Published in final edited form as:

Nat Methods. 2018 November ; 15(11): 847–848. doi:10.1038/s41592-018-0187-8.

Striped UniFrac: enabling microbiome analysis at unprecedented scale.

Daniel McDonald¹, Yoshiki Vázquez-Baeza¹, David Koslicki², Jason McClelland², Nicolai Reeve¹, Zhenjiang Xu¹, Antonio Gonzalez¹, Rob Knight^{1,3,4,5}

¹Department of Pediatrics; University of California San Diego; La Jolla; California; USA

²Pediatrics; University of California

³Department of Computer Science and Engineering; University of California San Diego; La Jolla; California; USA

⁴Center for Microbiome Innovation; University of California San Diego; La Jolla; California; USA

⁵Department of Bioengineering; University of California San Diego; La Jolla; California; USA

To the editor:

UniFrac is a metric frequently used in microbiome research, but which does not scale to today's large datasets. We propose a new algorithm, Striped UniFrac, which produces identical results to prior algorithms, but with dramatically reduced memory and compute requirements. We highlight its utility by computing UniFrac on 113,721 samples in 48 hours using 256 CPUs. A BSD-licensed implementation is available that produces a C shared library linkable by any programming language (Supplementary Software and <https://github.com/biocore/unifrac>).

UniFrac¹ is a phylogenetic distance metric used to compare pairs of microbiome profiles. Microbiome studies now span tens of thousands of samples, such as the 27,751 sample Earth Microbiome Project (EMP)² or the 15,096 sample American Gut Project³. Existing algorithms for computing UniFrac cannot scale in time or space to these study designs. For example, using Fast UniFrac with the EMP was projected to take months. Striped UniFrac produces identical results to existing algorithms, exhibits greater than 30-fold improvement in single-threaded performance, shows near linear parallel scaling (Supplementary fig. 1A,B), and can process the EMP dataset on a laptop in under 24 hours. It enables new biological insights, as shown by a meta-analysis³ of the American Gut and Earth

Corresponding author: Rob Knight (robknight@ucsd.edu).

Present addresses: Nicolai Reeve is now with: Biota Technology Inc.; La Jolla; California; USA

Contributions

D.M. designed Striped UniFrac, planned the study, analyzed data and wrote the manuscript. Y.V.B. integrated Striped UniFrac with QIIME 2 and contributed to the manuscript. D.K. contributed to the proof. J.M. contributed to the proof. N.R. contributed language interface code. Z.X. contributed to the manuscript. A.G. integrated Striped UniFrac with Qiita. R.K. planned the study and wrote the manuscript.

Completing Financial Interests Statement

No, I declare that the authors have no competing interests as defined by Nature Research, or other interests that might be perceived to influence the results and/or discussion reported in this paper.

Microbiome Projects. We highlight the algorithm here by computing UniFrac on 113,721 public samples in Qiita ⁴ in under 48 hours using 256 CPUs (interactive plot available at <https://bit.ly/2LHMDFC>).

The key insights with Striped UniFrac are improved space complexity by aggregating metric constituents in post-order traversal, and rotating proportion vectors for pairwise comparisons (see Supplementary Note for pseudo-code and complexity analysis). The post-order aggregation removes the dominant scaling factor for space complexity in Fast UniFrac (Supplementary fig. 1C,D). Vector rotation (expressed by embedding; fig. 1A) allows compilers to utilize single instruction multiple data (SIMD) operations. As a consequence of rotation, pairwise distances are computed along diagonals of the distance matrix (fig. 1B) resulting in more cache utilization, task-level parallelism, and hardware-level prefetch. We also introduce an optional heuristic that reduces compute by 50% by ignoring tips of the phylogeny (correlations with exact calculations, fig. S1E). Empirical scaling results show Striped UniFrac outperforms popular implementations of UniFrac (time fig. 1C; space fig. 1D); detail on the benchmark and algorithm can be found in Supplemental Notes.

Code Availability

An optimized implementation of Striped UniFrac is available on GitHub (<https://github.com/biocore/unifrac>) under the BSD-license, is part of QIIME 2 ⁵ and is the UniFrac implementation enabling meta-analyses in Qiita ⁴ (see Supplementary Note for implementation details). The source code builds a command line version, a C shared library, and a Python API. All known variants of UniFrac are implemented ^{1, 6, 7}.

Reporting Summary

Further information on the experimental design is available in the Nature Research Reporting Summary linked to this article.

Data Availability

The datasets analyzed during the current study are available in the Qiita repository with the specific study accessions in supplementary dataset 1, and were extracted using Qiita's redbiom interface.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

D.M. was supported by NSF DBI-1565100, Alfred P Sloan Foundation G-2017-9838 and ONR N00014-15-1-2809. Y.V.B. was supported by NSF DBI-1565100 and Alfred P Sloan Foundation G-2017-9838. D.K and J.M. were supported by NSF Award 1664803. Z.X. was supported by NSF DBI-1565100. A.G. was supported by NIH-NIDDK P01DK078669, NSF DBI-1565100, Alfred P Sloan Foundation G-2017-9838, Alfred P Sloan Foundation G-2015-13933, and ONR N00014-15-1-2809. R.K. was supported by NSF DBI-1565100, Alfred P Sloan Foundation G-2017-9838, Alfred P Sloan Foundation G-2015-13933, NIH-NIDDK P01DK078669 and ONR N00014-15-1-2809. This work was supported in part by CRISP, one of six centers in JUMP, a Semiconductor Research Corporation (SRC) program sponsored by DARPA.

References

1. Lozupone C & Knight R UniFrac: a new phylogenetic method for comparing microbial communities. *Appl. Environ. Microbiol* 71, 8228–8235 (2005). [PubMed: 16332807]
2. Thompson LR et al. A communal catalogue reveals Earth’s multiscale microbial diversity. *Nature* 551, 457–463 (2017). [PubMed: 29088705]
3. McDonald D et al. American Gut: an Open Platform for Citizen Science Microbiome Research. *mSystems* 3 (2018).
4. Gonzalez A et al. Qiita: rapid, web-enabled microbiome meta-analysis. *Nature Methods* (in press) (2018).
5. Caporaso JG et al. QIIME allows analysis of high-throughput community sequencing data. *Nat. Methods* 7, 335–336 (2010). [PubMed: 20383131]
6. Chang Q, Luan Y & Sun F Variance adjusted weighted UniFrac: a powerful beta diversity measure for comparing communities based on phylogeny. *BMC Bioinformatics* 12, 118 (2011). [PubMed: 21518444]
7. Chen J et al. Associating microbiome composition with environmental covariates using generalized UniFrac distances. *Bioinformatics* 28, 2106–2113 (2012). [PubMed: 22711789]
8. McMurdie PJ & Holmes S phyloseq: an R package for reproducible interactive analysis and graphics of microbiome census data. *PLoS One* 8, e61217 (2013). [PubMed: 23630581]
9. Amir A et al. Deblur Rapidly Resolves Single-Nucleotide Community Sequence Patterns. *mSystems* 2 (2017).

Figure 1.

Algorithm description and empirical performance results. **(A)** For a given node in a phylogeny, sample proportions are embedded, duplicating the proportion information. This duplication mimics rotating the sample proportions to compute many pairwise distances at once in contiguous blocks of memory. **(B)** A schematic of the two stripes in a four sample logical distance matrix; the labels above the stripes denote the pairwise comparison represented (e.g., “AB” indicates the distance between samples “A” and “B”). **(C-D)** Empirical time and space comparisons of weighted unnormalized UniFrac¹ between phyloseq⁸, QIIME v1.9.1⁵, scikit-bio, Striped UniFrac in exact mode, and Striped UniFrac in fast mode. Each data point represents n=10 independent experiments (random subsets) of the Earth Microbiome Project 90nt (317,314 unique Deblur⁹ sub-OTUs) dataset at increasing numbers of samples. All methods were run single threaded on non-shared compute nodes which were not running other compute tasks. A job was killed if it exceeded 24 hours walltime or 256GB of memory (system max). Error bars are 95% confidence intervals around the mean. The data for panels C and D can be found in Supplemental Dataset 2.