-art

On-the-fly (re)computation (e.g. [41])
- **Do not store** any of the computed distances
- It induces **redundant computations**
- Limited to **sketched distances**

Application-driven simplification
- **Application specific**
- E.g. [32] only stores the distances to the 10 nearest neighbors, which is sufficient for the subsequent clustering task

Main challenge

Dimensionality loss

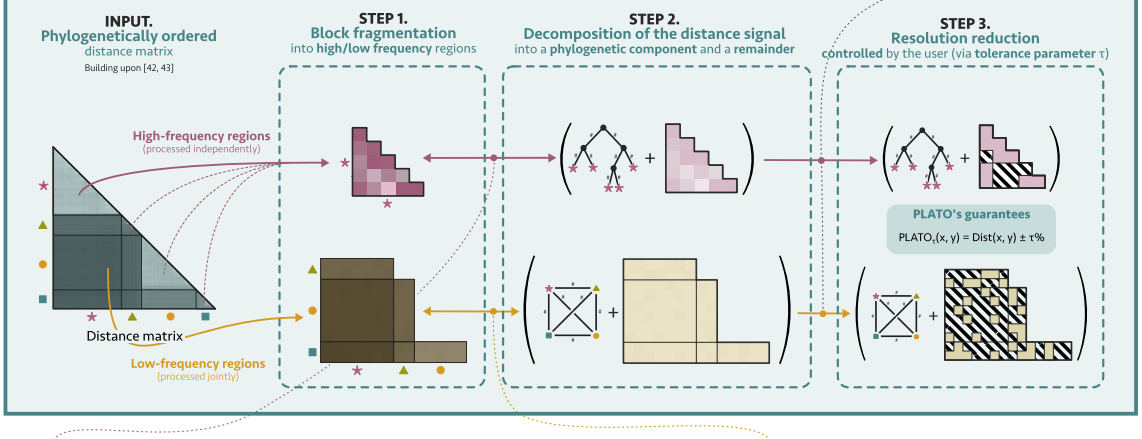
As scalar information, **distances tear down** the multidimensional genome comparison **structure** (e.g. alignment).

PLATO: Compression of microbial distance matrices with constant-time random access and controlled loss

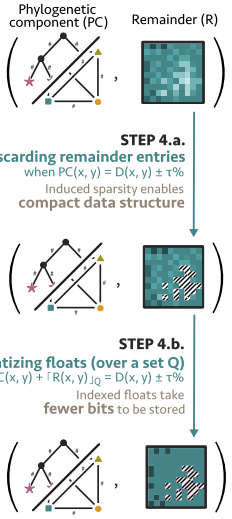
M.1 Key ideas

A **significant part** of the distance signal is **explained by** restricted classes of **phylogenies** that can be **efficiently estimated and stored** (steps 1 and 2)
Resolution needed by comparison-based downstream analyses is **relative**, allowing **efficient compression of the remaining part** of the signal (step 3)

M.2 Workflow overview



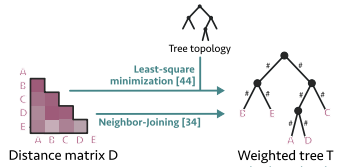
M.3 Remainder compression



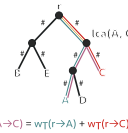
M.4 Estimation and storage of the tree phylogeny component for high-frequency regions

Rationale. In high-frequency regions, where **genomes are closely related**, the underlying **phylogeny is dominated by vertical descent**

Tree phylogeny component inference



Retrieving distance between leaves in O(1)-time



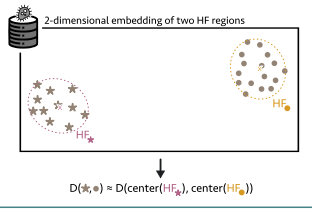
Observation. $w_T(A \to C) = w_T(r \to A) + w_T(r \to C) - 2 \cdot w_T(r \to \mathrm{lca}(A, C))$

- Store $w_T(r \to \cdot)$ at the level of each node
- Precompute a RMQ datastructure to answer lca queries

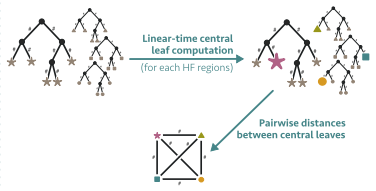
→ O(#leaves) data structure with constant-time interleaf-distance retrieval

M.5 Estimation and storage of the complete-graph phylogeny component for low-frequency regions

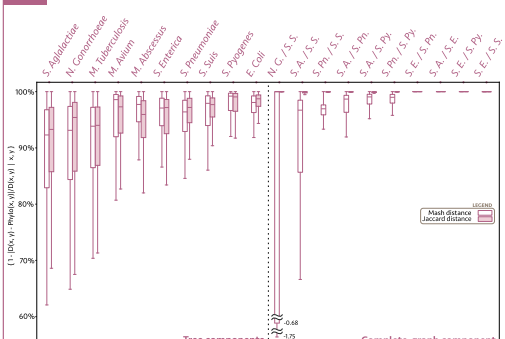
Rationale. In low-frequency regions, the distance signal **variations** are **orders of magnitude smaller** than the average **amplitude**



Complete-graph phylogeny component inference

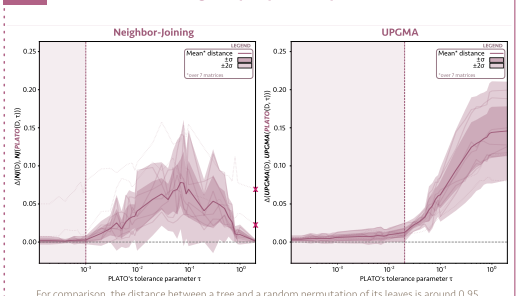


R.1 Phylogenetic components capture most of the distance signal



Details. Genomes used were 1k-large collections. Distance matrices were computed with Mash (k=21, s=10k), and then expressed as the corresponding Jaccard distances. Phylogenetic components (trees or complete graph) were computed using PLATO with τ=0. Box plots show extremal values, as well as first quartiles, medians, and third quartiles.

R.2 Tree inference is marginally impacted by small relative errors

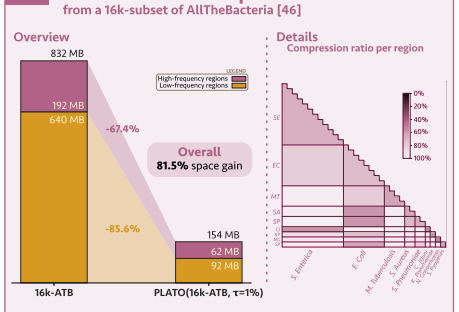


For comparison, the distance between a tree and a random permutation of its leaves is around 0.95.

- For tolerance $\tau \leq 0.1\%$, the trees inferred by **NeighborJoining** from PLATO(D, τ) and D are almost the same
- For tolerance $\tau \leq 1\%$, the trees inferred by **UPGMA** from PLATO(D, τ) and D are almost the same

Details. Genomes used were 1k-large collections of: E. Coli, M. Abscessus, M. Avium, M. Tuberculosis, N. Gonorrhoeae, S. Enterica, S. Pneumoniae, S. Pyogenes, S. Suis. Distance matrices were computed with Mash (k=21, s=10k). Tree inference was performed with FastMe. Tree distance is the complement of the Mutual Clustering Information metric [45], computed with TreeDist's ClusteringInfoDistance function. Distances for each collection are shown in grey. Mean statistics and standard deviation intervals are shown in purple (E. Coli and S. Pyogenes collections were excluded because of ambiguous topology, caused by 0-distances).

R.3 APPLICATION: Compression of distances from a 16k-subset of AllTheBacteria [46]



Details. The collection is made of multiple species: S. Enterica (6k genomes), E. Coli (4k), M. Tuberculosis (2k), S. Aureus (1k), S. Pneumoniae (1k), C. Jejuni (500), K. Pneumoniae (500), N. Gonorrhoeae (500), S. Pyogenes (500). The distribution of genomes is inspired by the distribution of genomes in the AllTheBacteria collection. Distance matrices were computed with Mash (k=21, s=10k). PLATO was run with τ=1%, and input data were clustered into regions using "species" metadata.

Conclusion

We introduce **PLATO**, a method to **compress any distance** matrix up to a user-chosen **relative precision** while keeping **constant-time random access to entries**.

PLATO's core idea is to **estimate** and **store phylogenies** from which the distances can be recovered, **instead of the distances themselves**.

We use it to **reduce by 81.5% the storage** needed by a typical small collection of genomes, as a proof of concept.

Perspectives

Full characterization of the compression capability of the workflow, e.g. as a function of the diameter (CS) or the mutation rate (Bio) of the region.

Scaling to larger microbial collections (e.g. 661k [47]). This requires new subsampling and clustering strategies to stay efficient, e.g. during the tree phylogeny inference step.

Acknowledgments

The authors thank Pavel Veselý for insightful discussions.
This work is supported by the ANR project REALL (ANR-24-CE45-1226).

References

1. Hamming, Bell System Technical Journal, 1950 2. Levenshtein, Soviet Physics Doklady, 1966 3. Needleman *et al.*, Journal of Molecular Biology, 1970 4. Goris *et al.*, IJSEM, 2007 5. Richter *et al.*, PNAS, 2009 6. Lee *et al.*, IJSEM, 2016 7. Jain *et al.*, Nature Communications, 2018 8. Hannenhalli *et al.*, FOCS, 1995 9. Friedberg *et al.*, Bioinformatics, 2008 10. Hilker *et al.*, Bioinformatics, 2012 11. Lin, IEEE Transactions on Information theory, 1991 12. Sims *et al.*, PNAS, 2011 13. Pearson, Proceedings of the Royal Society of London, 1895 14. Jeffrey, Nucleic Acids Research, 1990 15. Liu *et al.*, BBRC, 2008 16. Gao *et al.*, Gene, 2012 17. Ulitsky *et al.*, Journal of Computational Biology, 2006 18. Jaccard, Bulletin de la Société vaudoise des sciences naturelles, 1901 19. Broder, Compression and Complexity of Sequences, 1997 20. Fajolet *et al.*, Discrete Mathematics and Theoretical Computer Science, 2007 21. Yu *et al.*, IEEE Transactions on Knowledge and Data Engineering , 2022 22. Nunes *et al.*, Knowledge Discovery and Data Mining, 2023 23. Ondov *et al.*, Genome Biology, 2016 24. Baker *et al.*, Genome Biology, 2019 25. Lees *et al.* Genome Research, 2019 26. Xu *et al.*, Bioinformatics, 2024 27. MacQueen, Symposium on Mathematical Statistics and Probability, 1967 28. Sibson, The Computer Journal, 1973 29. Ester *et al.*, Knowledge Discovery and Data Mining, 1996 30. Zhou *et al.*, Bioinformatics, 2021 31. Fu *et al.*, Bioinformatics, 2012 32. Lees *et al.*, Philos Trans R Soc Lond B Biol Sci, 2022 33. von Wachsmann *et al.*,Preprint, 2025 34. Saitou *et al.*, Molecular Biology and Evolution, 1987 35. Sokal *et al.*,University of Kansas Science Bulletin, 1958 Felsenstein 36. Felsenstein, Journal of Molecular Evolution, 1981 37. Corander *et al.*, Bulletin of Mathematical Biology, 2007 38. Price *et al.*, PLoS ONE, 2010 39. Lefort *et al.*, Molecular Biology and Evolution, 2015 40. Goldfarb *et al.*, Nucleic Acids Research, 2025 41. Ingels *et al.*, RECOMB, 2026 42. Břinda *et al.*, Nature methods, 2025 43. Hendrychová, MsC Thesis, 2025 44. Legendre, 1805 45. Smith, Bioinformatics, 2020 46. Hunt *et al.*, Preprint, 2025 47. Blackwell *et al.*, PLOS Biology, 2021.

– This poster was presented during the RECOMB and RECOMB-Seq conferences | Thessaloniki, Greece | May 24-29 2026 –